universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Decoding the cis-regulatory information of enhancer sequences“

verfasst von / submitted by

Bernardo Lucas Carvalho Pereira de Almeida, Licenciatura Mestre

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 620 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molecular Biosciences
Betreut von / Supervisor:	Dipl.-Biochem. Dr. Alexander Stark

“The people who are crazy enough to think they can change the world are the ones who do!”

Steve Jobs

Acknowledgments

I would like to express my sincere gratitude to Alex Stark for his exceptional mentorship in all aspects. Since the beginning that it was clear that we share the same excitement about *understanding the information encoded in the genome*, and it was a real joy to work together on this challenge throughout my PhD journey. His guidance, expertise, and constant feedback have been invaluable to my scientific development and the successful completion of this thesis.

I want to thank the members of my PhD committee – Daniel Gerlich, Fyodor Kondrashov and Christa Buecker – for their insightful comments and suggestions, which have helped me improve the quality of this work.

I want to thank all the members from the great Stark Lab (past and present). I feel fortunate to have been part of this extraordinary, diverse, and fun group of people and scientists. Your unique blend of personalities, backgrounds, and perspectives have enriched my personal and professional growth throughout this journey. Thank you for all the cool science, I learned a lot from everyone. Thank you for the amazing time we spent together, including all the breaks, lunches (waiting for Jelle to finish…), beer hours (always the first to arrive and last to leave), dinners, parties and travels.

Special thanks to Fanny with whom I collaborated closely in different projects throughout my PhD – it was a pleasure to work with you.

I would like to thank all my colleagues and friends at the VBC campus. In particular my PhD buddy Catarina. We have done this whole journey together, helping each other in every moment, sharing the good and less good experiences, and now we are graduating together. Your friendship and encouragement have been a constant source of motivation for me.

I want to thank all my friends back at home, Portugal. Despite the distance, we have been always together.

I want to thank all my family for their understanding and unconditional support. Specially my parents, grandparents and brother.

Finally, I want to thank Sara, my soulmate, for all your love.

Big thanks to all of you.

TABLE OF CONTENTS

Summary

The instructions for when and where each of the approximately 20,000 human protein-coding genes is to be expressed are encoded in the DNA sequences of transcriptional enhancers. Enhancers are genomic non-coding cis-regulatory elements that act as on-off switches of gene transcription. The vast majority of disease-associated mutations fall into the non-coding part of the genome and appear to be particularly enriched in enhancers and affect gene regulation. However, despite the importance of enhancers for development and disease, deciphering the link between the sequence of an enhancer and its regulatory activity has remained one of the greatest challenges in biology, and neither predicting enhancer activity nor designing synthetic enhancers with specific properties has been achieved.

The aim of this PhD thesis was to achieve a better understanding of the cis-regulatory information encoded in enhancer sequences by integrating deep learning algorithms with high-throughput enhancer testing and systematic enhancer sequence perturbation assays, using *Drosophila melanogaster* S2 cells as the main model system.

First, I developed a deep learning model, DeepSTARR, that predicts the enhancer activity of any DNA sequence, its critical nucleotides, and enables the design of synthetic enhancers de novo. I applied this approach to *Drosophila* S2 cells and trained DeepSTARR to learn its enhancer sequence code with increased accuracy. In a second step, I interpreted the model and revealed long-sought-after sequence rules for enhancers, including the importance of motif-flanking nucleotides and transcription factor motif-motif distances. We validated these rules experimentally and demonstrated their conservation in human enhancers. Finally, we also designed and functionally validated synthetic enhancers with desired activities, not only demonstrating the validity of the model and its rules but also illustrating the power of such approaches for synthetic biology.

To further understand the rules of enhancer sequence syntax, we designed a large-scale enhancer mutagenesis project. The resultant enhancer activity changes validated the predictive sequence features of DeepSTARR and revealed that enhancers display constrained sequence flexibility – only a specific but still diverse set of sequences and TF motifs can function at a given position. This activity of motifs at specific positions is strongly determined by the enhancer sequence context, namely the flanking sequence, presence and diversity of other motif types, and distance between motifs.

Altogether, my work could provide the basis of current and future efforts to understand the regulatory information encoded in the human genome, predict the impact of genomic variation on function and disease, and of engineering synthetic enhancers for biotechnological applications, especially gene therapy.

Zusammenfassung

Die genauen Informationen darüber wann und wo jedes der etwa 20 000 menschlichen proteinkodierenden Gene exprimiert werden soll, sind in den DNA-Sequenzen der sogenannten „Enhancer"-Elemente kodiert. Enhancer sind genomische, nicht kodierende cis-regulierende Elemente, die als Ein- und Aus-Schalter der Gentranskription fungieren. Die überwiegende Mehrheit der krankheitsassoziierten Mutationen fällt in den nichtkodierenden Teil des Genoms und scheint sich besonders in Enhancern anzureichern und die Genregulation zu beeinträchtigen. Trotz der Bedeutung von Enhancern für Entwicklung und Krankheit, ist die Entschlüsselung des Zusammenhangs zwischen der Sequenz eines Enhancers und seiner regulatorischen Aktivität eine der größten Herausforderungen in der Biologie geblieben. Weder die Vorhersage der Enhancer-Aktivität noch die Entwicklung synthetischer Enhancer mit spezifischen Eigenschaften ist bisher gelungen.

Ziel dieser Doktorarbeit war es, ein besseres Verständnis der in Enhancer-Sequenzen kodierten cis-regulatorischen Informationen zu erlangen, indem Deep-Learning-Algorithmen mit Hochdurchsatz-Enhancer-Tests und systematischen Enhancer-Sequenz-Perturbationsexperimenten kombiniert wurden, wobei *Drosophila* melanogaster S2-Zellen als Hauptmodellsystem verwendet werden.

Zunächst entwickelte ich ein Deep-Learning-Modell - DeepSTARR, das die Enhancer-Aktivität einer beliebigen DNA-Sequenz und ihre kritischen Nukleotide vorhersagt und die Entwicklung synthetischer Enhancer de novo ermöglicht. Ich wandte diesen Ansatz auf S2-Zellen von *Drosophila* an und trainierte DeepSTARR, um den Code der Enhancer-Sequenz mit erhöhter Genauigkeit zu lernen. In einem zweiten Schritt habe ich das Modell interpretiert und Sequenzregeln für Enhancer ermittelt, wie zum Beispiel die Bedeutung von motivflankierenden Nukleotiden und Transkriptionsfaktor-Motiv-Abständen. Wir validierten diese Regeln experimentell und konnten ihre Erhaltung in menschlichen Enhancern nachgeweisen. Schließlich haben wir auch synthetische Enhancer mit den gewünschten Aktivitäten entworfen und funktionell validiert, was nicht nur den Nachweis für die Gültigkeit des Modells und seiner Regeln erbringt, sondern auch das Potenzial solcher Ansätze für die synthetische Biologie verdeutlicht.

Des weiteren entwickelten wir ein groß angelegtes Enhancer-Mutagenese-Projekt, um die Regeln der Enhancer-Sequenzsyntax besser zu verstehen. Die daraus resultierenden Veränderungen der Enhancer-Aktivität bestätigten die vorhergesagten Sequenzmerkmale von DeepSTARR und zeigten, dass Enhancer eine eingeschränkte Sequenzflexibilität aufweisen. Nur eine bestimmte, aber dennoch vielfältige Gruppe von Sequenzen und TF-Motiven kann an einer bestimmten Position funktionieren. Diese Aktivität von Motiven an bestimmten Positionen wird

stark durch den Kontext der Enhancer-Sequenz bestimmt, d. h. durch die flankierende Sequenz, das Vorhandensein und die Vielfalt anderer Motivtypen und den Abstand zwischen den Motiven.

Insgesamt hat meine Arbeit das Potenzial als Grundlage für aktuelle und künftige Bemühungen zu dienen, die im menschlichen Genom kodierte regulatorische Information zu verstehen, die Auswirkungen genomischer Variationen auf Funktion und Krankheit vorherzusagen und synthetische Enhancer für biotechnologische Anwendungen, insbesondere die Gentherapie, zu entwickeln.

Introduction

Genome as the "Life's Code"

Although DNA was discovered in 1869 by Friedrich Miescher[1], only in the middle of the 20th century it was demonstrated that DNA, not protein as previously thought, is the hereditary molecule that carries the genetic information[2]. Not even a decade later Watson and Crick deciphered the double helical structure of DNA[3], and yet another decade later they cracked the *genetic code* that dictates the encoding of the proteins' amino-acid sequences[4]. These and other findings at that time showed how the information for creating an organism could be encoded in a DNA sequence made of only four different letters (nucleotides) – also called *genome* – changing our view of life.

With the important advance of DNA sequencing technologies[5] and the ability to read the DNA four-letter code, it was possible to start reading the genomes of different organisms, including ourselves[6,7]. The Human Genome Project is one of the greatest technological feats in history and revealed that the human genome is approximately 3 billion base-pairs long, containing only around 20,000 protein-coding genes (1% of the genome). These were surprisingly fewer than the 100,000 genes expected at the time[8,9], not more genes than a worm[10], and a lot less than a tomato[11]. But it turned out that the eukaryotic genome contains more than just protein-coding genes and has evolved other mechanisms for generating complex multicellular organisms comprising a large variety of cell types and organs.

Much of this complexity derives from how the same genome is differentially interpreted by different cell types to express specific sets of genes and proteins that define their identity[12]. For example, despite containing the exact same DNA sequence, skin cells express several structural proteins, such as keratin, whereas red blood cells have high levels of the oxygen-carrying protein hemoglobin. These differential gene expression patterns are regulated predominantly at the level of gene transcription – the copying of DNA into RNA – and governed by multiple types of non-coding cis-regulatory elements[13,14], such as promoters[15], enhancers[16,17], insulators[18], tethering elements[19,20] and silencers [21–23]. In analogy to the well-understood *genetic code*, deciphering the *cis-regulatory code* of the genome is critical for understanding the developmental programs during organism development and how genetic variants and mutations affect development and disease, since over 80% of genetic variants associated with human diseases and traits fall in non-coding regulatory regions[24].

Gene expression is encoded in enhancers' DNA sequences

The most abundant cis-regulatory sequences are *enhancers*, sequence elements that activate transcription of their target genes in specific cell types and conditions, independent of their relative distance, location, or orientation to the cognate promoter[25] (Fig 1A). According to its definition, the activity of an enhancer must reside within its DNA sequence. Indeed, early experiments using functional reporter assays showed that enhancer DNA is sufficient to drive cell-specific enhancer activity even when placed outside of its endogenous genomic context[25–28] (Fig 1B). In addition, the enhancer DNA sequence also recapitulates endogenous TF binding, DNA methylation and histone modifications[29,30]. Enhancer sequences are associated with rapid evolutionary changes[31,32] and contain most of the known disease-associated variants[24] (Fig 1C,D). However, despite the crucial role of enhancers in development and disease[16], understanding how the gene-regulatory information is "encoded" in their sequence and interpreted within the cell has remained one of the greatest challenges in biology.

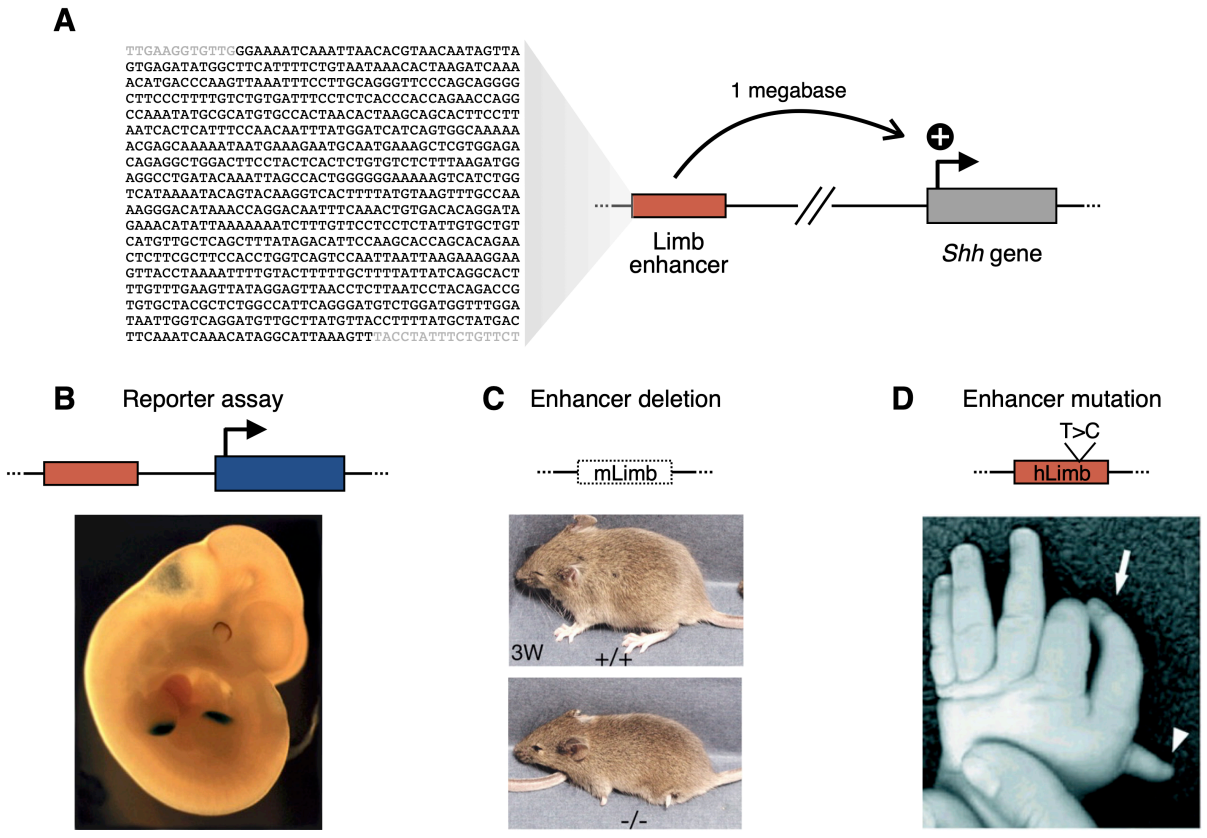


Figure 1. Enhancer activity is encoded in the enhancer's DNA sequence. A) Enhancers are DNA elements that activate transcription of a target gene. This is illustrated by the limb enhancer of the *Sonic hedgehog* (*Shh*) gene, located 1 megabase upstream of *Shh*, that is composed by the DNA sequence shown and activates the gene in the limbs. B) The limb enhancer drives expression of a reporter gene in the mouse limbs, recapitulating its endogenous activity. Image retrieved from the VISTA Enhancer database[33]. C) Enhancers are critical in organism development. For example, deletion of this limb enhancer in mice causes truncation of limbs. Image retrieved from ref.[34]. D) Mutations in enhancer sequences are associated with disease, as illustrated in this enhancer where a single nucleotide mutation (T/C transition at position 323) causes preaxial polydactyly in humans. Note the triphalangeal thumb on the left hand (arrow). Image retrieved from ref.[35].

Identification of enhancers

The identification of large sets of enhancers with similar functions has been instrumental for the study of enhancers and their sequence features. The enhancer activity of a DNA sequence is typically assessed by functional reporter assays that measure the abundance of reporter transcripts or proteins[36] (Fig 1B). However, classical reporter assays (e.g. based on luciferase) suffered from low throughput, as candidates needed to be tested one by one, and thereby prevented systematic studies of enhancer sequences. Therefore, other enhancer properties have been used for genome-wide enhancer predictions across different cell types, such as certain properties of chromatin (e.g. open chromatin flanked by histones carrying post-translational modifications on lysine 27 (acetylation, H3K27ac) and/or lysine 4 (monomehylation, H3K4me1)), transcription factor (TF) or co-activator (e.g. P300) binding, as well as bidirectional transcription[17,37–41]. Using these biochemical annotations, large-scale projects like ENCODE[39] and Roadmap Epigenomics[38] have annotated around two million candidate regulatory elements in the human genome as potential enhancers in one or more cell types, covering 13% of the genome cumulatively[40]. However, using these correlative features to identify enhancers yields both false positive and false negative candidate sequences, and cannot quantitatively assess the activity of putative enhancers[42].

To address this challenge and identify enhancers in a genome-wide manner through direct functional tests of enhancer activity, two main approaches have been developed[36,42]: they measure (1) the ability of candidate sequences to drive transcription in standardized reporter assays (sufficiency; e.g. massively parallel reporter assays or MPRAs[43–46]) or (2) the requirement of candidate regions for endogenous gene expression (necessity; e.g. CRISPR screens to directly perturb enhancers in their native genomic context[47–49]). One of such MPRA methods, self-transcribing active regulatory region sequencing (STARR-seq), has been pioneered by my host lab to measure the enhancer activity of millions of DNA fragments in parallel using a constant, tractable sequence environment. This method successfully identified enhancer sequences genome-wide and fostered the systematical measurement of large-scale libraries of enhancer variants in both *Drosophila* and human cells (refs.[46,50–54] and Publication 1 and 2). The use of STARR-seq to identify and characterize enhancer sequences is one topic of this thesis (Publication 1 and 2).

The availability of high-throughput approaches to identify and characterize enhancer sequences genome-wide in different cell types (discussed above) combined with the recent development of novel computational methods (discussed below in “Decoding the cis-regulatory code with deep learning”) offers a unique and timely opportunity to decipher the link between an enhancer’s DNA sequence and its regulatory activity in the cell – the main focus of this thesis.

The cis-regulatory code of enhancer sequences

Enhancers are 200-1000 base pairs long and contain combinations of short sequence motifs that are recognized and bound by different sequence-specific TFs[55] (Fig 2A,B). The combined regulatory cues of all bound TFs determine an enhancer's activity. TFs can have various roles at enhancers, such as acting as pioneer factors, triggering the repositioning of nucleosomes and promoting accessibility for other factors, or they may recruit or facilitate the recruitment of co-activators and -repressors that do not directly bind to the DNA[55]. Since TFs are differently expressed in different cell types during development or in response to signaling cascades, they provide the means for the cell to regulate the spatiotemporal activity of enhancers and their target genes[56]. Understanding how exactly TFs bind and regulate the activity of specific enhancers and how this information is encoded in each enhancer's sequence has become a key goal towards understanding the cis-regulatory code.

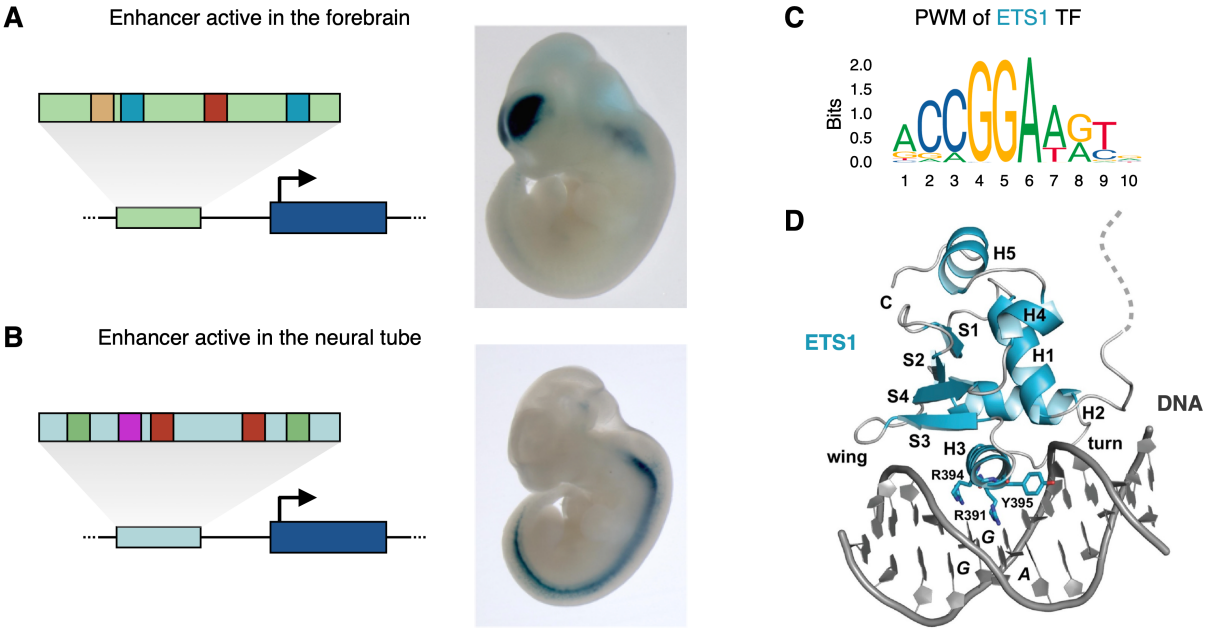


Figure 2. The combination of all TFs bound to an enhancer determines its spatio-temporal enhancer activity. A-B) Shown are cartoons of two different enhancers with different combinations of TF motifs (colored boxes) and thus bound by different sets of TFs. Due to the specific combinations of TFs, they drive expression in different tissues, such as (A) the forebrain and (B) the neural tube. Images from mouse embryos were retrieved from the VISTA Enhancer database[33]. C) Position weight matrix (PWM) of the human ETS1 TF displayed as a motif logo. The height of each position is proportional to its information content and the height of individual letters to their relative frequencies. Motif logo retrieved from the JASPAR database[57]. D) Example of how the binding domain of the ETS1 TF (blue) recognizes and binds in the major groove of DNA (grey), resulting in the sequence affinity represented in (C). Source: [58].

TF motifs

TFs typically recognize small 6–12 bp-long degenerate DNA sequences (motifs) through physical interactions between their amino acid side chains and the accessible edges of the DNA base-pairs, including direct hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic contacts[59] (Fig 2C,D). Due to these chemical interactions, TFs bind DNA in a sequence-specific manner. The binding preference of each TF is usually represented in the form of Position Weight Matrices (PWM) that are then used to identify instances of each motif in a DNA sequence[60,61] (Fig 2C).

TF motifs are often found in enhancer sequences and their importance for enhancer activity was initially demonstrated through systematic mutational analyses of individual enhancers, such as the *even-skipped* stripe 2[62]. Such mutational tests have recently been scaled up by the development of transcriptional reporter assays that can assess the activity of thousands of enhancer variants in parallel (MPRAs). This technology has been used to systematically measure the importance of nucleotides and TF motifs to enhancer activity by saturation mutagenesis of individual enhancer sequences[43–45,63] or the mutation of TF motifs in thousands of enhancers (Publication 1 and refs.[64]).

In addition to the genetic and experimental approaches, many computational approaches have been developed for the discovery of important TF motifs. Statistical sequence analyses of large sets of enhancers with similar activity have revealed over-represented TF motifs that are important in different cell types[28,38–40,51]. A complementary approach relies on sequence conservation to identify functional motifs within enhancer sequences[65–69].

Due to their short length, TF motifs occur very frequently throughout the genome (for example, each 6-bp long motif would be expected to occur every 4^6 bp = 4,096 bp on each DNA strand). However, only a small fraction of all matches in a genome are typically bound and fall in enhancer sequences, and these differ between cell types or conditions[70–76], suggesting that the TF motif alone is not sufficient to direct *in vivo* binding. Indeed, enhancer sequences typically contain a combination of motifs for different TFs that cooperate to drive enhancer activity, for example by facilitating the binding of each other[77] or by recruiting cofactor proteins[78]. This motif combinatorics combined with the differential expression of the respective TFs can explain cell type-specific binding and enhancer activities[55] (Fig 2A,B). This is also demonstrated by the ability to predict with good accuracy cell type-specific enhancer function based on their motif content[28,51].

Models of how enhancers encode function

In addition to the presence of multiple TF motifs, several studies have shown that enhancer activity depends on additional sequence constraints related to the motifs' flanking sequences, affinities and arrangements (their number, order, orientation and spacing), termed here '*motif syntax*' (Fig 3; Publication 1 and 2 and refs.[79–98]). However, the importance and characterization of such syntax features and constraints within enhancers are still debated and remain outstanding questions in gene regulation.

Two main models have been proposed to explain how the enhancer sequence relates to function. The *enhanceosome* model assumes very strict syntax rules with precise positioning of motifs required for cooperative TF binding and enhancer activation[84]. This model was derived from the Interferon-*β* enhancer, where even small sequence changes can affect the binding of all TFs and subsequent enhancer activity[94]. In contrast, the *billboard* model proposes that TFs bind largely independently to enhancers and that there are no constraints on how TFBSs are arranged, exemplified by the *even-skipped* stripe 2 enhancer[82,99]. Yet very few enhancers fit these models, having either invariant syntax or no constraints at all. Recent studies, including this thesis, support a model where most enhancers fall in between these two extremes, depending on flexible but still important motif syntax features (Publication 1 and 2 and refs.[80,82,100]).

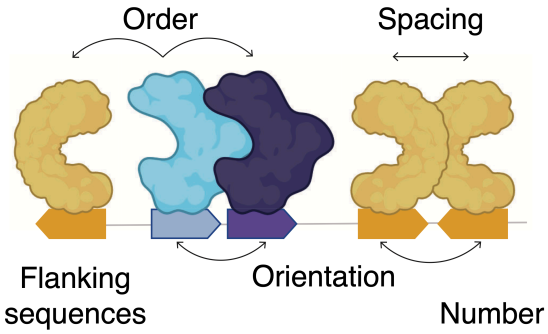


Figure 3. Enhancer syntax. The formation of a functional enhancer relies on both physical interactions between TFs and DNA, as well as protein-protein interactions. Given the physical properties of proteins and the enhancer DNA, such interactions are thought to depend on syntax features such as the flanking sequences, affinities, number, order, orientation and spacing of TF motifs. Figure adapted from ref.[80].

Enhancer syntax

While the contribution of motif syntax features in enhancer sequences is still not completely clear, several experiments have provided insights into existing syntax constraints. Below, I summarize studies that assessed the importance of such sequence features (for an extensive review see ref.[80]).

Motif affinity

TFs are believed to control gene expression primarily by binding to specific motifs on DNA which have a strong affinity for the protein, or "cognate" sites. Therefore, most genomic studies have focused on the highest affinity motif instances for each TF. However, work over the past decade has questioned this assumption and showed that both high- and low-affinity TF motifs within enhancers are important to direct precise patterns of gene expression[83,88–90,101–106]. Low-affinity motifs are prevalent in developmental enhancers and increasing their affinity can lead to loss of tissue-specific expression, developmental defects and diseases[89,90,102]. There appears to exist a tradeoff between affinity and specificity. Low-affinity motifs might confer tissue specificity by favoring cooperative binding, making the enhancer only active where all factors are present at the right concentration, thus ensuring combinatorial control of gene expression[88,90]. Conversely, higher motif affinity would make the enhancer active even in cells that have low levels of the TF and therefore less dependent on the combinatorial control, explaining the loss of tissue-specificity.

While it is straightforward to identify high-affinity motif instances, distinguishing between functional low-affinity instances and non-functional ones is more challenging, as low-affinity motif instances occur more frequently just by chance. Functional low-affinity instances are likely dependent on additional syntax features, such as a specific distance to neighboring TF motifs[97,102]. For example, the spacing and orientation of motifs can compensate for low-affinity motifs in notochord enhancers, presumably by stabilizing the TF interactions required for a functional enhancer complex[89]. Thus, the higher affinity motifs are not necessarily the most functionally significant ones and exploring the complex interplay between low-affinity motifs and enhancer syntax is crucial for a better understanding of enhancer sequences[105].

Motif number

Besides heterotypic combinations of different types of TF motifs, enhancers often contain multiple instances of the same motif type, with varying affinities. MPRAs using both natural and synthetic enhancers have consistently shown that higher number of instances of a TF motif correlates with stronger enhancer activity, with varying effects between different TFs[79,92,93,107]. The number of instances of each motif type has indeed been used as an important feature in computational predictive models of enhancer activity[28,51,107].

Motif order

Since TFs interact both with the enhancer DNA but also with each other via protein-protein interactions, the order in which their motifs are arranged within an enhancer can influence such interactions. These include hetero- or homodimers and higher-order interactions. Indeed, the

order of motifs has been shown to influence enhancer activity of synthetic enhancers in MPRAs[79,91,92] and to be important in encoding developmental expression patterns *in vivo*[85,86,108].

An interesting observation within the Notch-regulated *ma* enhancer in *Drosophila* showed that changing the order of motifs leads to loss of activity in the original cells and ectopic activity in other tissues. This shows that motif order is not only necessary for enhancer activity in the tissue of interest but can also prevent activity in ectopic tissues – described as *preventative syntax*[86]. However, a second Notch-regulated *Drosophila* enhancer in the same study (ASE5) was largely resistant to perturbations of its syntax, with similar activity upon changes in the order of its required motifs[86]. This highlights the challenge of understanding the importance of motif order for enhancer activity. This is further complicated by the need of first defining the functional features or motifs before assessing the importance of changes in their order, and because any changes in order can create or ablate other functional motifs confounding the resultant effects.

Motif orientation

Similar to the order of motifs, the orientation or relative direction of motifs can also affect the interaction of TFs with the enhancer DNA and other bound proteins and the formation of a functional complex. When the motif orientation is altered, the TF will bind to the opposite strand of DNA in a reversed orientation. Several examples have been described where the relative direction of motifs for two TFs is important for *in vivo* enhancer activity[89,109–112]. The importance of motif orientation was also observed for the pluripotency TFs in mouse synthetic enhancers[79,91].

Motif spacing

Cooperation between TFs can occur via multiple mechanisms and at different ranges[113]. Constraints on spacing between motifs may ensure that all motifs are accessible for the respective TFs and can modulate both TF-DNA interactions along with protein-protein interactions such as binding of dimers[114]. For example, facilitating the interaction between TFs that bind to adjacent motifs can enable cooperativity of activators[115,116] or short-range repression[117,118]. Other arrangements might inhibit such interactions between activators, preventing ectopic expression without the need for transcriptional repression[86]. Helical 10 bp periodicity of motifs within enhancers has also been shown to be important for gene expression, likely by promoting cooperative binding by having motifs on the same side of the DNA[44,94,95,97,110,117,119,120]. Even single base-pair[89,121] or two base-pair[122] changes in spacing can affect expression levels, demonstrating how subtle changes in motif spacing can impact enhancer activity.

Dependency syntax

Despite extensive evidence that enhancer activity may depend on the syntax of TF motifs, the extent to which such syntax features might constrain enhancer activity remains difficult to define, as it differs from one enhancer to another. Indeed, such rules are rarely found in genome-wide analyses and perturbing them affects the function of some enhancers but not others. In addition, although enhancer sequences evolve rapidly, their function can be conserved despite significant sequence changes[31,32,81,90,106,123–133]. This suggests that enhancer sequences are highly flexible and that the maintenance of function-defining features rather than overall sequence similarity is important for enhancer activity.

To explain how most enhancers fall in between the two extreme models of the billboard and the enhanceosome, Jindal & Farley proposed a *dependency syntax* model where enhancers encode enhancer activity through the dependency and interplay between such sequence syntax features, which in turn are shaped by evolutionary, biological, and mechanistic constraints[80] (Fig 4). For example, in *Ciona* development, previous analyses of notochord enhancers regulated by Zic and ETS TF motifs found that there is an interplay between affinity and organization of motifs, such that organization could compensate for poor affinity and vice versa[89]. The intricate syntax of enhancers has posed a long-standing challenge in identifying and generalizing sequence-rules, thus hindering the establishment of unified principles governing the regulatory code.

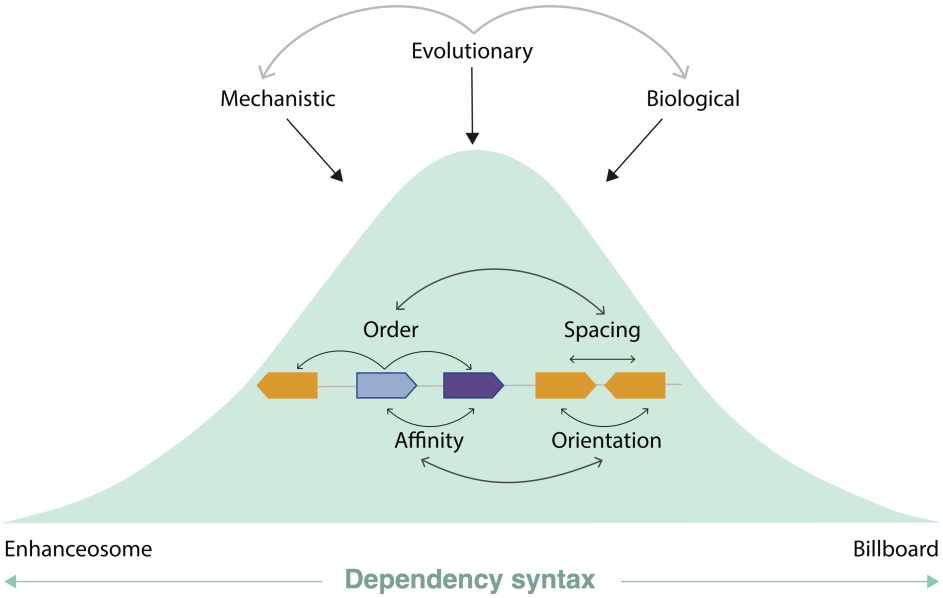


Figure 4. Dependency syntax. The enhanceosome and billboard models represent two opposite ends of a spectrum. Most enhancers fall at different points along this spectrum, encoding enhancer activity through the dependency and interplay between motif syntax features that are shaped by evolutionary, biological, and mechanistic constraints. Figure adapted from ref.[80].

Characterizing the cis-regulatory syntax features and constraints within enhancer sequences is one topic of this thesis (Publication 1 and 2).

Decoding the cis-regulatory code with deep learning

Although many features of enhancer sequences and their syntax have been described over the past 40 years, there is still no clear understanding of how the combination of such features encodes specific enhancer activity patterns. Moreover, the prediction of enhancers from sequence alone remains challenging, as well as de novo enhancer design. In addition to the importance and characterization of the described motif-based rules, there might be requirements that are not well represented by motifs, or for which we do not know the necessary motifs yet. Furthermore, there might be multiple codes intertwined in a single enhancer sequence, or enhancers with similar activity might be a mix of two or more enhancer codes, such as cell type-specific and ubiquitous. The complexity of the enhancer sequence space highlights the need for approaches that can identify the different sequence features and model their interdependencies.

Many modeling approaches based on thermodynamics or machine learning methods have been used to predict enhancer activity and to pursue a quantitative and mechanistic understanding of enhancer function[51,64,79,107,134–147]. However, such approaches have modeled enhancer sequences explicitly via predefined sets of features that were informed by previous biological and biophysical knowledge, such as motif dictionaries or de novo *k*-mers (Fig 5A). Despite important successes, no clear rules of the cis-regulatory code have emerged[55,59]. This is mainly due to the difficulty of computationally defining and encoding those features (e.g. TF motif affinity) and their interdependencies (e.g. motif organization can compensate for motif low affinity). There is hence a critical need for more general, unbiased and flexible computational methods that can take advantage of the existing large amount of accurate, high resolution genomics datasets to model the information encoded in enhancer sequences.

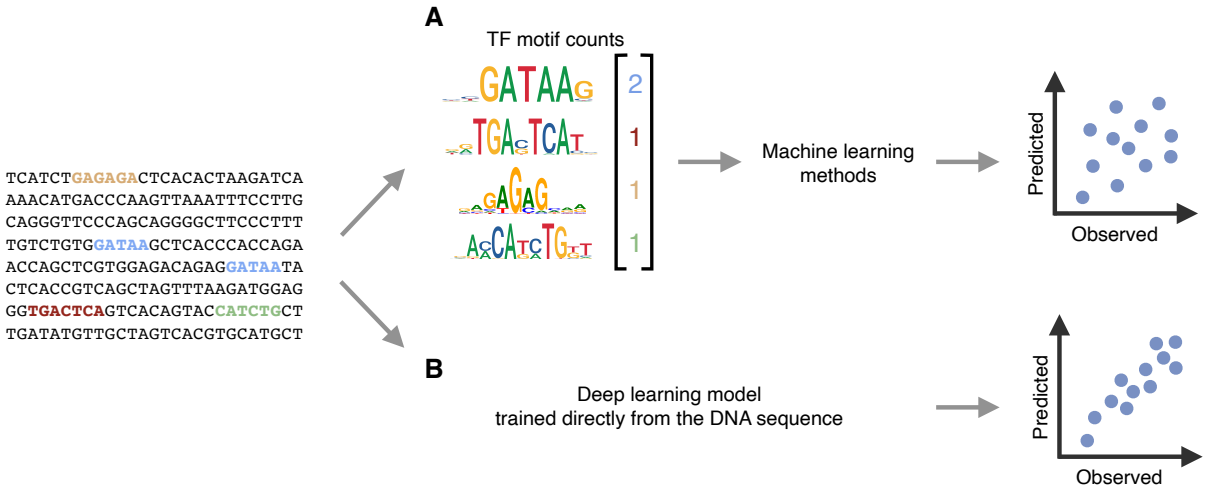


Figure 5. Predicting enhancer activity directly from the DNA sequence rather than via predefined sets of features. A) A common approach for predicting the enhancer activity of DNA sequences is to first extract specific sets of features (like TF motif counts) and use them in machine learning frameworks. B) Using deep learning algorithms, it is possible to skip the feature extraction step and use the DNA sequence directly as input to the model, which should discover the relevant features during training in order to predict the enhancer activity of DNA sequences.

Deep learning: a new paradigm

Deep learning – a sub-discipline of machine learning – bypasses the need for predefined known features by embedding the computation of features into the automated learning process, yielding so-called end-to-end models[148]. This was made possible by the development of deep neural networks, that are machine learning models consisting of multiple, consecutive layers performing successive elementary operations and nonlinear transformations. Since each layer operates on the results of the preceding layers, the model is able to improve prediction accuracy by learning increasingly complex features and modeling nonlinear relationships during model training.

One key feature of deep learning models is their capacity to extract higher-order features from the raw input data. Combined with the explosion of the amount of available data and computing capacity in the past years, deep learning models have led to multiple performance breakthroughs in a diverse set of tasks, including computer vision[149–151], speech recognition[152], machine translation[153], playing computer games[154] and self-driving cars[155]. Deep learning's success already reached numerous scientific fields, such as chemistry, physics, biology, and materials science, where it has outperformed other machine learning techniques[156–161]. More recently, in structural biology, the deep-learning-based methods AlphaFold2[162] and RoseTTAfold[163] achieved unprecedent accuracy in the long-standing problem of predicting protein structures from their genetic sequence, representing a true paradigm shift in the way we study biology.

I illustrate the conceptual difference to other, more classical machine learning approaches with the following example. To classify a tumor as malign or benign from a microscopy image, a preprocessing algorithm could be first used to detect cells, identify the cell type and estimate the cell counts for each cell type, which will be used by a machine learning model to classify the tumor. However, this makes the classification performance highly dependent on the quality and relevance of the handcrafted features used as input (i.e. estimated cell counts), and is inherently limited to the use of known features, missing other currently unknown features. An alternative approach would be to use the image directly as input to a deep neural network that would learn all the steps required to classify the tumor, including features of high complexity as the cell morphology and spatial organization of cells that are not captured in cell counts. Using the same analogy for predicting enhancer activity from the DNA sequence, rather than first extracting specific sets of features (like TF motif counts) and use them in a machine learning framework (Fig 5A), one could instead use the enhancer's DNA sequence directly as input to a deep learning model and let the model discover the relevant features in order to predict the enhancer's activity (Fig 5B).

Modelling regulatory genomics with convolutional neural networks

The applicability of deep neural networks in genomics was first demonstrated in 2015 by two studies that used deep learning to predict the sequence specificities of DNA- and RNA-binding proteins (DeepBind model[164]) as well as chromatin features (DeepSEA[165]) from DNA sequence, surpassing more established (non-deep learning) machine learning algorithms, such as support vector machines. This was followed by Basset[166], an open-source package developed to apply deep neural networks to learn the functional activity of DNA sequences from genomics data that facilitated the use of such models by the community. These seminal studies also demonstrated the utility of deep learning models in non-coding variant effect prediction and their promise to better annotate and interpret the non-coding genome. Since then, deep neural networks have been applied to predict diverse molecular phenotypes (see also Publication 1) and have emerged as the leading type of predictive models in regulatory genomics[167].

There are three major classes of neural networks: fully connected, convolutional and recurrent. Due to the importance of local and spatial dependencies between nucleotides in DNA sequences (e.g. TF motif sequences and the distance between motifs), the most widely used architecture to model DNA sequences is convolutional neural network (CNN). The three pivotal models mentioned above were indeed CNNs. CNNs are composed of one or more convolutional layers, each scanning a set of weight matrices (also called filters) across the input, therefore learning to recognize relevant local patterns[149]. The first convolutional layer applied to the input DNA sequence can also be viewed as scanning the sequence with several position weight matrices, such as TF motifs (Fig 6B-D). Since the same filter is scanned across all positions, it allows to detect that feature anywhere in the sequence (such as positions not seen during training) while keeping the total number of parameters small regardless of the sequence length. Subsequent convolutional layers allow a hierarchical decomposition of the input and are able to detect higher-order features, such as two TF motifs that are present within some distance range (Fig 6E-G). CNNs usually apply convolutional layers to extract the relevant features and patterns from the input sequence and combine them with fully connected layers that learn non-linear combinations of these features to perform the final prediction task (Figure 6H). Importantly, all the parameters of the convolutional (i.e. filters or features) and the fully connected layers (i.e. importance and combinations of features) are learned during model training.

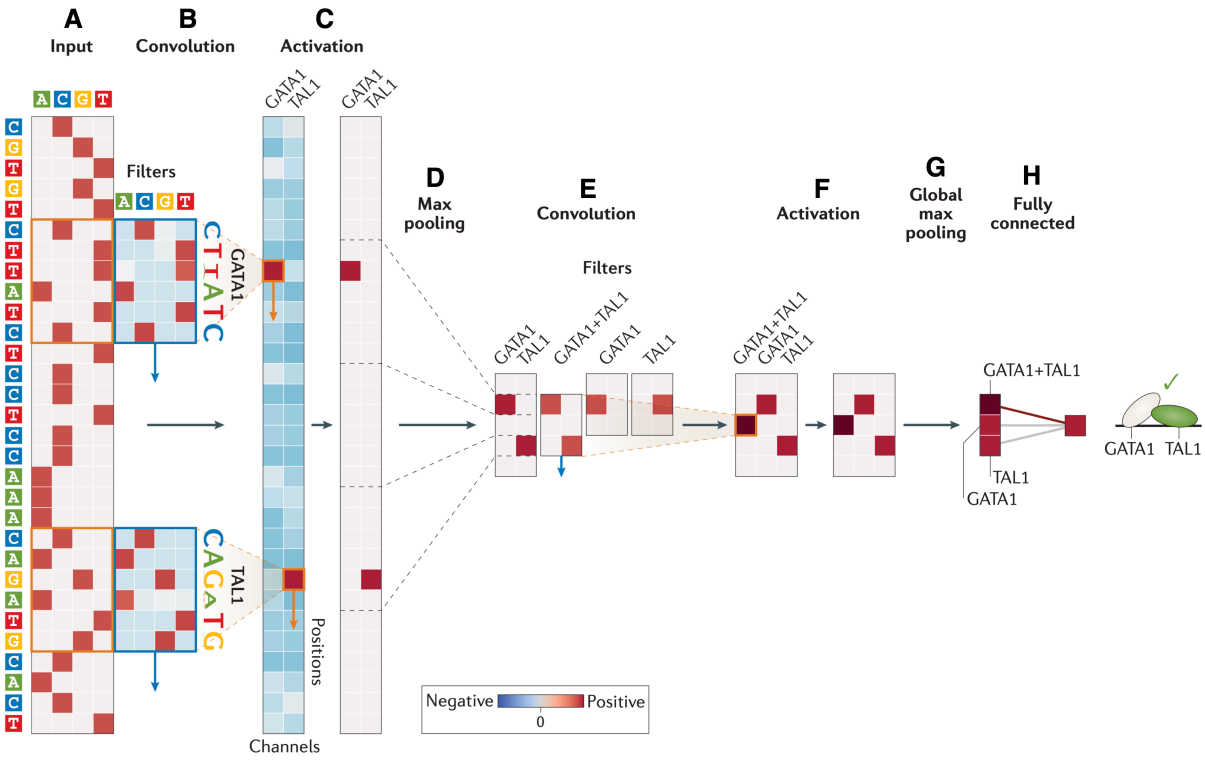


Figure 6. Modelling transcription factor binding and syntax with convolutional neural networks. The illustration shows a convolutional neural network that predicts the binding of the TAL1-GATA1 TF complex. A) The input is a one-hot encoded representation of the DNA sequence. B) The first convolutional layer scans the input sequence using filters to identify local sequence features, which are demonstrated using position weight matrices for the GATA1 and TAL1 TFs. C) An activation function (e.g. rectified-linear unit or ReLU) is used to set negative values to 0. D) Contiguous sections of the activation map can be summarized using a max pooling operation by taking the maximum value for each channel in each bin. E) The second convolutional layer scans the sequence for more complex patterns such as pairs of motifs and individual motifs. F) As in the first convolution, the ReLU activation function is applied. G) The maximum value across all positions for each channel is selected. H) A fully connected layer is used to make the final prediction. Figure retrieved from ref.[167].

Recurrent neural networks (RNNs)[168,169] are an alternative to CNNs for processing sequential data, such as DNA sequences. This type of network is composed of nodes arranged in a chain and uses the memory of the previous sequence element to influence the output, being able to model long-range dependencies in sequences. Both architectures are often used together, such that a sequence is passed through convolutional layers before entering a recurrent layer[170–174]. To increase the receptive field and integrate information from long-range interactions in the genome, CNNs have been also combined with dilated convolutions (achieving a receptive field up to 20 kb[175,176]) and transformers and self-attention layers (up to 100 kb[177]). Overall, CNNs and their different variant architectures have been used to predict with great accuracy several molecular phenotypes[167,178,179], including TF binding sites[97,164,180,181], DNA methylation[182,183], chromatin features[165,172,173,175,181,184–188], promoter[133] and enhancer activity (see Publication 1), 3D genome folding[189–193], splicing[194,195], gene expression[175,177,196], polyadenylation[197–200], mRNA stability[174], RBP binding[201–203], microRNA targets[204] and translation effiency[205].

Interpretation of deep learning models

Despite the impressive performance of deep learning models, their complexity and high capacity to encode latent feature representations make them particularly challenging to interpret. This explanatory information is usually of great value and can provide new insights into the biological mechanisms. For example, interpreting a trained model that predicts enhancer activity could reveal the importance of DNA motifs and their interactions in an enhancer sequence.

Over the past years, different approaches for model interpretation have been developed and show promise in identifying the features and feature combinations learned by the models. These can be categorized into[167,178]:

- model-based interpretation: interpreting first-layer convolutional nodes and visualizing attention weights;
- mathematical propagation of influence: in-silico mutagenesis and backpropagation-based methods;
- identification of interactions between features: examining deeper layer neurons or attention matrices, in-silico mutagenesis for specific combinations of features, and deep feature interaction maps;
- use of prior knowledge for transparent models, where the hidden nodes are constructed to be inherently interpretable.

There is still no consensus regarding which approaches are most effective and few studies have been able to uncover new biological features learned by this type of models. Only recently, model interpretation techniques have provided biological insights into the cis-regulatory code and the syntax of enhancer sequences (Publication 1 and refs.[97,173,188,198,206–210]). With the increasing number and accuracy of deep learning models, interpreting them and investigating complex relationships between features will have a central role in genomics.

Characterizing the cis-regulatory syntax of enhancer sequences using deep learning and model interpretation techniques is one topic of this thesis (Publication 1).

Aims of the thesis

The information that determines when and where a gene is going to be expressed resides in the DNA sequence of enhancers. Deciphering their cis-regulatory code has remained one of the greatest challenges in biology. This includes being able to quantitatively predict the activity of a given sequence, identify the important nucleotides, understand the motif syntax rules, and ultimately the de novo design of synthetic enhancers. Understanding the regulatory information encoded in enhancer sequences would unlock an enormous amount of regulatory information in the genome and allow us to interpret the impact of genetic variants involved in development and disease, which typically affect enhancers and gene regulation.

The general aim of this PhD thesis is to advance the understanding of the cis-regulatory information encoded in the genome. In particular, the main goals are the developmental of models to predict enhancers from the DNA sequence and the characterization of their sequence rules.

Specifically:

1. Develop a sequence-based deep learning model to predict enhancer activity directly from the DNA sequence (Publication 1),
2. Interpret the model to reveal relevant TF motifs and higher-order syntax rules in enhancer sequences, and use the model to design synthetic enhancers with specific activities (Publication 1),
3. Functionally characterize the rules of enhancer syntax using large-scale enhancer sequence perturbation assays (Publication 2).

Results and discussion of the publications

In this chapter I present the relevant publications I co-authored in the course of my PhD study.

Publication 1 – DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers

Bernardo P. de Almeida, Franziska Reiter, Michaela Pagani, Alexander Stark.

Nature Genet 54, 613–624 (2022). https://doi.org/10.1038/s41588-022-01048-5, REF: [211]

Summary and discussion

Identifying enhancers and characterizing their sequence determinants – the *cis-regulatory code* – is a long-standing problem. 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 its regulatory activity in the cell remains poorly understood, and de novo enhancer design with specific properties has been challenging.

In this paper, we combined a high-throughput enhancer testing technology with artificial intelligence to develop an innovative deep learning model, DeepSTARR. DeepSTARR predicts the enhancer activity of any DNA sequence, its critical nucleotides, and enables the design of synthetic enhancers de novo. We have applied this approach to *Drosophila melanogaster* S2 cells and trained DeepSTARR to learn its enhancer sequence code for two different transcriptional programs with remarkable accuracy. 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 inter-motif distances. We validated these rules experimentally and demonstrated their conservation in human enhancers. This revealed important insights about the regulatory principles of enhancer sequences in different species as distant as flies and human.

Finally, we further designed and functionally validated synthetic enhancers with desired activities de novo, demonstrating the validity of the model and its rules but also illustrating the power of such approaches for synthetic biology.

In summary, DeepSTARR allows us to predict, understand, and create enhancer sequences. Our work is complementary to recent efforts modeling other aspects of enhancer biology using deep learning, such as DNA accessibility, histone modifications or TF binding. I anticipate that these models will be further combined with models for other cis-regulatory elements (for example, promoters, insulators or silencers) as well as models that predict gene transcription from enhancer activities (for example the ABC model[48]) or the wider genomic sequence context (for example, Enformer[177]) towards ultimately understanding how our genomes store gene-regulatory information to dictate gene expression and development.

Author contribution

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.

nature genetics

ARTICLES

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

Check for updates

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

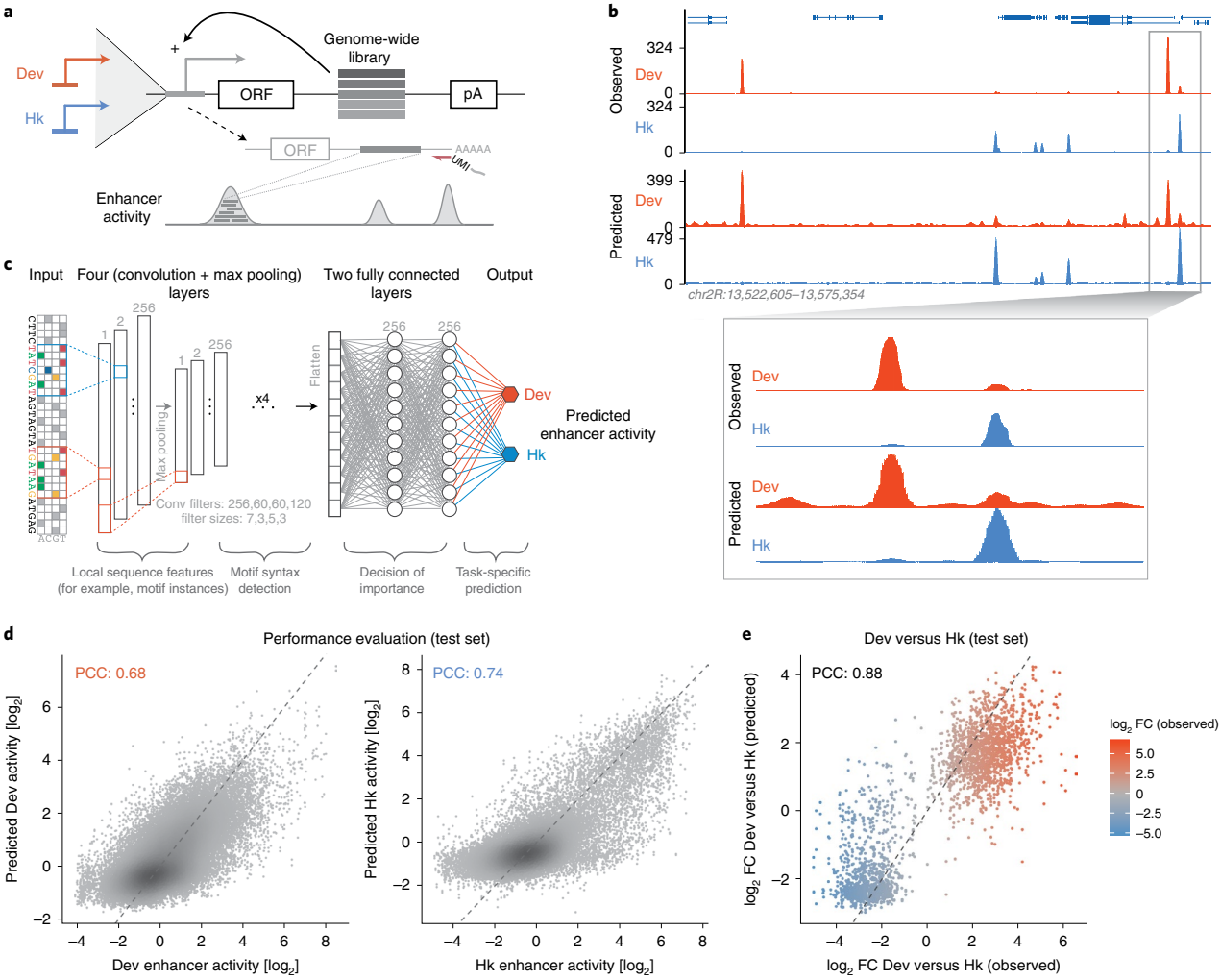


Fig. 1 | DeepSTARR quantitatively predicts enhancer activity genome wide from DNA sequence. a, Schematics of genome-wide UMI-STARR-seq using developmental (Dev) (DSCP; red) and housekeeping (Hk) (RpS12; blue) promoters. **b**, DeepSTARR predicts enhancer activity genome wide. Genome browser screenshot 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 $\log_2$FC 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

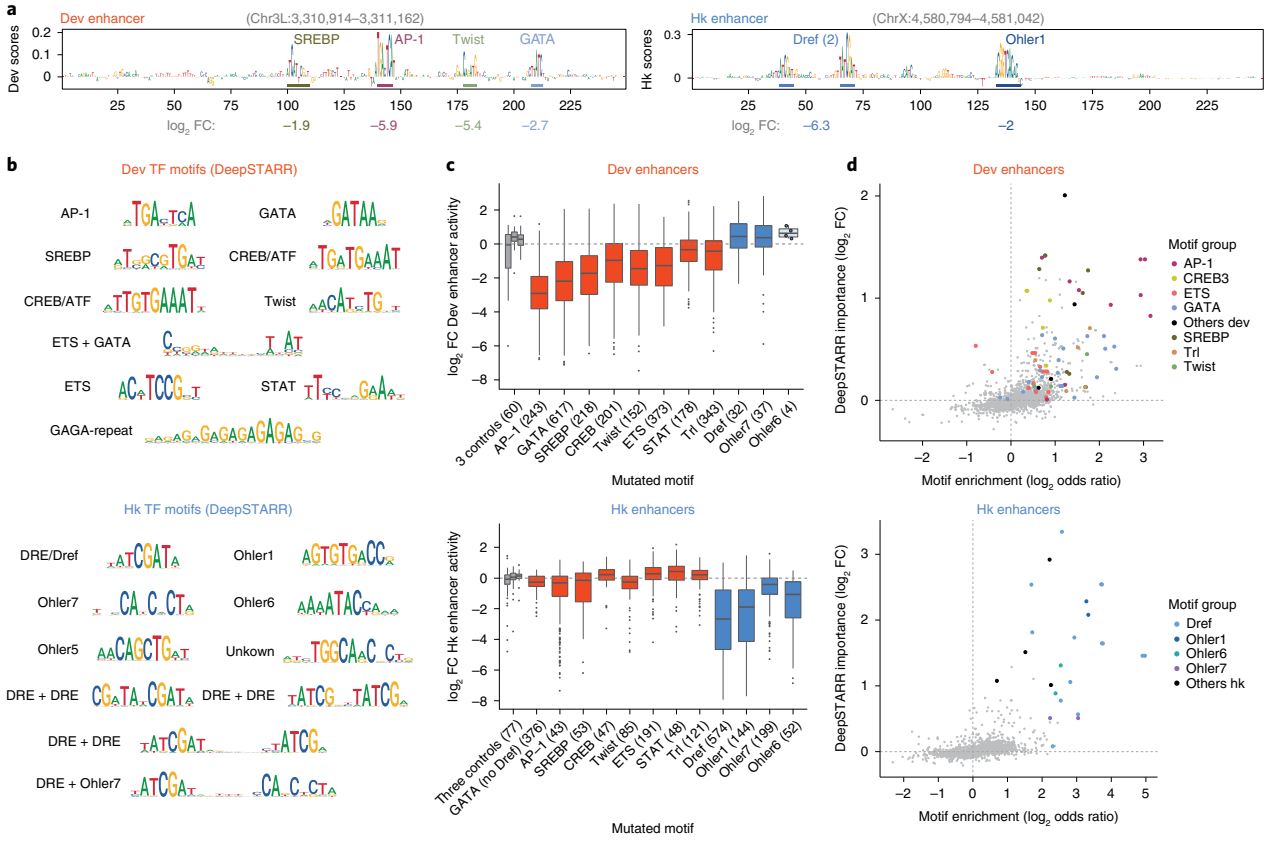


Fig. 2 | DeepSTARR reveals important TF motif types that validate experimentally. a, DeepSTARR-derived Dev and Hk nucleotide contribution scores for strong Dev (left) and Hk (right) enhancer sequences, respectively. Regions with high scores resembling known TF motifs are highlighted. log_2FC values (bottom) 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 two-fold 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.

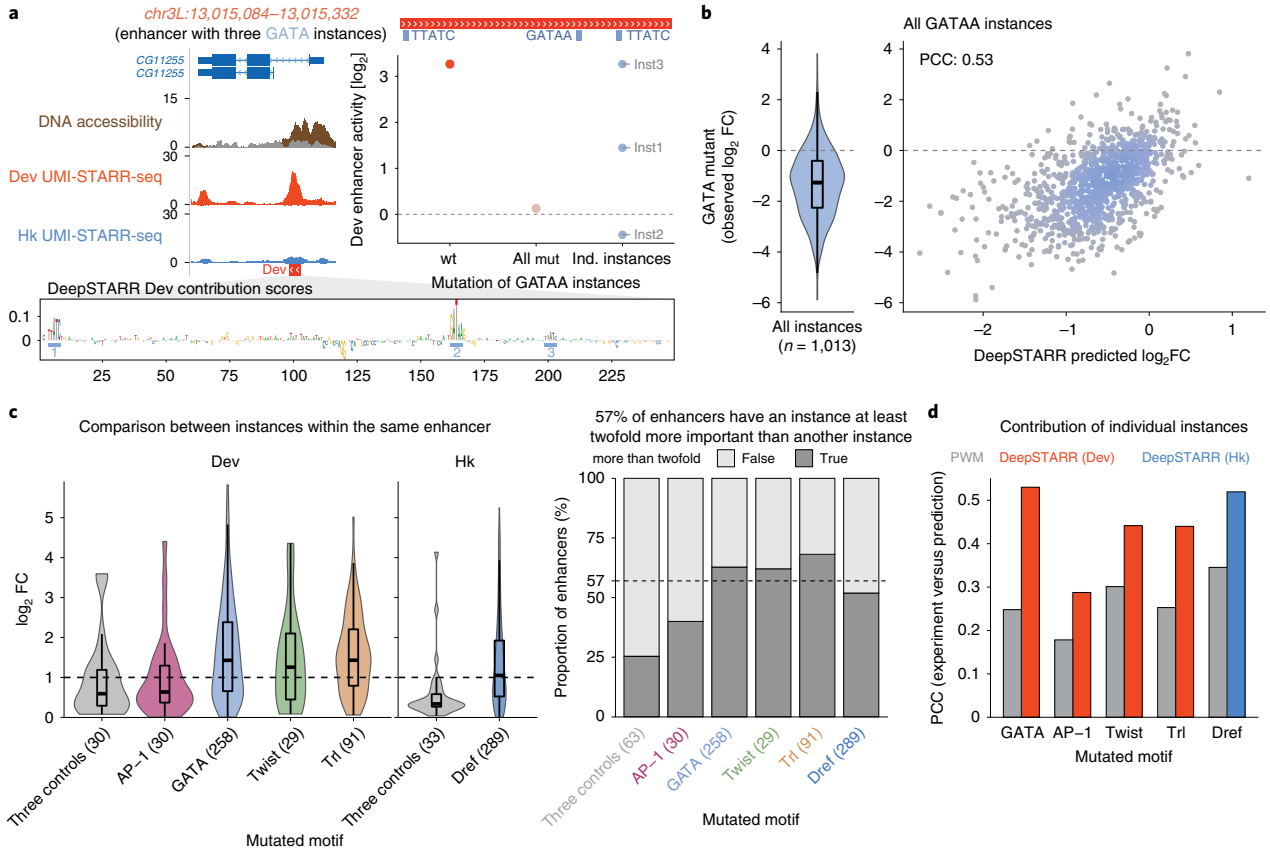


Fig. 3 | Instances of the same TF motif have nonequivalent contributions to enhancer activity. a, Developmental enhancer with three nonequivalent GATA instances. Left: genome browser screenshot showing tracks for DNA accessibility[63] and Dev and Hk UMI-STARR-seq for the *CG11255* locus. The designed 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

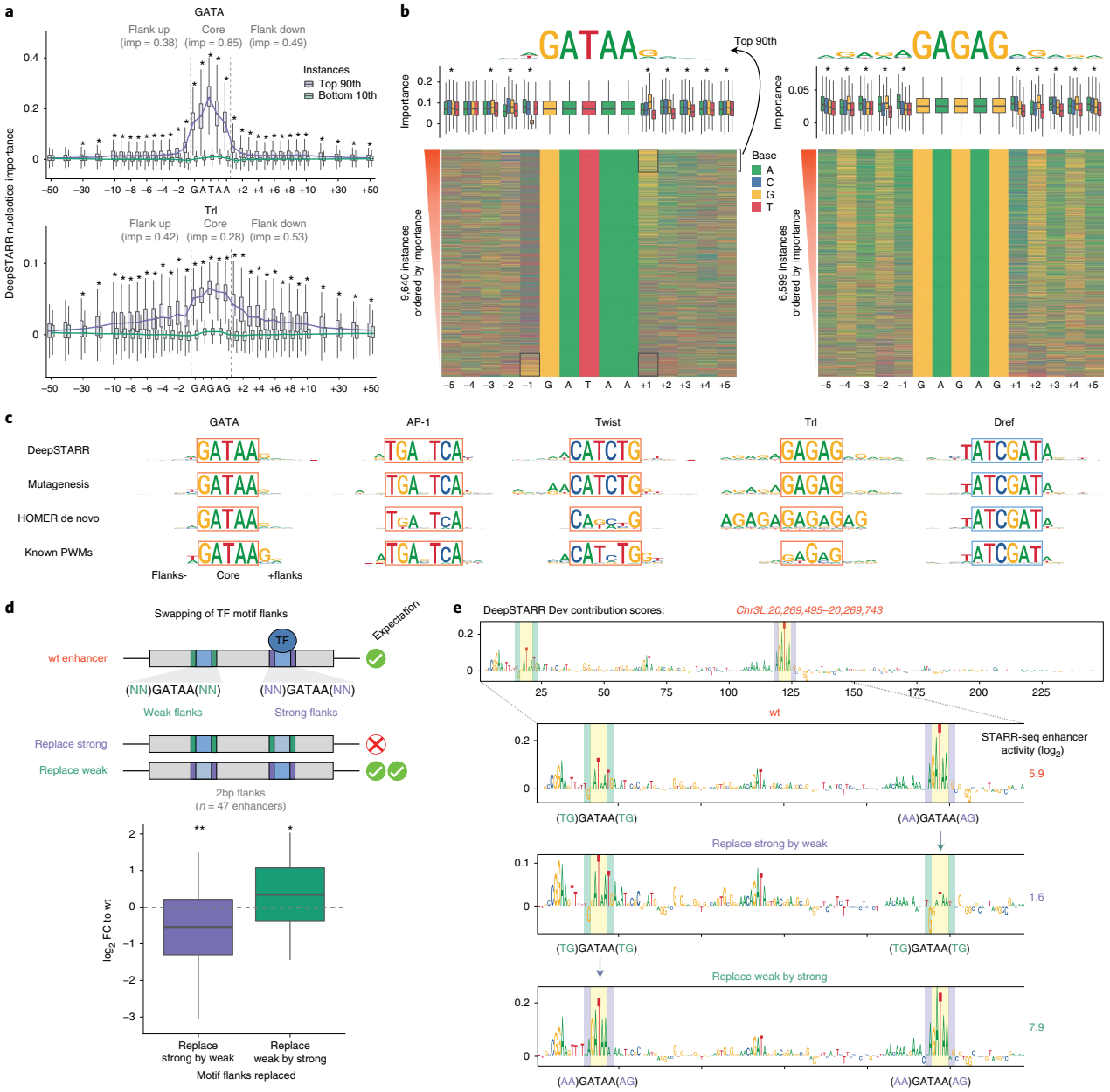


Fig. 4 | Contribution of TF motifs depends on the flanking sequence. a, DeepSTARR-predicted importance for ±50 flanking nucleotides of top 90th (purple) and bottom 10th (green) percentile GATA ($n = 992$ instances per box) and Trl ($n = 680$) motif instances selected based on DeepSTARR scores for core 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).

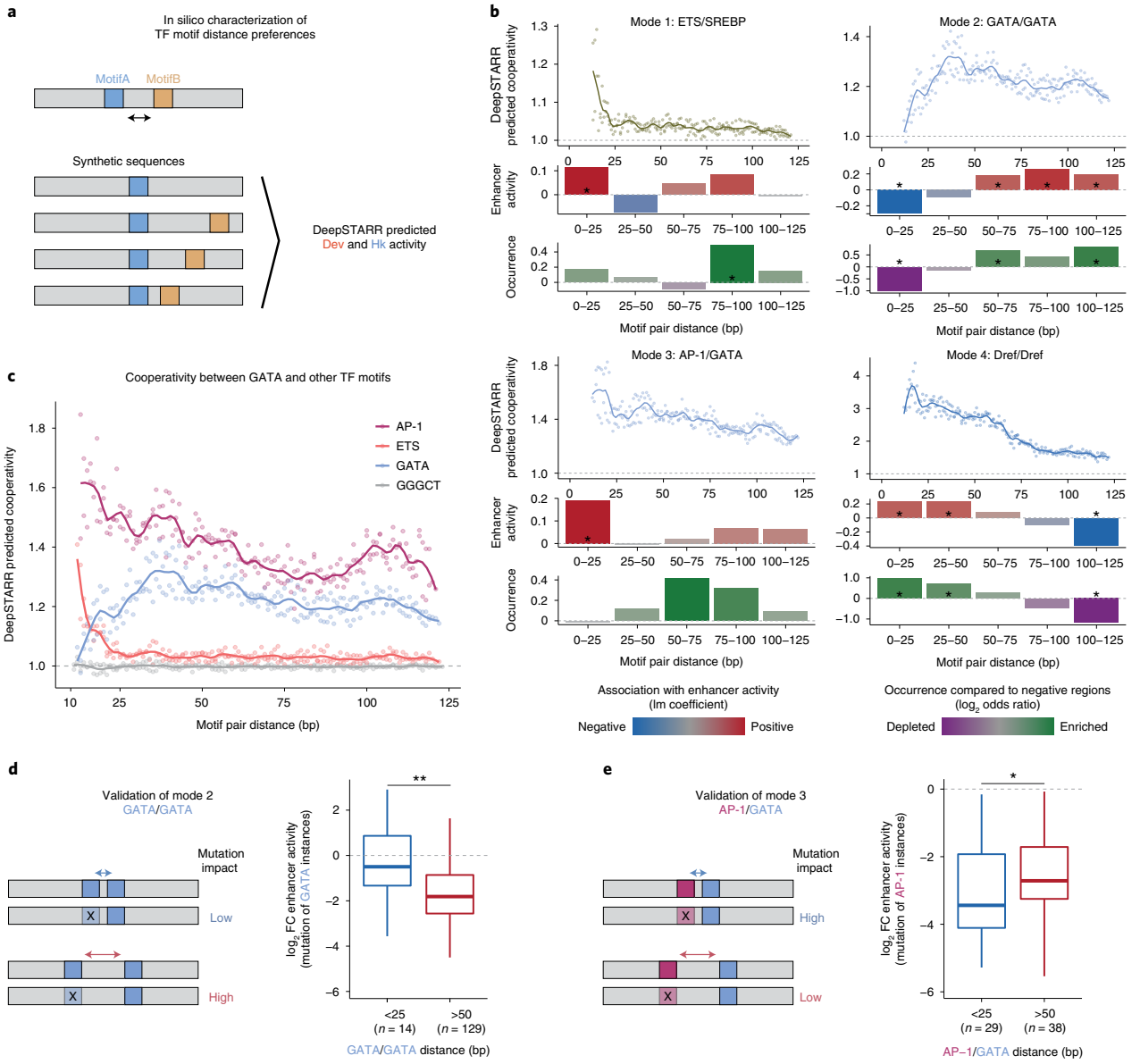


Fig. 5 | In silico analysis reveals distinct modes of motif cooperativity. a, Schematic of in silico characterization of TF motif distance preferences. MotifA was embedded in the center of 60 random sequences and MotifB at a range of distances from MotifA. The Dev and Hk enhancer activity was predicted 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, DNase I 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 DNase I 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.

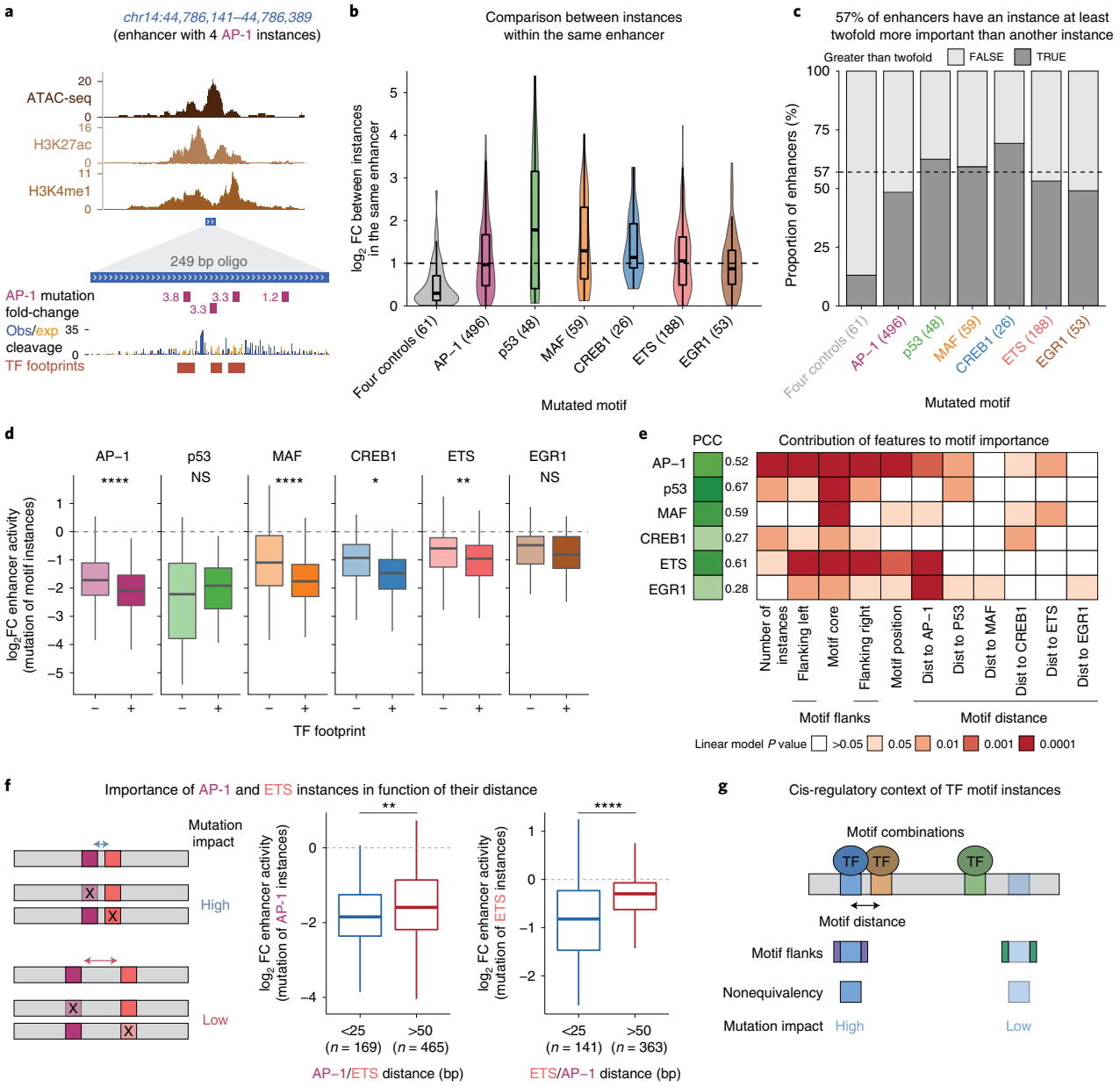


Fig. 6 | Motif syntax rules dictate the contribution of TF motif instances in human enhancers. a, Top: ATAC-seq[96], H3K27ac[97] and H3K4me1[97] signals for an enhancer with four AP-1 instances. Bottom: oligonucleotide used for motif mutagenesis containing four AP-1 instances and their mutation impact 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.

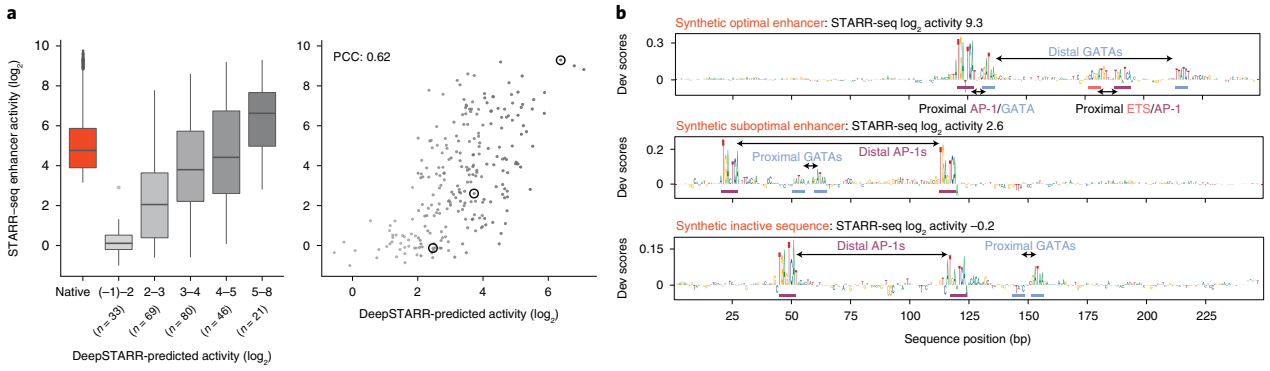


Fig. 7 | DeepSTARR designs synthetic enhancers using optimal sequence rules. a, Comparison between DeepSTARR predicted and experimentally measured enhancer activity (log_2) for 249 synthetic sequences binned (left) or not (right). The 'Native' category contains all *Drosophila* S2 Dev enhancer sequences. 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 DNase I 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

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

Supplementary information

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

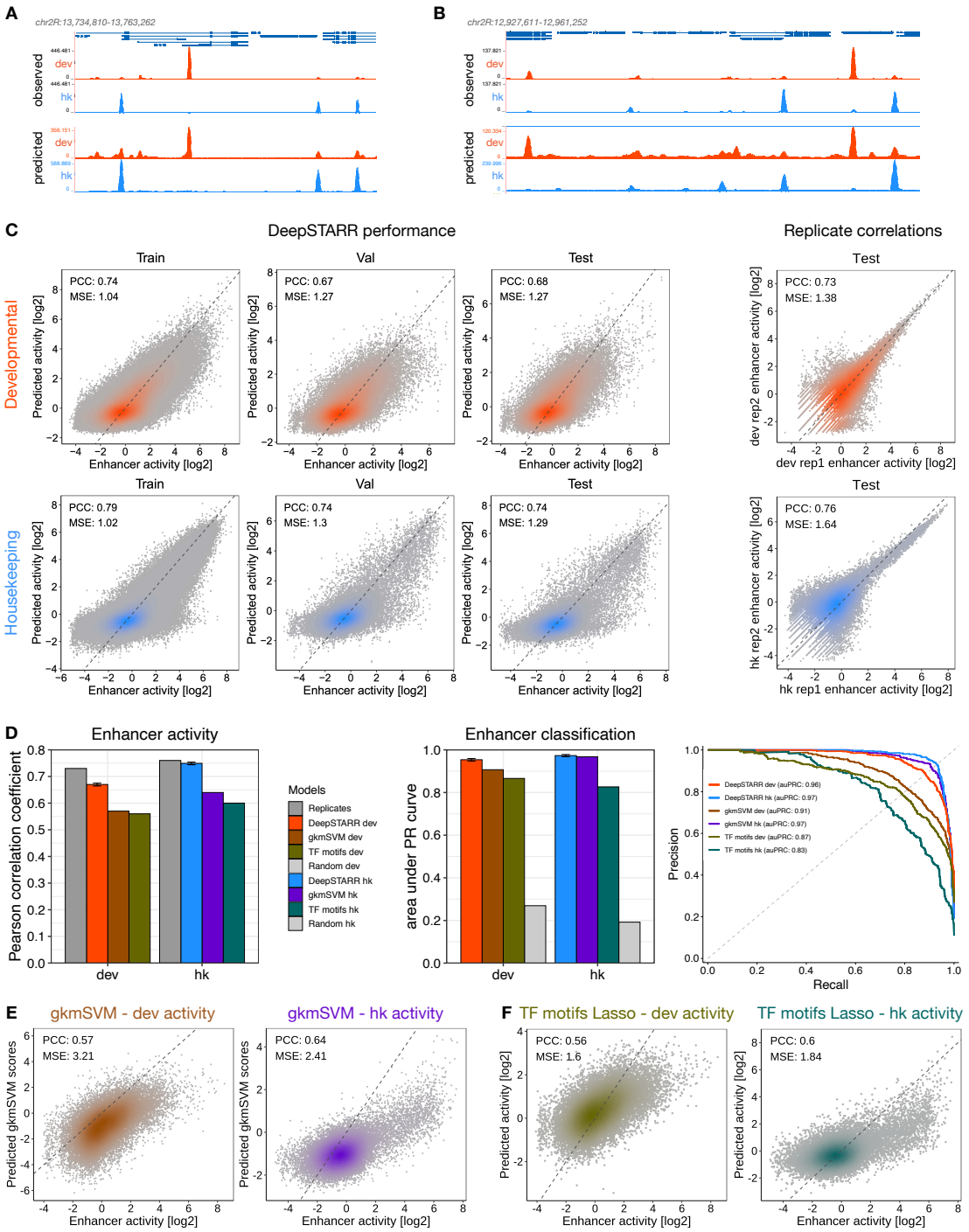
In the format provided by the authors and unedited

Supplementary Information

Table of Contents

Supplementary Figures

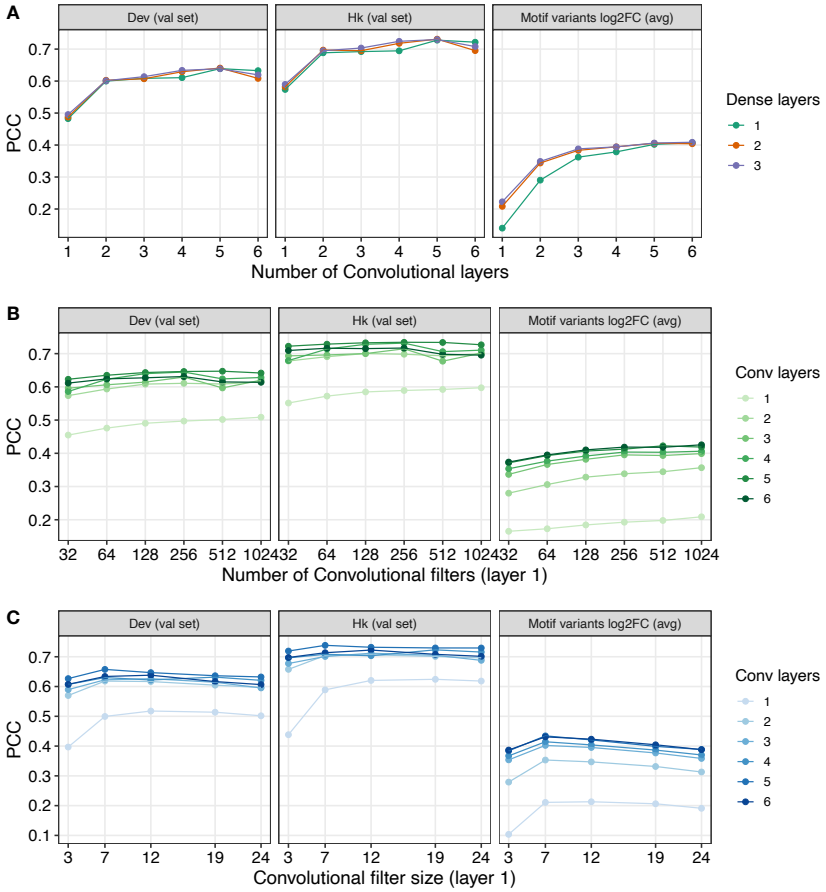
Supplementary Figure 1. Additional performance evaluation of DeepSTARR predictions.



A-B) DeepSTARR predicts enhancer activity genome-wide. Genome browser screenshot depicting UMI-STARR-seq observed (top) and predicted (bottom) profiles for both promoters (development, red; housekeeping, blue) for two loci located on held-out test chromosome 2R. **C)** DeepSTARR predicts enhancer activity quantitatively. Left: Scatter plots of predicted vs.

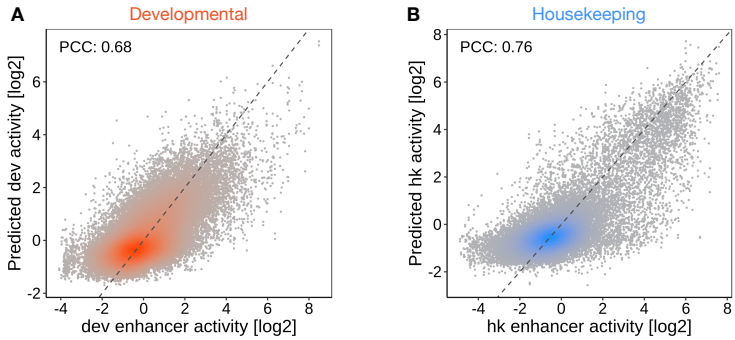
observed developmental (top) and housekeeping (bottom) enhancer activity signal across all DNA sequences in the train, validation and test set chromosomes. Right: Scatter plots of developmental (top) and housekeeping (bottom) enhancer activity signal between two biological replicates across all DNA sequences in the test set chromosome. Color reflects point density. The Pearson correlation coefficient (PCC) and mean squared error (MSE) are denoted for each comparison. **D)** DeepSTARR performed better than methods based on known TF motifs or unbiased k-mers. Left: Comparison of different models for predicting enhancer activity. Bar-plots with the PCC between observed and predicted activities for both developmental and housekeeping enhancer types across all DNA sequences in the test set chromosome. PCC between replicates is also shown. Middle: Bar-plots with the auPRC for the classification of enhancer sequences from the test set for the different models, compared with the expected by a random model. Right: precision-recall curve for the different models on test data. Error bars represent the mean values +/- 5^{th} and 95^{th} percentiles of the performance of 1000 DeepSTARR models. PCC: Pearson correlation coefficient, R2: R-squared, auPRC: area under precision-recall curve. **E-F)** Scatter plots of predicted (gkm-SVM **(E)** and TF motifs Lasso **(F)**) vs. observed developmental (left) and housekeeping (right) enhancer activity signal across all DNA sequences in the test set chromosome. The PCC and MSE are denoted for each comparison.

Supplementary Figure 2. Comparison of different model architecture choices.



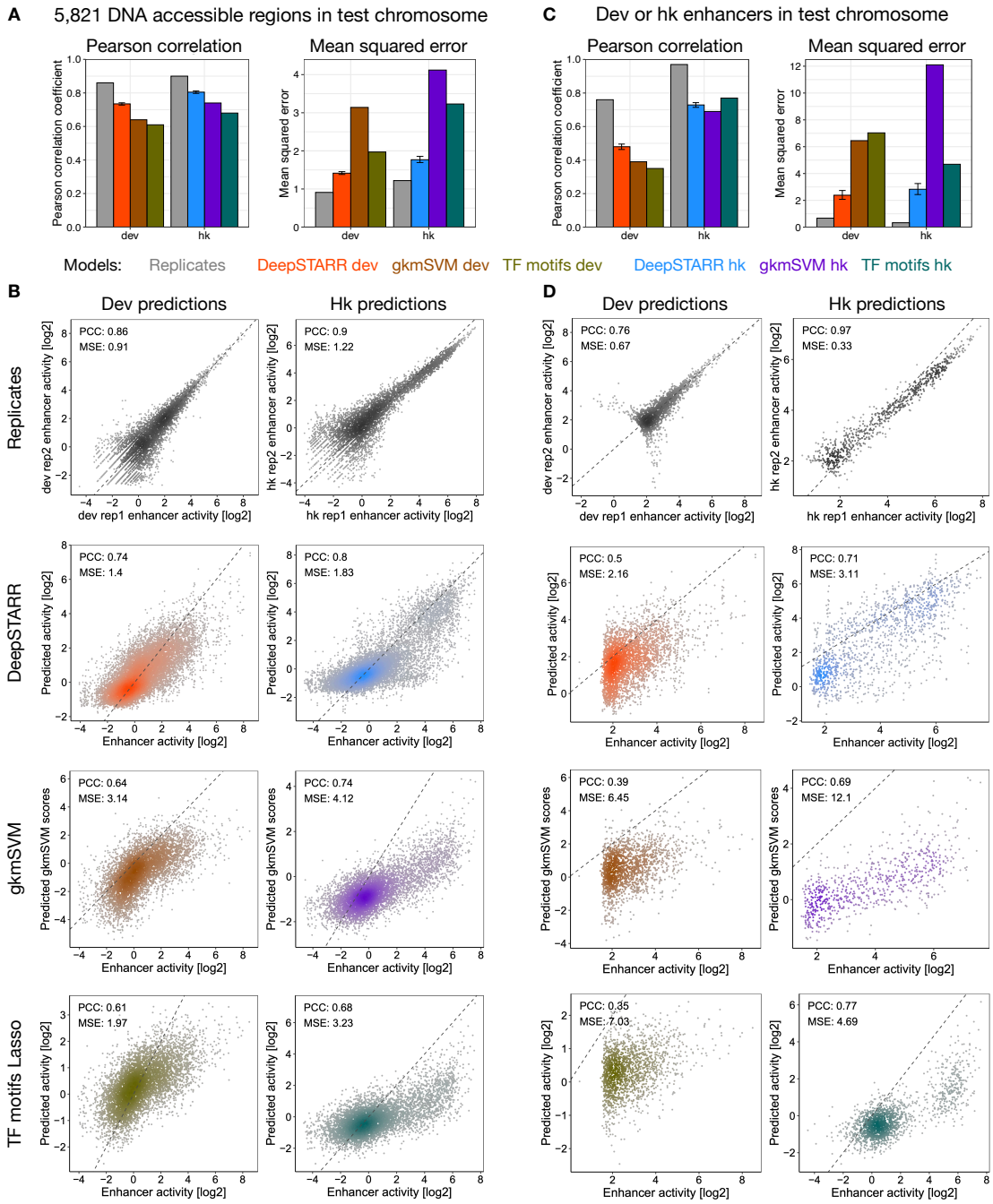
Performance comparison between models with varying number of convolutional and dense layers **(A)**, number of convolutional filters of the first layer **(B)**, and size of the convolutional filter of the first layer **(C)** (see Methods). For each combination of parameters, we trained at least 10 different models and assessed their performance (average PCC) on predicting enhancer activity (validation set, separately for developmental and housekeeping) and motif importance (motif mutation fold-changes, average across TF motifs). This revealed that 2 convolutional layers are minimally required to accurately predict enhancer activity, while 3 are minimally required to predict motif importance (DeepSTARR has 4) **(A)**. The number of dense layers has little impact in predicting enhancer activity, while 2 are required for better prediction of motif importance **(A)**. 256 or 512 convolutional filters **(B)** of size 7 **(C)** in the first layer are optimal but these parameters showed overall lower importance.

Supplementary Figure 3. Performance evaluation of DeepSTARR's predictions in test set excluding repeats.



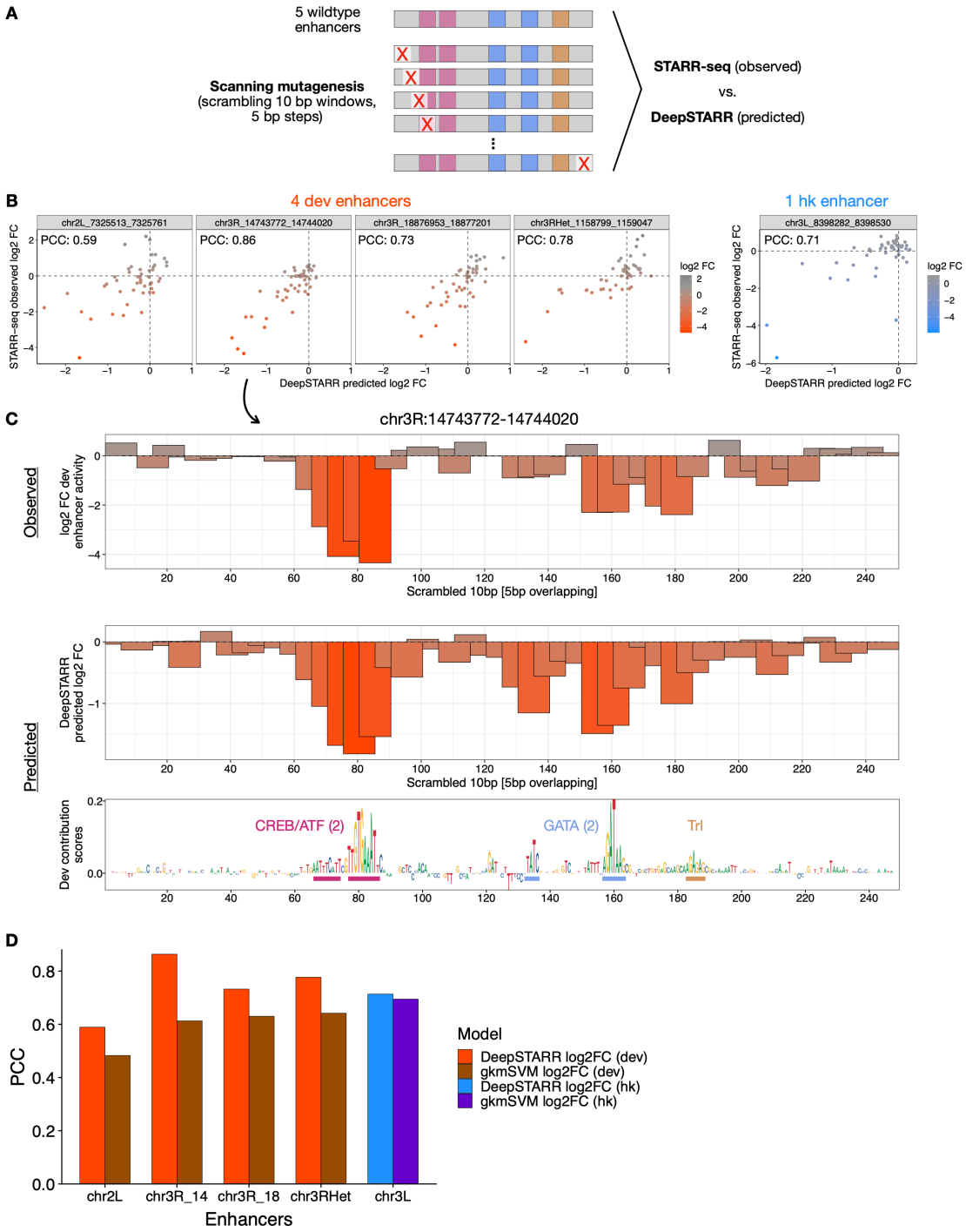
Scatter plots of DeepSTARR predicted vs. observed developmental **(A)** and housekeeping **(B)** enhancer activity signal across 32,036 DNA sequences in the test set chromosome not overlapping with repeats. The PCC is denoted for each comparison.

Supplementary Figure 4. Additional performance evaluation of DeepSTARR's predictions.



Comparison of different models for predicting enhancer activity. Bar-plots with the PCC (left) and MSE (right) between observed and predicted activities for both developmental and housekeeping enhancer types across DNA accessible regions **(A)** or enhancers **(C)** in test chromosome. Error bars represent the mean values +/- 5th and 95th percentiles of the performance of 1000 DeepSTARR models. Respective scatter plots shown in **(B,D)**.

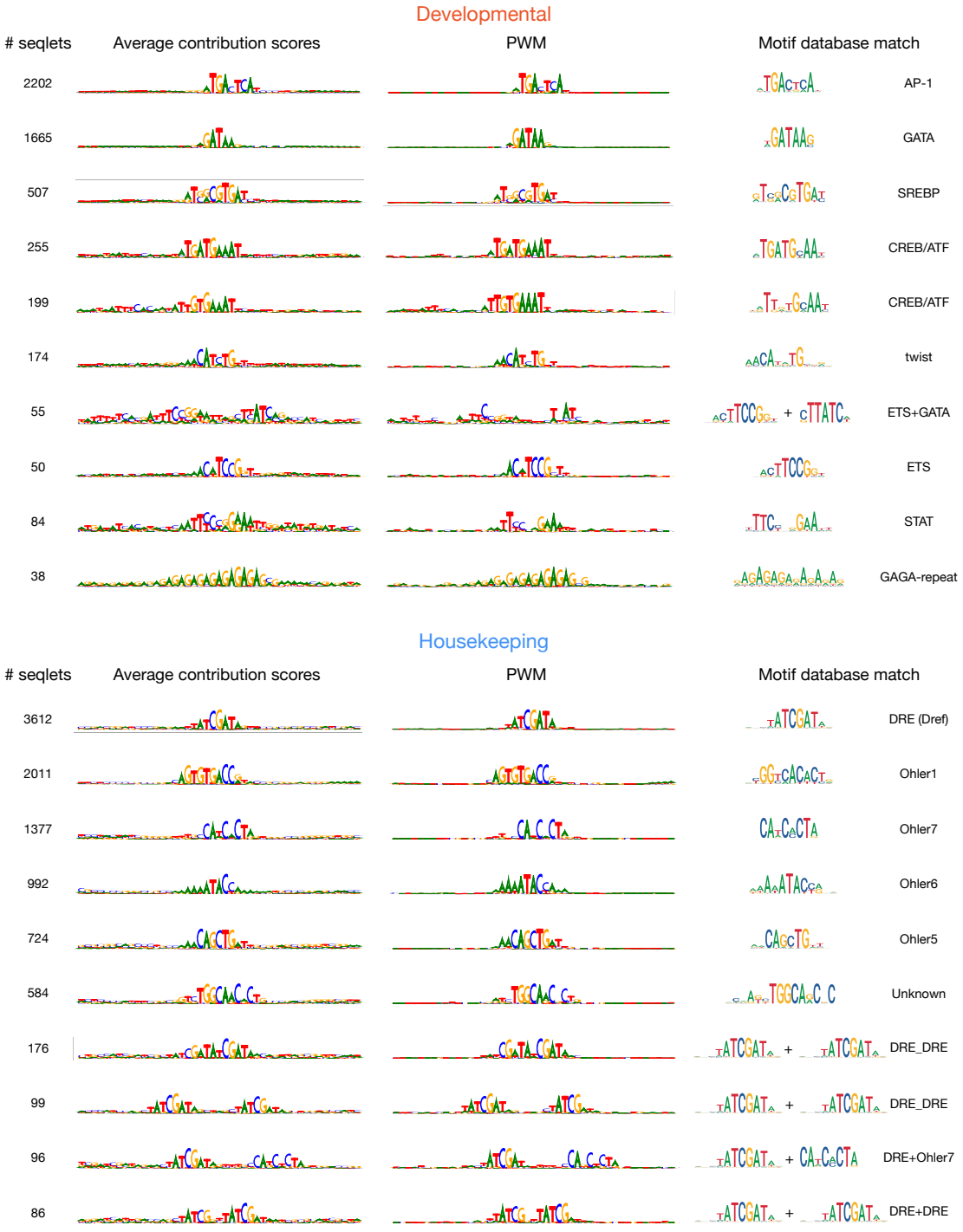
Supplementary Figure 5. DeepSTARR predicts important nucleotides in enhancers as measured by scanning mutagenesis experiments.



A) Scanning mutagenesis was performed by scrambling the nucleotides within 10 bp windows with 5 bp steps. The activity of the different variants was measured experimentally by UMI-STARR-seq and predicted by DeepSTARR. **B)** Scatter plots of predicted vs. observed log2 fold-change (log2 FC) enhancer activity (color scale) for each variant of four developmental and one housekeeping enhancer. The PCC is denoted for each comparison. **C)** Observed (UMI-STARR-seq; top) and predicted (middle) log2 FC for each 10 bp scrambled windows, together with the DeepSTARR derived nucleotide contribution scores (bottom) for the developmental enhancer chr3R:14743772-14744020. CREB/ATF, GATA and Trl motif

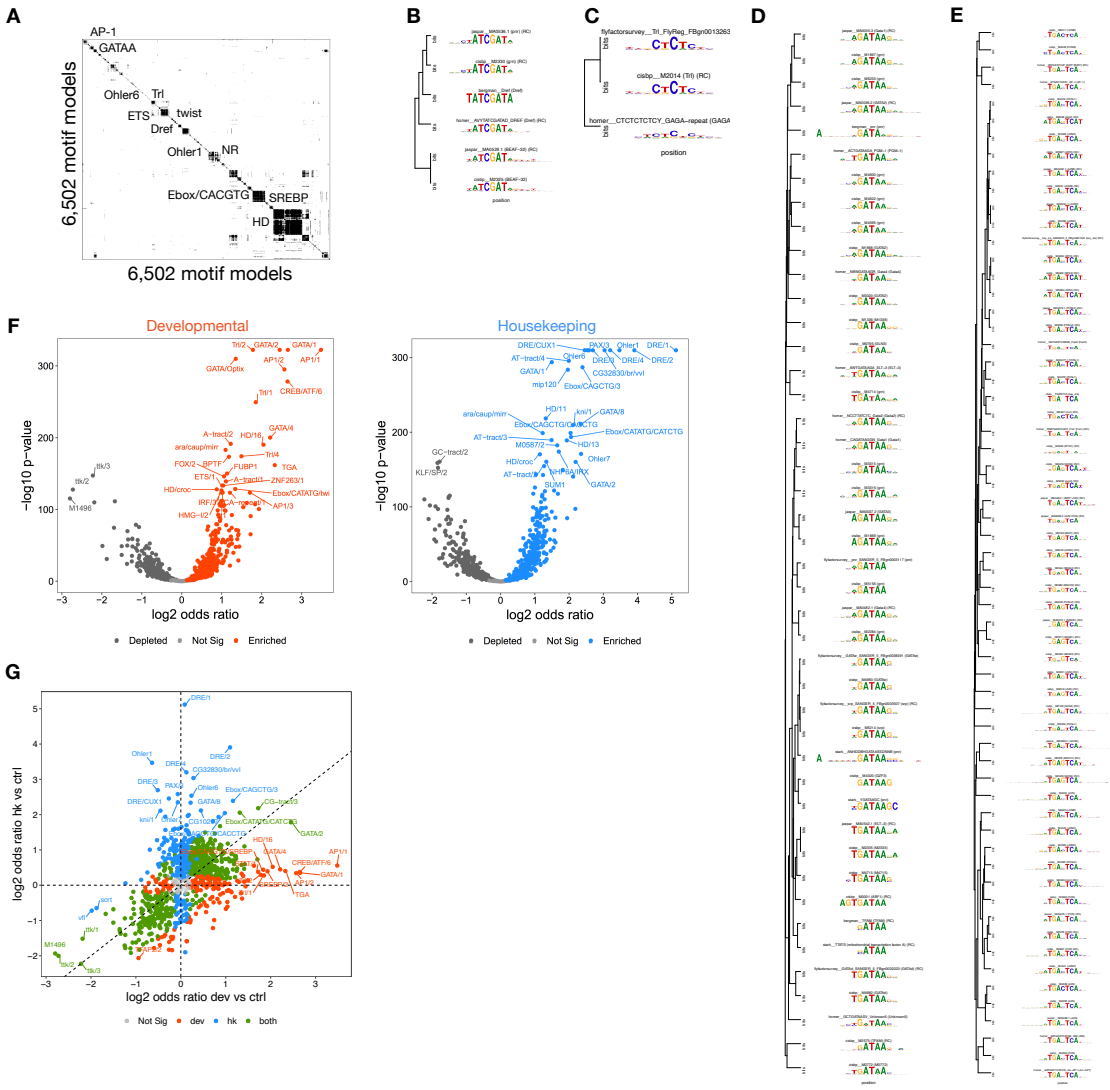
instances are highlighted. **D)** Bar-plots with the PCC between observed and predicted mutagenesis effects (log2 FC) for DeepSTARR and gkm-SVM models.

Supplementary Figure 6. Overview of motifs discovered by TF-Modisco.



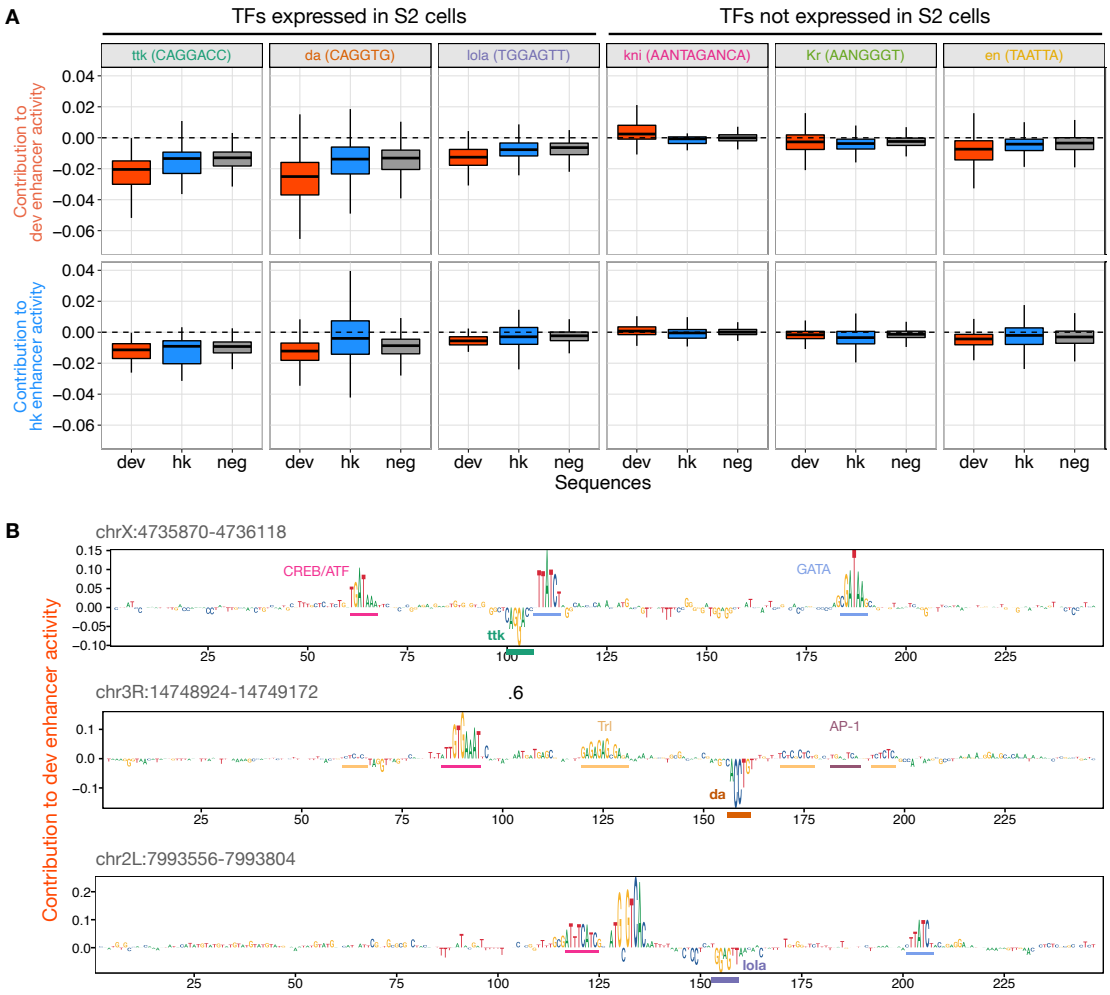
All discovered developmental (top) and housekeeping (bottom) motifs are shown from left to right with: number of seqlets supporting the motif, average contribution scores, converted PWM logo, their closest database match and respective motif name.

Supplementary Figure 7. Developmental and housekeeping enhancers are enriched in different TF motifs.



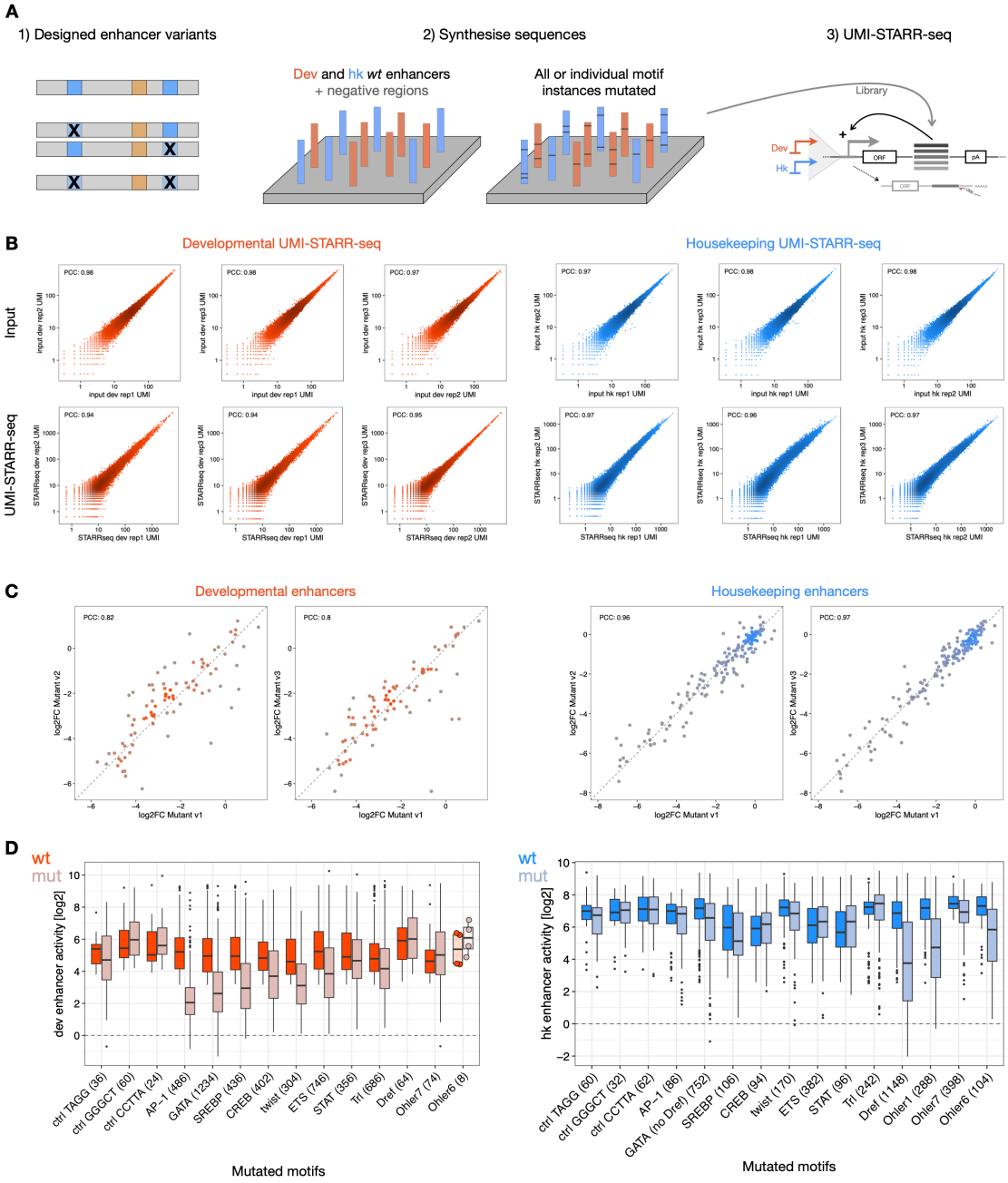
A) Hierarchically clustered heat map of the pairwise similarity scores between 6,502 TF motifs. The cluster dendrogram was cut at height 0.8, resulting in 901 non-redundant motif clusters that were manually annotated. **B-E)** Exemplar TF motif clusters. **F)** Enrichment of TF motifs in developmental (left) and housekeeping (right) enhancers over negative genomic regions. Log2 two-sided Fisher's exact test odds ratio compared with significance (-log10 p-value) for the most significant TF motif per motif cluster, to remove motif redundancy. Motifs significantly (FDR-corrected p-value <0.05) enriched or depleted are highlighted. **G)** Scatter plot comparing the motif enrichment (log2 odds ratio) in developmental and housekeeping enhancers. To remove motif redundancy, only the most significant TF motif per motif cluster was shown. Motifs significantly (FDR<0.05) enriched or depleted in each or both enhancer types are highlighted.

Supplementary Figure 8. DeepSTARR identifies candidate repressor motifs of expressed TFs.



A) Distributions of the DeepSTARR predicted developmental (top) and housekeeping (bottom) contribution scores of instances (average over all its nucleotides) of different repressive TF motif types across developmental enhancers (red), housekeeping enhancers (blue) and negative genomic regions (grey). Six motifs from TFs expressed (left: ttk, da, lola) or not expressed (right: kni, Kr, en) in S2 cells are shown, with their respective motif strings shown in parentheses. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). ttk, n = 36/24/724 independent instances per box; da, n = 362/461/3296; lola, n = 124/137/1169; kni, n = 98/48/253; Kr, n = 769/430/2903; en, n = 921/119/3760. **B)** DeepSTARR derived developmental nucleotide contribution scores for three enhancers with the activator and repressive TF motifs highlighted.

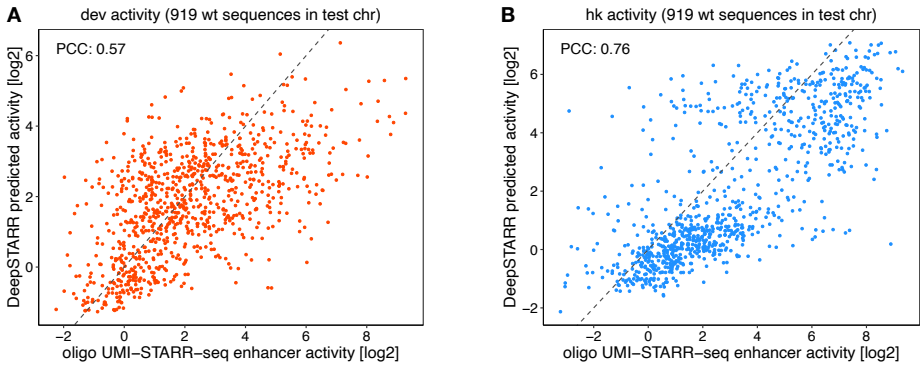
Supplementary Figure 9. Large-scale systematic TF motif mutagenesis.



A) Overview of the (1) design, (2) synthesis and (3) UMI-STARR-seq screen of the mutagenesis oligo library. UMI-STARR-seq was performed with a developmental (red) and a housekeeping (blue) promoter in *D. melanogaster* S2 cells. **B)** Pairwise comparisons of input (top) and UMI-STARR-seq (bottom) signal between three independent biological replicates across all oligos included in the library with a developmental (left) or housekeeping (right) promoter. Axes show counts per million in logarithmic scale. The PCC is denoted for each comparison. **C)** Motif requirements are independent of motif mutant variants. Pairwise comparisons of log2 fold-change (log2 FC) to wildtype activity between the three motif-mutant shuffled versions across developmental (left) and housekeeping (right) enhancers. The PCC is denoted for each comparison. **D)** Activity (log2) of wildtype and motif-mutant developmental (left) and housekeeping (right) enhancers that were used to derive the log2 fold-changes from Fig 2C. Number of enhancers mutated for each motif type are shown. The

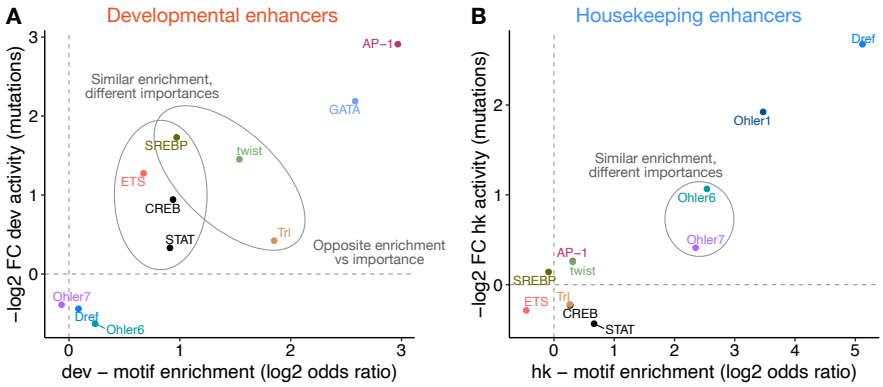
box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers); outliers are shown individually.

Supplementary Figure 10. DeepSTARR predicts enhancer activity of wildtype sequences in oligo UMI-STARR-seq.



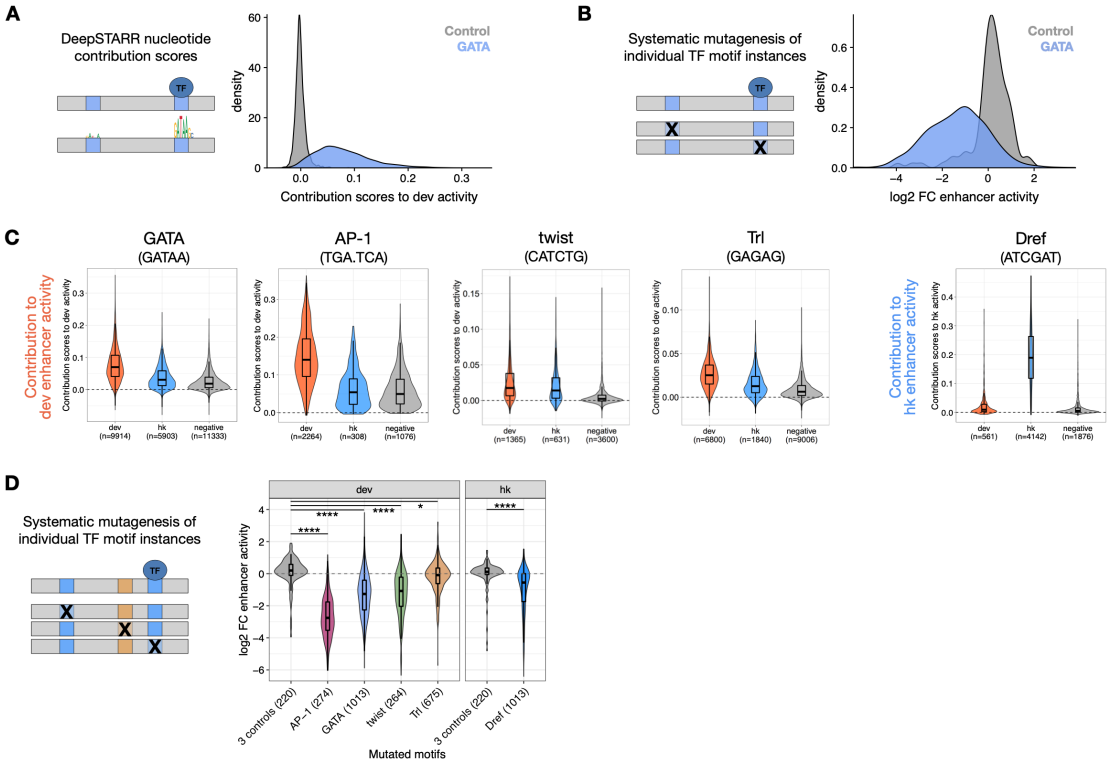
Scatter plots of predicted vs. observed developmental **(A)** and housekeeping **(B)** enhancer activity signal across wildtype sequences from the test set chromosome tested as individual oligos in oligo UMI-STARR-seq. The PCC is denoted for each comparison. Since DeepSTARR was trained and evaluated on UMI-STARR-seq data from the genome-wide screens using randomly sheared size-selected fragments (Fig S1), this result serves as baseline for the performance of DeepSTARR for the mutated oligos in oligo UMI-STARR-seq.

Supplementary Figure 11. Motif importance in native sequences compared with motif enrichment.



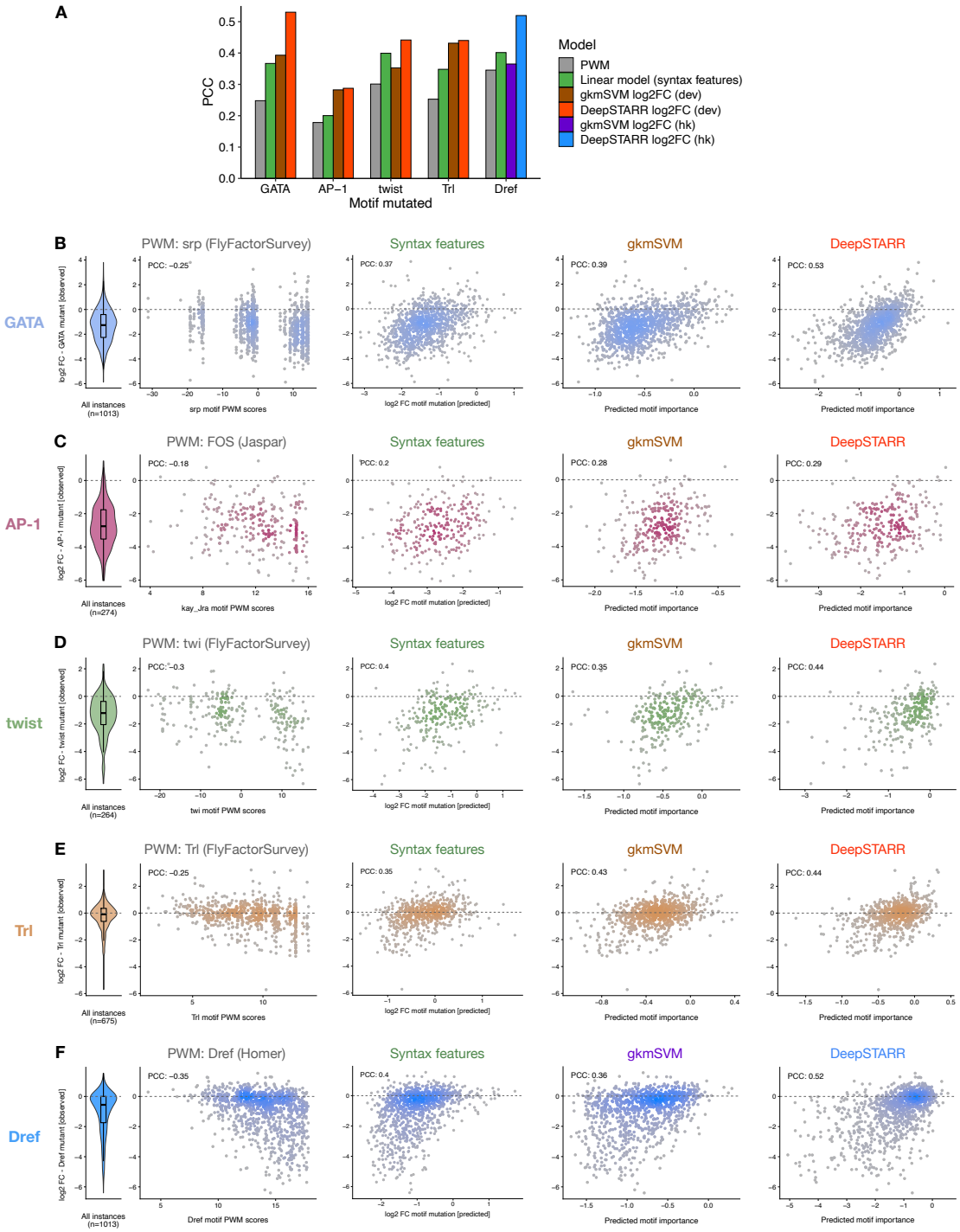
Scatter plots comparing motif enrichment (log2 odds ratio, Fig S7F; x-axis) with the results from experimental motif mutagenesis (median log2 fold-change values, Fig 2C; y-axis) in native developmental **(A)** and housekeeping **(B)** enhancers. Although motif enrichment is a good predictor of average motif importance, motifs similarly enriched can have very different (e.g. STAT, CREB, ETS and SREBP) or even opposite (SREBP vs Trl) importance values (see panel A).

Supplementary Figure 12. Instances of the same TF motif do not have equivalent contribution to enhancer activity.



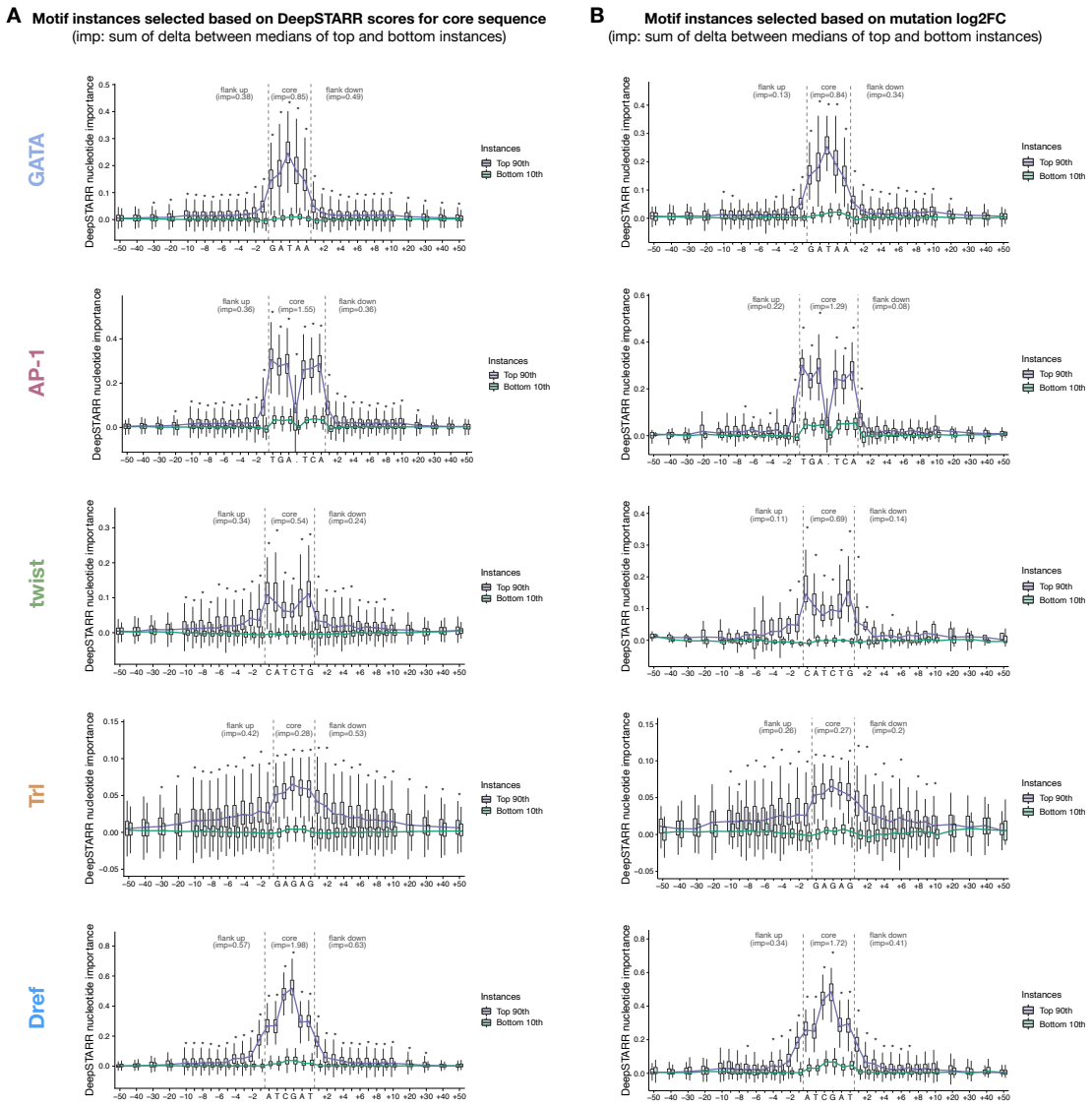
A) DeepSTARR predicts that instances of the same TF motif do not have equivalent contribution. Density distributions of the DeepSTARR predicted contribution scores (average over all its nucleotides) of GATA (blue) or GGGCT (as control; grey) instances in developmental enhancers. **B)** Systematic mutagenesis of individual TF motif instances validates motif non-equivalency. Density distributions of the experimentally derived (oligo UMI-STARR-seq) log2 FC in enhancer activity after mutation of GATA (blue) or control (grey) individual instances in developmental enhancers. **C)** DeepSTARR predicts that instances of the same TF motif are not equivalent. Distributions of the DeepSTARR predicted contribution scores (average over all its nucleotides) of instances of different TF motif types across developmental enhancers (red), housekeeping enhancers (blue) and negative genomic regions (grey). Number of instances for each motif type are shown. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **D)** Motif mutagenesis validates motif non-equivalency. Distributions of the experimentally derived (oligo UMI-STARR-seq) log2 FC in enhancer activity after mutation of individual instances of different TF motif types or control motifs in developmental or housekeeping enhancers. Note that the core sequence of different instances of the same motif type are identical, despite the different log2 FC. Number of instances for each motif type are shown. The two-sided Fligner-Killeen test of homogeneity of variances was used to compare the distributions of each TF motif type with the one from control motifs: AP-1, P = 1.2e-07; GATA, P = 2.8e-09; twis, P = 6.4e-09; Trl, P = 0.04; Dref, P = 1.4e-15. Box plots as in (C).

Supplementary Figure 13. Prediction of motif contribution by PWM scores, motif syntax features, gkm-SVM and DeepSTARR.



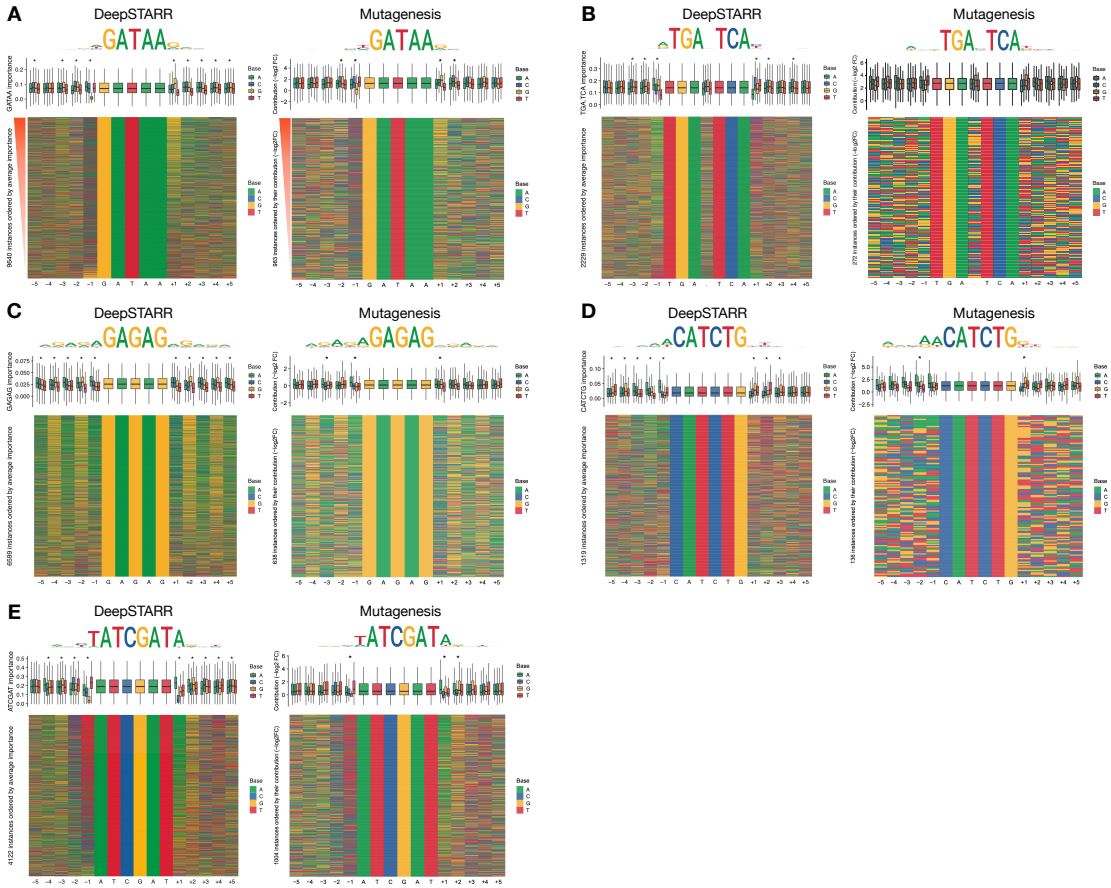
A) Bar-plots with the PCC between observed and predicted motif mutation effects (log2 FC) by PWM scores, a linear model with motif syntax features, and the gkm-SVM and DeepSTARR models. **B-F)** Distribution of experimentally measured fold-change (log2 FC) enhancer activity after mutating individual motif instances of the GATA **(B)**, AP-1 **(C)**, twist **(D)**, Trl **(E)** and Dref **(F)** motifs (violin plots), compared with the respective TF motif PWM scores and the log2 FC predicted by the models above. The PCC from 10-fold cross-validation is denoted for each comparison. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers).

Supplementary Figure 14. Flanking nucleotides of important motif instances contribute to enhancer activity.



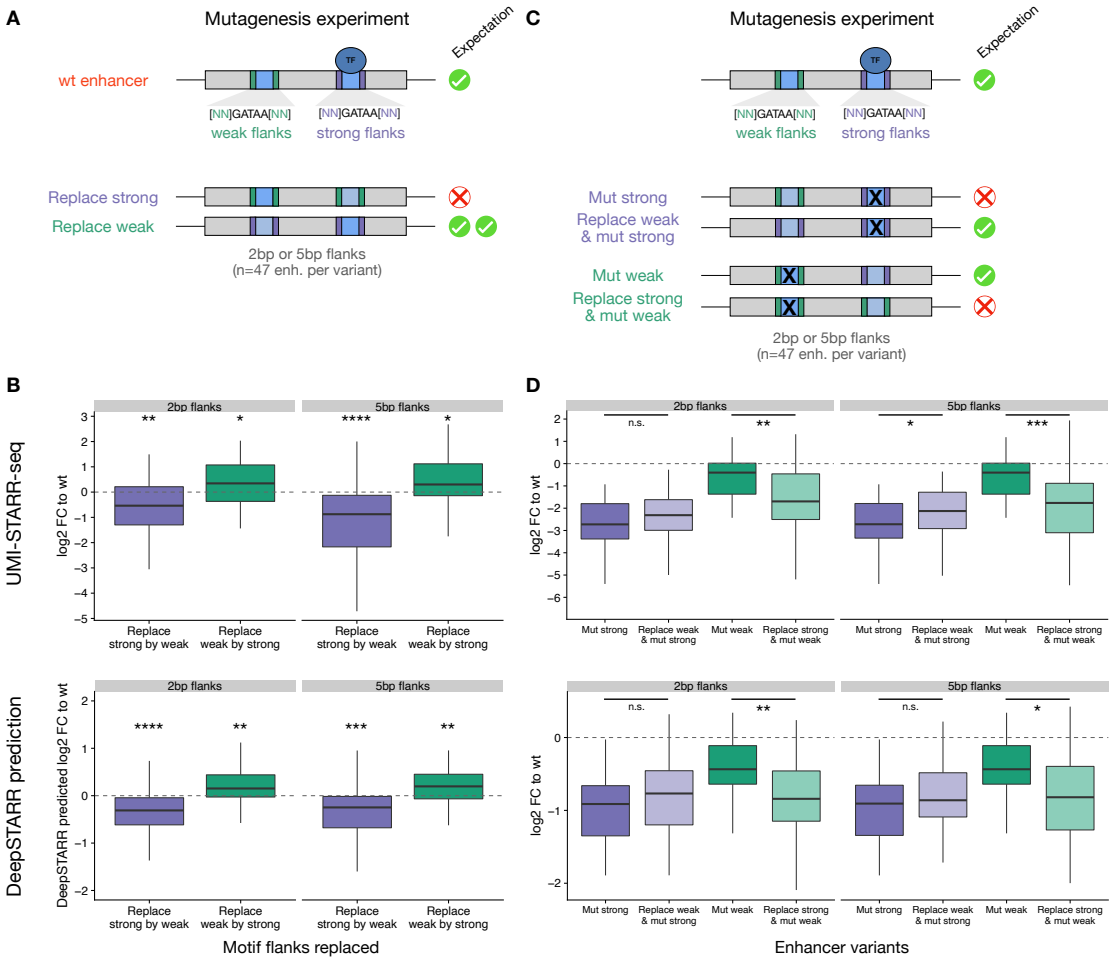
DeepSTARR predicted importance for +/- 50 flanking nucleotides of top 90th (purple) and bottom 10th (green) percentile motif instances selected based on DeepSTARR scores for core motif sequence **(A)** or its importance assessed by mutagenesis **(B)**. * marks positions with significant differences (two-sided Wilcoxon rank-sum test p-value < 0.001). **(A)** GATA, n = 992 independent instances per box; AP-1, n = 227; twist, n = 137; Trl, n = 680; Dref, n = 415. **(B)** GATA, n = 102; AP-1, n = 28; twist, n = 14; Trl, n = 70; Dref, n = 102. 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/downstream flanking sequences corresponds to the sum of delta between medians of top and bottom instances for the positions with significant differences.

Supplementary Figure 15. Contribution of TF motifs depend on their flanks.



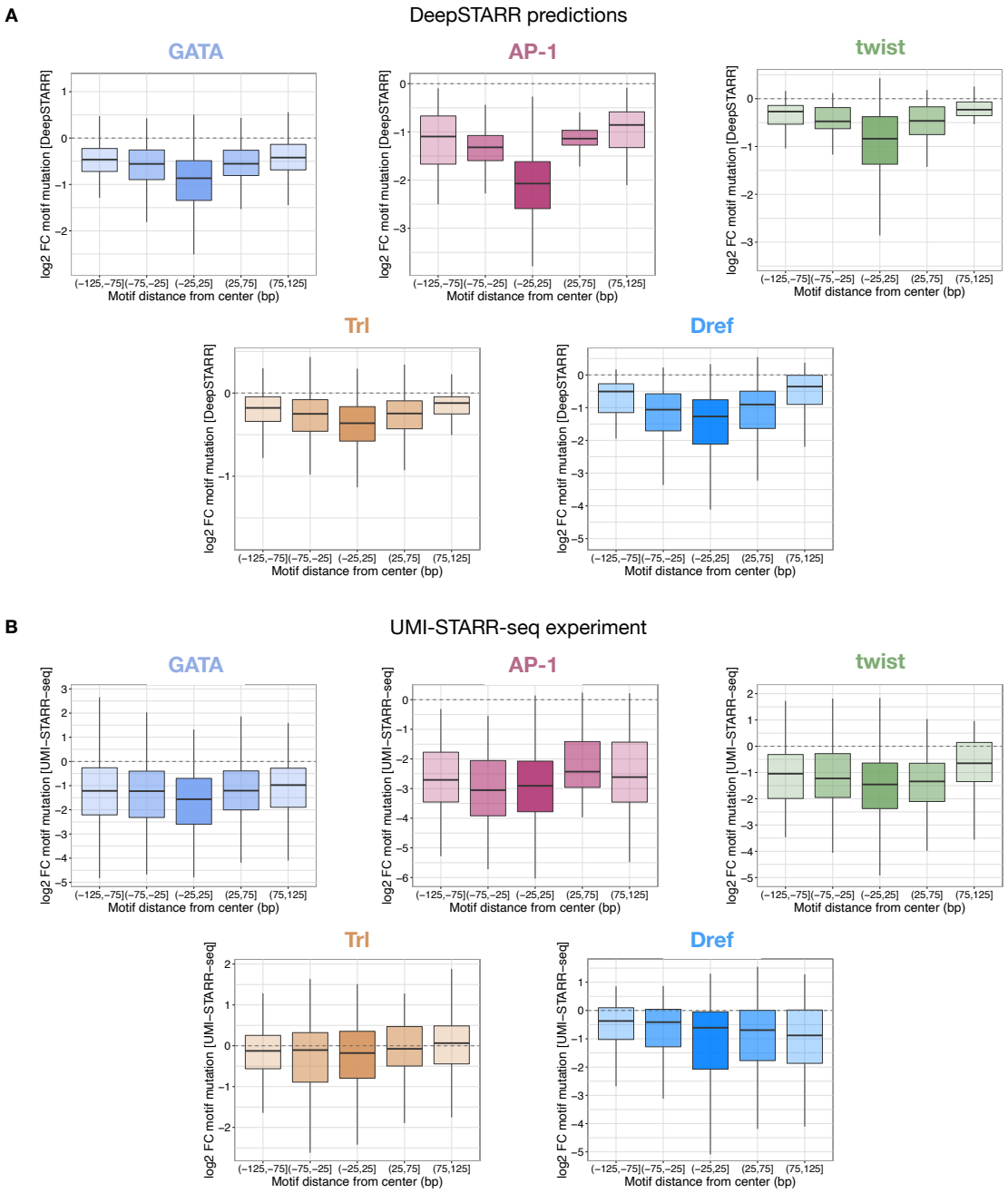
Motif contribution correlates with flanking base-pairs. Heatmaps: Flanking nucleotides of instances of different TF motif types across developmental (GATA: GATAA **(A)**, AP-1: TGA.TCA **(B)**, Trl: GAGAG **(C)**, twist: CATCTG **(D)**) or housekeeping (Dref: ATCGAT **(E)**) enhancers sorted by their DeepSTARR predicted contribution (left) or the experimentally derived (oligo UMI-STARR-seq) log2 fold-change in enhancer activity after mutation (right; minus log2 fold-change, -log2 FC). Box plots: Importance of motif instances according to the different bases at each flanking position. * marks positions with significant differences between the four nucleotides (FDR-corrected Welch One-Way ANOVA test p-value < 0.01). The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). Number of instances for each motif type are shown. Top: logos of the top 90th percentile motif instances for each sorting method.

Supplementary Figure 16. GATA flanking nucleotides are sufficient to switch motif contribution.



47 developmental enhancers containing both one strong (purple) and one weak (green) GATA instance (≥ 2-fold difference between instances) were selected. **B)** Observed (UMI-STARR-seq, top) and predicted (DeepSTARR, bottom) log2 FC enhancer activity to wildtype for sequences where the 2 or 5 bp flanks of strong instances were replaced by the ones of weak instances (purple) and vice versa (green) (cartoon in **(A)**). **D)** Observed (UMI-STARR-seq, top) and predicted (DeepSTARR, bottom) log2 FC enhancer activity to wildtype of mutating the strong instance (purple) compared to mutating this instance and additionally replacing the 2 or 5 bp flanks of the weak instance by the flanks of the strong instance (light purple) (cartoon in **(C)**). The same for the log2 FC of mutating the weak instance (green) compared to mutating this instance and additionally replacing the 2 or 5 bp flanks of the strong instance by the flanks of the weak instance (light green) (cartoon in **(C)**). **** p-value < 0.0001, *** < 0.001, ** < 0.01, * < 0.05, n.s. non-significant (two-sided Wilcoxon signed rank test). The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers).

Supplementary Figure 17. Motif importance in function of relative position in *Drosophila* enhancers.



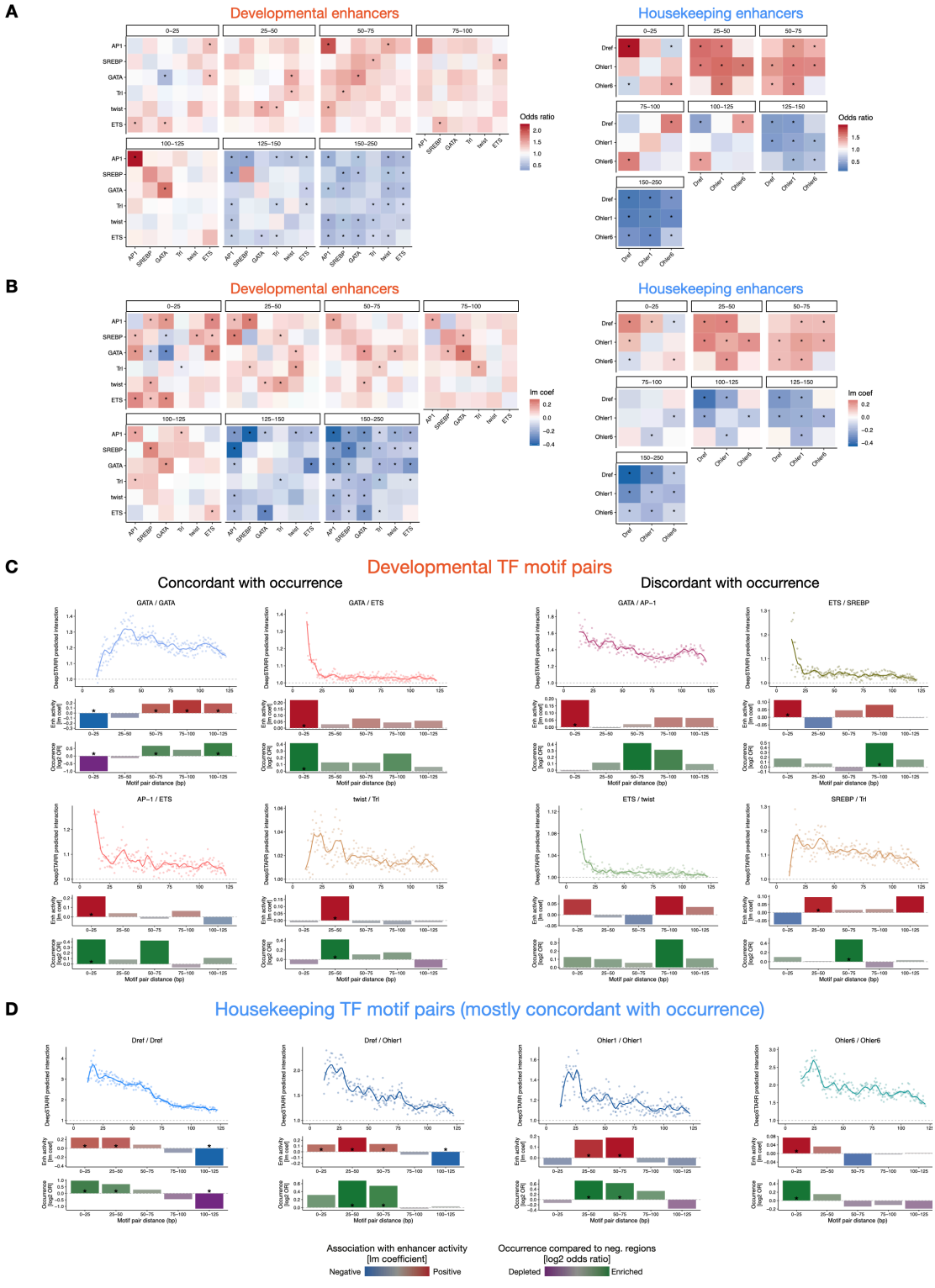
DeepSTARR predicted **(A)** and experimentally measured **(B)** log2 fold-change in enhancer activity after mutation of motifs at different positions relative to the enhancer center. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **(A, B)** GATA, n = 171/225/275/186/165 independent instances per box; AP-1, n = 35/57/108/42/36; twist, n = 41/49/74/60/43; Trl, n = 143/162/155/135/97; Dref, n = 75/211/509/178/45.

Supplementary Figure 18. Interpretation of DeepSTARR reveals TF motif distance preferences.

A) *In silico* characterization of TF motif distance preferences. *MotifA* was embedded in the center of 60 synthetic random DNA sequences and *MotifB* at a range of distances from *MotifA*, both up- and downstream. Both the average developmental and housekeeping enhancer activity is predicted by DeepSTARR. The cooperativity (residuals fold-change) between *MotifA* and *MotifB* as a function of distance is quantified as the activity of *MotifA+B* divided by the sum of the marginal effects of *MotifA* and *MotifB* (*MotifA* + *MotifB* – backbone (b)). **B)** Heatmaps showing the pairwise cooperativity (residuals) between different TF motif types in developmental (left) or housekeeping (right) enhancers. **C-D)** Cooperativity between motif

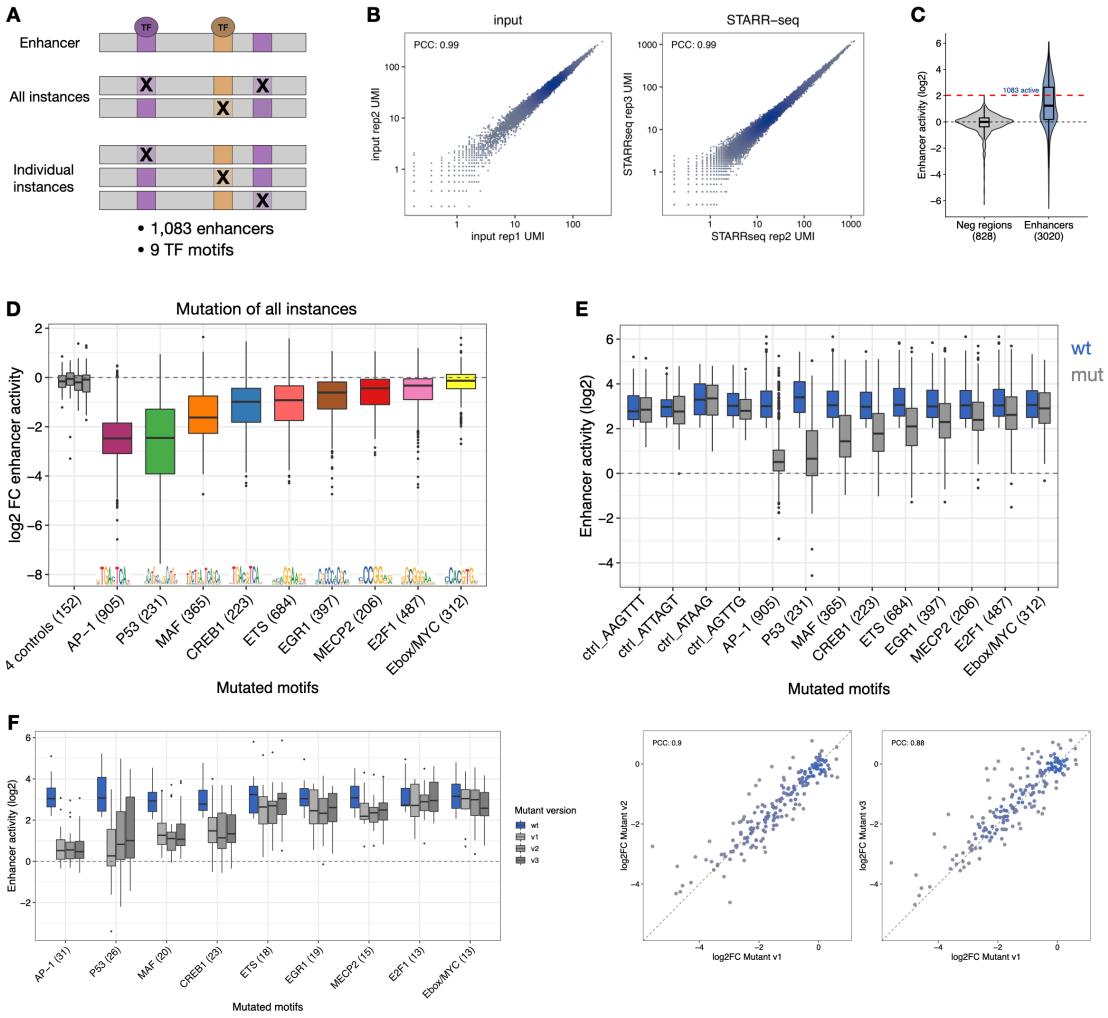
pairs at different distances in (C) developmental and (D) housekeeping enhancers. Points and smooth lines show the median cooperativity across all 60 backbones for each motif pair distance up- and downstream. The *MotifA* in the center is mentioned in each plot's title and tested with all *MotifB* motifs (different colours). GGGCT motif was used as control (grey). Dashed line at 1 represents no interaction.

Supplementary Figure 19. Motifs are not often at optimal distances in developmental enhancers, but enhancer activity follows optimal spacing rules.



A) Occurrence of motif pairs at different distances in genomic enhancers. Heatmaps showing the enrichment (Fisher's odds ratio) of motif pairs at different distance bins in developmental (left) or housekeeping (right) enhancers. * represents significant enrichment or depletions (two-sided Fisher's exact test FDR-corrected p-value < 0.05). **B)** Validation of optimal spacing rules for enhancer activity. Heatmaps showing the association between enhancer activity and the presence of motif pairs at different distance bins in developmental (left) or housekeeping (right) enhancers using a multiple linear regression. The multiple linear regression included, as independent variables, the number of instances for the different developmental or housekeeping TF motif types. Linear model coefficients are shown and * represents significant positive or negative associations (linear regression FDR-corrected p-value < 0.05). **C-D)** Top: Same as in Fig S8C,D (but with up- and downstream distances combined) per (C) developmental or (D) housekeeping motif pair. 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, as independent variables, the number of instances for the different developmental or housekeeping TF motif types. Bottom: Odds ratio (log2) by which the two motifs are found within a specified distance from each other in enhancers compared with negative genomic regions. Color legend is shown. Example motif pairs where optimal spacing preferences are concordant or discordant with their occurrence in enhancers are shown. * FDR-corrected two-sided Fisher's Exact test p-value < 0.05.

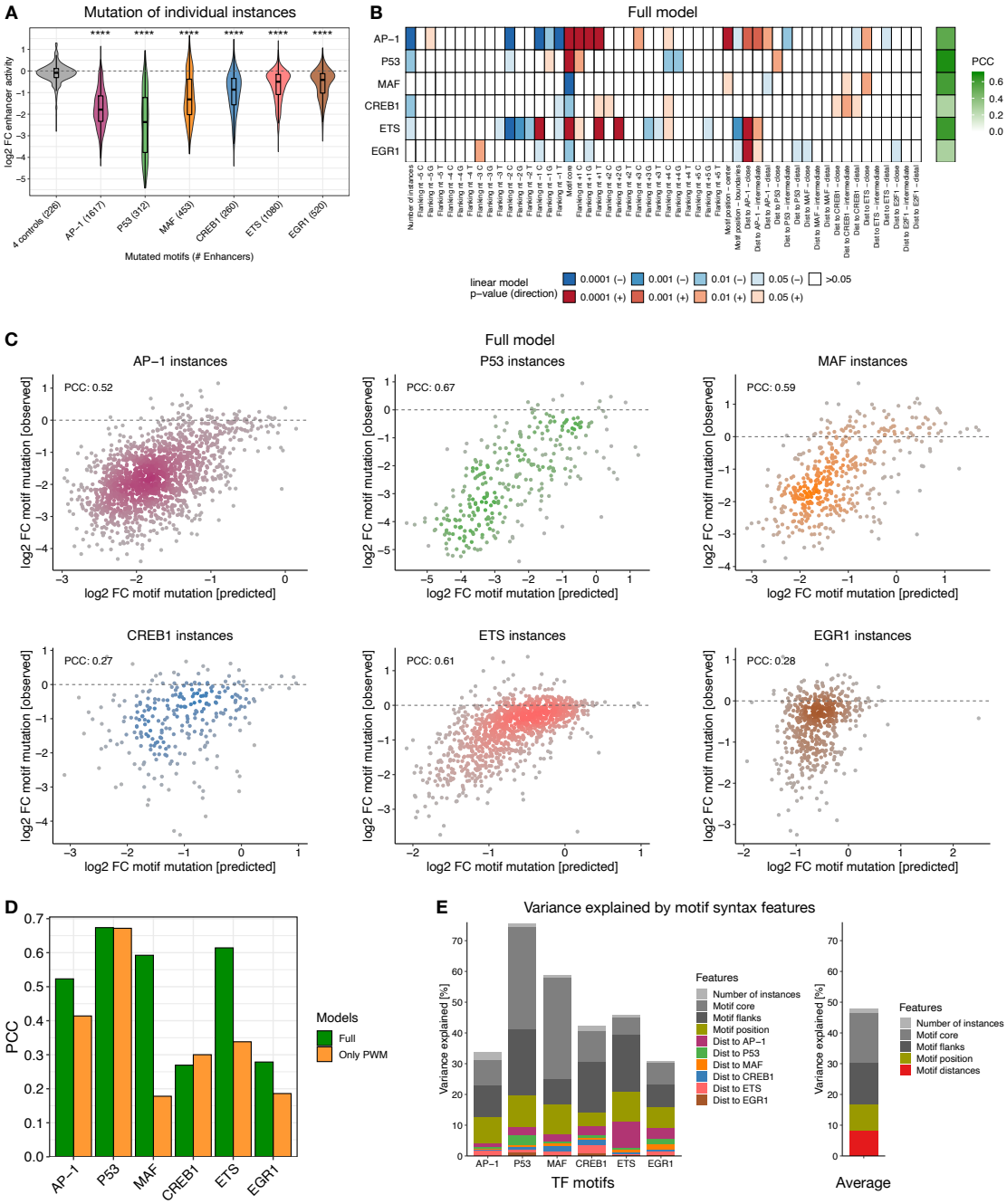
Supplementary Figure 20. Systematic TF motif mutagenesis in human HCT116 enhancers.



A) Systematic TF motif mutagenesis in human HCT116 enhancers. We selected 1,083 strong human enhancers and 9 TF motif types and mutated all instances of the same motif simultaneously or each instance individually. The activity of the wildtype and mutant sequences were measured through UMI-STARR-seq. **B)** Pairwise comparisons of input and STARR-seq signal between two independent biological replicates across all oligos included in the human oligo library. Axes are in logarithmic scale. The PCC is denoted for each comparison. **C)** Identification of 1,083 active short human enhancers. Distribution of log2 enhancer activity for oligos selected from negative regions (grey) or enhancer sequences (blue). 1,083 active short human enhancers (log2 wildtype activity in oligo UMI-STARR-seq >= 2.03, the strongest negative region, red dashed line; see Methods) were selected for subsequent motif mutation analyses. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **D)** TF motif requirements of human HCT116 enhancers. Log2 FC enhancer activity for hundreds of human enhancers after mutating all instances of four control (grey) and nine candidate human TF motifs. Number of enhancers mutated for each motif type and respective motif PWM logos are shown. Box plots as in (C); but outliers are shown individually. **E)** Activity (log2) of wildtype and motif-mutant enhancer sequences that were used to derived the log2 fold-changes from Fig S10D. Number of enhancers mutated is shown. Box plots as in (C); but outliers are shown individually. **F)** Motif requirements are independent of motif mutant variants. Left: Distribution of enhancer

activity for wildtype or motif-mutant enhancer sequences for the different TF motifs. The activity of sequences where the motifs were mutated to different motif shuffled versions is shown. Number of enhancers mutated for each motif type are shown. Box plots as in (C); but outliers are shown individually. Right: Pairwise comparisons of log2 FC to wildtype activity between the three motif-mutant shuffled versions across all enhancers. The PCC is denoted for each comparison.

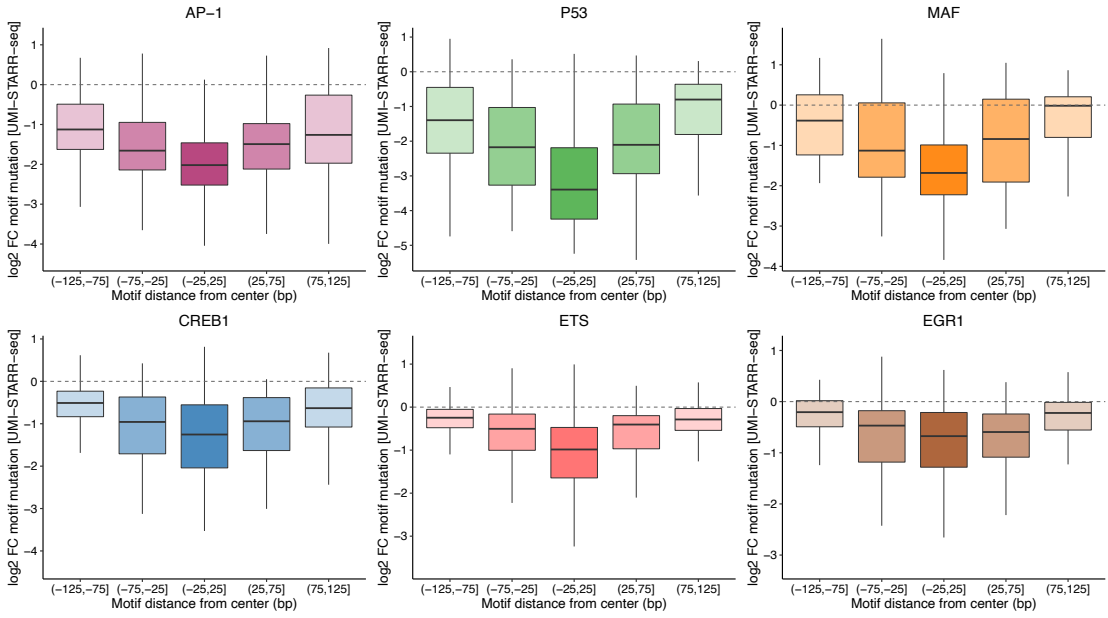
Supplementary Figure 21. Motif syntax rules dictate the contribution of motif instances.



A) TF motif non-equivalence is widespread in human enhancers. Distributions of the log2 FC in enhancer activity after mutation of individual instances of different TF motif types or control motifs. Number of instances for each motif type are shown. The two-sided Fligner-Killeen test of homogeneity of variances was used to compare the distributions of each TF motif type with the one from control motifs: AP-1, P = 7.5e-27; P53, P = 1.96e-41; MAF, P = 3.1e-33; CREB1, P = 7.8e-21; ETS, P = 7.8e-15; AP-1, P = 1.1e-10. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). **B)** Motif syntax rules dictate the contribution of TF motif instances in human enhancers. For each TF motif type (rows), we built a linear model containing the number of instances, 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, and the distance to all

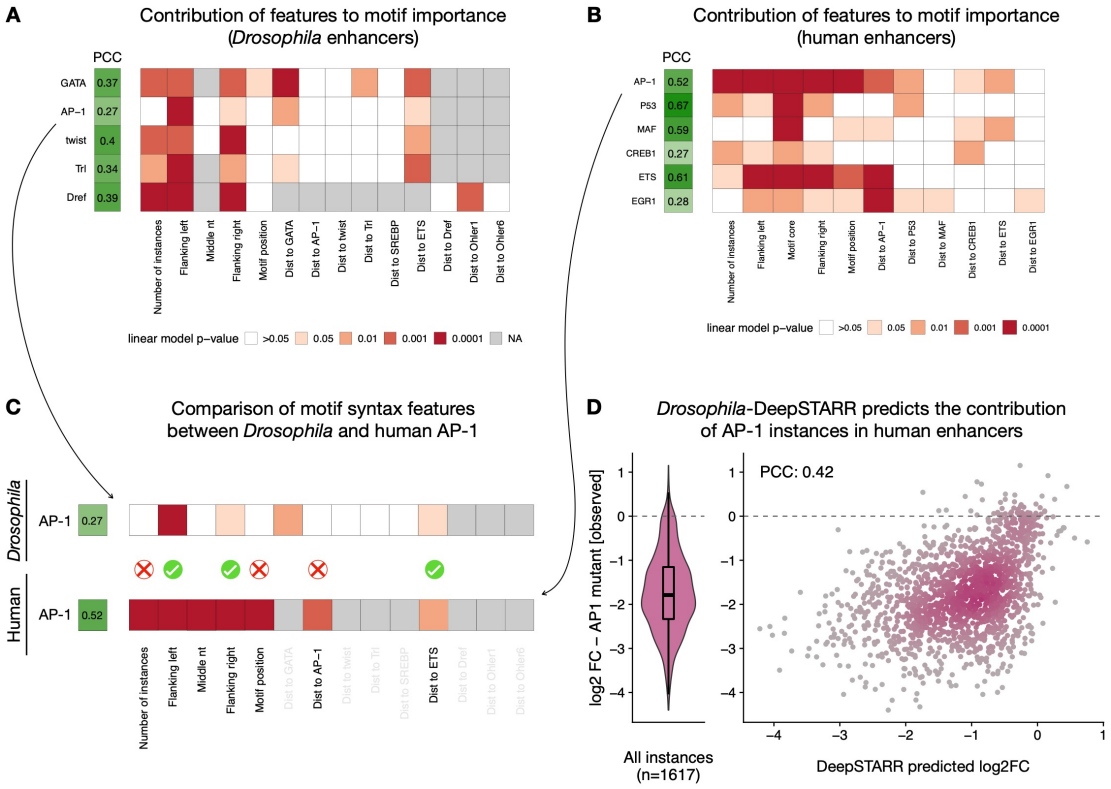
other TF motifs (close: < 25 bp; intermediate: ≥ 25 bp and ≤ 50 bp; distal: >50 bp) to predict the contribution of its individual instances (mutation log2 FC, from Fig S11A) across all enhancers. Heatmap shows the contribution of each feature (columns) for each model, colored by the direction (positive: red, negative: blue) and linear regression p-value. The PCC between predicted and observed motif contribution (using 10-fold cross-validation) is shown with the green color scale. **C)** Scatter plots comparing the measured contribution of individual instances of each TF motif type (log2 FC in enhancer activity after mutation) with the one predicted by the models from (B). The PCC is denoted for each comparison. **D)** Models taking into the motif syntax features predict better the contribution of motif instances than solely the PWM scores. Bar-plots comparing the PCC from the full models (from (B); green) and the same just using existing PWM scores (orange), assessed using 10-fold cross-validation. **E)** Variance explained by each motif syntax feature in the linear models built for each TF motif. Average across TF models is shown on the right.

Supplementary Figure 22. Motif importance in function of relative position in human enhancers.



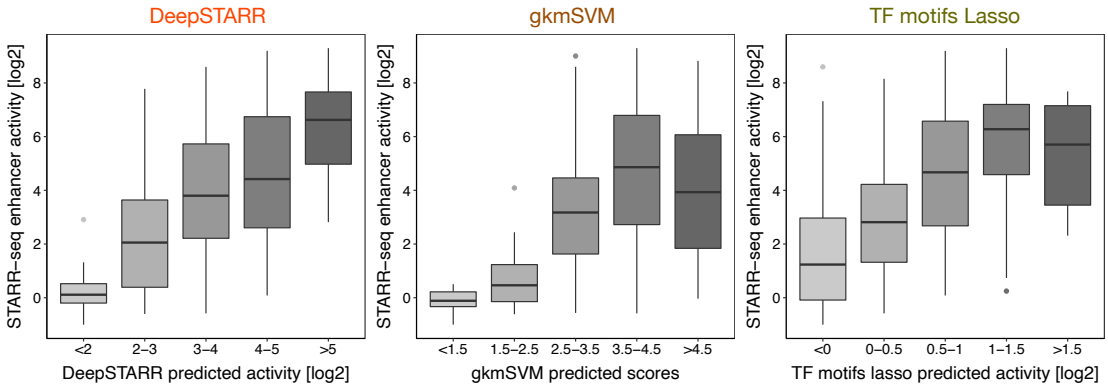
Experimentally measured log2 fold-change in enhancer activity after mutation of motifs at different positions relative to the enhancer center. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers). AP-1, n = 127/204/928/230/109 independent instances per box; P53, n = 34/47/138/54/29; MAF, n = 44/45/257/61/44; CREB1, n = 40/60/70/55/32; ETS, n = 146/202/342/206/173; EGR1, n = 88/128/103/124/70.

Supplementary Figure 23. Comparison of motif syntax features between *Drosophila* and human AP-1.



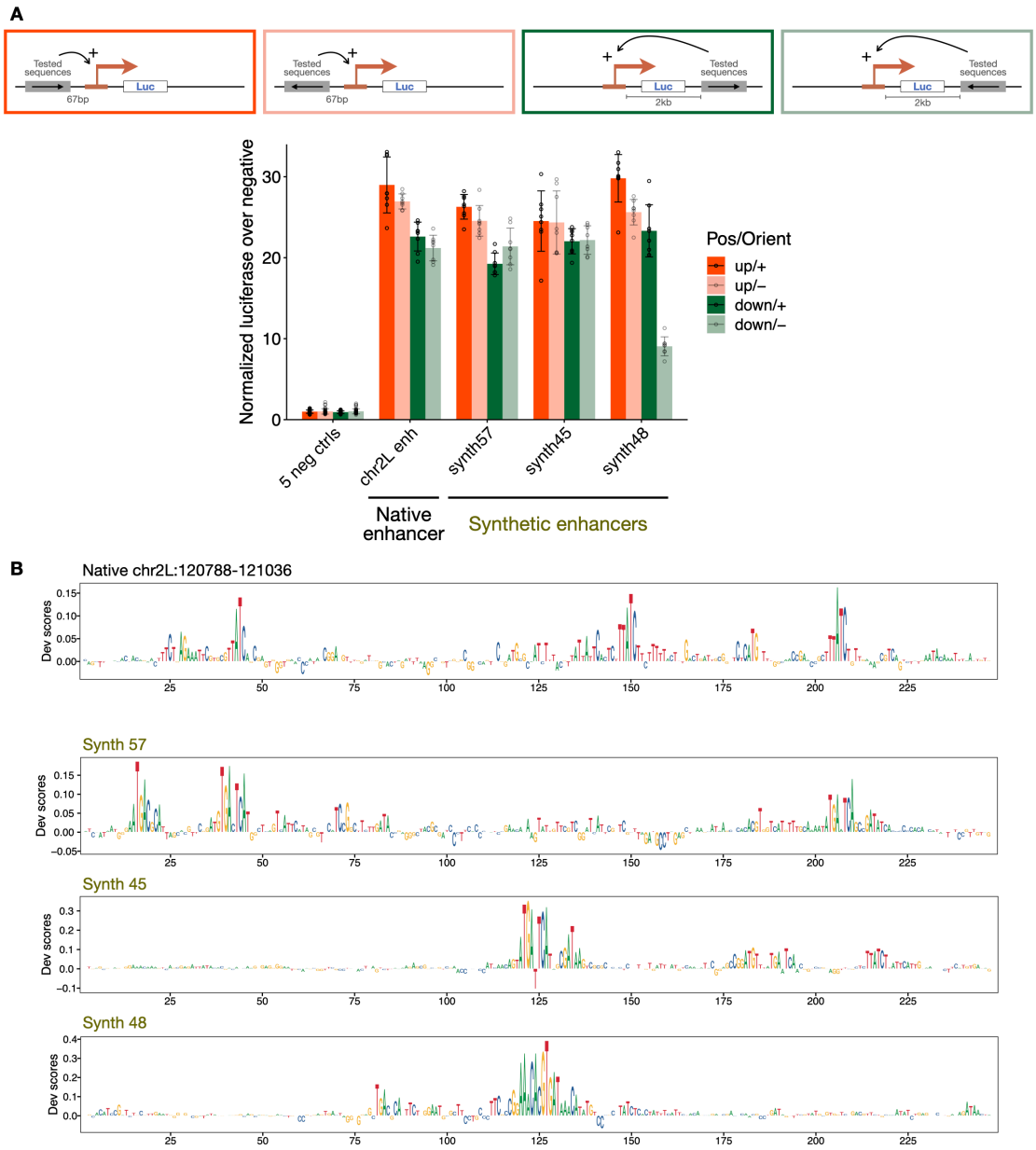
A-B) For each *Drosophila* **(A)** and human (**(B)**; see also Fig 6E) TF motif type (rows), we built a linear model containing the number of instances, the motif core and flanking nucleotides, the motif position relative to the enhancer center, and the distance to all other TF motifs to predict the contribution of its individual instances (mutation log2 fold-change) across all enhancers. The PCC between predicted and observed motif contribution (using 10-fold cross-validation) is shown with the green color scale on the left. Heatmap shows the contribution of each feature (columns) for each model, colored by the linear regression p-value (red scale). Grey denotes features not included in the respective models. **C)** Comparison of motif syntax features between *Drosophila* and human AP-1 models. Grey denotes features not shared between the respective models. **D)** DeepSTARR predicts the contribution of AP-1 instances in human enhancers. Distribution of experimentally measured log2 fold-change (log2 FC) enhancer activity after mutating 1,617 different AP-1 instances across HCT116 enhancers (left), compared with the log2 FC predicted by DeepSTARR (right). The PCC is denoted. The box plot marks the median, upper and lower quartiles and 1.5× interquartile range (whiskers).

Supplementary Figure 24. Prediction of synthetic sequences' activity by different methods.



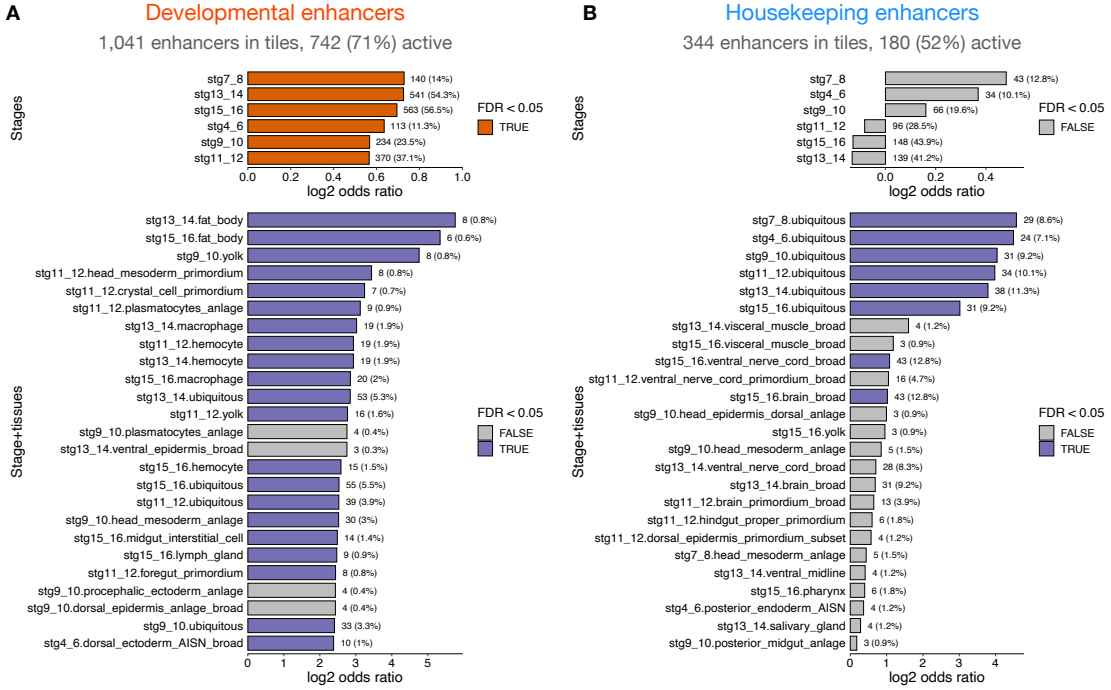
Comparison between experimentally measured enhancer activity (log2) for 249 synthetic sequences binned according to predicted activities by DeepSTARR, gkm-SVM and TF motifs Lasso models. The box plots mark the median, upper and lower quartiles and 1.5× interquartile range (whiskers)

Supplementary Figure 25. Synthetic enhancers function independent of their orientation and position.



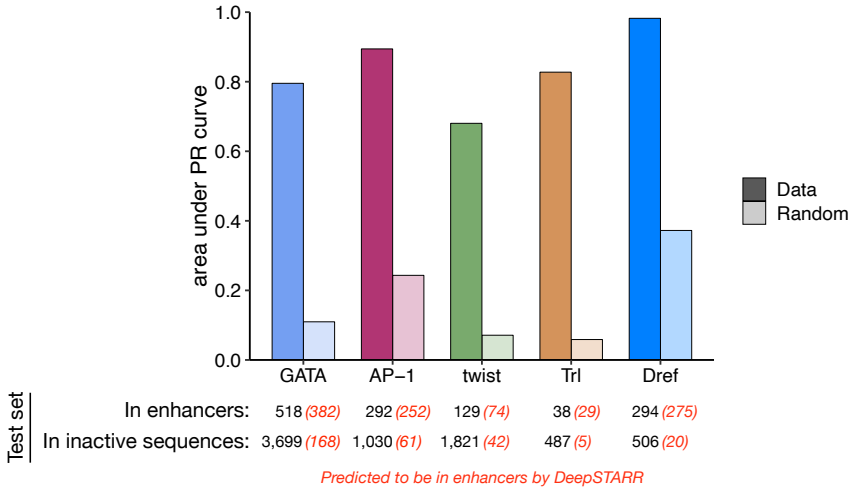
A) Candidate sequences were cloned in both orientations upstream (red) and downstream (green) of the minimal DSCP promoter and their enhancer activity assessed in luciferase assays. Bar plots show the average luciferase signal per sequence and construct across replicates (fold-change over the average signal of the five negative control sequences) +/- standard deviation. N=8 biologically independent samples. **B)** DeepSTARR derived developmental nucleotide contribution scores for the native and synthetic enhancers tested by luciferase.

Supplementary Figure 26. *In vivo* spatiotemporal activity of S2 enhancers.



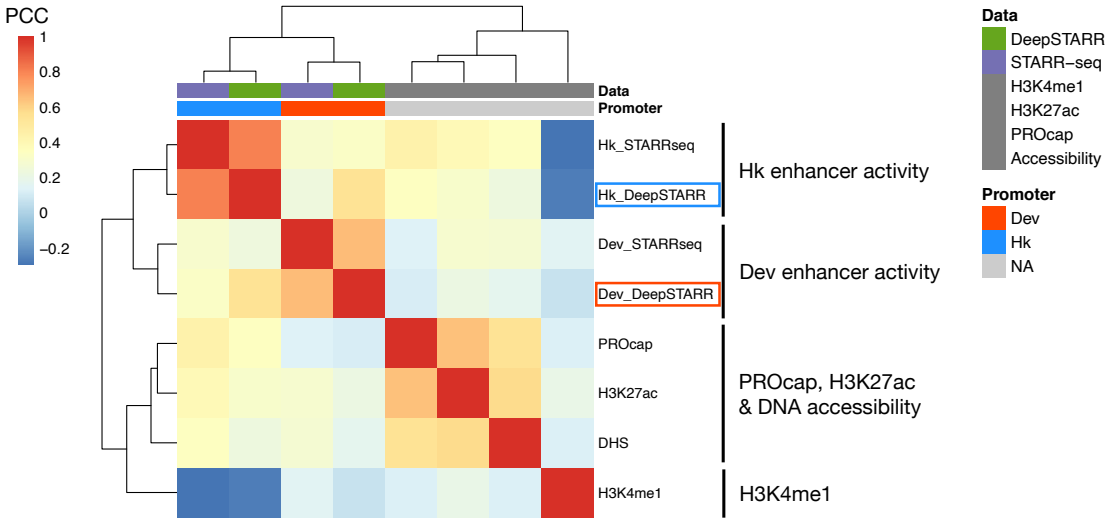
Bar plots show the enrichment (log2 Fisher's odds ratio) of the S2 developmental **(A)** and housekeeping **(B)** enhancer-overlapping tiles in different *Drosophila* embryonic stages (top) and tissues (bottom, only top 25 shown) (data from ref.[5]). For each bar, the number of active tiles and respective percentage of total enhancers is shown. Developmental enhancers **(A)** are enriched in embryonic enhancers active in hemocytes but also in other mesoderm-derived tissues, such as fat body, crystal cells, plasmatocytes and macrophages. In contrast, housekeeping enhancers **(B)** are enriched in regions ubiquitously active in the embryo.

Supplementary Figure 27. DeepSTARR discriminates motifs within enhancers from those outside enhancers among all instances selected to have favorable syntax context.



The *Drosophila* linear models based on motif syntax features (from Fig S13, S23A) were used to select instances of each TF motif in the test chromosome with a favorable syntax context: motif number, flanks, position and inter-motif distances (see Methods). This approach still overpredicts instances that are not in enhancers (numbers in black shown for each motif type). From these, DeepSTARR correctly predicts sites within enhancers: bar-plots with the area under precision-recall (PR) curve compared with the expected by a random model; number of instances predicted by DeepSTARR to be in enhancers in red.

Supplementary Figure 28. Comparison of DeepSTARR and STARR-seq with native chromatin and enhancer features.



Hierarchical clustering of DeepSTARR, STARR-seq and native chromatin and enhancer features on the basis of Pearson's correlation (PCC) of normalized read coverage over a merged set of all (1 kb) peaks. Developmental and housekeeping DeepSTARR and STARR-seq are colored. DeepSTARR models STARR-seq very precisely and both correlate only moderately with DNA accessibility (DHS; PCC dev: 0.21 and hk: 0.26), H3K27ac (PCC dev: 0.25 and hk: 0.33), H3k4me1 (PCC dev: 0.11 and hk: -0.21) and nascent RNA expression (PRO-cap; PCC dev: 0.15 and hk: 0.29).

Supplementary Tables

Supplementary Table 1. Primers used for UMI-STARR-seq library cloning and luciferase assay.

Supplementary Table 2. Genome-wide and oligo UMI-STARR-seq mapping statistics.
Summary of total sequenced reads, mapped reads and unique fragments (after collapsing by UMIs) for two developmental and two housekeeping genome-wide UMI-STARR-seq screens in S2 cells, three developmental and three housekeeping oligo UMI-STARR-seq screens in S2 cells, and three oligo UMI-STARR-seq screens in human HCT-116 cells. Counts mapping to the dm3 genome or oligo libraries are reported. The lower mapping rate for the *Drosophila* oligo libraries is because the library contained other oligos not used in this work.

Supplementary Table 3. 11,658 developmental and 7,062 housekeeping Drosophila S2 enhancers.
P-value from hypergeometric test.

Supplementary Table 4. Motif enrichment of developmental and housekeeping Drosophila S2 enhancer sequences.
Nominal and FDR-corrected p-values from two-sided Fisher's exact test.

Supplementary Table 5. Library of *Drosophila* S2 enhancers, motif-mutant, and motif flank swapping sequences.
Table of all 23,959 *Drosophila melanogaster* enhancer sequences and their motif-mutant sequences included in the oligo library, with genomic coordinates, oligo sequence, experiment, mutated motif, read counts for each screen and final developmental and housekeeping enhancer activity (log2).

Supplementary Table 6. Mutation of all motif instances in *Drosophila* S2 enhancers.
Table with all oligos used in the analysis of motif requirements (oligos with all motif instances mutated) with their DNA sequence, enhancer type, motif type mutated, motif mutant version, activity of mutant and wildtype sequences and respective log2 fold-change.

Supplementary Table 7. Comparison of DeepSTARR predicted motif importance and motif enrichment.
Data used in Fig. 2D.

Supplementary Table 8. Mutation of individual motif instances in *Drosophila* S2 enhancers.
Table with all oligos used in the analysis of mutations of individual motif instances with their DNA sequence, enhancer type, motif type mutated, experimentally measured activity of mutant and wildtype oligos and respective log2 fold-change, coordinates of

motif instance in enhancer oligo, sequence of wildtype and mutant motif instance, PWM scores and DeepSTARR developmental and housekeeping predicted log2 fold-change.

Supplementary Table 9. DeepSTARR-predicted contribution of activator motif instances in *Drosophila* S2 enhancers.
Data used in Fig S12A,C, S14 and S15. Motif instances mapped by string-matching.

Supplementary Table 10. PWM models used for the selected *Drosophila* TF motifs.
PWM logos shown in Fig 4C and used in Fig 3D, 5B, S13, S19.

Supplementary Table 11. Swapping of GATA motif flanks.
Data used in Fig 4D and S16.

Supplementary Table 12. PWM models used for the selected human TF motifs.
TF motif PWM models were retrieved from Vierstra et al., 2020. PWM logos are shown in Fig S20D.

Supplementary Table 13. Library of human HCT-116 enhancers and motif-mutant sequences.
Table of all 22,900 human enhancer sequences and their motif-mutant sequences included in the oligo library, with genomic coordinates, oligo sequence, experiment, mutated motif, read counts for each screen and final enhancer activity (log2).

Supplementary Table 14. Mutation of all motif instances in human HCT-116 enhancers.
Table with all oligos used in the analysis of motif requirements (oligos with all motif instances mutated) with their DNA sequence, motif type mutated, motif mutant version, activity of mutant and wildtype sequences and respective log2 fold-change. Data used in Fig S20D-F.

Supplementary Table 15. Mutation of individual motif instances in human HCT-116 enhancers.
Table with all oligos used in the analysis of mutations of individual motif instances with their DNA sequence, motif type mutated, experimentally measured activity of mutant and wildtype oligos and respective log2 fold-change, coordinates of motif instance in enhancer oligo, sequence of wildtype and mutant motif instance, and PWM scores. Data used in Fig 6 and S21.

Supplementary Table 16. DeepSTARR predicted contribution of AP-1 instances in human HCT-116 enhancers.
Data used in Fig S23.

Supplementary Table 17. Experimentally measured and DeepSTARR predicted activity of 249 synthetic enhancers in *Drosophila* S2 cells.
Data used in Fig 7.

Supplementary Table 18. Luciferase assay sequences and results.

Sequences selected for validation in luciferase assay with raw and normalized luciferase signals (eight replicates each divided in two different plates).

Supplementary Methods

UMI-STARR-seq

Cell culture

Drosophila S2 cells

Schneider 2 cells were grown in Schneider's Drosophila Medium (Gibco; 21720-024) supplemented with 10% heat inactivated FBS (Sigma; F7524) at 27ºC. Cells were passaged every 2-3 days.

Human HCT116 cells

Human HCT116 cells were cultured in DMEM (Gibco; 52100-047) supplemented with 10% heat inactivated FBS (Sigma; F7524) and 2mM L-Glutamine (Sigma; G7513) at 37ºC in a 5% $C0_2$-enriched atmosphere. Cells were passaged every 2-3 days.

Electroporation

The MaxCyte-STX system was used for all electroporations. S2 cells were electroporated at a density of 50 x 10^7 cells per 100µL and 5µg of DNA using the "Optimization 1" protocol. HCT116 cells were electroporated at a density of 1 x 10^7 cells per 100µL and 20µg of DNA using the preset "HCT116" program.

UMI-STARR-seq experiments

Library cloning

Drosophila genome-wide libraries were generated by shearing genomic DNA from the sequenced *D.mel* strain (y; cn bw sp) to an average of 200 bp fragments, decided to match the length of the oligonucleotide libraries (below) and increase the resolution of the enhancer sequences. Inserts were cloned into the standard *Drosophila* STARR-seq vector[1] containing either the DSCP or Rps12 core-promoters, and libraries grown in 6l of LB-Amp. *Drosophila* and human oligo libraries were synthesized by Twist Bioscience including 249 bp enhancer sequence and adaptors for library cloning. Fragments from the *Drosophila* library were amplified (primers see Supplementary Table 1) and cloned into *Drosophila* STARR-seq vectors containing either the DSCP or Rps12 core-promoters using Gibson cloning (New England BioLabs; E2611S). The oligo library for human STARR-seq screens was amplified (primers see **Supplementary Table 1**) and cloned into the human STARR-seq plasmid with the ORI in place of the core promoter[2]. Libraries were grown in 2l LB-Amp.

All libraries were purified with Qiagen Plasmid *Plus* Giga Kit (cat. no. 12991).

Drosophila S2 cells

UMI-STARR-seq was performed as described previously[1,3]. In brief, the screening libraries were generated from genomic DNA isolated of the sequenced *D.mel* strain (y; cn bw sp) or synthesized as oligo pools by Twist Bioscience (see above). We transfected 400 × 10^6 S2 cells total per replicate with 20 µg of the input library using the MaxCyte electroporation system. After 24 hr incubation, poly-A RNA was isolated and processed as described before[3]. Briefly: after reverse transcription and second strand synthesis a unique molecular identifier (UMI) was added to each transcript, allowing the counting of individual RNA molecules. This is followed by two nested PCR steps, each with primers that are specific to the reporter transcripts such that STARR-seq does not detect endogenous cellular RNAs.

Human HCT116 cells

STARR-seq was performed as described previously[1–3]. Screening libraries were generated from synthesized oligo pools by Twist Bioscience (see above). We transfected 80 × 10^6 HCT116 cells total per replicate with 160 µg of the input library using the MaxCyte electroporation system. After 6 hr incubation, poly-A RNA was isolated and further processed as described before[3].

Illumina sequencing

Next-generation sequencing was performed at the VBCF NGS facility on an Illumina HiSeq 2500, NextSeq 550 or NovaSeq SP platform, following manufacturer's protocol. Genome-wide UMI-STARR-seq screens were sequenced as paired-end 36 cycle runs (except the developmental input library, as paired-end 50 cycle runs) and Twist-oligo library screens were sequenced as paired-end 150 cycle runs, using standard Illumina i5 idexes as well as unique molecular identifiers (UMIs) at the i7 index. Deep sequencing base-calling was performed with CASAVA (v.1.9.1).

Genome-wide UMI-STARR-seq data analysis

Paired-end genome-wide UMI-STARR-seq RNA and DNA input reads (36 bp; except the developmental input library that was 50 bp) were mapped to the *Drosophila* genome (dm3), excluding chromosomes U, Uextra, and the mitochondrial genome, using Bowtie v.1.2.2[4]. 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**). We further computationally selected both RNA and input fragments of length 150-250 bp to only capture active sequences derived from short fragments. 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[1]. 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 (same length as used in oligo libraries) (**Supplementary Table 3**). Non-corrected enrichment over input was used as enhancer activity metric. Enhancers were classified as developmental or housekeeping based on the screen with the highest activity.

In vivo spatiotemporal activity of developmental and housekeeping enhancers

Drosophila ~2 kb genomic tiles tested for enhancer activity in different embryonic stages and tissues were retrieved from ref.[5] and overlapped with the S2 developmental and housekeeping enhancers (minimum overlap of 200 bp). 1,041 developmental and 244 housekeeping S2 enhancers were included in the tiles tested of which 742 (71%) and 180 (52%), respectively, were active in at least one stage/tissue. We assessed the enrichment of the S2 enhancer-overlapping tiles in each stage and tissue by two-sided Fisher's exact test (Fig S26). Obtained P-values were corrected for multiple testing by Benjamini-Hochberg procedure and considered significant if FDR ≤ 0.05.

Oligo library UMI-STARR-seq data analysis

Oligo library UMI-STARR-seq RNA and DNA input reads (paired-end 150 bp) 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[4]. For the *Drosophila* library we demultiplexed reads by the i5 and i7 indexes and oligo identity. Mapping reads with the correct length, strand and with no mismatches (to identify all sequence variants) were kept. Both DNA and RNA reads were collapsed by UMIs (10 bp) as above (**Supplementary Table 2**).

We excluded oligos with less than 10 reads in any of the input replicates and added one read pseudocount to oligos with zero RNA counts. The enhancer activity of each oligo in each screen was calculated as the log2 fold-change over input, using all replicates, with DESeq2[6]. We used the counts of wildtype negative regions in each library as scaling

factors between samples. This normalization only changes the position of the zero and consequently does not affect the calculation of log2 fold-changes between different sequences or the p-values for the statistical tests used.

Deep Learning

Data preparation

We selected all windows at the summit of developmental and housekeeping enhancers, in addition to three windows on either side of the regions (stride 100 bp). The remaining part of the genome was binned into 249 bp windows with a stride of 100 bp, excluding chromosomes U, Uextra, and the mitochondrial genome. We only included bins with more than five reads in the input and at least one read in the RNA of both developmental and housekeeping screens. To have a diversity of inactive sequences, we selected (1) 20,000 random bins overlapping accessible regions in different *Drosophila* cell types (S2, kc167 and OSC[1,7]) and embryogenesis stages[8], as well as all bins overlapping (2) enhancers from different *Drosophila* cell types (OSC and BG3[9]) and (3) inducible enhancers in S2 cells for two different stimuli (ecdysone[10] and Wnt signaling[11]). Lastly, we added 59,081 random windows with a range of enhancer activity levels. 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 post-augmentation). 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 multi-task 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[12] and built DeepSTARR with four 1D convolutional layers (filters=246,60,60,120; size=7,3,5,3), each followed by batch normalization, a ReLU non-linearity, 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 non-linearity, and dropout where the fraction is 0.4. The final layer mapped to both developmental and housekeeping outputs. Hyperparameters were manually adjusted to yield best performance on the validation set. The model was implemented and trained in Keras (v.2.2.4[13]) (with TensorFlow v.1.14.0[14]) using the Adam optimizer[15] (learning rate = 0.002), mean squared error (MSE) as loss function, a

batch size of 128, and early stopping with patience of ten epochs. Model training, hyperparameter tuning and performance evaluation were performed on different sets of genomic regions in distinct chromosomes.

We also explored how different architecture choices affect the model performance. We built different models changing the number of convolutional and fully connected layers, number of convolutional filters of the first layer, and the size of the convolutional filter of the first layer, and assessed their performance on predicting enhancer activity (validation set sequences) and motif importance (motif mutation fold-changes) (Fig S2). For each combination of parameters, we trained at least 10 different models.

Performance evaluation

The performance of the model was evaluated separately for developmental and housekeeping predictions on the held-out test sequences. We used the Pearson correlation coefficient (PCC) across all bins for a quantitative genome-wide evaluation and the area under the precision-recall curve (AUPRC; calculated using pr.*curve* from R package *PRROC* v.1.3.1[15]) for enhancer classification (enhancers vs. 2,685 negative control regions from the test set). We also report the model performance across bins from the test set not overlapping with repeats (from RepeatMasker dm3; Fig S3), or only the ones overlapping accessible elements[7] and active enhancers (Fig S4).

To test the robustness of the model, we trained 1,000 DeepSTARR models with the same set of hyperparameters and compared their performance. This accounted for the stochastic heterogeneity due to the random initialized weights in the neural network.

Prediction on full *Drosophila* genome

We extracted 249 bp sequences tiled across the *Drosophila* dm3 genome with a stride of 20 bp using ‘‘bedtools makewindows’’ (parameters -w 249 -s 20’) and ‘‘bedtools getfasta”[16]. We next predicted the developmental and housekeeping enhancer activity of each genomic window with DeepSTARR and averaged these per nucleotide to obtain genome-wide coverage. The DeepSTARR predicted coverage tracks are shown as examples in Fig 1B and S1A,B and are available at https://genome.ucsc.edu/s/bernardo.almeida/DeepSTARR_manuscript.

Models for comparison

The performance of DeepSTARR in the test set sequences was compared with two different methods: (1) a gapped k-mer support vector machine (gkm-SVM)[17] and (2) a lasso regression model based on TF motif counts (Fig S1D, S4).

(1) We used a 10-fold cross-validation scheme to train a developmental and a housekeeping gkm-SVM model to classify 249 bp DNA sequences into enhancers. Training was performed using developmental or housekeeping enhancers and a set of 21,463 negative control regions from the training set. The gkm-SVMs were done using LS-GKM[18] and the following parameters: (dev) *gkmtrain -t 0 -l 8 -k 5 -x 10*; (hk) *gkmtrain -t 0 -l 11 -k 7 -x 10*. We used the resulting support vectors of each trained model to score the DNA sequences of the test set by running *gkmpredict* and used these scores for the PCC and AUPRC analysis.
(2) We trained lasso regression models for developmental and housekeeping enhancer activity using the counts of 6,502 known TF motifs (see "Reference compendium of non-redundant TF motifs" below) as features across 40,000 random selected bins from the training set. Motif counts were calculated using the *matchMotifs* function from R package *motifmatchr* (v.1.4.0[19]) with the following parameters: *genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg="even"*. The model was trained using the optimal *lambda* retrieved from 10-fold cross-validation and the *glmnet* function from R package *glmnet* (v.2.0-16[20]).

Nucleotide contribution scores

We used DeepExplainer (the DeepSHAP implementation of DeepLIFT, see refs. [21–23]; 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 in 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, which were visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1[24]).

Motif discovery using TF–Modisco

To consolidate motifs, we ran TF–Modisco (v.0.5.12.0[25]) on the nucleotide contribution scores for each enhancer type separately using all developmental or housekeeping enhancers (Fig 2B). We specified the following parameters: *sliding*_window_size=15, *flank*_size=5, max_seqlets_per_*metacluster*=50000 and *TfModiscoSeqletsToPatternsFactory(trim_to_window_size=15, initial_flank_to_add=5)*. Motifs supported by less than 35 seqlets were discarded. The TF-Modisco discovered motifs are detailed in Fig S6, including the average contribution scores, converted PWM

logo and the closest match from the TF motif database (referenced below; similarity assessed using TOMTOM[26]). We trimmed the PWM motifs by removing flanking positions with an information content lower than 0.5 and show the final consolidated motifs in Fig 2B.

Comparison with native chromatin and enhancer features

We compared DeepSTARR predicted and experimental UMI-STARR-seq developmental and housekeeping enhancer activities with the endogenous DNA accessibility[1], nascent transcription[27] and H3K4me1 and H3K27ac chromatin marks[28]. We collected peaks from all datasets, extended to 1 kb, and computed the log average of the read coverage over the entire 1 kb. For each dataset, except nascent transcription, we normalized the signal over each respective input. We calculated pairwise PCCs between the different measures and performed hierarchical clustering ("complete" method) using the correlation values as similarities and the *pheatmap* R package (v.1.0.12[29]; Fig S28).

Reference compendium of non-redundant TF motifs

Reference compendium of non-redundant TF motifs

6,502 TF motif models were obtained from iRegulon (http://iregulon.aertslab.org/collections.html [30]) covering the following databases: Bergman (version 1.1[31]), CIS-BP (version 1.02[32]), FlyFactorSurvey (2010[33]), HOMER (2010[34]), JASPAR (version 5.0_ALPHA[35]), Stark (2007[36]) and iDMMPMM (2009[37]). We systematically collapsed redundant motifs by similarity by a previously described approach[38]. Specifically, we computed the distances between all motif pairs using TOMTOM[26] and performed hierarchical clustering using Pearson correlation as the distance metric and complete linkage using the *hclust* R function. The tree was cut at height 0.8, resulting in 901 non-redundant motif clusters that were manually annotated (Fig S7A-E). Clustering of motifs from each cluster and their logos were visualized using the *motifStack* R package (v.1.26.0[39]). The code and TF motif compendium are available from https://github.com/bernardo-de-almeida/motif-clustering.

TF motif enrichment analyses in developmental and housekeeping enhancers

We tested the enrichment of each TF motif in developmental or housekeeping enhancers (based on UMI-STARR-seq data, independent of their DeepSTARR predictions) over negative genomic regions (Fig S7F,G, **Supplementary Table 4**). Counts for each motif in each sequence were calculated using the *matchMotifs* function from R package

motifmatchr (v.1.4.0[19]) with the following parameters: *genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 1e-04, bg="genome"*. For each enhancer type, we assessed the differential distribution of each motif between the enhancers and negative regions by two-sided Fisher's exact test. Obtained P-values were corrected for multiple testing by Benjamini-Hochberg procedure and considered significant if FDR ≤ 0.05. To remove motif redundancy, only the most significant TF motif per motif cluster was shown.

TF motif mutagenesis in *Drosophila* S2 enhancers

Oligo library design

Selection of enhancer regions

A comprehensive library of 5,082 wildtype enhancer sequences in *D. melanogaster* S2 cells was compiled by selecting previously published developmental[1], housekeeping[40] and inducible (ecdysone[10] and Wnt signaling[11]) enhancers. 249 bp sequences centered on the enhancers' summit in both forward and reverse orientation were retrieved. We added 524 249-bp negative genomic regions in both orientations as controls (**Supplementary Table 5**).

Mapping of TF motif instances and generation of motif mutations

We selected eight predicted developmental motifs (GATA, AP-1, twist, Trl, SREBP, CREB, ETS, STAT), four predicted housekeeping motifs (Dref, Ohler1, Ohler6, Ohler7) and three control motifs (length-matched random motifs to control for enhancer-sequence perturbation). For each motif type, we mapped all instances using string-matching (shorter motifs – GATA: GATAA; AP-1: TGA.TCA; twist: CATCTG/CATATG; ETS: CCGGAA; Trl: GAGAG; Dref: ATCGAT; Ohler1: GTGTGACC; Ohler6: AAAATACCA; Ohler 7: CA.C.CTA; control: TAGG, GGGCT, CCTTA) or PWM-matching (longer motifs – SREBP, CREB, STAT, and also ETS, using TF-Modisco PWMs and the *matchMotifs* function from R package *motifmatchr* (v.1.4.0[19]) with the following parameters: *genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg="genome"*) in 2,375 enhancers (both motif orientations) and mutated all instances simultaneously to a motif shuffled variant (**Supplementary Table 5**; Fig S9A). In addition, for the GATA, AP-1, twist, Trl, and Dref motifs we mutated each instance individually to assess their importance. Here, we used string-matching not to identify motif instances per se but to specifically select instances with identical cores in order to assess their importance and important features outside the core. Each instance for a given motif was mutated always to the same shuffled

variant to allow the comparison of effects between instances of the same motif type. We designed motif-mutant sequences for each enhancer only for the orientation with the strongest wildtype enhancer activity. In addition, for GATA, AP-1, twist, Trl, and Dref motifs, we repeated mutations with two other different shuffled variants in 50 enhancers to control for the impact of the selected shuffled variant (**Supplementary Table 5**; Fig S9C).

Scanning mutagenesis of five enhancers

We selected four developmental and one housekeeping enhancer from above and scrambled the nucleotides within 10 bp windows with 5 bp steps, meaning 5 bp overlap between 10 bp windows, resulting in 49 mutant variants per enhancer (**Supplementary Table 5**; Fig S5). The effect of scrambling each window on the enhancer activity reveals the importance of the respective sequences.

Enhancers with swapped GATA motif flanks

We selected 100 developmental enhancers from above that contain 2 GATA instances (*inst1* and *inst2*) with different importance as predicted by DeepSTARR and swapped the flanking nucleotides (both 2 bp and 5 bp separately) between both instances (Fig 4D, S16). For each enhancer, we designed sequences where the flanks of *inst1* were replaced by the flanks of *inst2* and vice versa, resulting in sequences where both the two GATA instances contained either the flanks of *inst1* or the flanks of *inst2*. In addition, when replacing the flanks of *inst1* by the flanks of *inst2*, we also mutated *inst2* to assess how the flanks of *inst2* affected the contribution of *inst1*. The opposite was also done, with the flanks of *inst2* being replaced by the flanks of *inst1* together with mutation of *inst1*. Note that all selected instances had identical core sequences (i.e. all GATA instances had the identical core GATAA), and thus the variance in motif mutation effects can only be explained by the flanking sequence. The mutated sequences are listed in Supplementary Table 5. 47 active enhancers that contained one strong and one weak GATA instances (≥ 2-fold difference as assessed afterwards by mutagenesis) were used for the analyses in Fig 4D and S16 (**Supplementary Table 11**).

Designing of synthetic S2 developmental enhancers

1 billion random 249 bp DNA sequences were generated in *bash* with the following code: *cat /dev/urandom | tr -dc 'ACGT' | fold -w 249 | head -n 1000000000*. Bowtie v.1.2.2 [4] was used to remove sequences that exist in the *D. melanogaster* genome, which were none. The developmental enhancer activity of these sequences was predicted using DeepSTARR

and 249 sequences spanning different activity levels were selected for the oligo library (**Supplementary Table 5** and **17**).

Oligo library synthesis and UMI-STARR-seq

The *Drosophila* enhancers' motif mutagenesis oligo library contained wildtype (both orientations) and mutant enhancers, enhancers with swapped GATA motif flanks and synthetic enhancer sequences (**Supplementary Table 5**). All sequences were designed using the dm3 genome version. The enhancer sequences spanned 249 bp total, flanked by the Illumina i5 (25 bp; 5′ -TCCCTACACGACGCTCTTCCGATCT) and i7 (26 bp; 5′ AGATCGGAAGAGCACACGTCTGAACT) adaptor sequences upstream and downstream, respectively, serving as constant linkers for amplification and cloning. The resulting 21,758-plex 300-mer oligonucleotide library was synthesized by Twist Bioscience. UMI-STARR-seq using this oligo library was performed ("UMI-STARR-seq experiments") and analyzed ("Oligo library UMI-STARR-seq data analysis") as described above. We performed three independent replicates for developmental and housekeeping screens (correlation PCC=0.94-0.98; Fig S9B).

TF motif mutation analysis and equivalency

From the candidate 249 bp enhancers, we identified 855 active developmental and 905 active housekeeping *Drosophila* enhancers (log2 wildtype activity in oligo 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 log2 fold-change enhancer activity between the respective mutant and wildtype sequences (**Supplementary Table 6** and **8**). This was done separately for developmental and housekeeping enhancer activities.

Motif non-equivalency across all enhancers (Fig 3B, S12B,D) or within the same enhancer (Fig 3A,C) was assessed by comparing the impact of mutating individual instances of the same TF motif, i.e. the log2 fold-changes of each instance (**Supplementary Table 8**). For the comparison between instances in the same enhancer, only enhancers that require the TF motif (> 2-fold reduction in activity after mutating all instances) and contain two or more instances were used. Motif instances with >2-fold different contributions in the same enhancer were considered as non-equivalent. The same comparison across enhancers or within the same enhancer was performed for the three control motifs.

Motif syntax features

DeepSTARR predicted global importance of motif types and comparison with motif enrichment

To quantify the global importance of all known TF motifs to enhancer activity *in silico* (see ref. [41]), 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. The 249 bp backbone sequences were generated by sampling the base at each position with equal probability. The five different locations were the same for all motifs, centered at positions 25, 75, 125 (middle of the 249 bp oligo), 175 and 225. 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 log2 fold-change was averaged across all 100 backbones to derive the final global importance of each TF motif. Using random sequences allows to reduce the influence of background noise or other confounding signals that may exist in a given sequence when assessing the global importance of a TF motif for the model predictions[41]. The global motif importance predicted by DeepSTARR was compared with the enrichment of TF motifs at developmental and housekeeping enhancers, measured as the two-sided Fisher's exact test log2 odds ratio (described in "TF motif enrichment analyses in developmental and housekeeping enhancers") (Fig 2D, **Supplementary Table 7**). To remove motif redundancy, only the TF motif with the strongest predicted global importance or the strongest motif enrichment per motif cluster are shown in Fig 2D.

DeepSTARR predictions for the contribution of motif instances

We used two complementary approaches to measure the predicted contribution of each motif instance by DeepSTARR.

First, we measured the predicted importance of all string-matched instances of each TF motif type in 9,074 developmental enhancers, 6,369 housekeeping enhancers or 26,938 negative genomic regions (Fig S8A, S12A,C; **Supplementary Table 9**). The predicted importance of an instance was calculated as the average developmental or housekeeping DeepSTARR contribution scores over all its nucleotides. These scores represent the global contribution of motif instances captured by the model and were used for the analyses of figures: 4A-C, S8A, S12A,C, S14A, S15.

Second, to compare with the experimentally derived motif importance through motif mutagenesis, we used DeepSTARR to predict the log2 fold-change between wildtype and the motif-mutant enhancer sequences included in the oligo library for all instances of the different motif types (Fig 3B,D, S13, S17A). This was done by calculating the log2 fold-change between the predicted activity of the wildtype and respective motif-mutant sequences. Since the experimentally derived importance can be dependent on the shuffled mutant variant selected, this provides a more direct evaluation of the capability of DeepSTARR to predict the importance of a motif instance assessed by experimental mutagenesis.

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 (extended 10 nucleotides on each flank to account for the flanking sequence) with the PWM models of the selected TF motifs (Supplementary Table 10). We used the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; *genome = "BSgenome.Dmelanogaster.UCSC.dm3", bg="even"*[19]) with a p-value cutoff of 1 to retrieve the PWM scores of all sequences. These PWM scores were compared with the experimental log2 fold-changes using Pearson correlation (Fig 3D). We tested different PWM models for each TF motif if available and reported always the one with the best correlation (Supplementary Table 10).

Linear model with motif syntax rules to predict motif importance

For each TF motif type, we built a multiple linear regression model to predict the contribution of its individual instances (log2 fold-change) using as covariates the number of instances of the respective motif type in the enhancer, the motif core and flanking nucleotides (5 bp on each side), the motif position relative to the enhancer center[42] (center: -/+ 25 bp, flanks: -/+25:75 bp, boundaries: -/+75:125 bp; Fig S17B), and the distance to all other TF motifs (close: < 25 bp; intermediate: ≥ 25 bp and ≤ 50 bp; distal: >50 bp). Only motif instances that start after position 5 and end before position 245 of the 249 bp oligos were used, in order to be able to retrieve their 5 bp flanking sequences. In addition, for the motif distance analyses only non-overlapping motif pairs were used. All models were built using the *Caret* R package (v. 6.0-80[43]) and 10-fold cross-validation. Predictions for each held-out test sets were used to compare with the observed log2 fold-changes and assess model performance (Fig S13). The linear model coefficients and respective p-values were used as metrics of importance for each feature (Fig S23A,C).

To assess if these syntax rules are sufficient to identify functional instances, we used these linear models to select instances of each TF motif in the test chromosome with a predicted favorable syntax context: motif number, flanks, position and inter-motif distances. For each motif type, we identified all putative instances with the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; *genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg="even"*[19]), extracted all syntax features for each instance, predicted their importance (log2 fold-change), and selected instances with a predicted fold-change after mutation >= 2 as functional instances. We then overlapped these instances with developmental (for GATA, AP-1, twist, Trl) or housekeeping enhancers (Dref) and all genomic negative sequences. We tested if DeepSTARR could discriminate which instances are in enhancers by predicting the enhancer activity of each sequence and assessing its performance using the area under the precision-recall curve (calculated using *pr.curve* from R package *PRROC* v.1.3.1101[15]). To provide the numbers of predicted instances of each motif in enhancers or negative sequences we used as cut-off the predicted activity of 1.5 (Fig S27).

Predicted contribution of motif flanking nucleotides

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 log2 fold-change) importance. The DeepSTARR contribution scores of their +/- 50 flanking nucleotides were shown using box plots (Fig 4A, S14). For each position, significant differences between top and bottom instances were assessed through a Wilcoxon rank-sum test (p-value < 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 log2 fold-change) importance. Their 5 flanking nucleotides were shown using heatmaps and the importance of each nucleotide at each flanking position summarized using box plots (Fig 4B, S15). Significant differences between the four nucleotides per position were assessed through Welch One-Way ANOVA test followed by FDR multiple testing correction. The motif logos represent the frequency of each nucleotide at each position among the top 90th percentile instances and were compared with the logos of existing PWM models (Fig 4C).

The motifs recovered by DeepSTARR were compared with PWM models discovered *de novo* by HOMER (Fig 4C). HOMER (v4.10.4[34]) 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 (Fig 5A, S18). Backbone sequences were generated by sampling the base at each position with equal probability. 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 $2^{DeepSTARR\ prediction}$. 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. This analysis was performed for all motif pair combinations of AP-1, SREBP, GATA, Trl, twist and ETS motifs for developmental enhancer activity, and Dref, Ohler1 and Ohler6 for housekeeping enhancer activity in both strand orientations. Pairs with a negative control motif (GGGCT) were also included.

Enrichment of motif pairs at different distances in genomic enhancers

We mapped all instances of the different TF motif instances across all 9,074 developmental enhancers, 6,369 housekeeping enhancers and 26,938 negative genomic regions. We used their annotated PWM models (Supplementary Table 10) and the *matchMotifs* function from R package *motifmatchr* (v.1.4.0[19]) with the following parameters: *genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg="genome"*. Overlapping instances (minimum 50%) for the same TF motif were collapsed and counted only once. 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/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 log2 odds ratio used as metric. Obtained P-values were corrected for multiple testing by Benjamini-Hochberg procedure and considered significant if FDR ≤ 0.05. We performed this analysis separately for all

developmental motif pairs in developmental enhancers and all housekeeping motif pairs in housekeeping enhancers (Fig 5B, S19A,C,D).

Association between motif pair distances and enhancer activity

We obtained the positions of the different TF motif instances across all 9,074 developmental enhancers, 6,369 housekeeping enhancers and 26,938 negative genomic regions as described above ("Enrichment of motif pairs at different distances in genomic enhancers"). 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. We performed this analysis separately for all developmental motif pairs in developmental enhancers and all housekeeping motif pairs in housekeeping enhancers (Fig 5B, S19B-D).

Validation of motif distance preferences by motif mutagenesis

To test how the importance of GATA and AP-1 instances associate with the absolute distance *d* to a second GATA instance, we compared the log2 fold-change in enhancer activity after mutating individual GATA (Fig 5D) or AP-1 (Fig 5E) instances at close (< 25 bp; n=14 and 29, respectively) or longer (> 50 bp; n=129 and 38) distance to a second GATA instance. Only pairs of non-overlapping motif instances were used. A Wilcoxon rank-sum test was used to test this association.

TF motif mutagenesis in human HCT116 enhancers

TF motif enrichment

We characterized the motif composition of 5,891 strong STARR-seq enhancers in human HCT116 cells[2] using the 501 bp sequence centered on the summit. We generated 5,891 negative GC-matched genomic regions using the *genNullSeqs* function from R package *gkmSVM*[44]. 1,689 TF motif PWM models and respective motif clustering information were retrieved from Vierstra et al.,[38] covering the following databases: JASPAR (2018), Taipale HT-SELEX (2013) and HOCOMOCO (version 11). Counts for each motif in each 501 bp enhancer and negative sequence were calculated using the *matchMotifs* function from R package *motifmatchr* (v.1.4.0[19]) with the following parameters: *genome = "BSgenome.Hsapiens.UCSC.hg19", p.cutoff = 1e-04, bg="genome"*. We assessed the

differential distribution of each motif between the enhancers and negative regions by two-sided Fisher's exact test. We selected the nine TF motifs with the strongest enrichment in enhancers: AP-1, P53, MAF, CREB1, ETS, EGR1, MECP2, E2F1 and Ebox/MYC (Supplementary Table 12).

TF motif mutagenesis oligo library design and synthesis

Generation of TF motif mutations

For UMI-STARR-seq of wild type and mutant enhancers, we selected 3,200 enhancer candidates, defining short 249 bp windows (the limits of oligo 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[19]) 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 (length-matched random motifs to control for enhancer-sequence perturbation) using string-matching. We then designed enhancer variants with all instances of each motif type mutated simultaneously or individually to a motif shuffled variant (Supplementary Table 13; Fig S20A). Each instance for a given motif was mutated always to the same shuffled variant to allow the comparison of effects between motif instances. We designed motif-mutant sequences for each enhancer only for the orientation with the strongest activity in the genome-wide STARR-seq. In addition, for each motif type we repeated mutations with two other different shuffled variants in 50 enhancers to control for the impact of the selected shuffled variant (Supplementary Table 13; Fig S20F).

Oligo library synthesis and UMI-STARR-seq

The final human enhancers' motif mutagenesis library contained 3,200 wildtype and 18,780 motif-mutant enhancer sequences that we combined with 920 249-bp negative genomic regions as controls (Supplementary Table 13). All sequences were designed using the hg19 genome version. 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 oligo library was performed ("UMI-STARR-seq experiments") and analyzed ("Oligo library UMI-STARR-seq data analysis") as described above. We performed two independent replicates (correlation PCC=0.99; Fig S20B).

TF motif mutation analysis

From the 3,200 designed candidate 249 bp enhancers, we identified 1,083 active short human enhancers (log2 wildtype activity in oligo UMI-STARR-seq >= 2.03, the strongest negative region; Fig S10C) 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 log2 fold-change enhancer activity between the respective mutant and wildtype sequences (Fig S20D,E, S21A; Supplementary Table 14 and 15). Motif non-equivalency across all enhancers (Fig S21A) or within the same enhancer (Fig 6B,C) 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 [38]), as a surrogate for TF occupancy (Fig 6D). Footprints detected at different FPR adjusted p-value thresholds and coverage tracks with observed and expected cleavage counts were downloaded from https://resources.altius.org/~jvierstra/projects/footprinting.2020/per.dataset/h.RKO-DS40362/, in hg38 coordinates. All coordinates were converted to hg19 coordinates using the *UCSC liftOver* tool[45] and the *hg38ToHg19.over.chain* chain file (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz). For each TF motif type, a Wilcoxon rank-sum test was used to determine whether the mutation log2 fold-change 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 within HCT116-accessible enhancers were used in the analysis. Enhancers were defined as accessible if they overlap any of the DNase-seq peaks from the following ENCODE[46] identifiers (hg19 coordinates) (https://www.encodeproject.org/): ENCFF001SQU, ENCFF001WIJ, ENCFF001WIK, ENCFF175RBN, ENCFF228YKV, ENCFF851NWR, ENCFF927AHJ, ENCFF945KJN and ENCFF360XGA.

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 (log2 fold-changes) 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[42] (center: -/+ 25 bp, flanks: -/+25:75 bp, boundaries: -/+75:125 bp; Fig S22), and the distance to all other TF

motifs (close: < 25 bp; intermediate: ≥ 25 bp and ≤ 50 bp; distal: >50 bp) (Fig 6E, S21B-E). Only motif instances that start after position 5 and end before position 245 of the 249 bp oligos were used, in order to be able to retrieve their 5 bp flanking sequences. In addition, for the motif distance analyses only non-overlapping motif pairs were used. All models were built using the *Caret* R package (v. 6.0-80[43]) and 10-fold cross-validation. Predictions for each held-out test sets were used to compare with the observed log2 fold-changes and assess model performance. For each TF motif type, we compared the main regression model with a simple linear model only using the PWM scores as covariate (Fig S21D).

The linear model coefficients and respective p-values were used as metrics of importance for each feature (Fig 6E, S21B, S23B,C). In addition, we calculated the percentage of variance explained by each covariate (motif syntax features) in the linear models built for each TF motif with one-way ANOVAs. For each TF motif we generated 100 different models randomizing the order of the covariates (since the variance explained depends on the order of covariates entered), quantified the percentage of variance explained of each covariate as its sum of squares divided by the total sum of squares, and used the average value across all 100 models as the final variance explained per covariate.

DeepSTARR prediction of the importance of AP-1 instances in human enhancers

We used the DeepSTARR model trained in *Drosophila* S2 enhancers to predict the importance of AP-1 instances in human HCT116 enhancers. This was done by predicting the activity of the wildtype and motif-mutant enhancer sequences included in the human oligo library for all AP-1 instances and further calculating the log2 fold-change. This predicted log2 fold-change was compared with the experimentally measured log2 fold-change and its association assessed through Pearson correlation (Fig S23D; Supplementary Table 16).

Luciferase reporter assays

Luciferase reporter assays

We constructed luciferase reporters by cloning candidate enhancers in both orientations in the pGL3_DSCP_luc+ plasmid either upstream of the DSCP promoter in the KpnI site (to create pGL3_candidate_DSCP_luc+) or downstream of the DSCP promoter in the SalI site (to create pGL3_DSCP_luc+_candidate) (Fig S25A). Candidate enhancer sequences (one native and the three strongest synthetic enhancers) and five negative controls were

amplified from the Twist oligo pools and plasmids verified by sanger sequencing (see Supplementary Table 1 for primers).

Luciferase assays were performed in quadruplicates as described previously[47]. In short, *Drosophila* S2 cells were transfected using jetPEI (peqlab 13-101-40N) in 384-well plates with 30.000 cells per well. For each transfection we used 30ng of the pGL3 firefly reporter and 3ng of a *Renilla* luciferase expressing Ubi-RL plasmid as transfection control. After transfection cells were incubated for 48h at 27°C in Schneider2 Medium supplemented with 10% FBS and 1% penicillin/streptomycin.

Luciferase assay data analysis

We first normalized firefly over *Renilla* luciferase values for each of the eight biological replicates (independent transfections) individually (controlling for transfection efficiency). 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 standard deviation (Fig S25A; Supplementary Table 18).

Statistics and data visualization

All statistical calculations and graphical displays have been performed in R statistical computing environment (v.3.5.1[48]) and using the R package *ggplot2* (v.3.2.1[49]). Coverage data tracks have been visualized in the UCSC Genome Browser[50] 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.

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 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[51] (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 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 [30]). DNase-seq and ATAC-seq data in *Drosophila* S2 cells were obtained from ref.[1] and [7], respectively; nascent transcription from ref.[27] and H3K4me1 and H3K27ac chromatin marks from ref.[28]. 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[46] (https://www.encodeproject.org/; ENCFF001SQU, ENCFF001WIJ, ENCFF001WIK, ENCFF175RBN, ENCFF228YKV, ENCFF851NWR, ENCFF927AHJ, ENCFF945KJN, ENCFF360XGA, ENCFF130JBP and ENCFF400KKD) and ATAC-seq data from ref.[52].

Code availability

Code used to process the genome-wide and oligo 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

1. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (80-.).* **339**, 1074–1077 (2013).
2. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
3. 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).
4. 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).
5. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).
6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
7. Albig, C. *et al.* Factor cooperation for chromosome discrimination in Drosophila. *Nucleic Acids Res.* **47**, 1706–1724 (2019).
8. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.* **12**, R43 (2011).
9. 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**, 1147–56 (2014).
10. Shlyueva, D. *et al.* Hormone-Responsive Enhancer-Activity Maps Reveal Predictive Motifs, Indirect Repression, and Targeting of Closed Chromatin. *Mol. Cell* **54**, 180–192 (2014).
11. Franz, A., Shlyueva, D., Brunner, E., Stark, A. & Basler, K. Probing the canonicity of the Wnt/Wingless signaling pathway. *PLoS Genet.* **13**, 1–18 (2017).
12. 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).
13. Chollet, F. & others. Keras. https://keras.io. (2015).
14. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **1603.04467**, (2016).
15. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *arXiv* **1412.6980**, (2015).
16. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
17. 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).
18. Lee, D. LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
19. Schep, A. motifmatchr: Fast Motif Matching in R. R package version 1.14.0. (2021).

20. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

21. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *arXiv* **1704.02685**, (2017).

22. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *31st Conf. Neural Inf. Process. Syst.* (2017).

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

24. Omar Wagih. ggseqlogo: A 'ggplot2' Extension for Drawing Publication-Ready Sequence Logos. R package version 0.1. https://CRAN.R-project.org/package=ggseqlogo. (2017).

25. Shrikumar, A. *et al.* TF-MoDISco v0.4.4.2-alpha: Technical Note. *arXiv* **1811.00416**, (2018).

26. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

27. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science (80-.).* **339**, 950–953 (2013).

28. Rickels, R. *et al.* An Evolutionary Conserved Epigenetic Mark of Polycomb Response Elements Implemented by Trx/MLL/COMPASS. *Mol. Cell* **63**, 318–328 (2016).

29. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap. (2019).

30. 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).

31. Down, T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Comput. Biol.* **3**, 0095–0109 (2007).

32. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

33. Zhu, L. J. *et al.* FlyFactorSurvey: A database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39**, 111–117 (2011).

34. 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).

35. Mathelier, A. *et al.* JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).

36. Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).

37. Kulakovskiy, I. V. & Makeev, V. J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics (Oxf).* **54**, 667–674 (2009).

38. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

39. Ou, J., Wolfe, S. A., Brodsky, M. H. & Zhu, L. J. MotifStack for the analysis of transcription factor binding site evolution. *Nat. Methods* **15**, 8–9 (2018).

40. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).

41. 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).

42. Grossman, S. R. *et al.* Positional specificity of different transcription factor classes within enhancers. *Proc. Natl. Acad. Sci.* **115**, E7222–E7230 (2018).

43. Kuhn, M. caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret. (2018).

44. Ghandi, M. *et al.* GkmSVM: An R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).

45. Kuhn, R. M., Haussler, D. & James Kent, W. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

46. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).

47. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).

48. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2020).

49. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.* (2016).

50. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

51. Avsec, Ž. *et al.* The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).

52. Ponnaluri, V. K. C. *et al.* NicE-seq: High resolution open chromatin profiling. *Genome Biol.* **18**, 1–15 (2017).

Publication 2 – Enhancers display constrained sequence flexibility and context-specific modulation of motif function

Franziska Reiter*, Bernardo P. de Almeida*, Alexander Stark.

Genome Research 33, 346-358 (2023). https://doi.org/10.1101/gr.277246.122, REF: [212]

* These authors contributed equally to this work.

Summary and discussion

The DeepSTARR project (Publication 1) revealed the importance of syntax rules in enhancers and how we still have a limited understanding about the flexibility of important motif positions and how the sequence context modulates the activity of TF motifs. For example, it remains unclear if motifs can work similarly in different enhancers and positions (acting as independent modules) or in contrast require specific sequence contexts and syntax (suggesting interactions and dependencies with other features).

Here, we investigated how many defined DNA sequences might functionally replace the wildtype sequence in various motif and control positions by exhaustively testing all possible 8-nucleotide-long sequence variants at these positions in two enhancers in *Drosophila melanogaster* S2 cells. At each position, hundreds of sequence variants corresponding to several different motif types could functionally replace the wildtype sequence (i.e. constitute solutions), suggesting that enhancer sequences display flexibility within and across motif types. However, at each position, these solutions constituted only a tiny fraction of the approximately 65,000 possible sequences, indicating that enhancer sequence flexibility is constrained. In addition, the solutions differed between positions and most TF motifs had highly context-dependent activities.

We systematically compared the contribution of prominent TF motif types to enhancer activity when placed into different positions along an enhancer to assess how their intrinsic strengths are modulated by the sequence context in both *Drosophila* as well as human enhancers. Indeed, TF motifs contribute with different intrinsic strengths that are strongly modulated by the enhancer sequence context, namely the flanking sequence, the presence and diversity of other motif types, and the distance between motifs.

Overall, these complementary strategies revealed that enhancers display constrained sequence flexibility and the context-specific modulation of motif function. These two general principles of enhancer sequences are important to understand and predict enhancer function during development, evolution and in disease.

Author contribution

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

Enhancers display constrained sequence flexibility and context-specific modulation of motif function

Franziska Reiter,[1,2,4] Bernardo P. de Almeida,[1,2,4] and Alexander Stark[1,3]

[1]*Research Institute of Molecular Pathology, Vienna BioCenter, Campus-Vienna-BioCenter 1, 1030 Vienna, Austria;* [2]*Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, 1030 Vienna, Austria;* [3]*Medical University of Vienna, Vienna BioCenter, 1030 Vienna, Austria*

The information about when and where each gene is to be expressed is mainly encoded in the DNA sequence of enhancers, sequence elements that comprise binding sites (motifs) for different transcription factors (TFs). Most of the research on enhancer sequences has been focused on TF motif presence, whereas the enhancer syntax, that is, the flexibility of important motif positions and how the sequence context modulates the activity of TF motifs, remains poorly understood. Here, we explore the rules of enhancer syntax by a two-pronged approach in *Drosophila melanogaster* S2 cells: we (1) replace important TF motifs by all possible 65,536 eight-nucleotide-long sequences and (2) paste eight important TF motif types into 763 positions within 496 enhancers. These complementary strategies reveal that enhancers display constrained sequence flexibility and the context-specific modulation of motif function. Important motifs can be functionally replaced by hundreds of sequences constituting several distinct motif types, but these are only a fraction of all possible sequences and motif types. Moreover, TF motifs contribute with different intrinsic strengths that are strongly modulated by the enhancer sequence context (the flanking sequence, the presence and diversity of other motif types, and the distance between motifs), such that not all motif types can work in all positions. The context-specific modulation of motif function is also a hallmark of human enhancers, as we demonstrate experimentally. Overall, these two general principles of enhancer sequences are important to understand and predict enhancer function during development, evolution, and in disease.

[Supplemental material is available for this article.]

Transcriptional enhancers are DNA sequence elements that control gene expression by modulating the transcription of their target genes in specific cell types and conditions (Banerji et al. 1981; Levine 2010). These elements contain short sequence motifs bound by different transcription factors (TFs), and the combined regulatory cues of all bound TFs determine an enhancer's activity (Spitz and Furlong 2012). Due to the critical role of enhancers in development, evolution, and disease (Levine 2010; Rickels and Shilatifard 2018), understanding how enhancer sequences encode function is a major question in biology. Previous studies have highlighted the importance of sequence constraints within enhancers, such as the presence of TF motifs and features related to the motifs' flanking sequences, affinities, and arrangements (their number, order, orientation, and spacing), termed here "motif syntax" (Jindal and Farley 2021). However, although mutations in enhancer sequences can change enhancer function and lead to morphological evolution and disease (Gompel et al. 2005; Visel et al. 2009; Levine 2010; Rickels and Shilatifard 2018), enhancers usually display only modest or no sequence conservation across species (Ludwig et al. 1998; Blow et al. 2010; Schmidt et al. 2010; May et al. 2012; Arnold et al. 2014; Villar et al. 2015; Fuqua et al. 2020) and even random DNA sequences can act as enhancers (de Boer et al. 2020; Galupa et al. 2023). Therefore, the importance of sequence constraints and motif syntax within enhancers remain outstanding questions in gene regulation.

Two main models have been proposed to explain how enhancer sequence relates to function. The *enhanceosome* model assumes very strict syntax rules with invariant motif arrangements required for cooperative TF binding (Thanos and Maniatis 1995; Panne 2008). In contrast, the *billboard* model proposes that TFs bind independently without constraints on how motifs are arranged within the enhancer (Kulkarni and Arnosti 2003; Arnosti and Kulkarni 2005). Yet very few enhancers fit these models, having either invariant syntax or no constraints at all, and most enhancers fall in between these two extremes, with a flexible syntax yet high degree of dependency between enhancer features (Kulkarni and Arnosti 2003; Vockley et al. 2017; Jindal and Farley 2021). This complexity in enhancer sequence has prevented the generalization of sequence rules derived from individual enhancers into unifying principles of the regulatory code, thus limiting our understanding of the sequence constraints related to motif syntax and TF activity in enhancers.

Although enhancer sequences evolve rapidly, their function, which is comprised of enhancer strength as well as cell type–specificity, can be conserved despite significant sequence changes (Ludwig et al. 1998, 2000; Rastegar et al. 2008; Blow et al. 2010; Schmidt et al. 2010; Weirauch and Hughes 2010; Swanson et al. 2011; Taher et al. 2011; May et al. 2012; Arnold et al. 2014; Villar et al. 2015; Wong et al. 2020; Vaishnav et al. 2022). This suggests that there is considerable flexibility within enhancer sequences, and that the maintenance of function-defining features rather than overall sequence similarity is important for enhancer activity. This is illustrated most clearly by the maintenance of TF motifs at invariant positions or at different relative positions

[4]These authors contributed equally to this work.
Corresponding author: stark@starklab.org
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277246.122.

within orthologous enhancer sequences (Ludwig et al. 1998, 2000; Rastegar et al. 2008; Arnold et al. 2014; Wong et al. 2020). However, how flexible or constrained motif positions within enhancers are at both, the DNA sequence and the TF motif level, that is, how many different sequence variants or motif types might functionally replace the wild-type sequence at important motif positions, has remained unknown. Similarly, even though TF motifs have been observed to move between different enhancer positions over the course of evolution (presumably a consequence of motif decay and de novo formation), and despite position independence being a key assumption of the billboard model, the influence of the position and sequence context on a motif's contribution to enhancer function is not understood. These knowledge gaps restrict our understanding of the functional and evolutionary flexibility of enhancer sequences and how many sequence variants, as they might arise by DNA mutagenesis, might lead to similar or different enhancer activities.

Here, we investigated how many defined DNA sequences might functionally replace the wild-type sequence in various motif and control positions by exhaustively testing all possible 8-nucleotide-long sequence variants at these positions in two enhancers in *Drosophila melanogaster* S2 cells. In addition, we systematically compared the contribution of prominent TF motif types to enhancer activity when placed into different positions along an enhancer to assess how their intrinsic strengths are modulated by the sequence context in both *Drosophila* as well as human enhancers. Overall, these complementary approaches emphasize the flexibility of enhancer sequences and how the activity of TF motifs is modulated by the enhancer sequence context, namely the flanking sequence, the presence and diversity of other motif types, and the distance between motifs.

Results

STARR-seq comprehensively assesses the activity of enhancer variants revealing constrained enhancer sequence flexibility

To systematically test what sequences function in a certain enhancer position, we used an approach inspired by studies that tested the activity of fully randomized regulatory sequences (Farley et al. 2015; de Boer et al. 2020; Vaishnav et al. 2022; Galupa et al. 2023) or the local fitness landscape of the green fluorescent protein (GFP; Sarkisyan et al. 2016; Somermeyer et al. 2022). We generated a comprehensive library of sequence variants by replacing a specific 8-nt stretch in an enhancer with randomized nucleotides (N_8) and assessed the enhancer activity of each variant by UMI-STARR-seq in *Drosophila* S2 cells (Fig. 1A; see Methods; Arnold et al. 2013; Neumayr et al. 2019). Enhancer activity as used here is a quantitative measure and is defined as the increase in transcription of the reporter by a given enhancer candidate. We tested the power of this approach in the position of a GATA TF motif within the *ced-6* developmental enhancer (*ced-6* position 241 nt, or *pos241*) that is required for its activity. We recovered all possible 8-nt variants (65,536) in the input library and obtained reliable enhancer activity measurements for each variant (Supplemental Fig. S1). This showed that the vast majority of all variants drive low activity levels, whereas only 374 (<1%) achieve similar activity to wild type (±10%) and 600 (1%) drive even higher activity, that is, constitute valid *solutions* at this motif position (Fig. 1B).

Although only a few hundred sequences functioned at this position, these were highly diverse (Fig. 1C,D) and included not only different variants of the GATA motif (Fig. 1B—in blue, and 1E,F) but also other TF motifs, such as SREBP and AP-1 (Fig. 1E,F; Supplemental Figs. S2A,B, S3A). The different levels of importance of motifs were independent of their orientation, with the possible exception of SREBP and STAT for which differences are apparent yet not significant and cannot be assessed reliably because of a small number of instances (Supplemental Fig. S3A). Most of the 600 variants stronger than wild type (94%) created TF motifs overrepresented in S2 developmental enhancers (PWM *P*-value 1×10^{-4}; Fig. 1F; Supplemental Fig. S3B), showing that there is flexibility in the DNA sequences but also in the motif types they encode. However, different TF motifs rescued enhancer activity to different levels (Fig. 1E; Supplemental Fig. S3A). Whereas AP-1 and SREBP achieved similar activity to the wild-type GATA motif, twist and ETS had lower activity at this enhancer position, despite being generally associated with strong enhancer activity in S2 cells (de Almeida et al. 2022). Therefore, the observed sequence flexibility is constrained to some TF motifs. In addition, even within each TF motif not all specific sequence variants functioned similarly, as apparent in the large differences between their activities (Fig. 1E). We observed a positive association between the activities of motif sequence variants and the TF motif affinities for most motifs, yet the correlation was typically modest, indicating that the PWM motif score does not explain the widely different activities (only SREBP has a PCC > 0.6 and twist and ETS even have PCCs < 0.1; Supplemental Fig. S3C).

We also observed TF motif types that had neutral or repressive functions at the tested 8-nt position: The Dref motif, previously shown to only be important for housekeeping enhancers (Zabidi et al. 2015; de Almeida et al. 2022), had no activity in this *ced-6* developmental enhancer, whereas the Ttk motif created the most inactive 8-nt variants consistent with Ttk's function as a repressor (Fig. 1E; Supplemental Fig. S2C; Xiong and Montell 1993). These results show that this approach can comprehensively assess the activity of all sequence variants in a specific region of the enhancer and identify activating, neutral, and repressive sequences. Moreover, our findings indicate that developmental enhancers exhibit *constrained flexibility*, in that many variants, but still a strongly restricted number, can function at a given enhancer position. This constrained sequence flexibility applies not only to individual DNA sequences but also TF motif types in that several different motif types work, but not many or all.

Activity of random variants in seven specific positions of two different enhancers

To evaluate if the same principles and the same specific solutions apply at different enhancer positions, we selected three additional positions of the *ced-6* enhancer and three positions of a strong enhancer in the *ZnT63C* locus (Fig. 2A). To probe enhancer sequence flexibility at important motif positions and nonimportant control positions, we used the deep learning model DeepSTARR (Fig. 2A; de Almeida et al. 2022) and previous experimental enhancer mutations (Supplemental Fig. S4F) to choose positions that should (*ced-6* pos110, pos241; *ZnT63C* pos142, pos180, pos210) or should not (*ced-6* pos182, pos230) be important for enhancer activity. We generated exhaustive libraries of all 8-nt sequence variants for each position and performed UMI-STARR-seq on the combined libraries of each enhancer (Supplemental Fig. S4A–E; see Methods). As observed for the GATA position in Figure 1 (pos241), only a restricted set of variants achieved wild-type activity at a second important GATA motif position in the same enhancer (pos110) or at the important motif positions in the *ZnT63C* enhancer (Fig.

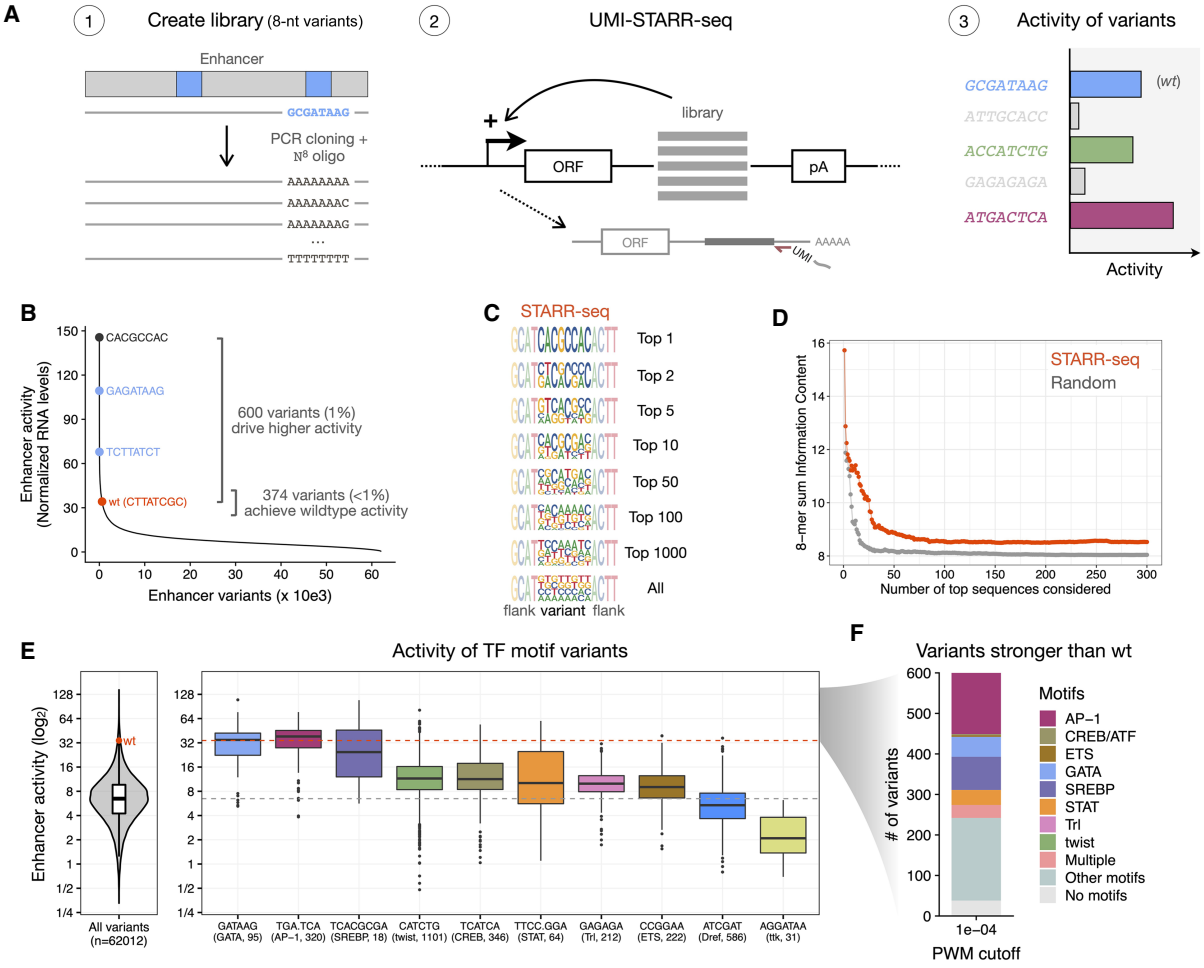


Figure 1. STARR-seq comprehensively assesses the activity of random variants in a specific enhancer position. (*A*) Schematics of STARR-seq for the analysis of random variants in an enhancer position: (1) A comprehensive library of sequence variants was generated by replacing the 8-nt stretch overlapping a GATA TF motif in the strong *ced-6* enhancer with all possible 65,536 randomized nucleotides; (2) the enhancer activity of each variant was measured by STARR-seq in *Drosophila* S2 cells; (3) expected outcomes include the wild-type sequence (wt, blue), inactive variants (gray), and variants that recover the wild-type activity (green) or are even stronger (purple). (*B*) Most sequence variants exhibit low activity levels. The distribution of enhancer activity for each of the 62,012 enhancer variants with confident activity is shown. The wild-type (wt, red) sequence, the strongest GATA variant in each orientation (blue), and the strongest sequence variant are highlighted, together with the number of variants that achieve similar activity to wild type (±10%) or drive even higher activity. (*C*) Strong sequence variants are highly diverse. Logos with nucleotide frequency of the most-active variants in STARR-seq (1, 2, 5, 10, 50, 100, 1000, and all) and flanking nucleotides. Please note that because variants are aligned this will smear out motifs that occur at different positions. Motif finding with HOMER for these variants is shown in Supplemental Figure S2. (*D*) Sum of information content within the most-active 8-mers in STARR-seq (red) compared with the same after randomly sorting the variants (gray), considering different number of top sequences. (*E*) Distribution of enhancer activity for all 62,012 enhancer variants (*left*) or variants creating each TF motif (*right*). The activity of the wild-type sequence (wt, red dot and dashed line) or median of all variants (gray dashed line) are shown. The string of each TF motif used for the motif matching and the number of variants matching to each motif are described in the *x*-axis in the format "motif string (TF motif name, number of variants)." (*F*) Number of variants among the 600 stronger than wild type that match to motifs enriched in S2 developmental enhancers (PWM *P*-value cutoff 1×10^{-4}).

2B), confirming that important positions in enhancers show constrained flexibility. This contrasted with the nonimportant positions (pos182 and pos230 of the *ced-6* enhancer) where most sequence variants were active at or near wild-type levels (Fig. 2B). This is expected, as these positions are predicted to not contain sequences associated with enhancer activity and are therefore less constrained. Thus, the importance of an enhancer position reflects its constraint, with nonimportant positions not being constrained (while they can still be modulated positively or negatively).

The most active sequences at each enhancer position were highly diverse and exhibited distinct nucleotide preferences (Supplemental Figs. S5–S7). For example, two positions located either in the *ced-6* (pos110) or the *ZnT63C* (pos210) enhancer showed distinct preferences among the strongest 100 variants, which preferentially match to an SREBP (GTCAC[flanked by GTC]) or an ETS motif (CCGGA[A]), respectively (Supplemental Fig. S5B). These results show that different enhancer positions require different motif types and thus are under different constraints.

Different TF motif types are active at different enhancer positions

Comparing the activity of the 8-nt sequence variants between the enhancer positions (scaled to the average activity of variants to be comparable across positions; see Methods) revealed that they

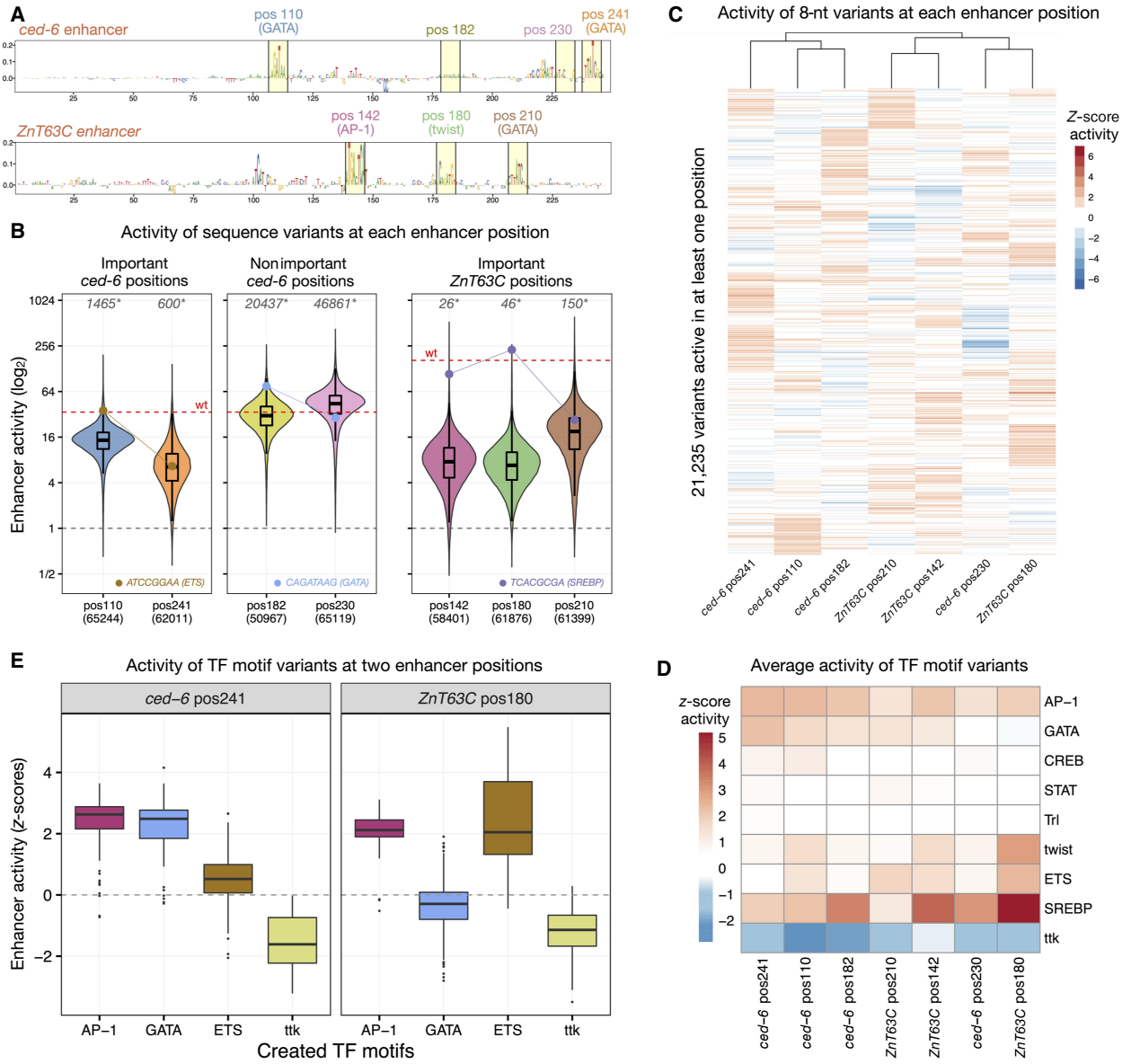


Figure 2. Sequence constraints at different enhancer positions. (*A*) DeepSTARR-predicted nucleotide contribution scores for the *ced-6* (*top*) and *ZnT63C* (*bottom*) selected enhancer sequences. Selected 8-nt motif positions and nonimportant control positions are highlighted in yellow with the respective numerical position, TF motif identity, and different colors. (*B*) Distribution of enhancer activity for all enhancer variants detected in each enhancer position. The activity of the wild-type sequence of each enhancer (wt, red dashed line) or of inactive sequences (gray dashed line) are highlighted, together with the activity of example sequence variants that create different TF motifs (ETS, GATA, and SREBP; dots and connected lines). Number of variants tested in each position are shown on the *x*-axis, whereas the number of variants with higher activity than wild type is shown on the *top* (gray, *). (*C*) Heatmap of *Z*-scores of $\log_2$ enhancer activity of 21,235 variants across all seven enhancer positions. Only variants assessed in all positions and active (*Z*-score > 1) in at least one are shown. Variants were clustered using hierarchical clustering and their activity is colored in shades of red (activating) and blue (repressing). (*D*) Heatmap of average *Z*-scores of $\log_2$ enhancer activity of variants creating each TF motif type (*y*-axis) across all enhancer positions (*x*-axis; sorted as in C). Motif activity is colored in shades of red (activating) and blue (repressing). (*E*) Distribution of *Z*-scores of $\log_2$ enhancer activity for variants creating each of four TF motifs (AP-1, GATA, ETS, ttk) in two selected enhancer positions (*ced-6* pos241 and *ZnT63C* pos180).

indeed functioned differently at different positions (Pearson correlation coefficients [PCCs] below 0.4 between positions; Fig. 2C; Supplemental Fig. S8A–C). Further consolidating the 8-nt into 6-nt variants to reduce the impact of the surrounding sequence of each position (averaged activity across the flanking nucleotides) showed higher correlations but still strong differences between positions (Supplemental Fig. S8A,B,D). The top variants and solutions of each position differed substantially, with each position revealing specific sequences with particularly high activity, matching to known TF motifs (Fig. 2C). For example, an ETS motif variant was among the strongest sequences at *ced-6* pos110 but not at pos241, a GATA variant was very active at *ced-6* pos182 but inactive at pos230, and an SREBP variant was active in all positions of the *ZnT63C* enhancer except at pos210 (Fig. 2B).

We next compared the activity of motifs between the seven positions of the two different enhancers, by consolidating the activity of all 8-nt variants (±4 nt flanks) creating each motif (Fig. 2D, E; Supplemental Fig. S9; see Methods). For each position the wild-type sequence as well as different variants of that motif were among the top variants. Whereas the repressor Ttk motif repressed in all positions and showed little specificity (similar to other known and novel repressor motifs; Supplemental Fig. S10), the activator motifs showed distinct profiles, such as motifs that are globally active in all positions (AP-1), motifs with low activity in all tested positions (STAT, CREB, and Trl), and motifs with highly context-dependent activities (GATA, twist, ETS, and SREBP) (Fig. 2D,E). For example, GATA was active at the *ced-6* pos110 but not at the *ZnT63C* pos180 position, whereas ETS motifs showed the opposite profile with the strongest activity at *ZnT63C* pos180 (Fig. 2E). For GATA motifs, we observed strong activity in all positions except on *ced-6* pos230 and *ZnT63C* pos180, which are positioned close to another GATA motif (Fig. 2A). This observation is in line with the previously observed negative interaction of GATA/GATA motif pairs at short distances (de Almeida et al. 2022) and suggests that the observed different activities of TF motifs at different enhancer positions depend on their interaction with other TFs and the sequence context.

In summary, testing thousands of sequence variants in different enhancer positions revealed that enhancer sequences display constrained flexibility, in that only a specific but still diverse set of sequences and TF motifs can function at a given position. However, these constraints and solutions differed between enhancer positions, with different TF motifs active at different positions, suggesting that their activity is modulated by the sequence context.

Systematic motif pasting shows that motifs work differently at different enhancer positions

To systematically test if and how the enhancer sequence context modulates the function of TF motifs, we selected eight TF motifs that showed distinct position-dependent preferences (GATA, Trl, SREBP, AP-1, Atf2, twist, Stat92E, and ETS) and pasted their optimal sequences into 763 positions in a total of 496 developmental enhancers (Fig. 3A; see Methods). These positions were selected to be TF motifs important for the activity of the respective enhancers, as assessed by motif mutagenesis, allowing the reliable measurement of the increase in enhancer activity after pasting each TF motif (here quantified as the $\log_2$ fold-change activity over the motif-mutated enhancer). UMI-STARR-seq experiments with these designed libraries produced highly reproducible and quantitative enhancer activity measurements (replicates PCC between 0.94 and 0.98; Supplemental Fig. S11). Disrupting the selected enhancer positions by shuffling the wild-type sequences substantially reduced the activity of the respective wild-type enhancers by an average of more than sixfold, and pasting the different TF motifs in these same positions rescued enhancer activity to different levels (Supplemental Fig. S12A). Because we pasted the same optimal sequence for each TF motif into all positions, the differences in activity can only be explained by their respective sequence context; the differences between TF motifs are also directly comparable, because we pasted them in the same set of positions.

Across all positions TF motifs had different median activities, which we interpret as different *intrinsic strengths*, with SREBP, ETS, and AP-1 being the strongest motifs and Trl the weakest (Fig. 3B; Supplemental Fig. S12A). However, enhancer positions had large effects on the motif activities that differed more than 100-fold for the same motif (Fig. 3B). For example, pasting a GATA motif activated enhancer activity more than 20-fold for 33 positions but not at all for 72 different positions. This position dependency was particularly strong for Trl, Stat92E, and GATA motifs, and weaker for AP-1, SREBP, and ETS (Supplemental Fig. S12B), which all had higher intrinsic strengths. Additionally, each TF motif showed differential activity across enhancer positions and activated in a unique set of positions. For example (Fig. 3C), GATA motifs activated enhancer1-position168 but not enh2-pos68, whereas ETS showed the opposite effect, and both motifs activated enh3-pos135. The different TF motifs showed different activity profiles across all positions, as revealed by global comparisons and hierarchical clustering (Fig. 3D; Supplemental Fig. S13). These results highlight the complexity of enhancer syntax and the difficulty of predicting and interpreting individual sequence manipulations.

The distinct preferences observed between pasted motifs were largely independent of the identity of the replaced wild-type motif across all positions, as revealed by the weak interaction scores between the wild type and the pasted motif identity in a multivariate linear regression analysis of all motif-pasting experiments (<1% explained variance, Supplemental Fig. S14). In contrast, the pasted motif identity (irrespective of the identity of the replaced motif) explains the most (23%) whereas 65% of variance remains unexplained and is likely due to surrounding enhancer sequence features affecting the motifs' activities. Thus, systematic pasting of TF motifs across hundreds of enhancer contexts shows that motifs have different intrinsic strengths but work differently at different enhancers and positions, suggesting that the enhancer sequence context constrains the activity of TF motifs.

TF motifs have different intrinsic strengths that are modulated by the enhancer sequence context

The observed differential activities of motifs in different enhancer positions (Fig. 3D) suggest that the enhancer sequence context modulates the function of TF motifs. We found no significant differences when comparing the motif activity between pairs of positions in the same enhancer or in different enhancers, suggesting that the local context immediately surrounding the motif is as important as enhancer identity (Supplemental Fig. S15).

More globally, the sequence context for a motif can be related to its position within the enhancer, the motif flanking sequence, and the presence and distance to other motifs. To characterize the importance of these features, we tested if they contribute to the performance of predicting enhancer activity following the pasting of a motif at different enhancer positions. We first built a baseline random forest model that only includes the importance of the wild-type motif and the identity of the wild-type and pasted motifs as features, thereby not taking any sequence context features into account. This model obtained a PCC of 0.59 in the whole data set using tenfold cross-validation and showed that the pasted motif and the wild-type motif importance are strong determinants for enhancer activity (Supplemental Fig. S16A). Training a second random forest model that also includes context features such as the motif position relative to the enhancer center, the motif flanking sequence (defined as ±5 bp around the optimal motif as in de Almeida et al. [2022]), and the presence and distance to other TF motifs, improved this performance to a PCC of 0.69 (Supplemental Fig. S16B). This shows that the enhancer sequence context, particularly the closest flanking nucleotides as well as the presence of other motifs at specific distances (e.g., GATA or ETS),

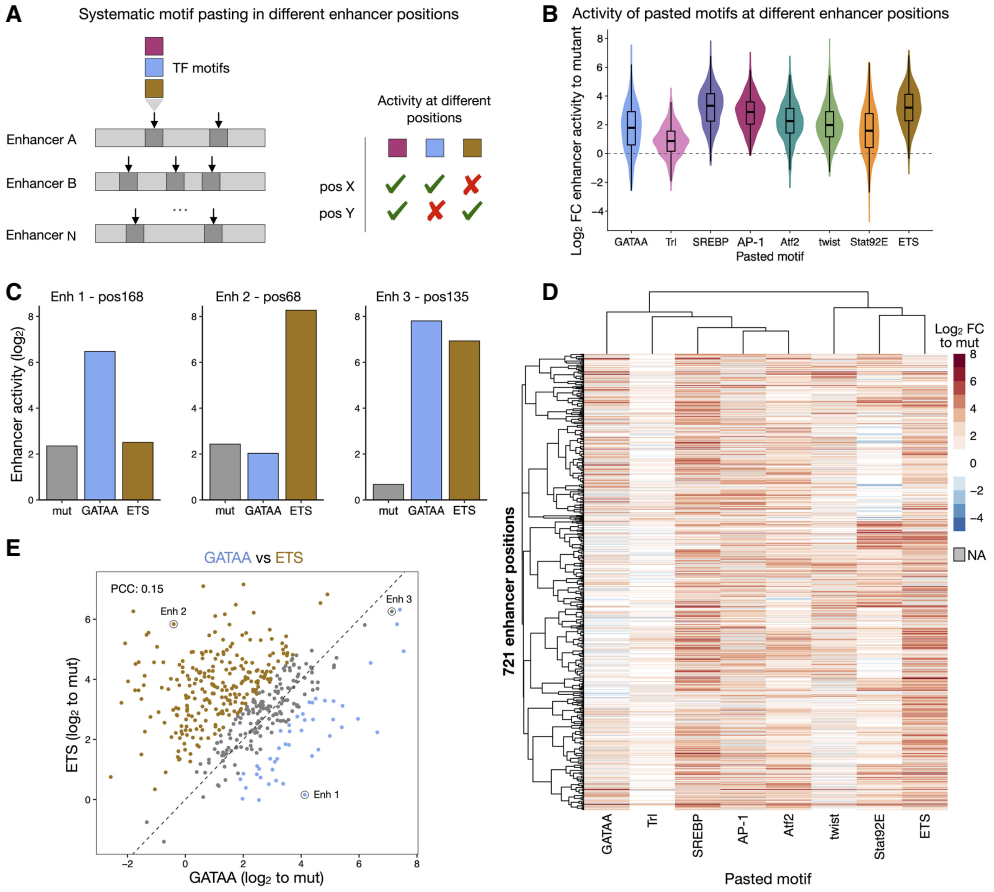


Figure 3. TF motifs work differently at different enhancer positions. (*A*) Schematics of systematic motif pasting in different enhancer positions. Eight TF motifs that showed distinct position-dependent preferences were selected and their optimal sequence was pasted in 763 positions distributed among 496 enhancers, representing different contexts. The enhancer activity of each variant was measured by STARR-seq in *Drosophila* S2 cells to quantify the activity of motifs at the different positions. (*B*) Distribution of enhancer activity changes ($\log_2$ FC to mutated sequence) across all enhancer positions for each pasted TF motif. (*C*) Bar plots with activity ($\log_2$) of variants of three different enhancers with a mutated sequence (gray), a GATA (blue), or a ETS (brown) motif pasted at the same position. (*D*) Heatmap of enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting each of the eight selected TF motifs in 721 enhancer positions (positions with data for at least six motifs). TF motifs and positions were clustered using hierarchical clustering and the activity is colored in shades of red (activating) and blue (repressing); missing values are colored in gray. (*E*) GATA and ETS motifs work differently at different enhancer positions. Comparison between enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting GATA (*x*-axis) or ETS (*y*-axis) across all enhancer positions. Positions with stronger activity of GATA or ETS (≥twofold with respect to the other motif) are colored in blue and brown, respectively. Enhancer positions shown in *C* are highlighted. PCC: Pearson correlation coefficient.

has an impact on the activity of TF motifs (Supplemental Fig. S16B).

To better characterize the importance of these sequence rules for each TF motif separately, we generated interpretable linear models based on these rules to predict the motif activities across all positions (Fig. 4A). These models were able to predict the motif pasting results, with PCCs to experimentally assessed $\log_2$ fold changes between 0.39 (ETS) and 0.64 (Stat92E) (Fig. 4A; Supplemental Fig. S17). The motif flanks and the presence of additional motifs explained on average 16.7% and 6.7% of the motif activities variance, respectively, whereas the motif position within the enhancer had lower importance (0.4%).

The TF motif type-specific models revealed how the sequence context rules differ between TF motif types, explaining the motif-specific enhancer position preferences. For example, GATA activity was strongly dependent on the flanking nucleotides and was modulated by the presence of a second GATA at close distance (negative interaction) or ETS motifs (positive interaction) (Fig. 4B; Supplemental Fig. S18A). We saw different associations for ETS activity, as expected by the different GATA and ETS activity profiles across all positions (Fig. 3E). ETS activity was only mildly influenced by the flanking nucleotides but strongly by neighboring motifs: it was stronger close to GATA motifs and weaker in enhancers with another ETS motif (Fig. 4C; Supplemental Fig. S18B). These sequence features, such as the negative GATA/GATA and the positive ETS/GATA interactions at close distances, were observed previously via computational models of wild-type S2 enhancer sequences (de Almeida et al. 2022).

In addition, the DeepSTARR-predicted importance of each nucleotide when pasting different TF motifs into the same position revealed their interaction with the sequence context (Fig. 4D,E; Supplemental Fig. S19): GATA but not ETS activated the Chr 3L enhancer in a position with additional distal GATA motifs, while ETS but not GATA activated the Chr X enhancer in a position with a GATA motif at close distance, and both activated the Chr 2L enhancer that contains multiple surrounding twist

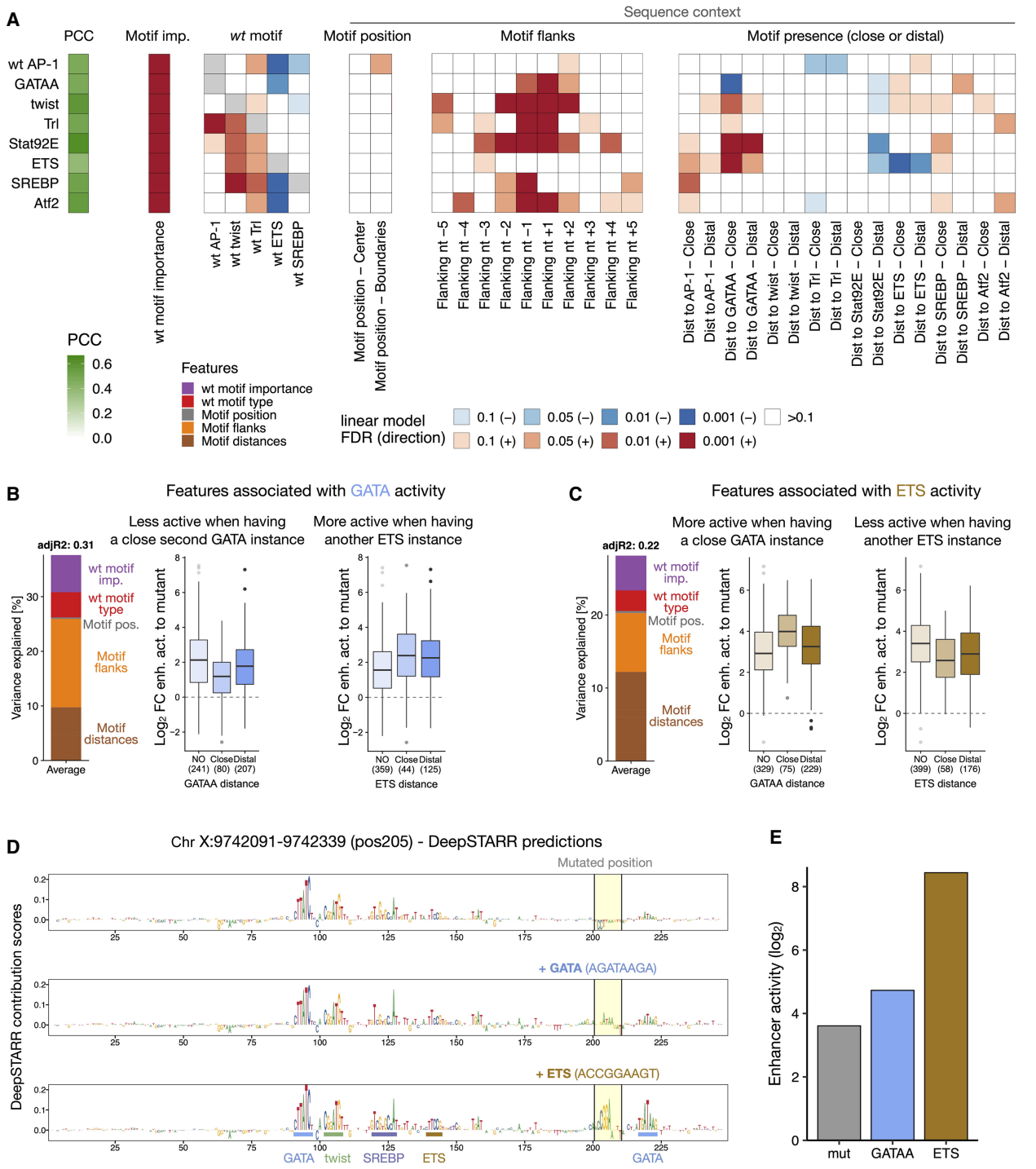


Figure 4. Characterization of preferred syntax features of each TF motif. (*A*) Motif syntax rules modulate TF motif function. For each TF motif type (rows), a linear model was built to predict its activity across all enhancer positions, using as covariates the number of instances, the wild-type TF motif importance and identity, and sequence context features such as the position within the enhancer, the flanking nucleotides, and the presence at close or distal distances to all other TF motifs. The PCC between predicted and observed motif activities is shown with the green color scale on the *left*. The heatmap shows the contribution of each feature (columns) for each model, colored by the FDR-corrected *P*-value (red or blue scale depending on positive or negative association, respectively). (*B*,*C*) Syntax features associated with GATA (*B*) or ETS (*C*) activity. *Left*: bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. *Middle* and *right*: enhancer activity changes ($\log_2$ FC to a mutated sequence) after pasting each TF motif in positions with no additional GATA (*middle*) or ETS (*right*) in the enhancer, or with additional GATA or ETS at close (≤25 bp) or distal (>25 bp) distances. Number of instances are shown. (*D*) DeepSTARR-predicted importance scores for pasting a mutant sequence (gray), GATA (blue), or ETS (brown) in a specific position (Chr X: 9,742,091–9,742,339, pos205). Motif sequences pasted are shown. (*E*) Bar plots with measured enhancer activity ($\log_2$) of variants from *D*.

motifs, all consistent with these motifs' respective distance preferences (de Almeida et al. 2022). Together these results demonstrate how the sequence context (e.g., the flanking sequence, and the presence and diversity of other motif types) modulates the function of TF motifs, constraining enhancer sequence flexibility.

Enhancer sequence context modulates the function of human TF motifs

To test whether TF motifs also work differently in different enhancer sequence contexts in other species, we performed the systematic motif pasting experiment in human HCT116 cells for eight previously characterized human TF motifs (P53, AP-1, ETS, CREB1, MAF, EGR1, E2F1 and MECP2; see Methods; de Almeida et al. 2022). Pasting of the motifs into 1354 important positions in 753 different HCT116 enhancers revealed that human TF motifs also have different intrinsic strengths and work differently in different enhancers and positions (Fig. 5A; Supplemental Figs. S20, S21). P53 was the strongest motif and the only one that showed globally strong activity across all enhancer positions, suggesting little dependence on the enhancer context, as has been suggested before (Verfaillie et al. 2016). AP-1, the second strongest motif, was strongly dependent on the enhancer positions, with activities ranging more than 50-fold across enhancer contexts. This position dependence was also observed for the other motifs, even though their overall activity was lower (Fig. 5A).

TF motifs preferred different enhancer contexts, with four groups of motifs showing characteristically different preferences: (1 – P53) strong activity in all positions; (2 – CREB1, AP-1, MAF, EGR1) and (3 – ETS) highly context-dependent activities; (4 – MECP2, E2F1) only active in few and highly specific enhancer positions (Fig. 5B,C; Supplemental Fig. S22). These distinct preferences were independent of the identity of the replaced motif (Fig. 5D; Supplemental Fig. S23) but correlated with sequence context features. Similar to *Drosophila* TF motifs, motif context features such as motif flanks and the presence and distance to other TF motifs were important to predict the activities of human motifs across the different enhancer positions (Supplemental Fig. S24). TF-specific linear models based on such syntax features were able to predict the motif activities across all positions (PCCs between 0.46 and 0.51; Supplemental Fig. S25) and revealed the context preferences of each TF motif (Fig. 5E).

All motif activities were influenced by the flanking nucleotides, which explained on average 8.2% of the motif activities' variance, whereas the presence of additional motifs and their distance explained 8.5% (Fig. 5E; Supplemental Figs. S25, S26). As expected by the weak context specificity of P53 (group 1, Fig. 5A), its activity was independent of the presence and distance to other TF motifs (Fig. 5E; Supplemental Fig. S26A). All the other motifs preferred contexts with an additional AP-1 instance (Fig. 5E). The AP-1 motif itself, as well as MAF, CREB1, and EGR1 (group 2), all preferred positions close to an ETS motif, concordant with previous studies showing direct protein–protein interactions between ETS and other TFs (Li et al. 2000; Burda et al. 2010), whereas the ETS motif (group 3) had a negative interaction with a second close ETS motif (Fig. 5E), as also observed in *Drosophila* enhancers (Fig. 4A). These findings are also concordant with the motif syntax rules found in a previous study (de Almeida et al. 2022). Altogether, this establishes that TF motifs require specific enhancer sequence contexts in species as divergent as fly and human, suggesting that this is a general principle of regulatory enhancer sequences.

Discussion

In this study, we used two complementary strategies to explore the flexibility of enhancers with regard to nucleotide and motif identity at specific enhancer positions as well as the position dependence of motif activity. Even though median enhancer activity drops significantly when randomizing an 8-nt stretch at important positions, many sequence variants, including variants of the wild-type motif but also other TF motifs, can achieve strong enhancer activity. The diverse solutions at each position show that enhancers exhibit some degree of flexibility. However, as only a few hundred out of the >65,000 tested sequences work, the flexibility at any given position is constrained. Similarly, systematically pasting different motifs into hundreds of enhancer positions revealed that motif activity is strongly modulated by the enhancer sequence context. Therefore, constrained sequence flexibility and the modulation of motif function by the sequence context seem to be key features of enhancers.

The observation that both *Drosophila* and human TF motifs require specific enhancer sequence contexts suggests that this is a general principle of enhancers. Even though motifs possess some intrinsic strengths, their potential to activate transcription strongly depends on the sequence context and follows certain syntax rules, including motif flanks, combinations, and distances. Although our study cannot assess the mechanistic causes for these rules, they might be related to local DNA shape (Dror et al. 2015; Mathelier et al. 2016; Samee et al. 2019) or to more general enhancer DNA properties such as DNA bending. Our observation that homotypic interactions of certain motifs at close distances (e.g., GATA or ETS) are negatively associated with enhancer activity is consistent with repressive homotypic interactions between pluripotency TFs found by thermodynamic modeling (Fiore and Cohen 2016); the mechanisms, however, are still unclear. Intermotif distances can impact the synergy between TFs at the level of DNA binding or after binding, such as cofactor recruitment and activation, which could explain both positive and negative TF-TF interactions (Reiter et al. 2017). Although these syntax rules seem to be stricter for some TF motifs (e.g., GATA) and more relaxed for others (e.g., P53), our results show that motifs are not simply independent modules. Instead, they interact with all enhancer features in a highly cooperative manner, which can modulate motif activity by more than 100-fold. This is an important result that supports a model where enhancer activity is encoded through a complex interdependence between motifs and context, rather than motifs acting independently and additively. Whereas tissue- or cell type–specificity can already be predicted by motif presence-absence patterns alone (Kvon et al. 2014; Janssens et al. 2022), the encoding of different enhancer strengths seems to depend on more complex *cis*-regulatory syntax rules (Jindal and Farley 2021; de Almeida et al. 2022). The functional implications of mutations in TF motifs or elsewhere within enhancer sequences can therefore only be assessed in the context of these syntax features.

The motif syntax rules described here agree well with the ones learned by DeepSTARR trained on genome-wide enhancer activity data (de Almeida et al. 2022) and the BPNet model trained on endogenous TF binding and cooperativity (Avsec et al. 2021), suggesting that these rules are important in wild-type enhancer sequences. As an ectopic reporter assay STARR-seq measures the potential of sequences to act as enhancers, even if the sequences might be repressed endogenously at the chromatin level (Arnold et al. 2013; Muerdter et al. 2018), making it a powerful tool to

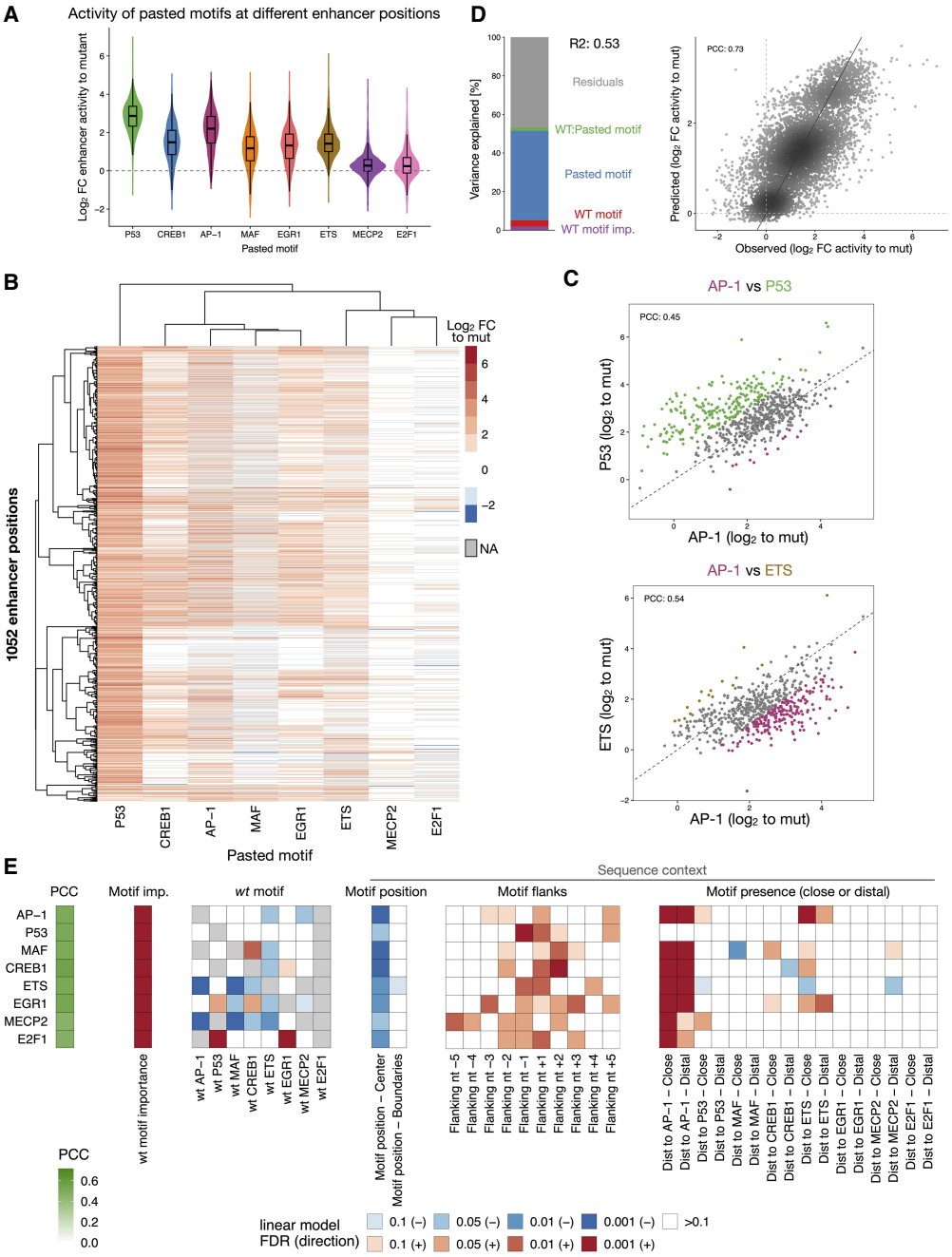


Figure 5. Human TF motifs require specific enhancer sequence contexts. (*A*) Distribution of enhancer activity changes ($\log_2$ FC to mutated sequence) across all enhancer positions for each pasted TF motif. (*B*) Heatmap of enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting each of the eight selected human TF motifs in 1052 enhancer positions (positions with data for at least six motifs). TF motifs and positions were clustered using hierarchical clustering and the activity is colored in shades of red (activating) and blue (repressing); missing values are colored in gray. (*C*) Human TF motifs work differently at different enhancer positions. Comparison between enhancer activity changes ($\log_2$ FC to mutated sequence) after pasting AP-1 (*x*-axis) and P53 (*top*) or ETS (*bottom*) (*y*-axis), across all enhancer positions. Positions with stronger activity of each motif (≥twofold in respect to the other motif in the scatter plot) are colored (P53: green, AP-1: purple, ETS: brown). PCC: Pearson correlation coefficient. (*D*) TF motif activity in function of wild-type and pasted motif identity. *Left*: Bar plot showing the amount of variance explained by the wild-type motif importance and identity, the pasted motif identity, and the interaction between the wild-type and pasted motifs, using a linear model fit on all motif pasting results. *Right*: Scatter plots of predicted (linear model) versus observed enhancer activity changes ($\log_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. (*E*) Motif syntax rules modulate the function of human TF motifs. For each TF motif type (rows), we built a linear model to predict their activity across all enhancer positions, using as covariates the number of instances, the wild-type TF motif importance and identity, and sequence context features such as the position within the enhancer, the flanking nucleotides, and the presence at close or distal distances to all other TF motifs. The PCC between predicted and observed motif activities is shown with the green color scale on the *left*. The heatmap shows the contribution of each feature (columns) for each model, colored by the FDR-corrected *P*-value (red or blue scale depending on positive or negative association, respectively).

uncover the sequence determinants for enhancer activity. It will be interesting to explore the sequence rules and mechanisms by which chromatin modulates endogenous enhancer activities and gene expression using complementary methods (Catarino and Stark 2018). In addition, DeepSTARR also predicted with good accuracy the activity of all randomized sequence variants and of motifs pasted in different enhancer contexts (Supplemental Figs. S27, S28). This supports the validity of computational models such as DeepSTARR and their use in in-silico-*like* experiments (e.g., motif pasting experiments with a larger set of TF motifs across many more genomic positions) to improve our understanding of the regulatory information encoded in enhancer sequences and the impact of mutations.

Our study shows that enhancer sequences are flexible enough for enhancer strength to be achieved by a small yet diverse set of sequence variants, and that mutations in information-poor positions have little impact on the enhancer activity in a single cell type. This flexibility allows many different sequences to achieve similar enhancer activities in a single cell type, which might be an important prerequisite for the evolution of developmental enhancers that operate under many additional constraints, for example, regarding the precise spatiotemporal control of enhancer activities. As the activity in a given cell can be achieved by many solutions, the specific solutions that fulfill additional requirements can be explored during evolution. Indeed, previous studies that have analyzed expression changes of enhancer mutations across different cell types in vivo have observed that the cell type–specific expression patterns of enhancers can change upon (minimal) sequence perturbations (Farley et al. 2015; Fuqua et al. 2020; Galupa et al. 2023). The fact that enhancer strength in any given cell type and enhancer specificity across cell types and developmental time are subject to different yet overlapping sequence constraints highlights the complexity of the regulatory code. We expect that the combination of quantitative enhancer-sequence-to-function models in individual cell types and qualitative predictions of enhancer activities across cell types will provide unprecedented progress in our understanding of enhancer biology and our ability to read and write enhancer sequences.

Methods

UMI-STARR-seq library cloning

Libraries of *Drosophila* enhancer variants with 8-nt randomized sequences were generated using a PCR approach with degenerate oligonucleotides. Forward primers (Supplemental Table S1) were designed to anneal directly downstream of the enhancer position of interest followed by 8 degenerate bp (creating 65,536 variants) and another 20 bp complementary stretch. Reverse primers were complementary to the 20 bp 5′ of the degenerate stretch. The STARR-seq vector containing the wild-type enhancer of interest (*ced-6* or *ZnT63C*) was used as a template for the PCR. The PCR was run across the whole STARR-seq plasmid, followed by DpnI digestion and a Gibson reaction that recircularizes the plasmid. *Drosophila* and human oligo libraries were amplified (for primers, see Supplemental Table S1) and cloned into *Drosophila* STARR-seq vectors containing the DSCP core promoter and into the human STARR-seq plasmid with the ORI in place of the core promoter (Muerdter et al. 2018), respectively. All libraries were grown in 2l LB-Amp (final ampicillin concentration 100 µg/mL). All libraries were purified with Qiagen Plasmid *Plus* Giga Kit (cat. no. 12991).

Cell culture, transfection, and UMI-STARR-seq

Drosophila S2 and human HCT116 cells were cultured as described previously (Arnold et al. 2013; Muerdter et al. 2018). Cells were electroporated using the MaxCyte-STX system at a density of 50×10^6 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 (Arnold et al. 2013; Muerdter et al. 2018; Neumayr et al. 2019). Further experimental details can be found in the Supplemental Methods.

Random variant UMI-STARR-seq data analysis

RNA and DNA input reads (paired-end 150 bp) were mapped to dedicated Bowtie indices using Bowtie v.1.2.2 (Langmead et al. 2009). Because the N_8 variants were all positioned in the last 150 nt of each enhancer, we allowed for flexible mapping in the beginning of the fragments to increase the number of mapped reads while keeping high sensitivity for the different enhancer variants. Specifically, we trimmed the forward reads to 36 bp and mapped them to the indices allowing for three mismatches; the full 150-bp-long reverse reads were mapped with no mismatches, to identify all sequence variants; paired-end reads with the correct position, length, and strand were kept. For paired-end DNA and RNA reads that mapped to the same variant, we collapsed those that have identical UMIs (10 bp, allowing one mismatch) to ensure the counting of unique molecules (Supplemental Table S2).

We excluded oligos with less than five reads in any of the input replicates and less than one read in any of the RNA replicates. The enhancer activity of each sequence in each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014).

Oligo library UMI-STARR-seq data analysis

As described previously (de Almeida et al. 2022), RNA and DNA input reads were mapped to a reference containing the 249-bp-long sequences from the fragments present in the *Drosophila* (dm3) or human (hg19) libraries using Bowtie v.1.2.2 (Langmead et al. 2009). We used these reference genomes to be able to integrate our results with older in-house and published data sets and made sure this choice does not affect the quantifications of enhancer activity. 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 (Supplemental Table S2).

We excluded oligos with less than 10 reads in any of the input replicates and added one read pseudocount to oligos with zero RNA counts. The enhancer activity of each oligo in each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014).

Random variant libraries of *Drosophila* enhancers and UMI-STARR-seq

Two strong S2 developmental enhancers with different TF motif compositions were selected to test a diversity of random 8-nt variants in different positions: *ced-6* (Chr 2R: 5,326,628–5,326,876) and *ZnT63C* (Chr 3L: 3,310,914–3,311,162) enhancers. We selected five positions important for the activity of the two enhancers (*ced-6* pos110 and pos241; *ZnT63C* pos142, pos180, pos210) and two nonimportant positions of the *ced-6* enhancer (pos182 and

pos230) and replaced each 8-nt stretch of the enhancer with randomized nucleotides (N_8), creating 65,535 enhancer variants in addition to the wild-type sequence per position. For each enhancer, we pooled the libraries of the different positions and combined them with an oligo library of thousands of wild-type enhancers and negative sequences (de Almeida et al. 2022) for normalization. UMI-STARR-seq using the *ced-6* or *ZnT63C* pooled libraries was performed and analyzed as described above (Supplemental Table S3). We performed two independent replicates per enhancer pooled library screen (Pearson correlation coefficient (PCC) = 0.85–0.91). To be able to compare the activity of variants and motifs between enhancer positions, we next scaled the enhancer activity of all variants per position (*Z*-scores). This allowed us to measure the change in activity of a given variant over the average of all variants, correcting for the importance of the different enhancer positions tested.

Diversity of top active variants and de novo motif discovery

The most-active 8-nt variants of each screen (1, 2, 5, 10, 50, 100, and 1000) were retrieved and consolidated into position probability matrices based on the nucleotide frequencies at each position. Logos were visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1; https://CRAN.R-project.org/package=ggseqlogo). The top 100 and 1000 or bottom 1000 variants (8 nt ± 4 nt flanks) of each screen were used for de novo motif discovery analyses using HOMER, taking all detected variants of the respective screen as background. HOMER (v4.10.4; Heinz et al. 2010) was run with the findMotifs.pl command and the arguments *fly -len 6,7,8*.

Activity of TF motifs created by sequence variants

To robustly assess the activity of a given TF motif, we retrieved the activity of all 16-nt variants (8 nt ± 4 nt flanks) creating each motif by string matching. For a more systematic comparison across all TF motif types, we matched variants to the optimal string from each TF motif PWM model in a motif database (de Almeida et al. 2022). The average activity across variants was defined as the motifs' intrinsic strength. To find how many active variants are explained by the creation of known motifs enriched in S2 developmental enhancers, we performed PWM-based motif scanning of those candidate motifs onto variants (8 nt ± 4 nt flanks). We used the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; genome = "BSgenome.Dmelanogaster.UCSC.dm3", bg="genome" [https://bioconductor.org/packages/release/bioc/html/motifmatchr.html]) with *P*-value cutoffs 1×10^{-4} and 1×10^{-5}.

Comparison of random variants activity across enhancer positions

We compared the activity of all 8-nt random variants across enhancer positions using their *Z*-score scaled activity (Supplemental Table S3). We calculated pairwise PCCs between the different libraries, performed hierarchical clustering ("complete" method) using the correlation values as similarities, and displayed heatmaps using the pheatmap R package (v.1.0.12; https://CRAN.R-project.org/package=pheatmap). To reduce the impact of the flanking sequence of each position when comparing the activity of variants between them, we repeated the same after consolidating the 8-nt into shorter variants by taking the centered sequence and averaging the activity across variants with different flanking nucleotides.

Drosophila and human TF motif mutagenesis oligo library synthesis and UMI-STARR-seq

For the *Drosophila* library, we selected 1172 motif positions (among 728 enhancers) that are required for the activity of the respective enhancers and designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of eight TF motifs (GATA, AP-1, twist, Trl, ETS, SREBP, Stat92E, and Atf2; one at a time; sequences in Supplemental Table S4) in each of these positions. For the human library, we selected 1456 motif positions important for the activity of 808 enhancers and designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of the same eight TF motifs (AP-1, ETS, E2F1, EGR1, MAF, MECP2, CREB1, P53; one at a time; sequences in Supplemental Table S4) in each of these positions. Each of the *Drosophila* and human libraries was synthesized and pooled with a previous library containing the respective wild-type enhancer sequences (de Almeida et al. 2022) to be screened together (Supplemental Tables S5, S6). All details can be found in the Supplemental Methods. The resulting 300-mer oligonucleotide *Drosophila* and human libraries were synthesized by Twist Bioscience. UMI-STARR-seq using these oligo libraries was performed and analyzed as described above (Supplemental Tables S5, S6). We performed three independent replicates for *Drosophila* (correlation PCC = 0.95–0.98) and human (PCC = 0.96–0.98) screens.

Quantification of motif activity at different enhancer positions

We used our enhancer activity measures of the wild-type and mutated sequences to stringently select important enhancer positions for further analyses: positions where mutation reduced the activity by at least twofold (Supplemental Figs. S12A, S21A). These resulted in 763 important positions distributed among 496 *Drosophila* enhancers and 1354 positions distributed among 753 human enhancers. Variability of activity of each motif across enhancer positions was quantified using the coefficient of variation (ratio of the standard deviation to the mean; Supplemental Fig. S12B). We compared the activity of motifs across enhancer positions by pairwise PCCs and performed hierarchical clustering ("complete" method) using the correlation values as similarities. Heatmaps were displayed using the pheatmap R package (v.1.0.12; https://CRAN.R-project.org/package=pheatmap).

Prediction of motif activities using motif syntax features

We extracted the following syntax features per tested enhancer position: the position relative to the enhancer center (center: −/+25 bp, flanks: −/+25:75 bp, boundaries: −/+75:125 bp), the position flanking nucleotides (5 bp on each side), and the presence and distance to other TF motifs (close: ≤25 bp; distal: > 25 bp; between motif centers). Instances of each TF motif type were mapped across all enhancers using their annotated PWM models (Supplemental Table S3) and the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; https://bioconductor.org/packages/release/bioc/html/motifmatchr.html) with the following parameters: genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg = "genome".

We used a 10-fold cross-validation scheme to train random forest models to predict *Drosophila* or human motif pasting activities ($\log_2$ fold-change to mutant) using as features the wild-type TF motif identity and importance ($\log_2$ fold-change activity between wild-type and motif-mutant sequence) and the pasted motif identity, together or not with the syntax features described above. All models were built using the caret R package (v. 6.0–80; https://CRAN.R-project.org/package=caret) and feature importance was calculated using its *varImp* function.

In addition, we trained a multiple linear regression model per TF motif type to predict its activity across different enhancer positions using as covariates the wild-type TF motif identity and importance together with the syntax features described above. All

models were built using the caret R package (v. 6.0–80; https://CRAN.R-project.org/package=caret) and 10-fold cross-validation. The linear model coefficients and respective FDR-corrected *P*-values were used as metrics of importance for each feature, using the red or blue scale depending on positive or negative associations (Figs. 4A, 5E). We calculated the percentage of variance explained by each covariate in the linear models built for each TF motif with one-way ANOVAs. Further details can be found in the Supplemental Methods.

DeepSTARR nucleotide contribution scores and predictions of enhancer sequence changes

Nucleotide contribution scores for wild-type enhancers or enhancer variants were calculated using DeepSTARR as described previously (de Almeida et al. 2022) and visualized using the *ggseqlogo* function from the R package ggseqlogo (v.0.1; https://CRAN.R-project.org/package=ggseqlogo). DeepSTARR was also used to predict the enhancer activity of N_8 variants in enhancers or the $\log_2$ fold-change enhancer activity of motif pasting sequences.

Statistics and data visualization

All statistical calculations and graphical displays have been performed in R statistical computing environment (v.3.5.1; R Core Team 2020) and using the R package ggplot2 (Wickham 2016). 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.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE211659 or Zenodo (https://zenodo.org/record/7010528#.ZAeEay1h2v4). Code used to process the UMI-STARR-seq data as well as to reproduce all analyses, results, and figures has been submitted to GitHub (https://github.com/bernardo-de-almeida/Variant_STARRseq) and is available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank V. Loubiere and T. Pachano (IMP) for comments on the manuscript and all members of the Stark group for discussions. Deep sequencing was performed at the Vienna Biocenter Core Facilities GmbH. F.R. is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Research Institute of Molecular Pathology. Research in the Stark group is supported by the Austrian Science Fund (FWF). Basic research at the IMP is supported by Boehringer Ingelheim GmbH and the Austrian Research Promotion Agency (FFG).

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

References

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077. doi:10.1126/science.1232542

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. *Nat Genet* **46:** 685–692. doi:10.1038/ng.3009

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94:** 890–898. doi:10.1002/jcb.20352

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, Mcanany C, Gagneur J, Kundaje A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53:** 354–366. doi:10.1038/s41588-021-00782-6

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27:** 299–308. doi:10.1016/0092-8674(81)90413-X

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-seq identification of weakly conserved heart enhancers. *Nat Genet* **42:** 806–810. doi:10.1038/ng.650

Burda P, Laslo P, Stopka T. 2010. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* **24:** 1249–1257. doi:10.1038/leu.2010.104

Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32:** 202–223. doi:10.1101/gad.310367.117

de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54:** 613–624. doi:10.1038/s41588-022-01048-5

de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38:** 56–65. doi:10.1038/s41587-019-0315-8

Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25:** 1268–1280. doi:10.1101/gr.184671.114

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of developmental enhancers. *Science* **350:** 325–328. doi:10.1126/science.aac6948

Fiore C, Cohen BA. 2016. Interactions between pluripotency factors specify *cis*-regulation in embryonic stem cells. *Genome Res* **26:** 778–786. doi:10.1101/gr.200733.115

Fuqua T, Jordan J, van Breugel ME, Halavatyi A, Tischer C, Polidoro P, Abe N, Tsai A, Mann RS, Stern DL, et al. 2020. Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587:** 235–239. doi:10.1038/s41586-020-2816-5

Galupa R, Alvarez-Canales G, Borst NO, Fuqua T, Gandara L, Misunou N, Richter K, Alves MRP, Karumbi E, Perkins ML, et al. 2023. Enhancer architecture and chromatin accessibility constrain phenotypic space during *Drosophila* development. *Dev Cell* **58:** 51–62.e4. doi:10.1016/j.devcel.2022.12.003

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* **433:** 481–487. doi:10.1038/nature03235

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004

Janssens J, Aibar S, Taskiran II, Ismail JN, Spanier KI, González-Blas CB, Quan XJ, Papasokrati D, Hulselmans G, Makhzami S, et al. 2022. Decoding gene regulation in the fly brain. *Nature* **601:** 630–636. doi:10.1038/s41586-021-04262-z

Jindal GA, Farley EK. 2021. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev Cell* **56:** 575–587. doi:10.1016/j.devcel.2021.02.016

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130:** 6569–6575. doi:10.1242/dev.00890

Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A. 2014. Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. *Nature* **512:** 91–95. doi:10.1038/nature13395

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi:10.1186/gb-2009-10-3-r25

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20:** R754–R763. doi:10.1016/j.cub.2010.06.070
Li R, Pei H, Watson DK. 2000. Regulation of Ets function by protein–protein interactions. *Oncogene* **19:** 6514–6523. doi:10.1038/sj.onc.1204035
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8
Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* **125:** 949–958. doi:10.1242/dev.125.5.949
Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567. doi:10.1038/35000615
Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW. 2016. DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst* **3:** 278–286.e4. doi:10.1016/j.cels.2016.07.001
May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, et al. 2012. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44:** 89–93. doi:10.1038/ng.1006
Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR, et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15:** 141–149. doi:10.1038/nmeth.4534
Neumayr C, Pagani M, Stark A, Arnold CD. 2019. STARR-seq and UMI-STARR-seq: assessing enhancer activities for genome-wide-, high-, and low-complexity candidate libraries. *Curr Protoc Mol Biol* **128:** e105. doi:10.1002/cpmb.105
Panne D. 2008. The enhanceosome. *Curr Opin Struct Biol* **18:** 236–242. doi:10.1016/j.sbi.2007.12.002
Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, Hadzhiev Y, Thies WG, Scherer G, Strähle U. 2008. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* **318:** 366–377. doi:10.1016/j.ydbio.2008.03.034
R Core Team. 2020. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Reiter F, Wienerroither S, Stark A. 2017. Combinatorial function of transcription factors and cofactors. *Curr Opin Genet Dev* **43:** 73–81. doi:10.1016/j.gde.2016.12.007
Rickels R, Shilatifard A. 2018. Enhancer logic and mechanics in development and disease. *Trends Cell Biol* **28:** 608–630. doi:10.1016/j.tcb.2018.04.003
Samee MAH, Bruneau BG, Pollard KS. 2019. A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst* **8:** 27–42.e6. doi:10.1016/j.cels.2018.12.001
Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature* **533:** 397–401. doi:10.1038/nature17995
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328:** 1036–1040. doi:10.1126/science.1186176
Somermeyer LG, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, Alaball Pujol M-E, Putintseva EV, Sarkisyan KS, Kondrashov FA. 2022. Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11:** e75842. doi:10.7554/eLife.75842
Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13:** 613–626. doi:10.1038/nrg3207
Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr Biol* **21:** 1186–1196. doi:10.1016/j.cub.2011.05.056
Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* **21:** 1139–1149. doi:10.1101/gr.119016.110
Thanos D, Maniatis T. 1995. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83:** 1091–1100. doi:10.1016/0092-8674(95)90136-1
Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. 2022. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603:** 455–463. doi:10.1038/s41586-022-04506-6
Verfaillie A, Svetlichnyy D, Imrichova H, Davie K, Fiers M, Atak ZK, Hulselmans G, Christiaens V, Aerts S. 2016. Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res* **26:** 882–895. doi:10.1101/gr.204149.116
Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160:** 554–566. doi:10.1016/j.cell.2015.01.006
Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* **461:** 199–205. doi:10.1038/nature08451
Vockley CM, McDowell IC, D'Ippolito AM, Reddy TE. 2017. A long-range flexible billboard model of gene activation. *Transcription* **8:** 261–267. doi:10.1080/21541264.2017.1317694
Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: The more things change, the more they stay the same. *Trends Genet* **26:** 66–74. doi:10.1016/j.tig.2009.12.002
Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York. ISBN 978-3-319-24277-4, http://ggplot2.org.
Wong ES, Zheng D, Tan SZ, Bower NI, Garside V, Vanwalleghem G, Gaiti F, Scott E, Hogan BM, Kikuchi K, et al. 2020. Deep conservation of the enhancer regulatory code in animals. *Science* **370:** eaax8137. doi:10.1126/science.aax8137
Xiong W-C, Montell C. 1993. *tramtrack* is a transcriptional repressor required for cell fate determination in the *Drosophila* eye. *Genes Dev* **7:** 1085–1096. doi:10.1101/gad.7.6.1085
Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518:** 556–559. doi:10.1038/nature13994

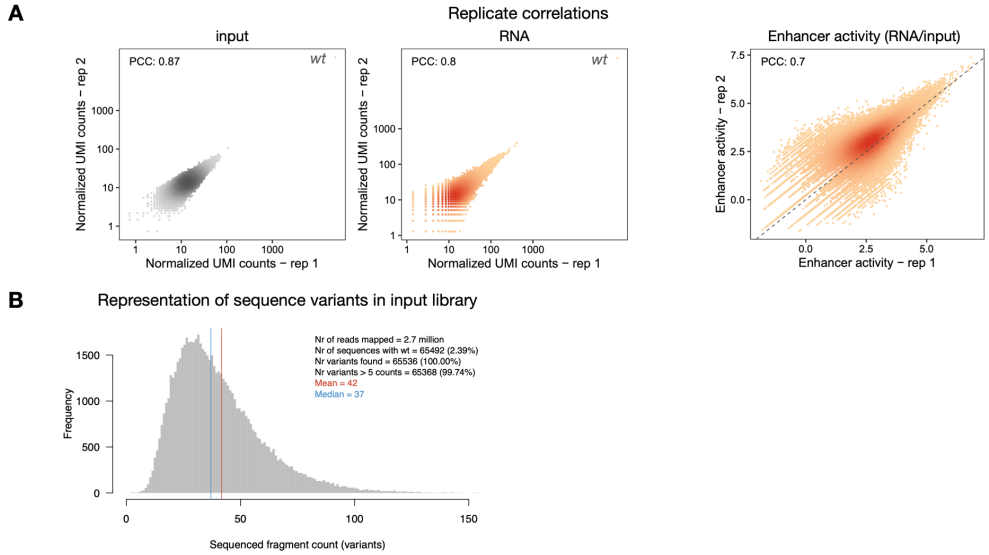
Received August 26, 2022; accepted in revised form February 14, 2023.

Supplemental Information

Table of Contents

Supplemental Figures

Supplemental Fig S1. STARR-seq comprehensively assesses the activity of random variants in a specific region of the enhancer.



A) Pairwise comparisons of normalized STARR-seq input (left) and RNA (middle) UMI read counts or enhancer activity (RNA/input; right) between two independent biological replicates across all sequence variants tested in the GATA position (pos241) in the *ced-6* enhancer. Color reflects point density. The PCC is denoted for each comparison. Note the overrepresentation of the wild-type sequence both in the input and RNA libraries (top right corner), since it was used as the template for the PCR cloning (see Methods). **B)** Representation of sequence variants in STARR-seq input library. Frequency of variants covered by different number of UMI read counts. Number of sequences matching to wild type and the number of variants recovered are shown, together with the mean and median counts sequenced per variant.

Supplemental Fig S2. *De novo* motif discovery with Homer of top and bottom variants at the GATA position (pos241) in the *ced-6* enhancer.

A Top 100 variants

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-25	-5.813e+01	15.00%	0.13%	GATA
2		1e-25	-5.784e+01	20.00%	0.50%	SREBP
3		1e-18	-4.300e+01	19.00%	0.93%	AP-1

B Top 1000 variants

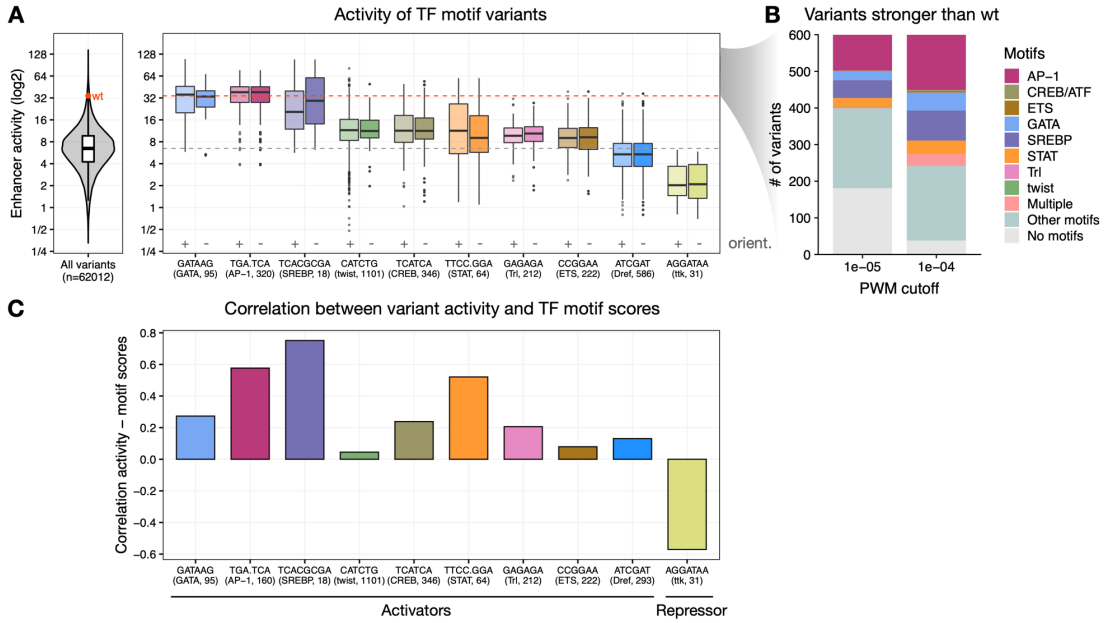
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-154	-3.554e+02	21.70%	1.72%	GATA
2		1e-144	-3.335e+02	19.70%	1.45%	AP-1
3		1e-123	-2.849e+02	17.70%	1.42%	sd
4		1e-51	-1.192e+02	9.00%	0.99%	SREBP
5		1e-25	-5.774e+01	2.10%	0.04%	Rfx
6		1e-24	-5.572e+01	5.10%	0.75%	Stat92E
7		1e-22	-5.249e+01	4.90%	0.74%	E-box/twist
8		1e-13	-3.110e+01	2.80%	0.41%	E-box/twist

C Bottom 1000 variants

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-86	-1.994e+02	22.40%	4.37%	ttk
2		1e-37	-8.739e+01	13.50%	3.51%	H15
3		1e-30	-6.983e+01	5.90%	0.77%	?
4		1e-25	-5.887e+01	14.40%	5.29%	ZEB1/zfh-1
5		1e-22	-5.096e+01	5.50%	0.99%	Nf-YB
6		1e-21	-5.004e+01	9.40%	2.86%	HNF
7		1e-13	-3.143e+01	1.90%	0.15%	?

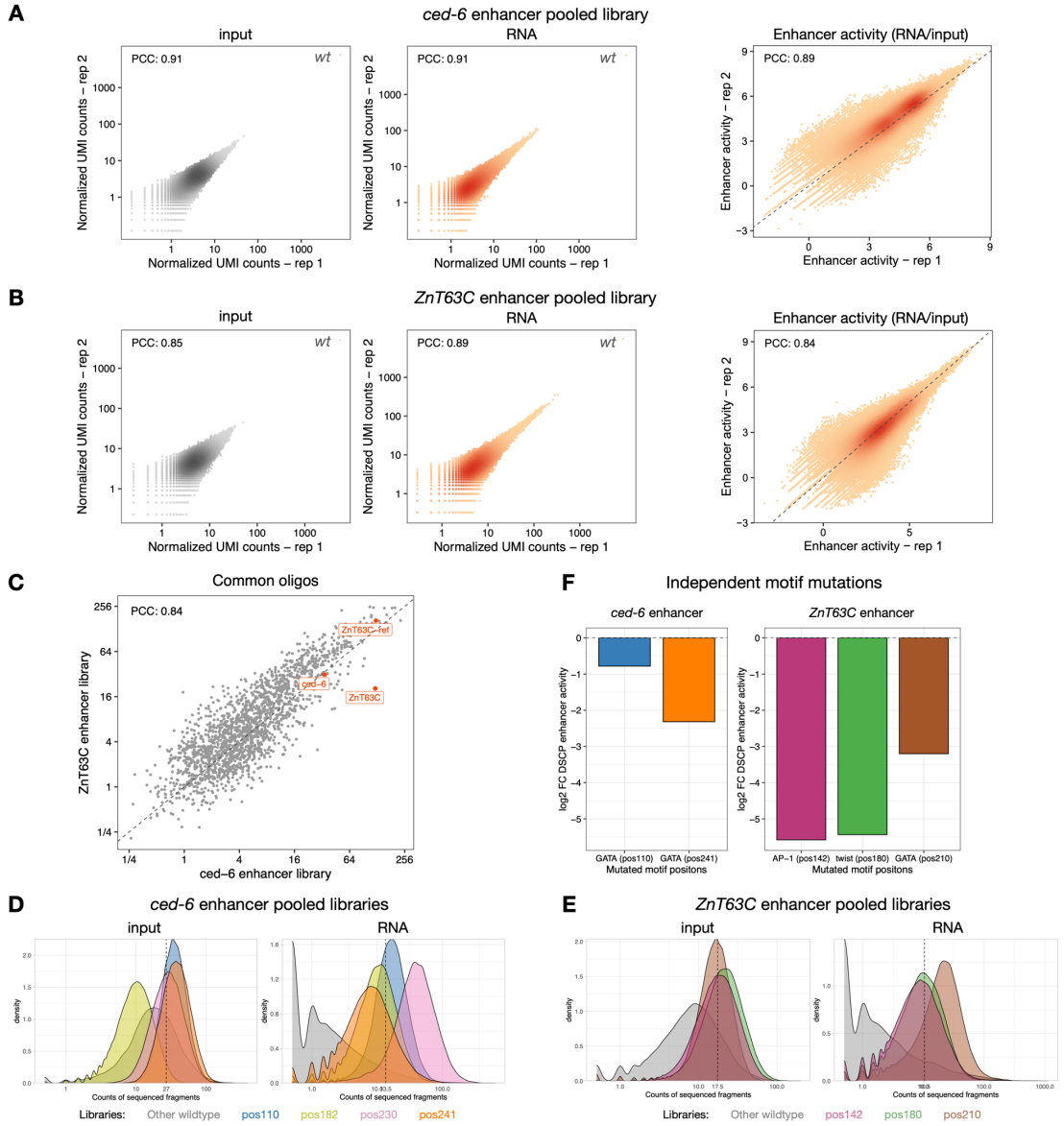
TF motifs found *de novo* (Homer) within the top 100 **(A)**, top 1,000 **(B)** or bottom 1,000 **(C)** variants. Motifs logo, statistics and predicted TF are shown.

Supplemental Fig S3. Activity of variants creating different TF motif types at the GATA position (pos241) in the *ced-6* enhancer.



A) Distribution of enhancer activity for all 62,012 enhancer variants (left) or variants creating each TF motif in either orientation (right; positive and negative orientation are shown in grey). The motif activities are independent of their orientation (Wilcoxon rank sum test p-value > 0.05). The activity of the wild-type sequence (wt, red dot and dashed line) or median of all variants (grey dashed line) are highlighted. The string of each TF motif used for the motif matching and the number of variants matching to each motif are described in the x-axis in the format "motif string (TF motif name, number of variants)". **B)** Number of variants among the 600 stronger than wild type that match to motifs enriched in S2 developmental enhancers, using two different PWM p-value cutoffs ($1e^{-05}$ and $1e^{-04}$). **C)** Pearson correlation coefficient between variant activity and TF motif PWM scores. Note that for repressors, as ttk, the correlation is expected to be negative.

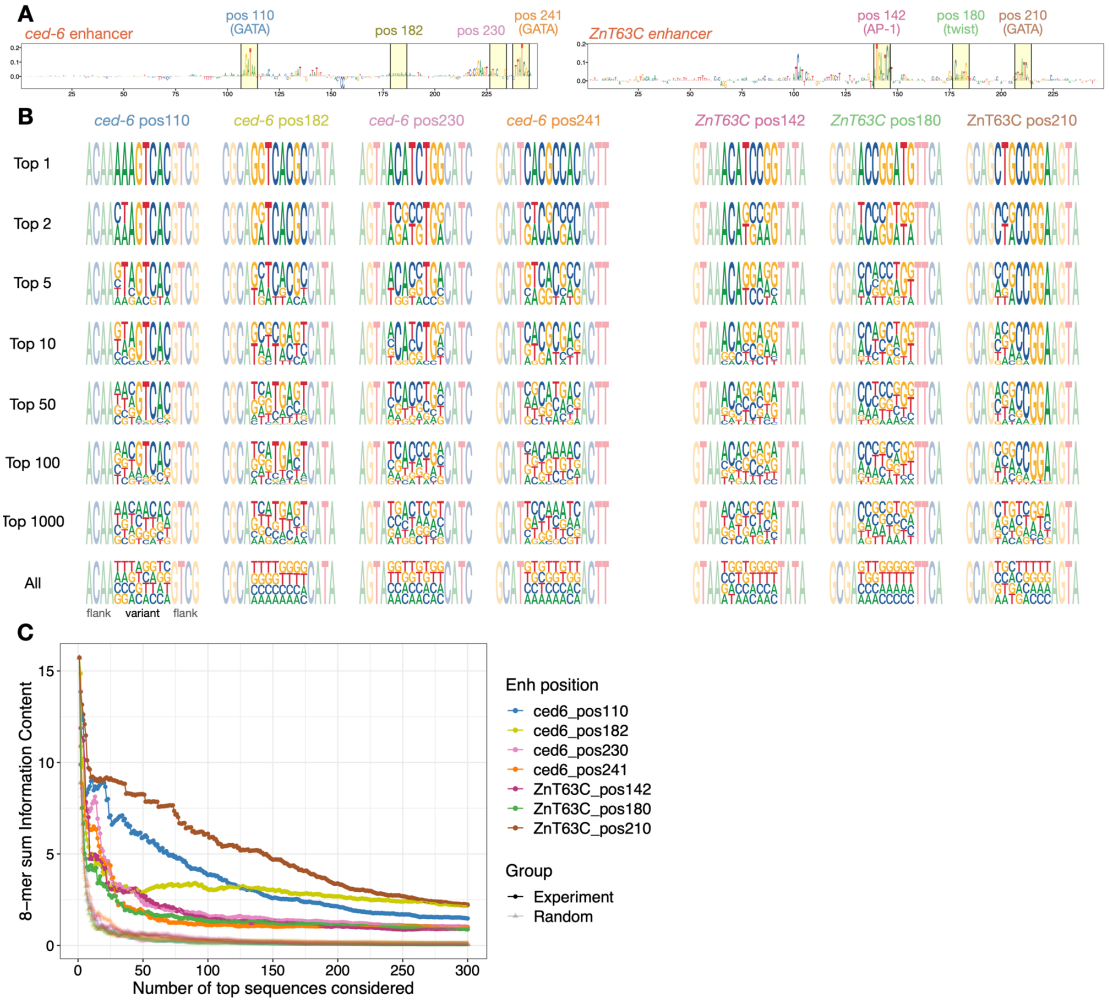
Supplemental Fig S4. STARR-seq screens with random variants in seven positions of two different enhancers.



A, B) Pairwise comparisons of normalized STARR-seq input (left) and RNA (middle) UMI read counts or enhancer activity (RNA/input; right) between two independent biological replicates across all sequence variants tested in positions of the *ced-6* **(A)** or *ZnT63C* **(B)** enhancer. Color reflects point density. The PCC is denoted for each comparison. Note the overrepresentation of the wild-type sequence both in the input and RNA libraries (top right corner), since it was used as the template for the PCR cloning (see Methods). **C)** Comparison of enhancer activity between the two different enhancer pooled libraries for the common oligos (a library of wild-type enhancer or negative sequences; see Methods). The PCC is shown. The respective wild-type enhancers are highlighted. Given the underestimation of the activity of the *ZnT63C* wild-type sequence in its pooled library, we used as reference wildt-ype activity the activity of another enhancer with similar activity that was conserved in both libraries (see Methods). **D, E)** Representation of sequence variants from each individual library (a library of wild-type enhancer and negative sequences, grey, or libraries with random variants in each enhancer position, different colors) in STARR-seq input and RNA pooled libraries of the *ced-6* **(D)** or *ZnT63C* **(E)** enhancer. The mean counts sequenced per variant is shown per pooled library with a dashed line. **F)** Importance of each motif position selected in the *ced-6* (Left) or *ZnT63C*

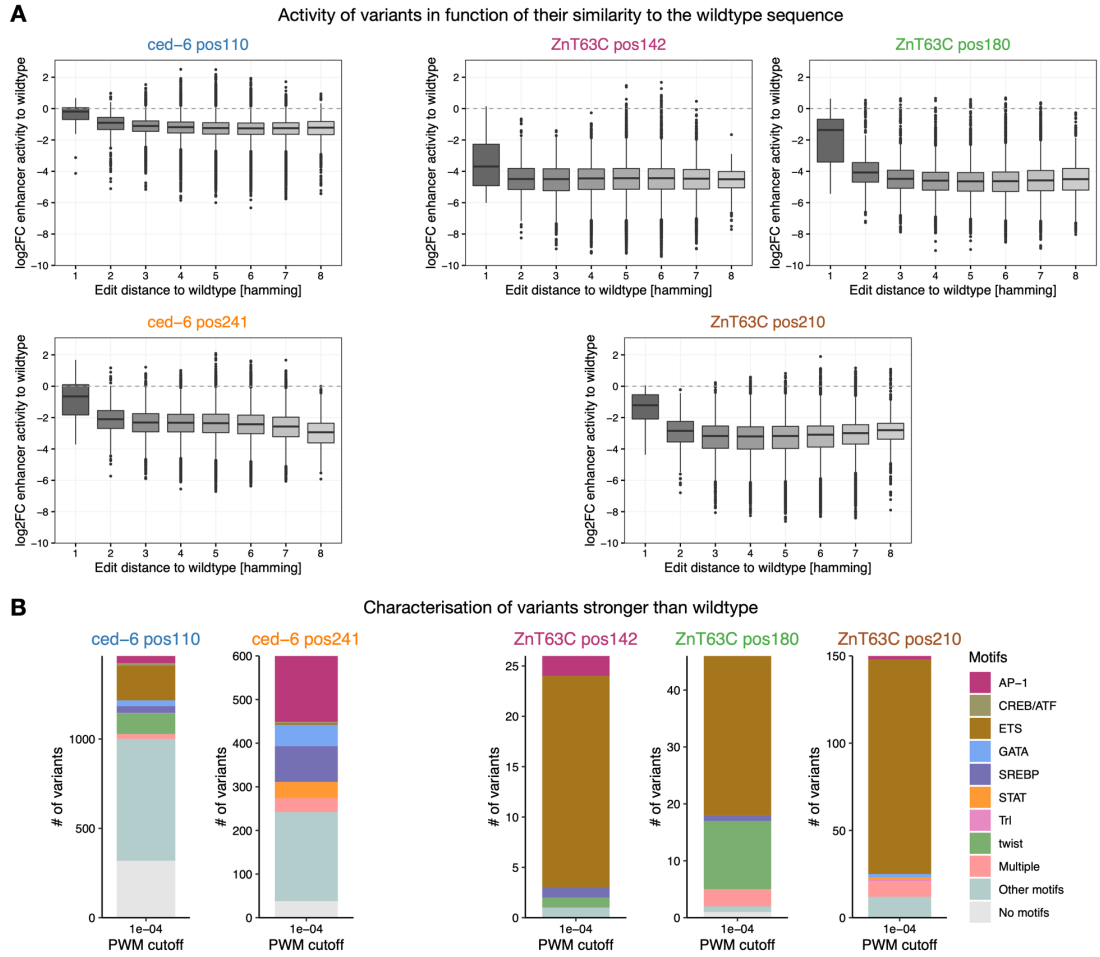
(Right) enhancer as judged by the impact of their individual mutation in enhancer activity (log_2 fold-change). Data retrieved from *de Almeida et al., 2022* (de Almeida et al. 2022).

Supplemental Fig S5. Top active variants at each enhancer position are highly diverse.



A) DeepSTARR-predicted nucleotide contribution scores for the *ced-6* (left) and *ZnT63C* (right) selected enhancer sequences. Selected 8nt motif positions and non-important control positions are highlighted in yellow with the respective numerical position, TF motif identity and different colors. **B)** Strong sequence variants are highly diverse. Logos with nucleotide frequency of the most-active variants in STARR-seq (1, 2, 5, 10, 50, 100, 1,000 and all) at each enhancer position (colored as in (A)). **C)** Sum of information content within the most-active 8-mers in STARR-seq (colored as in (A)) compared with the same after randomly sorting the variants (grey) for each enhancer position, considering different number of top sequences.

Supplemental Fig S6. Characterization of active variants.



A) Log_2 fold-change enhancer activity over the wild-type activity for all enhancer variants grouped by their edit distance (hamming distance) to the wild-type sequence, per enhancer position. **B)** Number of variants stronger than wild type that match to motifs enriched in S2 developmental enhancers is shown (PWM p-value cutoff $1e^{-04}$), per enhancer position.

Supplemental Fig S7. *De novo* motif discovery with Homer of the top 1000 variants at the different enhancer positions.

A *ced-6* pos110

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-133	-3.076e+02	21.10%	2.07%	SREBP
2		1e-106	-2.459e+02	16.50%	1.52%	ETS
3		1e-88	-2.040e+02	14.70%	1.52%	GATA
4		1e-71	-1.653e+02	12.90%	1.51%	AP-1
5		1e-66	-1.527e+02	12.10%	1.45%	E-box/twist
6		1e-23	-5.449e+01	3.10%	0.21%	ERR
7		1e-14	-3.376e+01	1.40%	0.04%	odd
8		1e-13	-3.162e+01	3.40%	0.62%	GATA

B *ced-6* pos182

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-215	-4.966e+02	35.60%	3.99%	AP-1
2		1e-55	-1.272e+02	9.70%	1.06%	E-box/twist
3		1e-43	-9.940e+01	8.30%	1.04%	GATA
4		1e-22	-5.226e+01	9.60%	2.86%	sd
5		1e-21	-5.017e+01	4.30%	0.57%	SREBP

C *ced-6* pos230

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-83	-1.914e+02	42.30%	16.09%	?
2		1e-82	-1.891e+02	8.80%	0.35%	ERR
3		1e-63	-1.467e+02	17.60%	3.64%	AP-1
4		1e-56	-1.312e+02	9.60%	1.00%	E-box/twist
5		1e-55	-1.280e+02	21.50%	6.16%	?
6		1e-55	-1.276e+02	5.10%	0.13%	ETS
7		1e-36	-8.407e+01	13.70%	3.74%	CREB/Atf
8		1e-33	-7.761e+01	8.80%	1.67%	SREBP
9		1e-14	-3.431e+01	1.80%	0.11%	SOX

D *ced-6* pos241

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-154	-3.554e+02	21.70%	1.72%	GATA
2		1e-144	-3.335e+02	19.70%	1.45%	AP-1
3		1e-123	-2.849e+02	17.70%	1.42%	sd
4		1e-51	-1.192e+02	9.00%	0.99%	SREBP
5		1e-25	-5.774e+01	2.10%	0.04%	Rfx
6		1e-24	-5.572e+01	5.10%	0.75%	Stat92E
7		1e-22	-5.249e+01	4.90%	0.74%	E-box/twist
8		1e-13	-3.110e+01	2.80%	0.41%	E-box/twist

E *ZnT63C* pos142

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-117	-2.694e+02	18.80%	1.85%	ETS
2		1e-72	-1.661e+02	12.90%	1.48%	E-box/twist
3		1e-49	-1.142e+02	21.20%	6.55%	Fkh
4		1e-44	-1.024e+02	7.00%	0.64%	GATA
5		1e-41	-9.605e+01	10.60%	1.95%	SREBP
6		1e-37	-8.690e+01	11.40%	2.53%	Fkh
7		1e-29	-6.759e+01	3.00%	0.10%	AP-1
8		1e-14	-3.369e+01	2.70%	0.33%	?
9		1e-14	-3.226e+01	1.40%	0.05%	ERR

F *ZnT63C* pos180

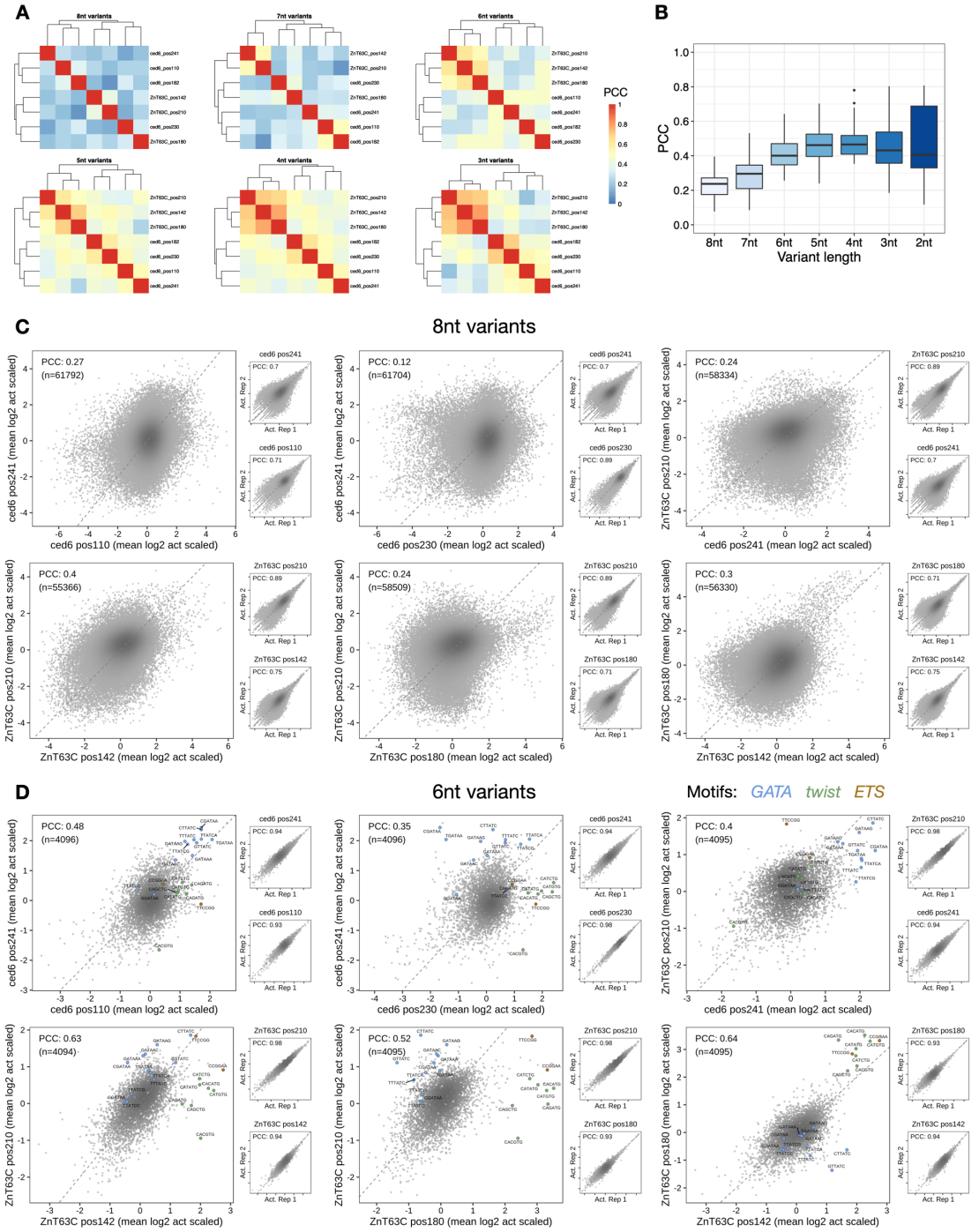
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-169	-3.911e+02	38.30%	6.89%	E-box/twist
2		1e-146	-3.378e+02	22.50%	2.08%	ETS
3		1e-78	-1.810e+02	16.20%	2.32%	AP-1
4		1e-27	-6.447e+01	3.90%	0.29%	SREBP
5		1e-27	-6.353e+01	10.50%	2.88%	?
6		1e-24	-5.554e+01	3.80%	0.36%	ERR

G *ZnT63C* pos210

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	
1		1e-178	-4.112e+02	25.70%	2.16%	ETS
2		1e-98	-2.264e+02	22.40%	3.77%	GATA
3		1e-97	-2.243e+02	18.40%	2.36%	NHLH1
4		1e-41	-9.623e+01	4.30%	0.15%	AP-1

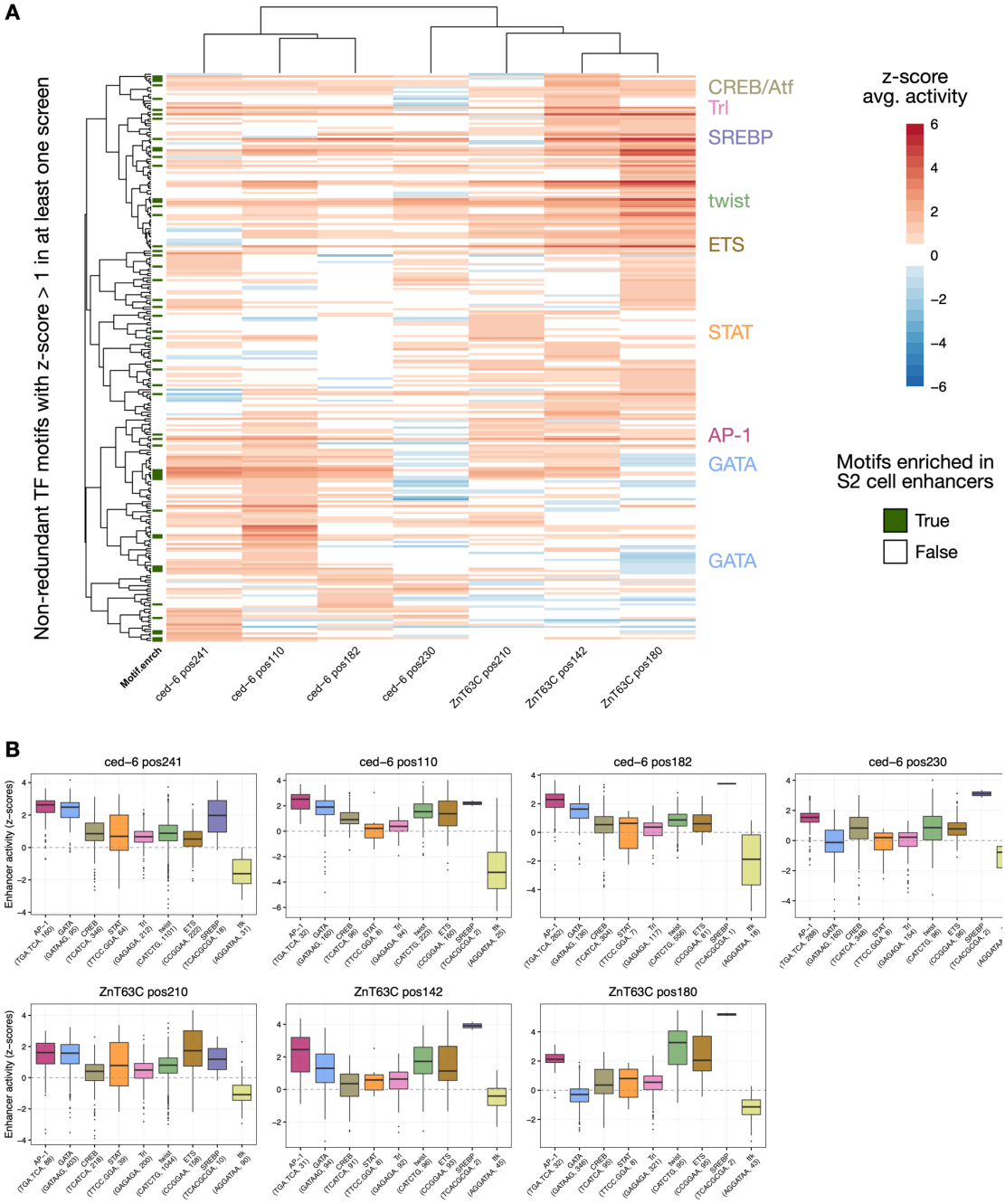
TF motifs found *de novo* (Homer) within the top 1,000 variants at each enhancer position. Motifs logo, statistics and predicted TF are shown.

Supplemental Fig S8. Comparison of all random variants across enhancer positions.



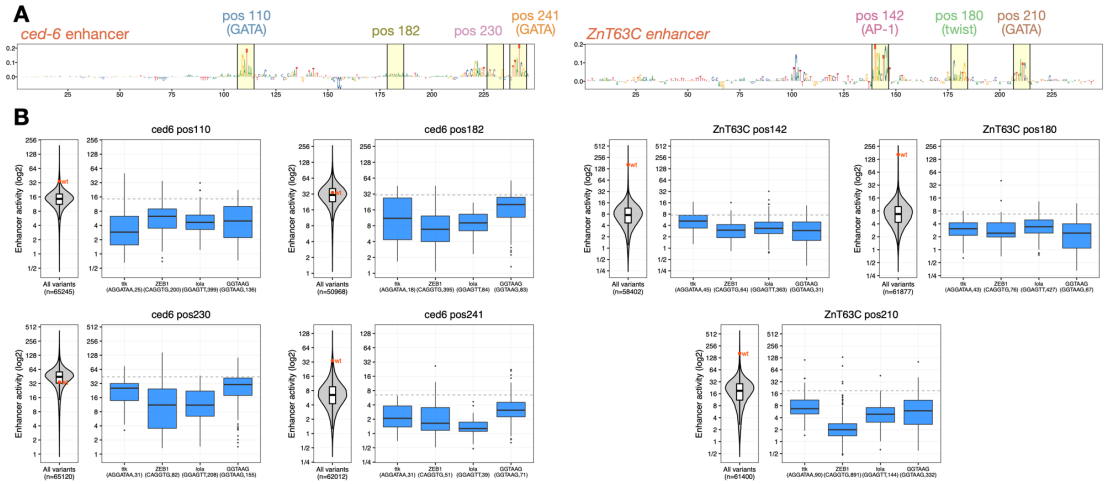
A) Hierarchical clustering of all enhancer positions based on PCC of variant enhancer activities in each position, when considering different lengths of sequence variants (see Methods). **B)** Distribution of PCCs from (A) in function of the length of sequence variants considered. **C,D)** Comparison of *z*-scores of $\log_2$ enhancer activity of all 8nt **(C)** or 6nt **(D**; see Methods**)** variants between enhancer positions (insets show activity for replicates (Act. Rep) 1 versus 2 for each position). Color reflects the enhancer position and point density. PCCs and number of sequence variants are shown. Variants matching to GATA, twist and ETS motifs are highlighted in (D).

Supplemental Fig S9. Activity of TF motif types at different enhancer positions.



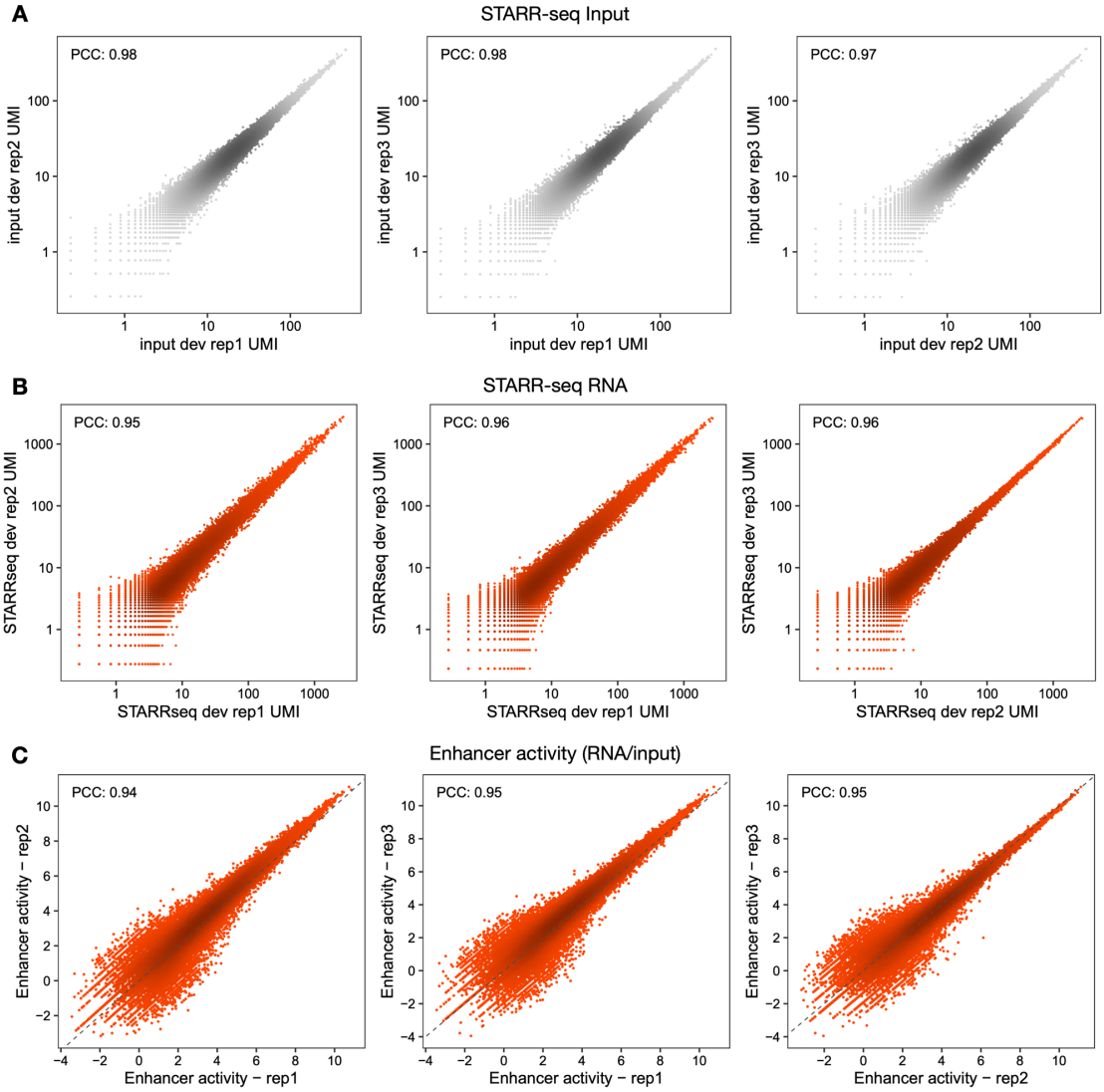
A) Heatmap of average *z*-scores of $\log_2$ enhancer activity of variants creating each TF motif type across all seven enhancer positions. Only motif types active (average *z*-score > 1) in at least one position are shown. Motifs and enhancer positions were clustered using hierarchical clustering and their activity is colored in shades of red (activating) and blue (repressing). Motifs enriched in S2 cell enhancers are labelled in green. Motif types used in the motif pasting experiment are highlighted. **B)** Activity of different TF motifs at each enhancer position. Distribution of *z*-scores of $\log_2$ enhancer activity for variants creating each TF motifs in *ced-6* and *ZnT63C* enhancer positions.

Supplemental Fig S10. STARR-seq identifies known and novel motifs that repress enhancer activity.



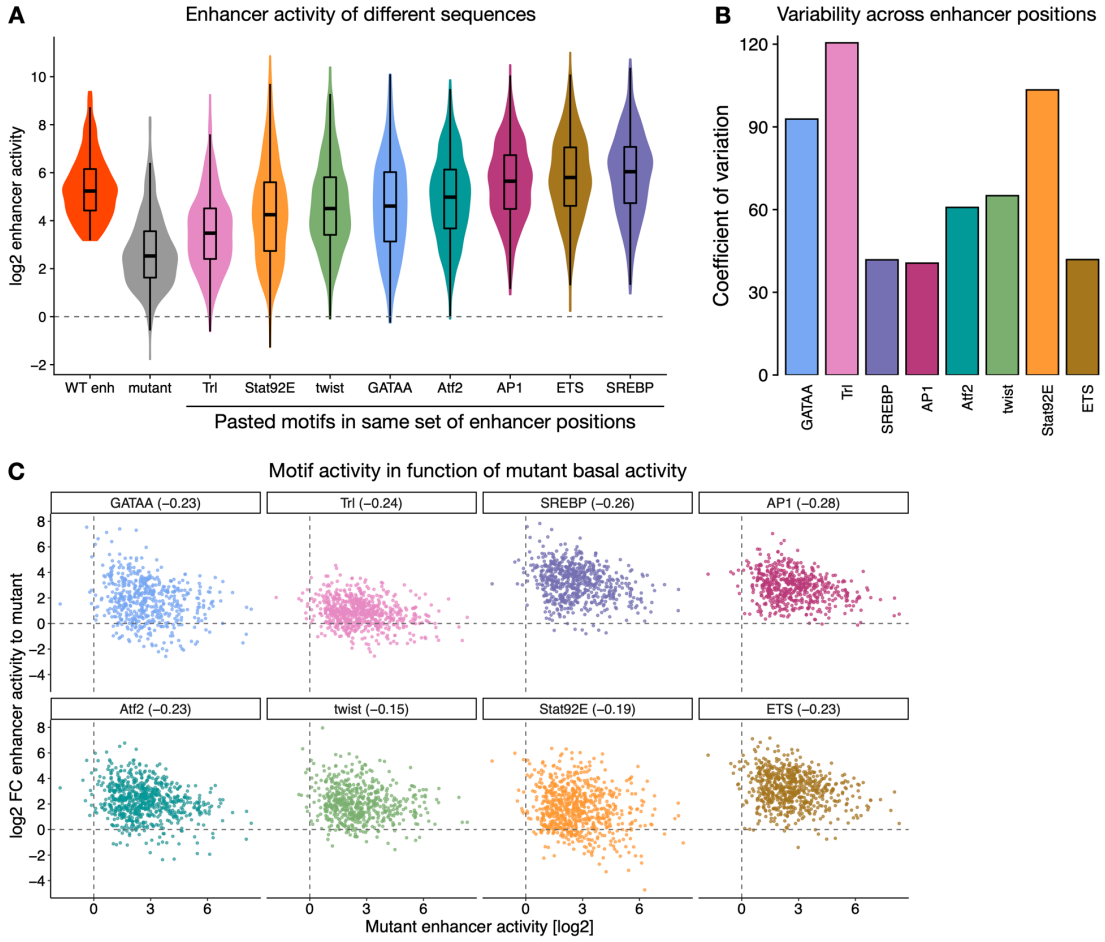
A) DeepSTARR-predicted nucleotide contribution scores for the *ced-6* (left) and *ZnT63C* (right) selected enhancer sequences. Selected 8nt motif positions and non-important control positions are highlighted in yellow with the respective numerical position, TF motif identity and different colors. **B)** Activity of different repressor motifs at each enhancer position. Distribution of enhancer activity for all enhancer variants (left) or variants creating each repressor TF motif (right), per enhancer position. The activity of the wild-type sequence (wt, red) or median of all variants (grey dashed line) are shown. The string of each TF motif used for the motif matching and the number of variants matching to each motif are described in the x-axis: in the format "motif string (TF motif name, number of variants)".

Supplemental Fig S11. Systematic motif pasting screens in *Drosophila* enhancers.



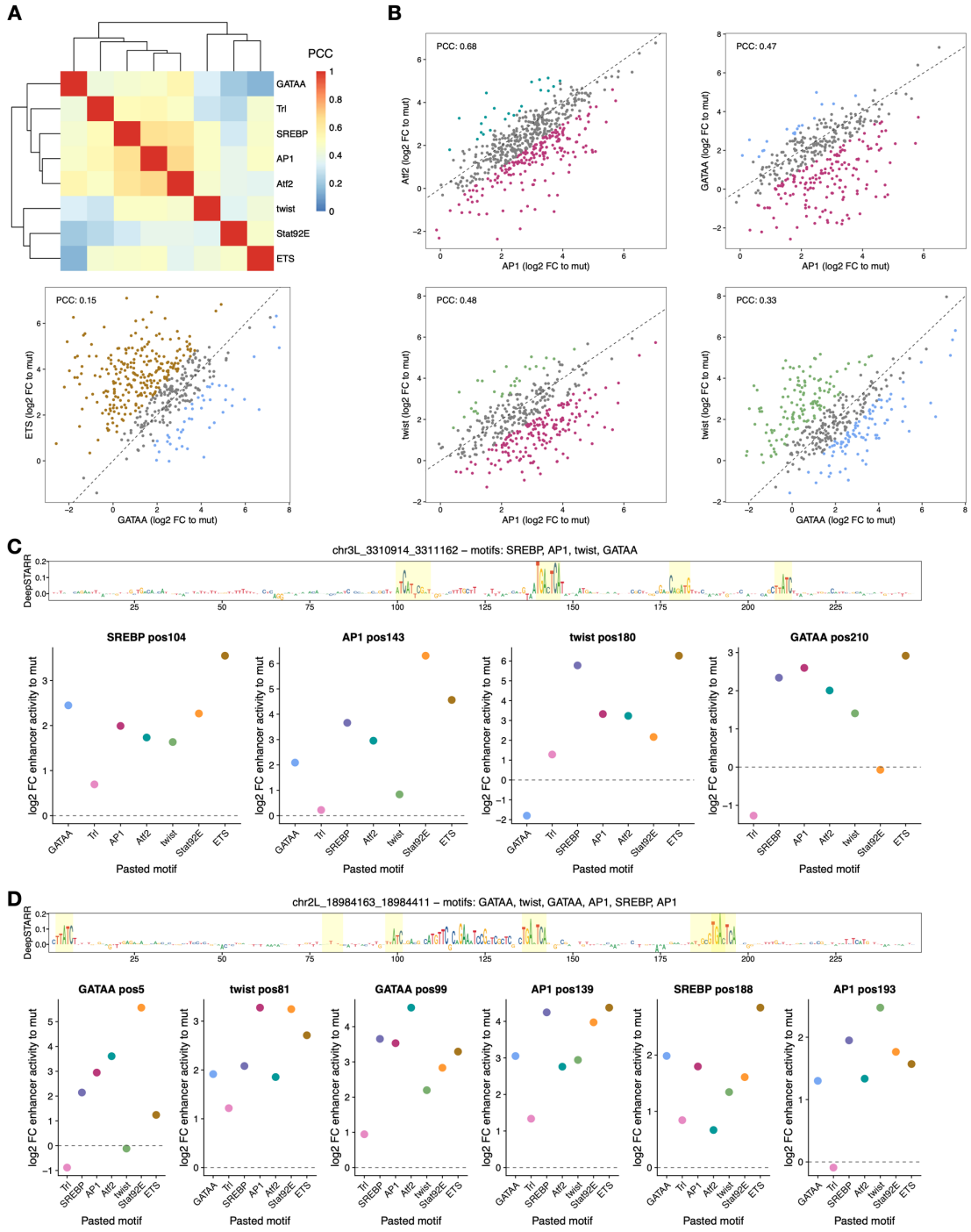
Pairwise comparisons of normalized STARR-seq input **(A)** and RNA **(B)** UMI read counts or enhancer activity (RNA/input) **(C)** between three independent biological replicates across all oligos tested. Color reflects point density. The PCC is denoted for each comparison.

Supplemental Fig S12. Enhancer activity of different sequences in Drosophila enhancers.



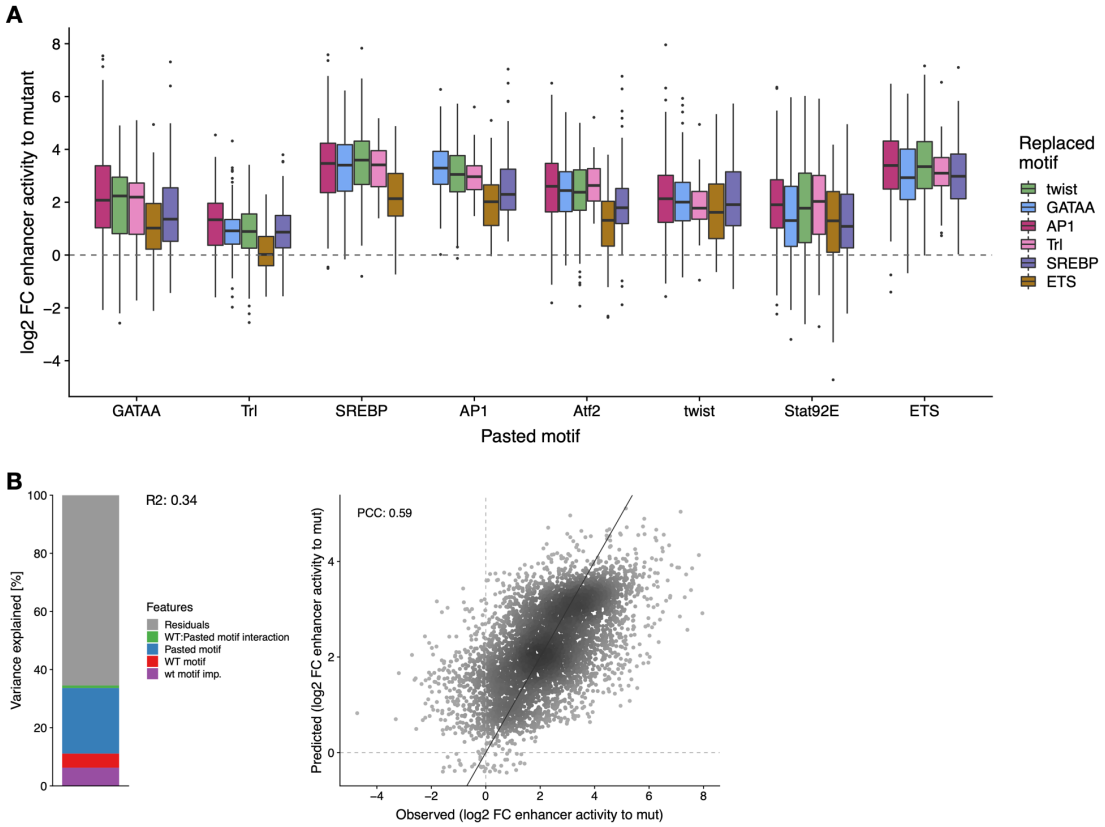
A) Activity of pasted motifs at different enhancer positions. Distribution of enhancer activity changes (log_2) of all wild-type enhancers used and their variants with either mutant sequences or different TF motifs pasted. Few instances show negative values: these are not dependent on the specific mutant sequence but rather correspond to the creation of a repressor motif at the flanks of the pasted motif and the backbone enhancer. **B)** Bar plots showing the coefficient of variation (ratio of the standard deviation to the mean) of the activity of each TF motif across all enhancer positions. **C)** Activity of pasting motifs (y-axis, log_2 fold-change activity over basal motif-mutated enhancer activity) in function of the basal activity (x-axis, activity of motif-mutated enhancer). The PCC is denoted for each motif.

Supplemental Fig S13. Motifs work differently at different enhancer positions.



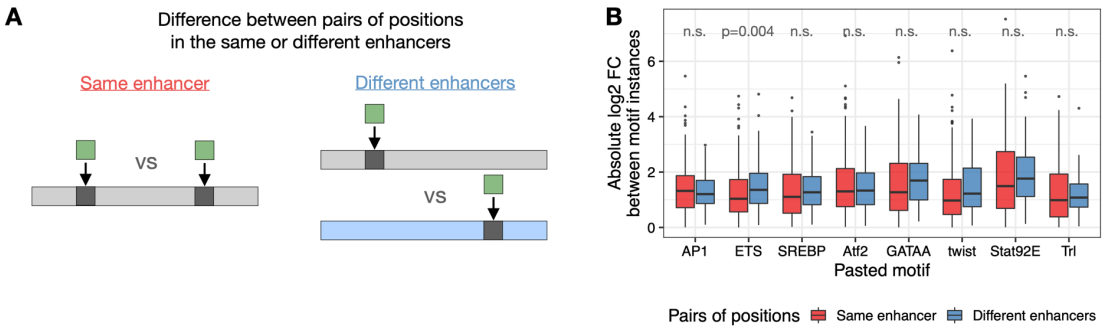
A) Hierarchical clustering of all TF motifs based on PCC of motif activities across all enhancer positions. **B)** Motifs work differently at different enhancer positions. Comparison between enhancer activity changes (log_2 FC to mutated sequence) after pasting different TF motifs across all enhancer positions. Positions with stronger activity of each motif (>= 2-fold in respect to the other motif) are colored with the respective colors. PCC: Pearson correlation coefficient. **C,D)** DeepSTARR-predicted nucleotide contribution scores for two enhancers and respective positions (highlighted in yellow, with wild-type motif types described on top) included in the screen. For each position, the enhancer activity changes (log_2 FC to mutated sequence) after pasting each TF motif are shown in dot plots (bottom).

Supplemental Fig S14. TF motif activity in function of wild-type motif identity.



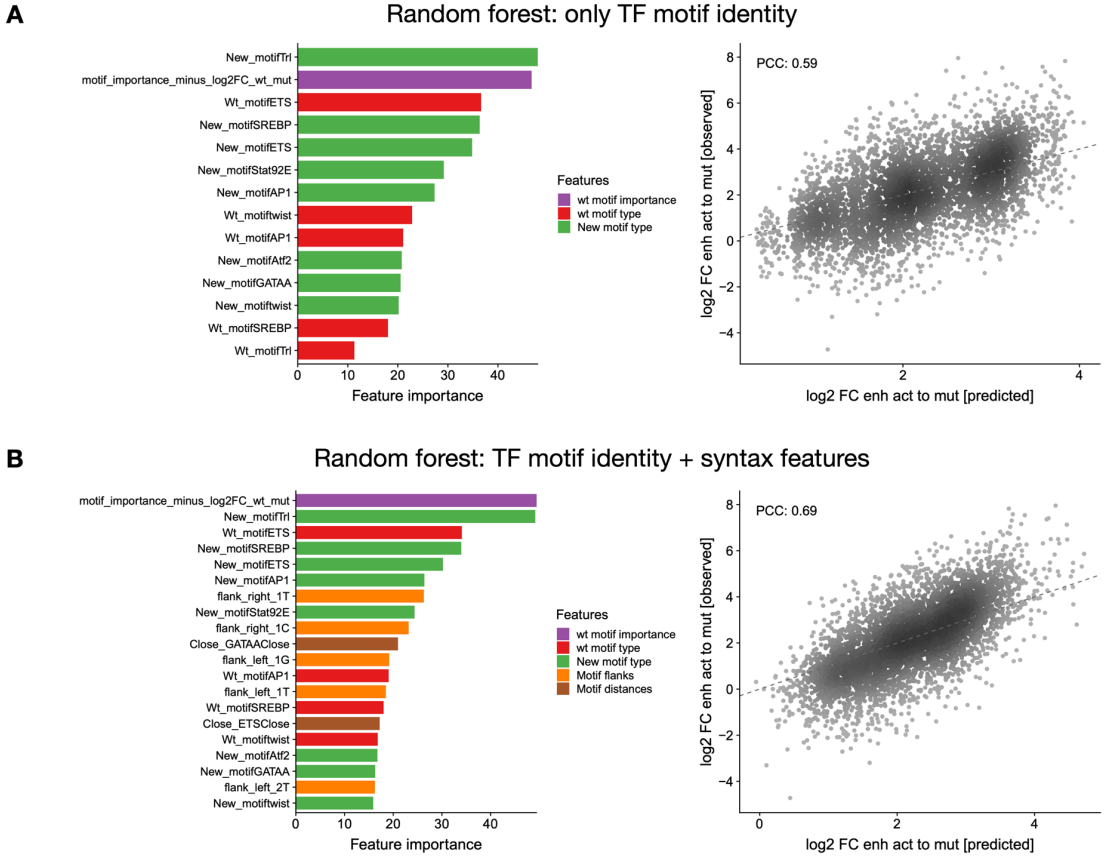
A) Distribution of enhancer activity changes (log_2 FC to mutated sequence) across all enhancer positions for each pasted TF motif, grouped by the identity of the wild-type motif. **B)** Left: Bar plot showing the amount of variance explained by the wild-type motif importance and identity, the pasted motif identity and the interaction between the wild type and pasted motifs, using a linear model fit on all motif pasting results. Right: Scatter plots of predicted (linear model) vs. observed enhancer activity changes (log_2 FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S15. Motif activity in different positions in the same or different enhancers.



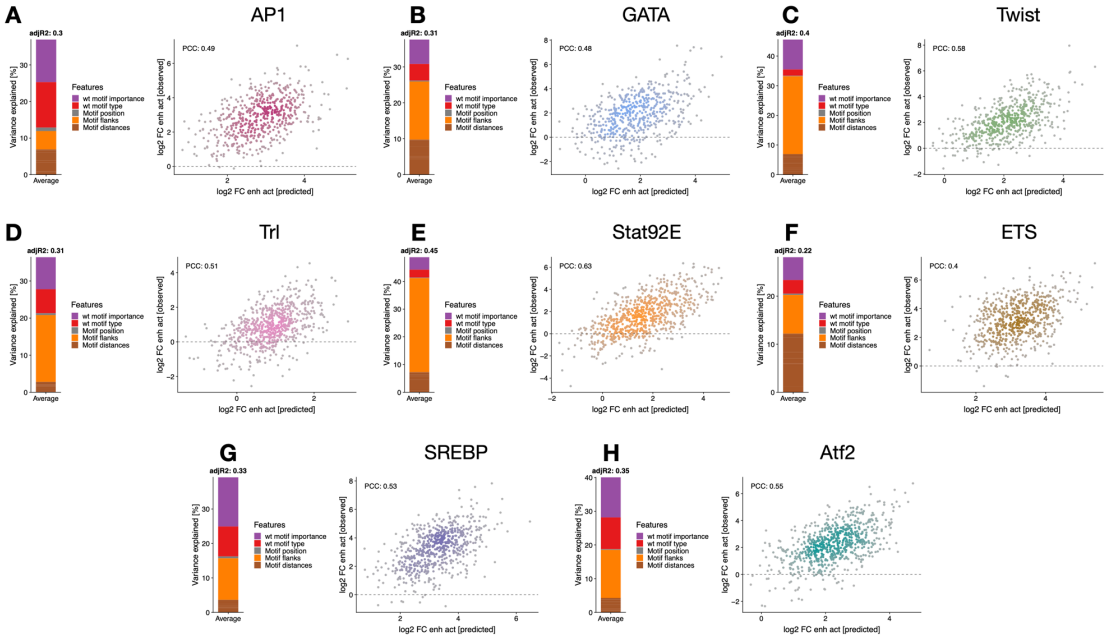
A) Schematics of comparison of motif activity between instances within the same enhancer or in different enhancers. **B)** Absolute log_2 fold-change in enhancer activity between instances within the same enhancer (red) or in different enhancers (blue) for each pasted TF motif type. n.s. non-significant (Wilcoxon signed rank test).

Supplemental Fig S16. Prediction of motif activities using motif syntax features in random forest model.



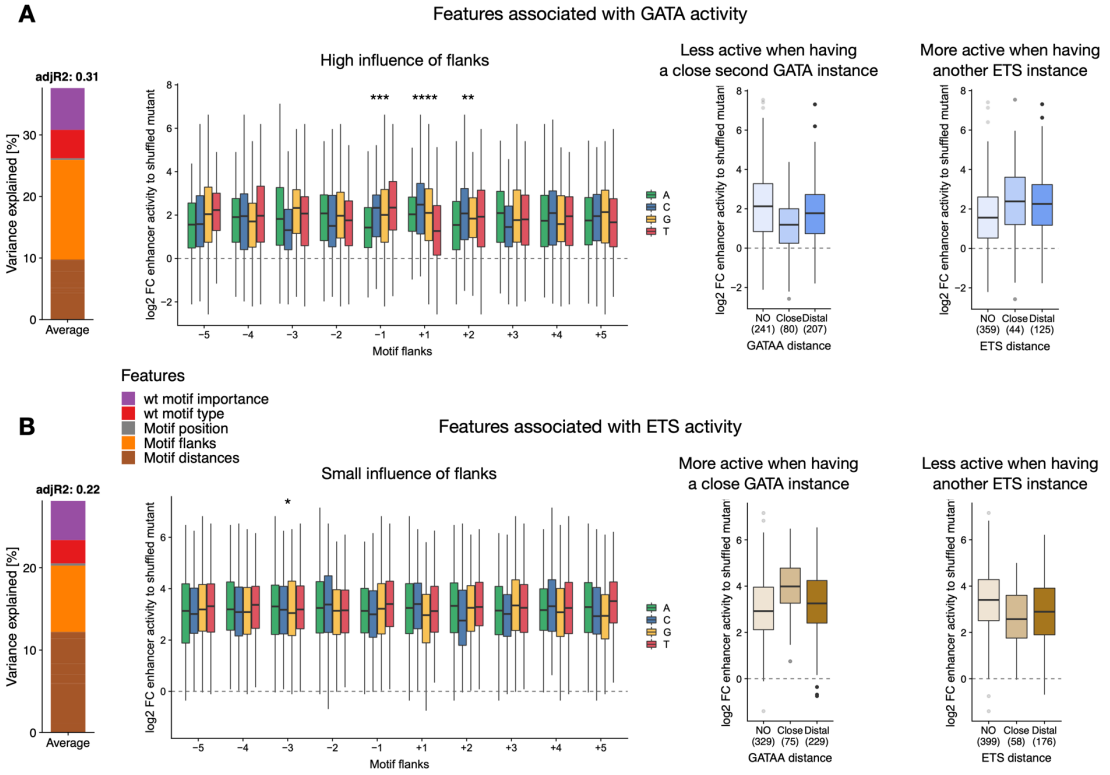
Left: Importance of all features **(A)** or only the top 20 **(B)** included in the random forest models with only TF motif identity **(A)** or also with syntax features **(B)**, sorted by importance and colored by feature type. Right: Scatter plots of predicted vs. observed enhancer activity changes ($\log_2$ FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S17. Linear models with syntax features to predict motif activities.



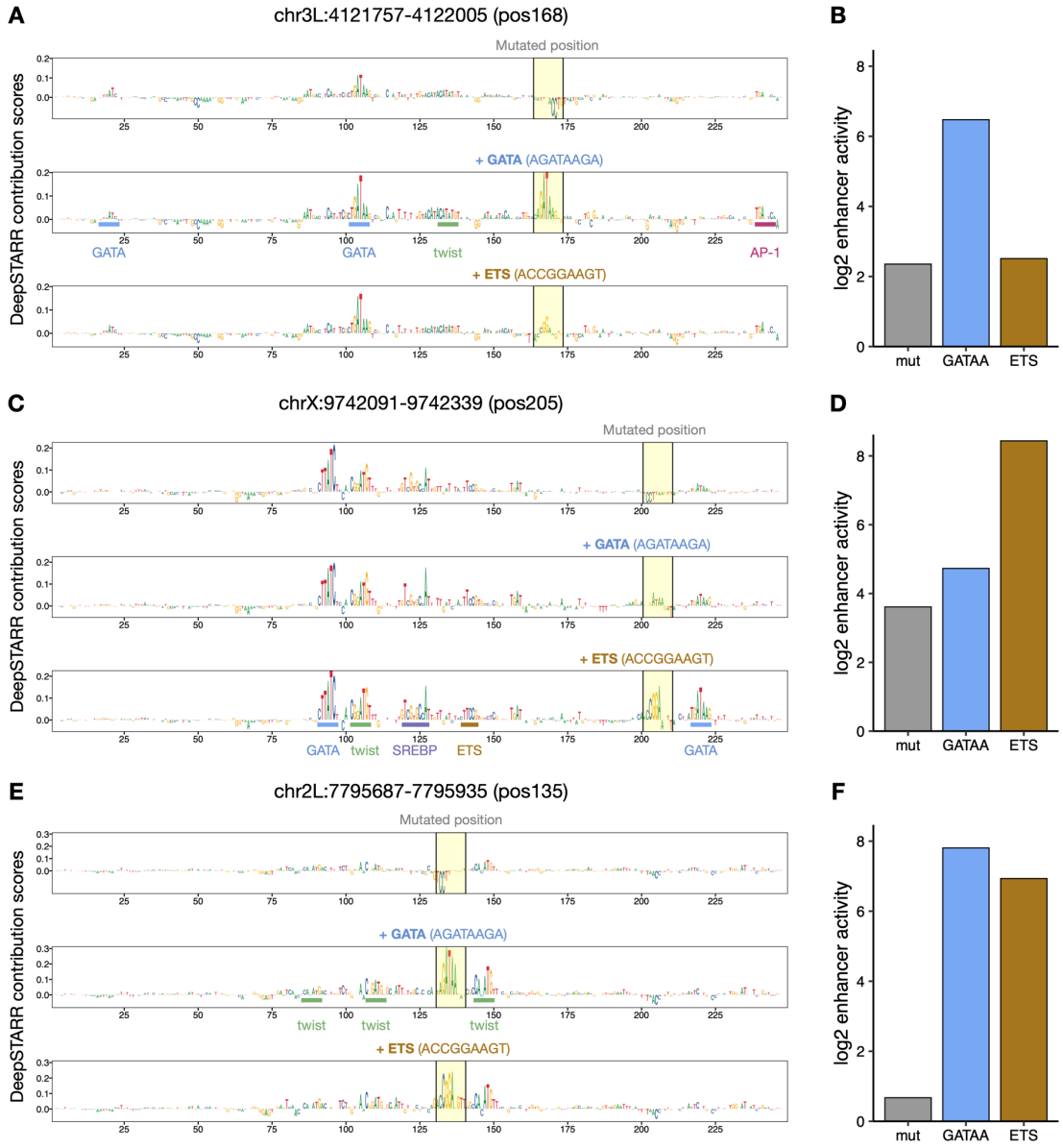
A-H) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Right: Scatter plots of predicted vs. observed enhancer activity changes (log_2 FC to mutated sequence) for motif pasting experiments per TF motif type. Color reflects point density. PCC is shown.

Supplemental Fig S18. Characterization of preferred syntax features of GATA and ETS motifs.



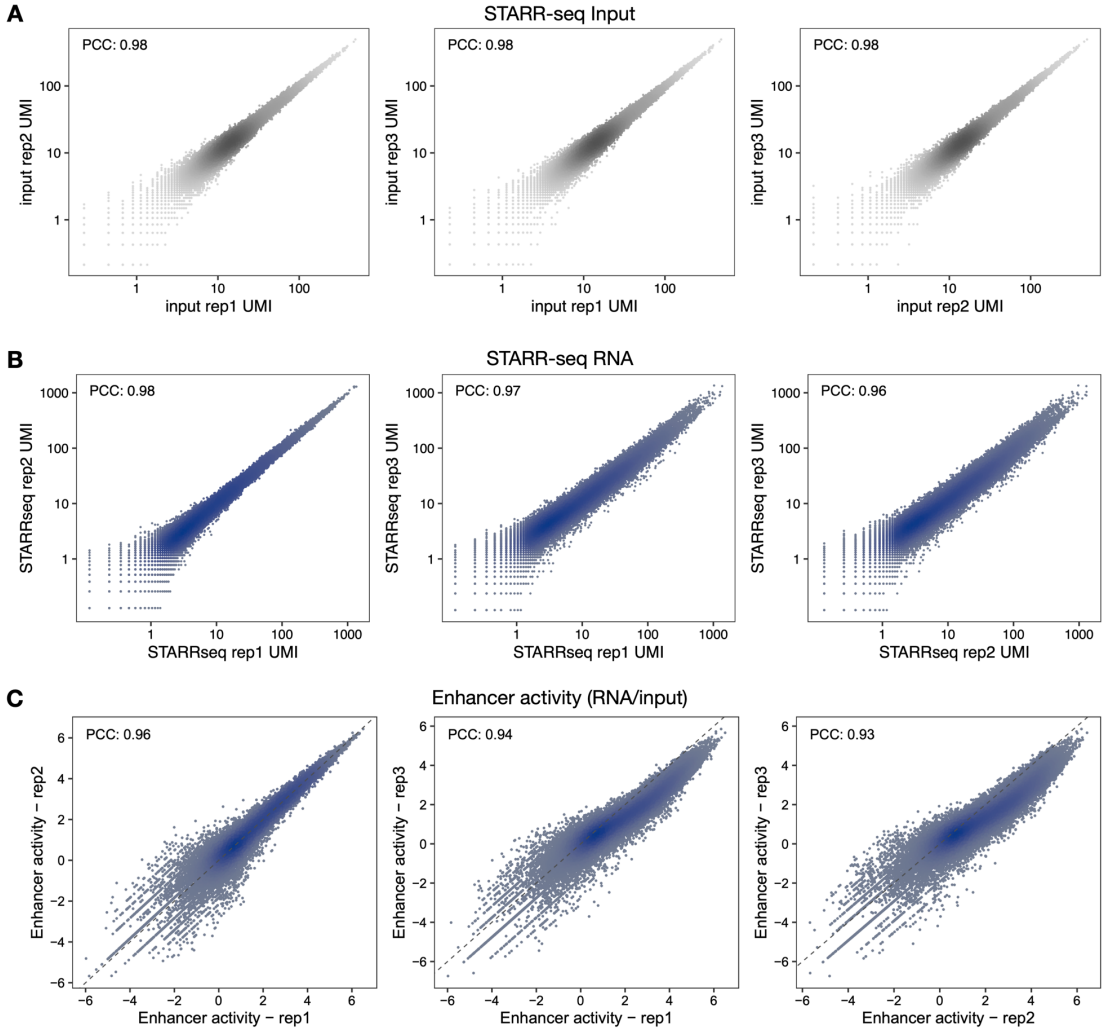
Syntax features associated with GATA **(A)** or ETS **(B)** activity. Left: bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Middle-left: motif activity according to the different bases at each flanking position, colored by nucleotide identity. Statistics from linear model in Fig 4A: ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05 (linear regression p-value). Middle-right and right: enhancer activity changes (log_2 FC to mutated sequence) after pasting each TF motif in positions with no additional GATA (middle-right) or ETS (right) in the enhancer, or with additional GATA or ETS at close (<= 25 bp) or distal (>25 bp) distances. Number of instances are shown.

Supplemental Fig S19. DeepSTARR-predicted importance scores for pasting GATA or ETS in the same positions.



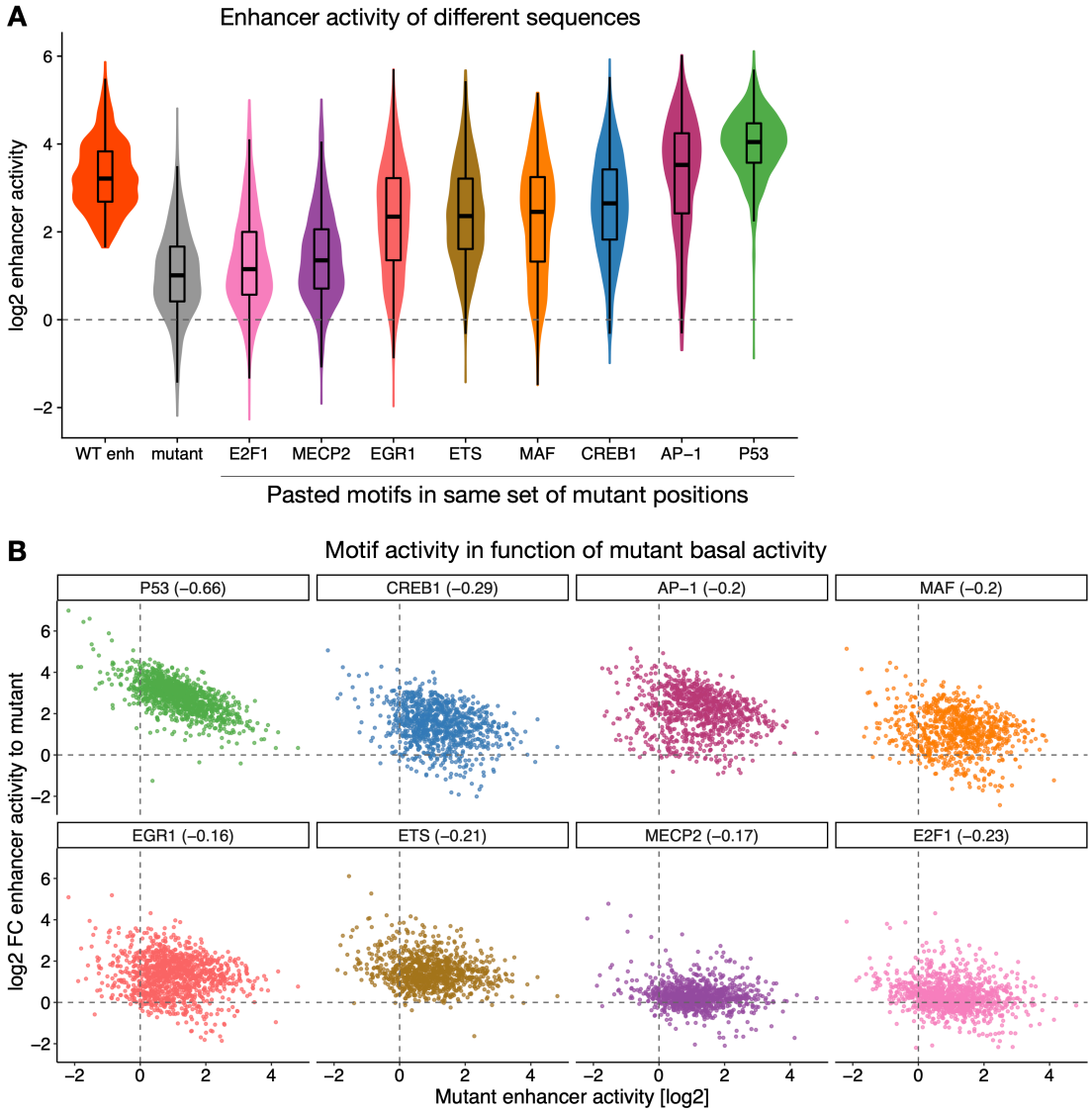
A,C,E) DeepSTARR-predicted nucleotide contribution scores for three different enhancers with a mutant sequence, GATA or ETS pasted at the highlighted positions. Motif sequences pasted are shown. **B,D,F)** Bar plots with enhancer activity (log_2) of variants from (A,C,E).

Supplemental Fig S20. Systematic motif pasting screens in human enhancers.



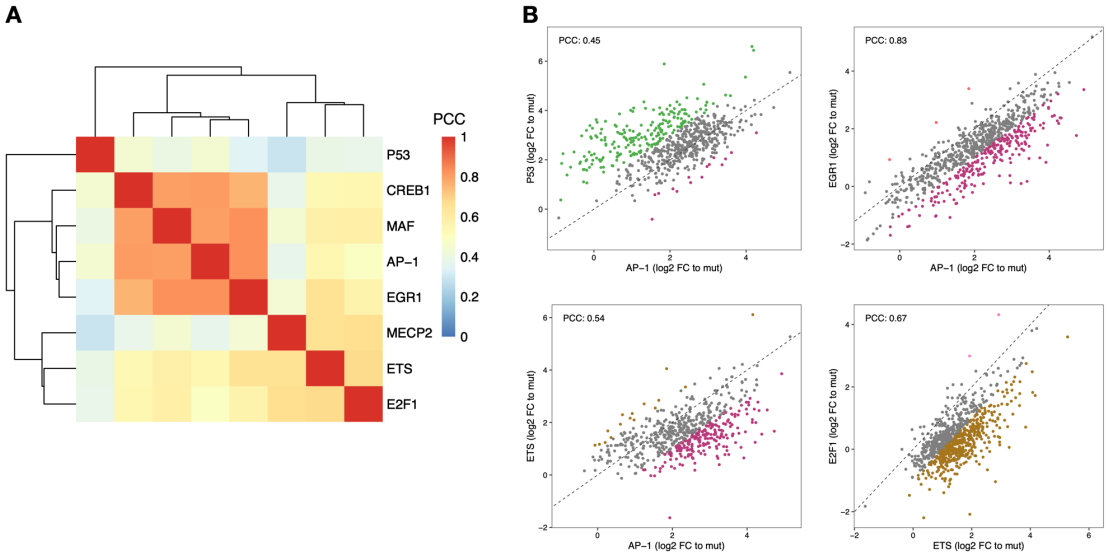
Pairwise comparisons of normalized STARR-seq input **(A)** and RNA **(B)** UMI read counts or enhancer activity (RNA/input) **(C)** between three independent biological replicates across all oligos tested. Color reflects point density. The PCC is denoted for each comparison.

Supplemental Fig S21. Enhancer activity of different sequences in human enhancers.



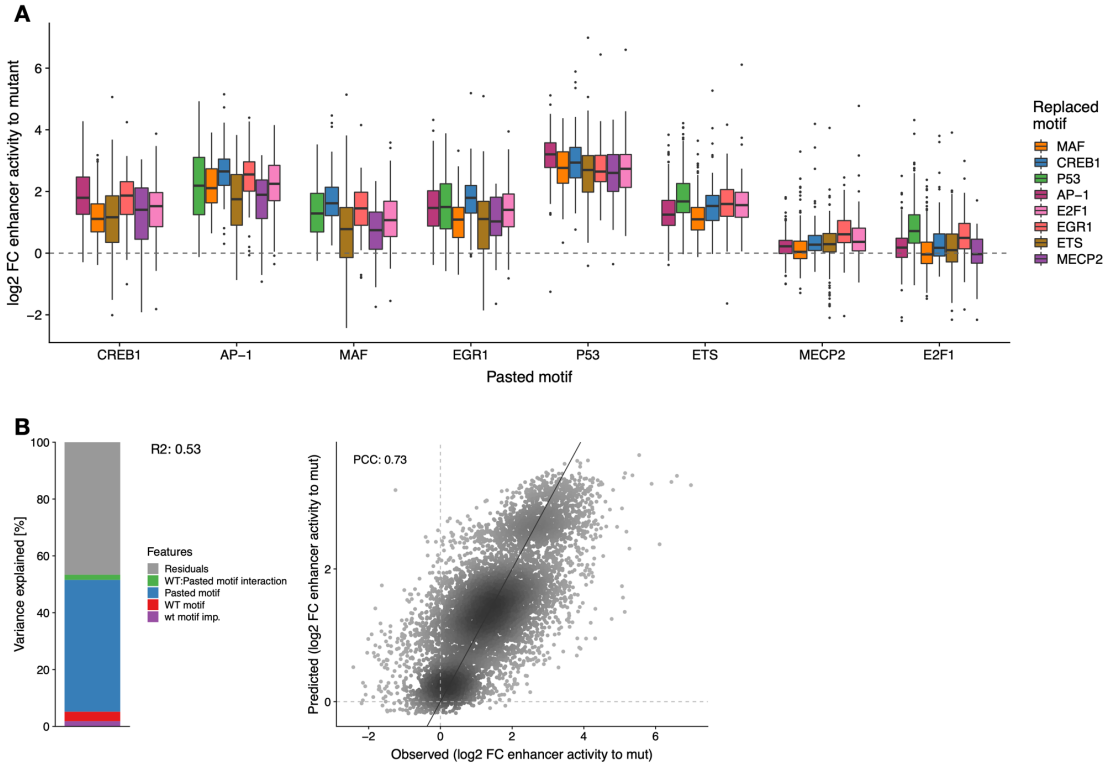
A) Activity of pasted motifs at different enhancer positions. Distribution of enhancer activity changes (log_2) of all wild-type enhancers used and their variants with either mutant sequences or different TF motifs pasted. Few instances show negative values: these are not dependent on the specific mutant sequence but rather correspond to the creation of a repressor motif at the flanks of the pasted motif and the backbone enhancer. **B)** Activity of pasting motifs (y-axis, log_2 fold-change activity over basal motif-mutated enhancer activity) in function of the basal activity (x-axis, activity of motif-mutated enhancer). The PCC is denoted for each motif.

Supplemental Fig S22. Human TF motifs work differently at different enhancer positions.



A) Hierarchical clustering of all TF motifs based on PCC of motif activities across all enhancer positions. **B)** Motifs work differently at different enhancer positions. Comparison between enhancer activity changes (log_2 FC to mutated sequence) after pasting different TF motifs across all enhancer positions. Positions with stronger activity of each motif (>= 2-fold in respect to the other motif) are colored with the respective colors. PCC: Pearson correlation coefficient.

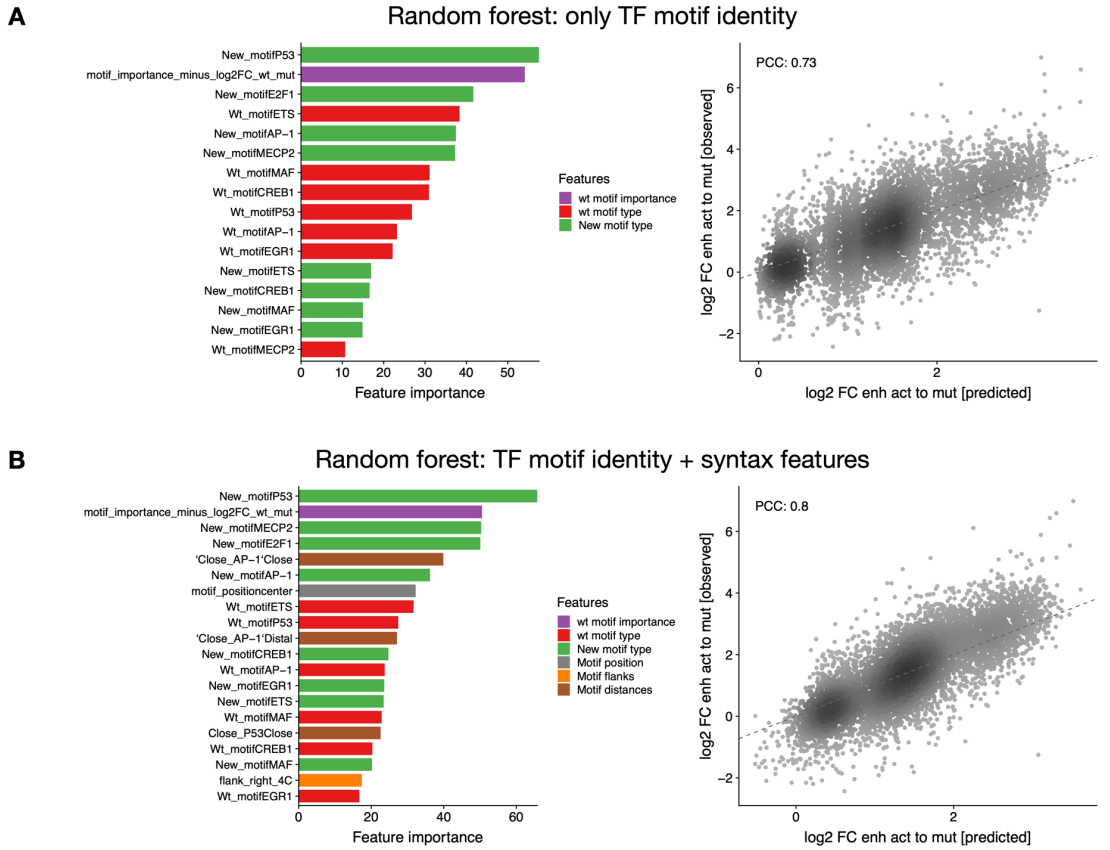
Supplemental Fig S23. TF motif activity in function of wild-type motif identity in human enhancers.



A) Distribution of enhancer activity changes (log_2 FC to mutated sequence) across all enhancer positions for each pasted TF motif, grouped by the identity of the wild-type motif.

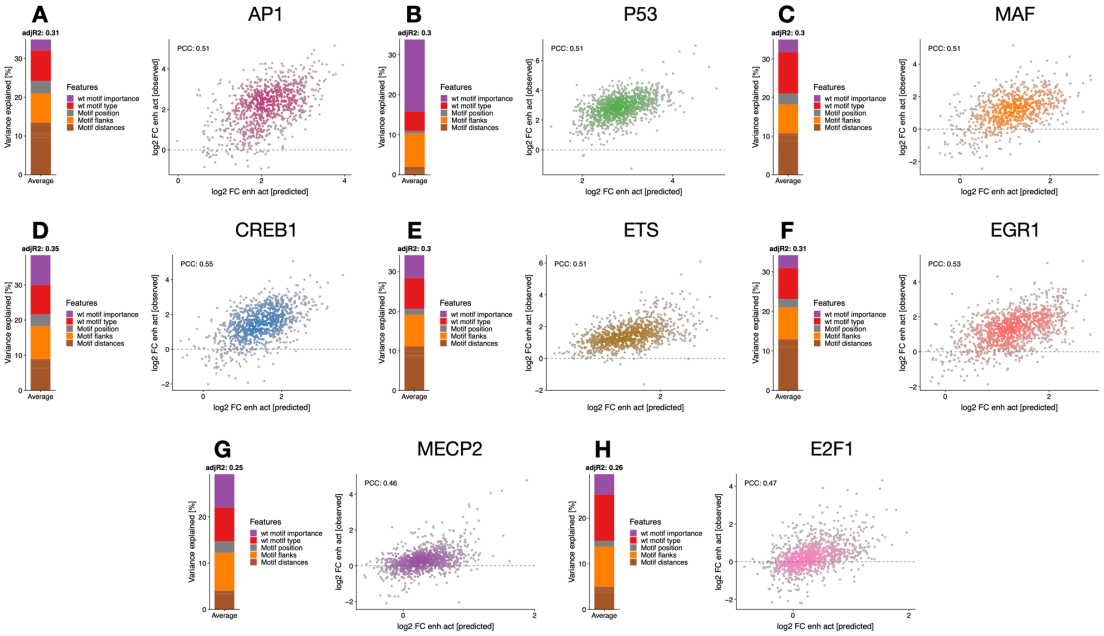
B) Left: Bar plot showing the amount of variance explained by the wild-type motif importance and identity, the pasted motif identity and the interaction between the wild type and pasted motifs, using a linear model fit on all motif pasting results. Right: Scatter plots of predicted (linear model) vs. observed enhancer activity changes (log_2 FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S24. Prediction of motif activities using motif syntax features in human enhancers.



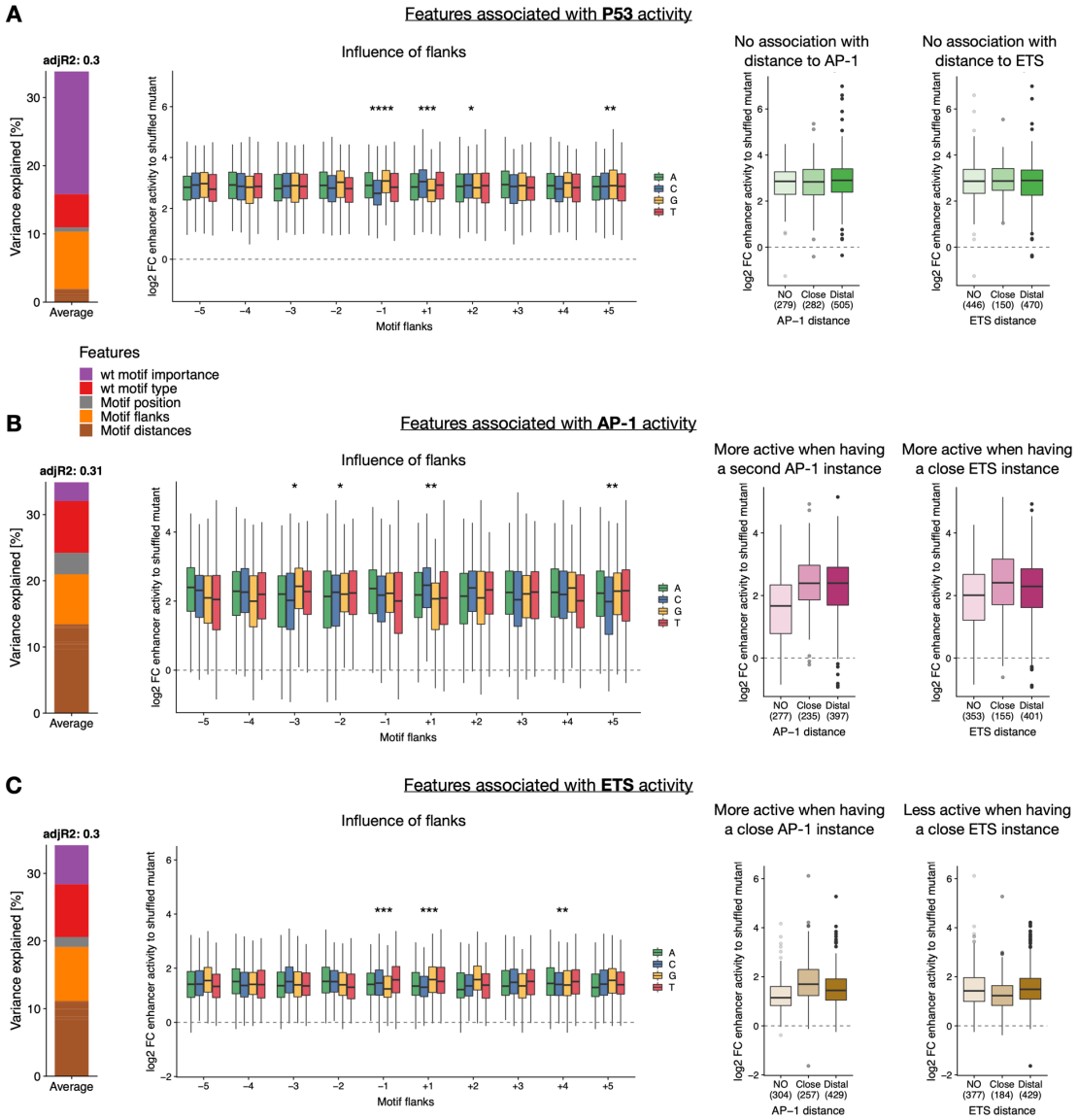
Left: Importance of all features **(A)** or only the top 20 **(B)** included in the random forest models with only TF motif identity **(A)** or also with syntax features **(B)**, sorted by importance and colored by feature type. Right: Scatter plots of predicted vs. observed enhancer activity changes (log_2 FC to mutated sequence) across all motif pasting experiments. Color reflects point density. PCC is shown.

Supplemental Fig S25. Linear models with syntax features to predict motif activities in human enhancers.



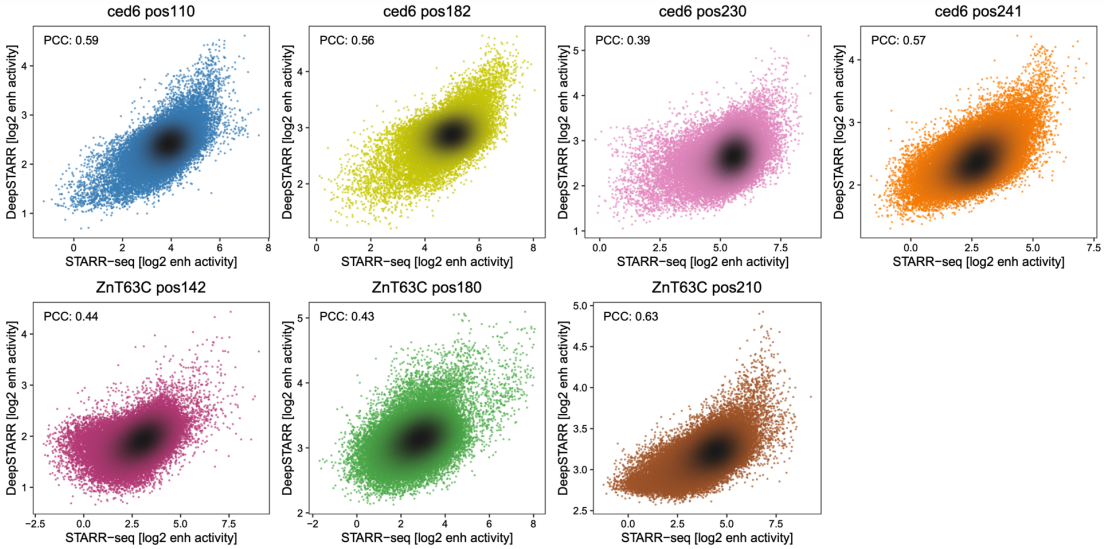
A-H) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Right: Scatter plots of predicted vs. observed enhancer activity changes (log_2 FC to mutated sequence) for motif pasting experiments per TF motif type. Color reflects point density. PCC is shown.

Supplemental Fig S26. Sequence features associated with activity of P53, AP-1 and ETS motifs in human enhancers.



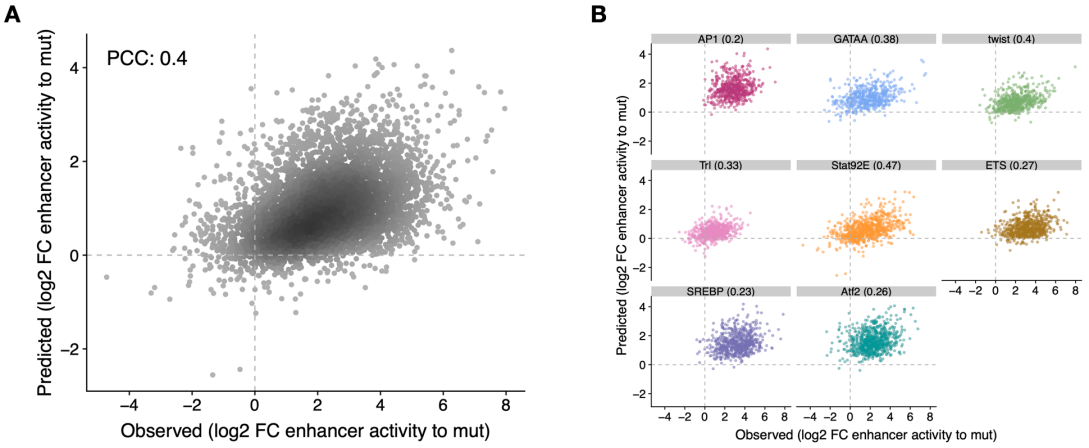
A-C) Left: Bar plot showing the variance explained by the different types of features (color legend) for each of the linear models. Middle-left: Motif activity according to the different bases at each flanking position, colored by nucleotide identity. Statistics from linear model in Fig 5E: ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05 (linear regression p-value). Middle-right and right: Enhancer activity changes (log_2 FC to mutated sequence) after pasting each TF motif in positions with no additional AP-1 (middle-right) or ETS (right) in the enhancer, or with additional AP-1 or ETS at close (<= 25 bp) or distal (>25 bp) distances. Number of instances are shown.

Supplemental Fig S27. DeepSTARR predicts enhancer sequence changes.



Comparison between DeepSTARR predicted (y-axis) and experimentally measured (x-axis) activity of random sequence variants tested at the different enhancer positions. Color reflects the enhancer position and point density. PCCs are shown.

Supplemental Fig S28. DeepSTARR predicts activity of motifs in different enhancer positions.



A) Comparison between DeepSTARR predicted (y-axis) and experimentally measured (x-axis) enhancer activity changes (log_2 FC to mutated sequence) for all motif pasting sequences. Color reflects the enhancer position and point density. PCCs are shown. **B)** Same as in (A) but per pasted TF motif.

Supplemental Tables

Supplemental Table S1. Primers used for UMI-STARR-seq library cloning.
Primers used for UMI-STARR-seq library cloning.

Supplemental Table S2. Random variants and oligo UMI-STARR-seq mapping statistics.
Summary of total sequenced reads, mapped reads and unique fragments (after collapsing by UMIs) for two random variants and three oligo UMI-STARR-seq screens in S2 cells, and three oligo UMI-STARR-seq screens in human HCT-116 cells.

Supplemental Table S3. Activity of random variants in seven enhancer positions.
8nt and 16nt forward and reverse sequences, activities and scaled activities in each of the seven enhancer positions.

Supplemental Table S4. Drosophila and human TF motif sequences used in the motif pasting experiments.
Drosophila and human TF motif sequences used in the motif pasting experiments.

Supplemental Table S5. Results of motif-pasting experiment in Drosophila S2 enhancers.
Table with all oligos used in the analysis of *Drosophila* motif pasting with their DNA sequence, wild-type motif information, pasted motif information, activity of respective enhancer variant, of the original wild type or motif-mutant enhancer, and respective $\log_2$ fold-changes.

Supplemental Table S6. Results of motif-pasting experiment in human HCT-116 enhancers.
Table with all oligos used in the analysis of human motif pasting with their DNA sequence, wild-type motif information, pasted motif information, activity of respective enhancer variant, of the original wild type or motif-mutant enhancer, and respective $\log_2$ fold-changes.

Supplemental Methods

UMI-STARR-seq

Cell culture and transfection

Drosophila Schneider 2 cells were grown in Schneider's *Drosophila* Medium (Gibco; 21720-024) supplemented with 10% heat inactivated FBS (Sigma-Aldrich; F7524) at 27°C. Human HCT116 cells were cultured in DMEM (Gibco; 52100-047) supplemented with 10% heat inactivated FBS (Sigma-Aldrich; F7524) and 2mM L-Glutamine (Sigma-Aldrich; G7513) at 37ºC in a 5% $C0_2$-enriched atmosphere. Both cell types were passaged every 2-3 days.

We used the MaxCyte-STX electroporation system for all library transfections. S2 cells were collected at 300 x g for 5min and washed once in 1:1 Schneider's Drosophila Medium and MaxCyte electroporation buffer (EPB-1). 50 x 10^6 cells were transfected with 5µg of DNA using the "Optimization 1" protocol, recovered for 30min at 27°C and resuspended in 10mL S2 Medium with 10% FBS. HCT116 cells were collected at 200 x g for 5min and washed once in MaxCyte electroporation buffer (EPB-1). Cells were electroporated at a density of 1 x 10^7 cells per 100µL and 20µg of DNA using the preset "HCT116" program, recovered for 20min at 37 °C and resuspended in 10mL DMEM 10% FBS and 2mM L-Glutamine.

Each replicate for a STARR-seq screen was transfected in 2 OC400 cuvettes with a total of 400 x 10^6.

UMI-STARR-seq experiments

Library cloning

Random 8nt variant libraries were generated using a PCR approach with degenerate oligonucleotides. Forward primers (primers see Supplemental Table S1) were designed to anneal directly downstream of the enhancer position of interested followed by 8 degenerate bp (creating 65,536 variants) and another 20 bp complementary stretch. Reverse primers were complementary to the 20 bp 5' of the degenerate stretch. The STARR-seq vector containing the wild-type enhancer of interest (either *ced-6* or *ZnT63C*) was used as a template for the PCR. The PCR was run across the whole STARR-seq plasmid, followed by DpnI digest and a Gibson reaction that re-circularizes the plasmid. Libraries were grown in 2l LB-Amp (final ampicillin concentration 100µg/mL). Variant libraries of the same enhancer i.e. *ced-6* enhancer pos110, pos182, pos230, pos241 and

ZnT63C enhancer pos142, pos180, pos210 were pooled to equimolar ratio, together with another synthetic oligo library containing wt enhancer sequences and negative regions. *Drosophila* and human oligo libraries were synthesized by Twist Bioscience including the 249 bp enhancer sequence and adaptors for library cloning. *Drosophila* library fragments were amplified (primers see Supplemental Table S1) and cloned into *Drosophila* STARR-seq vectors containing the DSCP core-promoters using Gibson cloning (New England BioLabs; E2611S). The oligo library for human STARR-seq screens was amplified (primers see Supplemental Table S1) and cloned into the human STARR-seq plasmid with the ORI in place of the core promoter (Muerdter et al. 2018). Libraries were grown in 2l LB-Amp (final ampicillin concentration 100µg/mL).
All libraries were purified with Qiagen Plasmid *Plus* Giga Kit (cat. no. 12991).

Drosophila S2 cells

UMI-STARR-seq was performed as described previously (Arnold et al. 2013; Neumayr et al. 2019). In brief, we transfected 400 × 10^6 S2 cells total per replicate with 20 µg of the input library (see libraries above). After 24 hr incubation, poly(A) RNA was isolated and processed as described before (Neumayr et al. 2019). Briefly: after reverse transcription and second strand synthesis a unique molecular identifier (UMI) was added to each transcript, allowing the counting of individual RNA molecules. This is followed by two nested PCR steps, each with primers that are specific to the reporter transcripts such that STARR-seq does not detect endogenous cellular RNAs.

Human HCT116 cells

UMI-STARR-seq was performed as described previously (Arnold et al. 2013; Muerdter et al. 2018; Neumayr et al. 2019). Screening libraries were generated from synthesized oligo pools by Twist Bioscience (see above). We transfected 80 × 10^6 HCT116 cells total per replicate with 160 µg of the input library. After 6 hr incubation, poly(A) RNA was isolated and further processed as described before (Neumayr et al. 2019).

Illumina sequencing

High-throughput sequencing was performed at the VBCF NGS facility on an Illumina NextSeq 550 or NovaSeq SP platform, following manufacturer's protocol. Random variants UMI-STARR-seq and Twist-oligo library screens were sequenced as paired-end 150 cycle runs, using standard Illumina i5 indexes as well as unique molecular identifiers (UMIs) at the i7 index. Deep sequencing base-calling was performed with CASAVA (v.1.9.1).

Random variants UMI-STARR-seq data analysis

Dedicated Bowtie indices were created for each enhancer position's N_8 library and combined with an oligo library of thousands of wild-type enhancers and negative sequences (de Almeida et al. 2022) for normalization, all 249 bp-long sequences. UMI-STARR-seq RNA and DNA input reads (paired-end 150 bp) were mapped to these dedicated Bowtie indices using Bowtie v.1.2.2 (Langmead et al. 2009). Since the N_8 variants were all positioned in the last 150 nt of each enhancer, we allowed for flexible mapping in the beginning of the fragments to increase the number of mapped reads while keeping high sensitivity for the different enhancer variants. Specifically, we trimmed the forward reads to 36 bp and mapped them to the indices allowing for 3 mismatches; the full 150 bp-long reverse reads were mapped with no mismatches, to identify all sequence variants; paired-end reads with the correct position, length and strand were kept. This mapping strategy was used for both DNA and RNA reads. For paired-end DNA and RNA reads that mapped to the same variant, we collapsed those that have identical UMIs (10 bp, allowing one mismatch) to ensure the counting of unique molecules (Supplemental Table S2).

We excluded oligos with less than 5 reads in any of the input replicates and less than 1 read in any of the RNA replicates. The enhancer activity of each sequence in each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014). We used the counts of wild-type negative regions in each library as scaling factors between samples.

Oligo library UMI-STARR-seq data analysis

As described previously (de Almeida et al. 2022), oligo library UMI-STARR-seq RNA and DNA input reads (paired-end 150 bp) were mapped to a reference containing the 249 bp-long sequences from the fragments present in the *Drosophila* (dm3) or human (hg19) libraries using Bowtie v.1.2.2 (Langmead et al. 2009). We used these reference genomes to be able to integrate our results with older in-house and published datasets and made sure this choice does not affect the quantifications of enhancer activity. For each library we demultiplexed reads by the i5 and i7 indexes and oligo identity. Mapping reads with the correct length, strand and with no mismatches (to identify all sequence variants) were kept. Both DNA and RNA reads were collapsed by UMIs (10 bp) as above (Supplemental Table S2).

We excluded oligos with less than 10 reads in any of the input replicates and added one read pseudocount to oligos with zero RNA counts. The enhancer activity of each oligo in

each screen was calculated as the $\log_2$ fold-change over input, using all replicates, with DESeq2 (Love et al. 2014). We used the counts of wild-type negative regions in each library as scaling factors between samples.

Analyses of random variants at different enhancer positions

Independent motif mutations

Two strong S2 developmental enhancers with different TF motif compositions were selected to test a diversity of random 8 nt variants in different positions: *ced-6* (chr2R:5326628-5326876) and *ZnT63C* (chr3L:3310914-3311162) enhancers. Experimental mutations of GATA, AP-1 and twist motifs in these enhancers were performed in a previous study (Supplemental Fig S4F; (de Almeida et al. 2022)) and used here to select important enhancer positions.

Enhancer random variants libraries and UMI-STARR-seq

We selected five positions important for the activity of the two enhancers (*ced-6* pos110 and pos241; *ZnT63C* pos142, pos180, pos210) and two non-important positions of the *ced-6* enhancer (pos182 and pos230). At each position, we experimentally replaced the respective 8nt stretch of the enhancer with randomized nucleotides (N_8), creating 65,535 enhancer variants in addition to the wild-type sequence per position. For each enhancer, we pooled the libraries of the different positions and combined them with an oligo library of thousands of wild-type enhancers and negative sequences (de Almeida et al. 2022) for normalization. UMI-STARR-seq using the *ced-6* or *ZnT63C* pooled libraries was performed ("UMI-STARR-seq experiments") and analyzed ("Random variants UMI-STARR-seq data analysis") as described above (Supplemental Table S3). We performed two independent replicates per enhancer pooled library screen (Pearson correlation coefficient (PCC)=0.85-0.91; Supplemental Fig S4A-E).

To be able to compare the activity of variants and motifs between enhancer positions, we next scaled the enhancer activity of all variants per position (*z*-scores). This allows to measure the change in activity of a given variant over the average of all variants, correcting for the importance of the different enhancer positions tested.

Comparison between pooled libraries using common oligos

The respective wild-type enhancer sequence was overrepresented in each N_8 library input since it was used as the template for the PCR cloning (Supplemental Fig S4A,B). We compared the activities of the *ced-6* and *ZnT63C* enhancer sequences and all other wild-

type enhancers and negative sequences present in both *ced-6* and *ZnT63C* pooled libraries (Supplemental Fig S4C). The activities of the common sequences were similar between both screens, except for the *ZnT63C* enhancer whose activity was underestimated in the *ZnT63C* pooled library, likely due to the technical overrepresentation in the input. We therefore selected another enhancer with the same activity as the *ZnT63C* enhancer (chrX:9273894-9274142) to be used as the reference wild-type activity for the *ZnT63C* enhancer variants (Supplemental Fig S4C, 2B).

Diversity of top active variants and *de novo* motif discovery

The most-active 8nt variants of each screen (1, 2, 5, 10, 50, 100 and 1,000) were retrieved and consolidated into position probability matrices based on the nucleotide frequencies at each position (Fig 1C, S5B). Logos were visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1; (Omar Wagih 2017)). The same was done after randomly sorting the variants of each screen for comparison. The information content of the top sequences at each position was calculated as described in https://bioconductor.org/packages/release/bioc/vignettes/universalmotif/inst/doc/IntroductionToSequenceMotifs.pdf (Schneider and Stephens 1990; Schneider et al. 1986) (Fig 1D, S5C).

The top 100 and 1,000 or bottom 1,000 variants (8nt +/- 4nt flanks) of each screen were used for *de novo* motif discovery analyses using HOMER, taking all detected variants of the respective screen as background (Supplemental Fig S2, S7). HOMER (v4.10.4; (Heinz et al. 2010)) was run with the findMotifs.pl command and the arguments *fly -len 6,7,8*.

Activity of TF motifs created by sequence variants

To robustly assess the activity of a given TF motif, we retrieved the activity of all 16nt variants (8nt +/- 4nt flanks) creating each motif by string-matching. The main motifs used were: GATA – GATAAG, AP-1 – TGA.TCA, SREBP – TCACGCGA, twist – CATCTG, CREB/ATF – TCATCA, STAT – TTCC.GGA, Trl – GAGAGA, ETS – CCGGAA, Dref – ATCGAT, ttk – AGGATAA, ZEB1 – CAGGTG, lola – GGAGTT (format: TF motif – string). For a more systematic comparison across all TF motif types, we matched variants to the optimal string from each TF motif PWM model in a motif database (Supplemental Fig S9A; (de Almeida et al. 2022)). The average activity across variants was defined as the motifs' intrinsic strength. These activities were used in Fig 1E, 2E,D, Supplemental Fig S3A, S6A, S9, S10.

To find how many active variants are explained by the creation of known motifs enriched in S2 developmental enhancers (from (de Almeida et al. 2022)), we performed PWM-

based motif scanning of those candidate motifs onto variants (8nt +/- 4 flanks) (Fig 1F, Supplemental Fig S3B, S6B). We used the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; genome = "BSgenome.Dmelanogaster.UCSC.dm3", bg="genome" (Schep 2021)) with p-value cutoffs $1e^{-04}$ and $1e^{-05}$.

Activity of variants in function of their similarity to the wild-type sequence

The similarity of each sequence variant to the wild-type sequence at each enhancer position was measured using the *stringdist* R package and *hamming* distance method (Supplemental Fig S6A).

Comparison of random variants activity across enhancer positions

We compared the activity of all 8nt random variants across enhancer positions using their *z*-score scaled activity (Fig 2C, Supplemental Fig S8; Supplemental Table S3). We calculated pairwise PCCs between the different libraries, performed hierarchical clustering ("complete" method) using the correlation values as similarities, and displayed heatmaps using the *pheatmap* R package (v.1.0.12; (Kolde 2019)). To reduce the impact of the flanking sequence of each position when comparing the activity of variants between them, we repeated the same after consolidating the 8nt into shorter variants by taking the centered sequence and averaging the activity across variants with different flanking nucleotides.

Analyses of motif pasting screens in *Drosophila* and human enhancers

Oligo library design

Drosophila motif pasting library

We selected 1,172 motif positions (among 728 enhancers) that are required for the activity of the respective enhancers, assessed by experimental mutagenesis in a previous study (de Almeida et al. 2022). These wild-type positions cover different contexts and TF motifs: GATA, AP-1, twist, Trl, ETS and SREBP. We next designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of eight TF motifs (GATA, AP-1, twist, Trl, ETS, SREBP, Stat92E and Atf2; one at a time; sequences in Supplemental Table S4) in each of these positions (Fig 3A). To reduce the influence of flanking nucleotides and different motif affinities and focus on differences due to the enhancer context we pasted an extended optimal sequence of each TF motif (as in de Almeida et al. (de Almeida et al. 2022)). This library (Supplemental Table S5) was

synthetized and pooled with a previous library containing the wild-type enhancer sequences (de Almeida et al. 2022) to be screened together.

Human motif pasting library

Similar to the *Drosophila* library, we selected 1,456 motif positions important for the activity of 808 enhancers, assessed by experimental mutagenesis in a previous study (de Almeida et al. 2022). These wild-type positions cover different contexts and TF motifs: AP-1, ETS, E2F1, EGR1, MAF, MECP2, CREB1, P53. We next designed sequences of enhancer variants where we pasted a mutant sequence or the optimal sequence of the same eight TF motifs (AP-1, ETS, E2F1, EGR1, MAF, MECP2, CREB1, P53; one at a time; sequences in Supplemental Table S4) in each of these positions. As for the *Drosophila* motifs, we pasted an extended optimal sequence of each TF motif to reduce the influence of flanking nucleotides and different motif affinities and focus on differences due to the enhancer context. This library (Supplemental Table S6) was synthetized and pooled with a previous library containing the wild-type enhancer sequences (de Almeida et al. 2022) to be screened together.

Oligo library synthesis and UMI-STARR-seq

The *Drosophila* and human enhancers' oligo libraries contained each sequences for the wild-type enhancers and enhancers with mutant variants or motifs pasted at the selected positions (Supplemental Table S5 and S6, respectively). All sequences were designed using the dm3 and hg19 genome versions, respectively. The enhancer sequences spanned 249 bp total, flanked by the Illumina i5 (25 bp; 5′ -TCCCTACACGACGCTCTTCCGATCT) and i7 (26 bp; 5′ AGATCGGAAGAGCACACGTCTGAACT) adaptor sequences upstream and downstream, respectively, serving as constant linkers for amplification and cloning. The resulting 300-mer oligonucleotide *Drosophila* and human libraries were synthesized by Twist Bioscience. UMI-STARR-seq using these oligo libraries was performed ("UMI-STARR-seq experiments") and analyzed ("Oligo library UMI-STARR-seq data analysis") as described above (Supplemental Table S5 and S6). We performed three independent replicates for *Drosophila* (correlation PCC=0.95-0.98; Supplemental Fig S11A,B) and human (PCC=0.96-0.98; Supplemental Fig S20A,B) screens.

Quantification of motif activity at different enhancer positions

We used our enhancer activity measures of the wild-type and mutated sequences to stringently select important enhancer positions for further analyses: positions where mutation reduced the activity by at least 2-fold (Supplemental Fig S12A, S21A). These

resulted in 763 important positions distributed among 496 *Drosophila* enhancers and 1,354 positions distributed among 753 human enhancers. This was important to select positions where we could reliably measure the increase in enhancer activity after pasting each TF motif – quantified as the $\log_2$ fold-change activity over the mutated enhancer (Fig 3B, 5A). Variability of activity of each motif across enhancer positions was quantified using the coefficient of variation (ratio of the standard deviation to the mean; Supplemental Fig S12B).

We compared the activity of motifs across enhancer positions by pairwise PCCs and performed hierarchical clustering ("complete" method) using the correlation values as similarities. Heatmaps were displayed using the *pheatmap* R package (v.1.0.12; (Kolde 2019)) (Fig 3D, 5B, Supplemental Fig S13A, S22A).

Importance of the wild-type motif

We fitted motif activity values ($\log_2$ fold-change enhancer activity after motif pasting) with linear models using the wild-type TF motif identity and importance ($\log_2$ fold-change activity between wild-type and motif-mutant sequence), the pasted motif identity, and the interaction between the wild-type and pasted motifs as covariates, using the *lm* function (v.3.5.1; (R Core Team 2020)). Variance explained by each covariate was calculated with one-way ANOVAs of the respective models (Fig 5D, Supplemental Fig S14B, S23B).

Difference between pairs of positions in the same or different enhancers

Drosophila enhancers with two positions tested in our assay were selected and the fold-change in motif activity between pairs of positions in the same enhancer was compared with the fold-change between pairs of positions in different enhancers (matched by similar position-mutant baseline activities). For each pasted TF motif, significant differences were assessed through a two-sided Wilcoxon signed rank test followed by FDR multiple testing correction (Supplemental Fig S15).

Prediction of motif activities using motif syntax features

Motif syntax features

To test how motif activities depend on motif syntax features we extracted the following features per tested enhancer position: the position relative to the enhancer center (center: -/+ 25 bp, flanks: -/+25:75 bp, boundaries: -/+75:125 bp), the position flanking nucleotides (5 bp on each side), and the presence and distance to other TF motifs (close: <= 25 bp; distal: >25 bp; between motif centers).

Instances of each TF motif type were mapped across all enhancers using their annotated PWM models (Supplemental Table S3) and the *matchMotifs* function from R package *motifmatchr* (v.1.4.0; (Schep 2021)) with the following parameters: genome = "BSgenome.Dmelanogaster.UCSC.dm3", p.cutoff = 5e-04, bg="genome". Overlapping instances (minimum 50%) for the same TF motif were collapsed and counted only once.

Random forest models

We used a 10-fold cross-validation scheme to train random forest models to predict *Drosophila* or human motif pasting activities (log_2 fold-change to mutant) using as features the wild-type TF motif identity and importance (log_2 fold-change activity between wild-type and motif-mutant sequence) and the pasted motif identity, together or not with additional syntax features (described above). All models were built using the *Caret* R package (v. 6.0-80; (Kuhn 2018)) and feature importance was calculated using its *varImp* function. Predictions for each held-out test sets were used to compare with the observed motif activities and assess model performance (Supplemental Fig S16, S24).

Linear model with motif syntax rules to predict motif activities

For each TF motif type, we built a multiple linear regression model to predict its activity (log_2 fold-change to mutant) across different enhancer positions using as covariates the wild-type TF motif identity and importance (log_2 fold-change activity between wild-type and motif-mutant sequence) together with additional syntax features (described above). All models were built using the Caret R package (v. 6.0-80; (Kuhn 2018)) and 10-fold cross-validation. Predictions for each held-out test sets were used to compare with the observed log_2 fold- changes and assess model performance (Supplemental Fig S17, S25). The linear model coefficients and respective FDR-corrected p-values were used as metrics of importance for each feature, using the red or blue scale depending on positive or negative associations (Fig 4A, 5E). For flanking positions, we used always red because the direction of the association is not relevant. In addition, we calculated the percentage of variance explained by each covariate in the linear models built for each TF motif with one-way ANOVAs. For each TF motif, we generated 100 different models, randomizing the order of the covariates (since the variance explained depends on the order of covariates entered), quantified the percentage of variance explained of each covariate as its sum of squares divided by the total sum of squares, and used the average value across all 100 models as the final variance explained per covariate (Supplemental Fig S17, S25).

DeepSTARR predictions

Nucleotide contribution scores

Nucleotide contribution scores for wild-type enhancers or enhancer variants (Fig 2A, S5A, S11C,D, S16) were calculated as described previously (de Almeida et al. 2022), using DeepExplainer (the DeepSHAP implementation of DeepLIFT, see refs. (Shrikumar et al. 2017; Lundberg and Lee 2017; Lundberg et al. 2020); update from https://github.com/AvantiShri/shap/blob/master/shap/explainers/deep/deep_tf.py) and visualized using the *ggseqlogo* function from R package *ggseqlogo* (v.0.1; (Omar Wagih 2017)).

DeepSTARR predictions of enhancer sequence changes

DeepSTARR (https://github.com/bernardo-de-almeida/DeepSTARR, (de Almeida et al. 2022)) was used to predict the enhancer activity of N_8 variants in enhancers (Supplemental Fig S27) or the $\log_2$ fold-change enhancer activity of motif pasting sequences (Supplemental Fig S28).

Statistics and data visualization

All statistical calculations and graphical displays have been performed in R statistical computing environment (v.3.5.1; (R Core Team 2020)) and using the R package *ggplot2* (Wickham 2016). 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.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE211659 or Zenodo at https://doi.org/10.5281/zenodo.7010528. Code used to process the UMI-STARR-seq data as well as to reproduce all analyses, results and figures has been submitted to GitHub (https://github.com/bernardo-de-almeida/Variant_STARRseq) and is available as Supplemental Code.

References

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (1979)* **339**: 1074–1077.

de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet*.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Kolde R. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.

Kuhn M. 2018. caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**: 56–67.

Lundberg SM, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems*.

Muerdter F, Boryn ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR, et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**: 141–149.

Neumayr C, Pagani M, Stark A, Arnold CD. 2019. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr Protoc Mol Biol* **128**: e105.

Omar Wagih. 2017. ggseqlogo: A "ggplot2" Extension for Drawing Publication-Ready Sequence Logos. R package version 0.1. https://CRAN.R-project.org/package=ggseqlogo.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schep A. 2021. motifmatchr: Fast Motif Matching in R. R package version 1.14.0.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**: 6097.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**: 415–431.

Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. *ArXiv* **1704.02685**.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.*

Conclusions and perspectives

Identifying enhancers and characterizing their sequence determinants – the cis-regulatory code – has remained one of the greatest challenges of modern biology. This thesis builds on recent advances in high-throughput enhancer testing assays and deep neural networks and further developed them to dissect the cis-regulatory information encoded in enhancer sequences. This included the development of a deep learning model, DeepSTARR, that predicts enhancer activity for two different transcriptional programs directly from DNA sequence and reveals important aspects of the enhancer code in *Drosophila* S2 cells (Publication 1). Additionally, we performed a large-scale enhancer mutagenesis screen to explore the flexibility of enhancer sequences with regards to nucleotide and motif identity at specific enhancer positions as well as the position-dependence of motif activity (Publication 2).

DeepSTARR predicts enhancer activity quantitatively for unseen sequences and reveals different coding features for the developmental and housekeeping programs, including specific TF motifs and higher-order syntax rules that we validated experimentally. DeepSTARR performed better than methods based on known TF motifs or unbiased k-mer counts, both at predicting continuous enhancer activity and at binary classification of enhancer sequences, supporting CNNs as the state-of-the-art methods for genomic prediction tasks. Since these models are not based on statistical over-representation, they can discover both abundant features but also features that are relatively rare in enhancers but still important for enhancer activity, as we demonstrate with DeepSTARR. Still, the motif syntax features described here (TF motif combinations, flanks and distances) likely capture less information than DeepSTARR: linear models using these features showed lower performance on identifying enhancers or important motif positions, 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.

The enhancer syntax rules learned by DeepSTARR agree well with the ones identified through the enhancer mutagenesis analyses and converged on a key aspect of the enhancer code: enhancers display constrained sequence flexibility where only a specific but still diverse set of TF motifs can function at a given position. This activity of motifs at specific positions is strongly modulated by the enhancer sequence context, namely the flanking sequence, presence and diversity of other motif types, and distance between motifs, such that motifs need to be analyzed in their cis-regulatory context. The observation that both *Drosophila* and human TF motifs require specific enhancer sequence contexts suggests that this is a general principle of enhancers.

The understanding of these motif syntax constraints is crucial for our ability to interpret the impact of disease-related sequence variants, which typically affect individual motif instances.

Finally, the sequence-rules uncovered by DeepSTARR allowed the design of synthetic enhancers with desired activity levels de novo. The synthetic enhancers were of similar complexity as endogenous enhancers in the training set, for example in terms of TF motif number and diversity. The observation that a vast number of different sequences can have similar enhancer strengths, as also demonstrated in the study of enhancer sequence variants, highlights the flexibility of regulatory sequences and the evolutionary opportunities that this provides. This flexibility might be an important pre-requisite for the evolution of developmental enhancers that operate under many additional constraints, for example regarding the precise spatiotemporal control of enhancer activities. Given that the activity in a given cell can be achieved by many solutions, the specific solutions that fulfill additional requirements can be explored during evolution. Indeed, the cell type–specific expression patterns of enhancers can change upon (minimal) sequence perturbations[90,213,214]. The fact that enhancer strength in a given cell type and enhancer specificity across cell types and developmental time are subject to different yet overlapping sequence constraints highlights the complexity of the regulatory code.

Future perspectives

Here, I will outline potential avenues for extending the research presented in this thesis, hoping to inspire further progress in this exciting and rapidly evolving field of computational genomics.

Understanding the enhancer code of different cell types

We and others have shown that deep learning models such as DeepSTARR can be applied to enhancer data from individual cell types to learn their regulatory code with remarkable accuracy. The next step would be to extend such models to learn the enhancer code of all main tissue types and specific cell types of an organism, taking advantage of the single-cell and tissue enhancer atlases that are being generated using genomic assays (mainly ATAC-seq)[38–40,173,215–219] and transcriptional reporter assays[28,33,220–222]. This should reveal the cell type-specific enhancer syntax rules, including key TF motifs, their arrangements, and the corresponding TFs. Improved interpretation tools will be required to further understand the models and the features they learn. Comparison of the sequence-rules between cell types and species will build a more complete understanding of the cis-regulatory code and its evolution.

Understanding how the enhancer code evolves during key developmental transitions

In addition to model different tissues and developmental stages as independent states, in the future it will be interesting to analyze continuous trajectories of successive cell states and build sequence-based models that predict enhancer activity changes during key developmental decisions and branch points. This could be done by modeling the differences ("delta") between temporally successive cell types or more fine-grained mini-clusters of cells ordered by pseudotemporal measures of individual cells, such as pseudotime[223,224] or predicted developmental age[217]. As single-cell technologies continue to advance, we will ultimately be able to train sequence-models to predict the state of each individual cell using data from various single-cell assays such as ATAC-seq and RNA-seq (see the scBasset model for an example in this direction[225]). Such models will reveal the enhancer codes of different cells and cell types and how these codes differ between adjacent cell states to directionally advance developmental progression or cell-state transitions, regarding how both activating and repressing cues are integrated at dynamic enhancers. This will revolutionize our understanding of cellular heterogeneity and of how enhancer sequences encode the complex patterns of temporal and spatial activity to drive the evolving gene expression profiles along cell differentiation trajectories.

Designing of cell type-specific enhancers

Deep learning sequence-models that predict cell type-specific enhancers will not only advance our understanding of the enhancer cis-regulatory code but enable the design of synthetic enhancers with particular regulatory properties, such as driving gene expression in specific cell types[226] or in response to signals from the cellular environment. Such tools will have great potential as specific markers for cell states and cell-state transitions, enabling the detection of such transient events as well as lineage-tracing experiments to determine the cells of origin for selected cell types and tissues. Furthermore, the engineering of synthetic enhancers with desired properties provides unanticipated opportunities for controlling gene expression, with future applications for cell and gene therapy.

Building large language foundation models for genomics

One promising approach to improve the prediction of molecular phenotypes from DNA sequences is the development of foundational models pre-trained on DNA sequences. This type of self-supervised language models has deeply transformed the artificial intelligence field with notable examples in natural language processing (NLP), including BERT[227] and GPT-4[228], and has already been successfully applied to the prediction of protein structure and sequence variant effects[162,229–232]. Given the complex sequence features embedded in non-coding DNA sequences

and the increasing amount of data generated by modern genomics research, large DNA language models trained on unlabeled genome sequences across diverse species and populations have the potential of developing general and transferable understandings of the structure, constraints, and function of genomic sequences[233,234]. Such pre-trained models and the knowledge encoded therein can be further applied to downstream tasks, in what is called transfer learning, to solve various sequence-related tasks such as predicting enhancers, promoters and ultimate gene expression. This "top-down" approach contrasts with traditional "bottom-up" approach that uses task-specific data (models such as DeepSTARR). While a recent foundational 2.5 billion-parameter model trained on genome sequences from 850 species and fine-tuned on the DeepSTARR training data did not surpass DeepSTARR (625 thousand parameters) on quantitative predictions of enhancer activity[234], it is still a promising approach in its early stages. As pre-training and fine-tuning techniques continue to improve, foundational models might become the state-of-the-art for genomic predictions. In addition, despite the resource-intensive nature of pre-training such large models, the trained models can be utilized for various downstream tasks at a significantly reduced cost.

Prediction of gene expression from sequence

In addition to predict enhancers and the activity of all cis-regulatory elements in the genome, the ultimate goal should be to predict gene expression levels in different cell types solely from the DNA sequence. Such models have the potential to enhance our understanding of how genes are regulated in different cell types and how their expression is affected by the numerous non-coding genetic variants linked to human diseases and traits.

One way of achieving this would be to combine task-specific models that predict all cis-regulatory elements (enhancers, promoters, insulators and silencers) from the DNA sequence (using genome-wide ATAC-seq or similar data) with models that predict their impact on gene transcription (for example the ABC model[48] for enhancers) trough ensemble learning. A related approach more focused on enhancers would be to first build enhancer-driven gene regulatory networks, for example from joint profiling of chromatin accessibility and gene expression of individual cells[235], and then train sequence-based models to predict such enhancers, thus generating a fully sequence-based model of those gene networks. The advantage of such models is that they can be more easily interpretable since different aspects are modeled separately.

A different path towards the same objective of predicting gene expression from genomic sequence would be to learn the sequence-to-expression relationship in an end-to-end and unbiased fashion, taking into account a large sequence receptive field that can account for interactions between regulatory elements. This has been attempted by combining CNNs with dilated convolutions (Basenji[175]) or transformer architectures (Enformer[177]) which model gene

expression through the encoding of regulatory elements up to 20 and 100 kb, respectively, away from the locus of interest. However, despite their improved performance, these models are still limited at capturing the causal effects of distal enhancers on expression and their sequence features[236]. One promising direction for the field would be to remove the earlier CNNs and directly build transformer models capable of handling lengthy inputs of up to 200 kb or more that could model the DNA language directly (similar to the foundational models above). Although the standard transformer architecture cannot handle such large inputs effectively because the self-attention operation scales quadratically with sequence length, alternative techniques such as sparse attention[237] could be explored to overcome the computational limitations.

When combined with the single-cell transcriptional atlases in health and disease that are being generated (e.g. by the Human Cell Atlas initiative[238]), improvements of these deep learning algorithms will allow to build sequence-based predictive models for all human cell states and behavior, ultimately understanding how our genomes store gene-regulatory information to dictate gene expression and development.

Predicting the effect of genetic variants

Improved deep learning methods that accurately predict various functional properties from genomic DNA, including cell type-specific chromatin states and gene expression, will play a crucial role in interpreting the full set of genetic variations in individual genomes. After training, such models can be used to process distinct alleles and compare predictions to score and prioritize genetic variants. This approach has shown good results on chromatin prediction tasks but still has a limited predictive value on variants that impact gene expression[239]. It is possible that this performance is limited by current models being trained on a single Reference genome, and that training on personalized genomes (more locus-specific data) could increase their sensitivity in predicting the cellular impact of genetic variants across the entire genome. In addition, current in-silico analyses have focused on predicting the effect of individual variants in isolation, thus not accounting for the genetic background of the individuals and potential interactions between variants[165,166,177,184,187,196]. As the number of personal genomes sequenced increases, we should move to predictions for individual-specific genomes to account for genetic interactions, even across different biological mechanisms.

Medical genomics

The future of medicine and our understanding of the human genome and cellular behavior will likely lie in the interface between single-cell omics and perturbation data and artificial intelligence and machine learning algorithms. Combining these approaches to build sequence-based predictive models of every human cellular state will allow to interpret an individual's

genome, including susceptibility to diseases and respective disease mechanisms, thus providing new tools and therapeutics for personalized medicine. Together with the emerging genome editing technologies, DeepSTARR-like and more complex sequence-based deep learning models have the potential to revolutionize medicine and the future of humanity.

Bibliography

1. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* **278**, 274–288 (2005).
2. Avery, O. T., Macleod, C. M. & Mccarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* **79**, 137–158 (1944).
3. Watson, J. & Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
4. Crick, F., Barnetr, L., Brenner, S. & Waits-Tobin, R. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
5. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* The sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
8. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
9. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V, Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2002).
10. The C. elegans sequencing consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
11. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
12. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911–920 (2010).
13. Lelli, K. M., Slattery, M. & Mann, R. S. Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annu Rev Genet* **46**, 43–68 (2012).
14. Levine, M. & Tijan, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
15. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**, 233–245 (2012).
16. Levine, M. Transcriptional enhancers in animal development and evolution. *Current Biology* **20**, R754–R763 (2010).
17. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat Rev Genet* **15**, 272–286 (2014).
18. Gaszner, M. & Felsenfeld, G. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**, 703–713 (2006).
19. Calhoun, V. C., Stathopoulos, A. & Levine, M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proceedings of the National Academy of Sciences* **99**, 9243–9247 (2002).

20. Levo, M., Raimundo, J., Bing, X. Y., Sisco, Z., Batut, P. J., Ryabichko, S., Gregor, T. & Levine, M. S. Transcriptional coupling of distant regulatory genes in living embryos. *Nature* **605**, 754–760 (2022).

21. Petrykowska, H. M., Vockley, C. M. & Elnitski, L. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Res* **18**, 1238–1246 (2008).

22. Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev* **22**, 2651–2663 (2008).

23. Ayer, S. & Benyajati, C. Conserved enhancer and silencer elements responsible for differential Adh transcription in Drosophila cell lines. *Mol Cell Biol* **10**, 3512–3523 (1990).

24. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am J Hum Genet* **102**, 717–730 (2018).

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

26. Doyle, H. J., Kraut, R. & Levine, M. Spatial regulation of zerknüllt: a dorsal-ventral patterning gene in Drosophila. *Genes Dev* **3**, 1518–1533 (1989).

27. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).

28. Kvon, E. Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J. O., Pagani, M., Schernhuber, K., Dickson, B. J. & Stark, A. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).

29. Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S. & Odom, D. T. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).

30. Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F. & Schübeler, D. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* **43**, 1091–1097 (2011).

31. Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C. & Stark, A. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**, 685–692 (2014).

32. Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

33. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88–D92 (2007).

34. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803 (2005).

35. Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. & de Graaff, E. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725–1735 (2003).

36. 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).

37. Andersson, R., Gebhard, C., Miguel-escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

38. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

39. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

40. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).

41. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A. & Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

42. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **40**, (2019).

43. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences* **109**, 19498–19503 (2012).

44. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271–277 (2012).

45. Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M., *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265–270 (2012).

46. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

47. Klein, J. C., Chen, W., Gasperini, M. & Shendure, J. Identifying novel enhancer elements with CRISPR-based screens. *ACS Chem Biol* **13**, 326–332 (2018).

48. Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).

49. Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390 (2019).

50. Shlyueva, D., Stelzer, C., Gerlach, D., Yáñez-Cuna, J. O., Rath, M., Boryń, Ł. M., Arnold, C. D. & Stark, A. Hormone-Responsive Enhancer-Activity Maps Reveal Predictive Motifs, Indirect Repression, and Targeting of Closed Chromatin. *Mol Cell* **54**, 180–192 (2014).

51. Yanez-Cuna, J. O., Arnold, C. D., Stampfel, G., Boryn, L. M., Gerlach, D., Rath, M. & Stark, A. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res* **24**, 1147–56 (2014).

52. Muerdter, F., Boryn, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**, 141–149 (2018).

53. Liu, Y., Yu, S., Dhiman, V. K., Brunetti, T., Eckart, H. & White, K. P. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol* **18**, 1–13 (2017).

54. 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).

55. Spitz, F. & Furlong, E. E. M. Transcription factors: From enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

56. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences* **102**, 4936–4942 (2005).

57. Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Perez, N. M., *et al.* JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**, D165–D173 (2022).

58. Perez-Borrajero, C., Lin, C. S. H., Okon, M., Scheu, K., Graves, B. J., Murphy, M. E. P. & McIntosh, L. P. The biophysical basis for phosphorylation-enhanced DNA-binding autoinhibition of the ETS1 transcription factor. *J Mol Biol* **431**, 593–614 (2019).

59. Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R. & Rohs, R. Absence of a simple code: How transcription factors read the genome. *Trends Biochem Sci* **39**, 381–399 (2014).

60. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).

61. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).

62. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* **11**, 4047–57 (1992).

63. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J. & Ahituv, N. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**, 1–15 (2019).

64. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S. & Kellis, M. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**, 800–811 (2013).

65. Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).

66. Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J. & Birney, E. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol* **6**, (2005).

67. Warner, J. B., Philippakis, A. A., Jaeger, S. A., He, F. S., Lin, J. & Bulyk, M. L. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Methods* **5**, 347–353 (2008).

68. Naval-Sańchez, M., Potier, D., Haagen, L., Sánchez, M., Munck, S., Van De Sande, B., Casares, F., Christiaens, V. & Aerts, S. Comparative motif discovery combined with comparative transcriptomics yields accurate targetome and enhancer predictions. *Genome Res* **23**, 74–88 (2013).

69. Zhu, Z., Shendure, J. & Church, G. M. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res* **15**, 848–855 (2005).

70. Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S. & Brown, M. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).

71. Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A. & Levine, M. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev* **21**, 385–390 (2007).

72. Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P. & Furlong, E. E. M. A temporal map of transcription factor activity: Mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**, 797–807 (2006).

73. Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V. & Furlong, E. E. M. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* **21**, 436–449 (2007).

74. Odom, D. T., Zizlsperger, N., Gordon, D. B. G., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).

75. Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., Reddy, J., Guenther, M. G., Dekoter, R. P. & Young, R. A. Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**, 565–576 (2011).

76. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65–70 (2009).

77. Yanez-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* 2018–2030 (2012).

78. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr Opin Genet Dev* **43**, 73–81 (2017).

79. King, D. M., Hong, C. K. Y., Shepherdson, J. L., Granas, D. M., Maricque, B. B. & Cohen, B. A. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *Elife* **9**, 1–24 (2020).

80. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev Cell* **56**, 575–587 (2021).

81. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).

82. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).

83. Zinzen, R. P., Senger, K., Levine, M. & Papatsenko, D. Computational Models for Neurogenic Gene Expression in the Drosophila Embryo. *Current Biology* **16**, 1358–1365 (2006).

84. Panne, D. The enhanceosome. *Curr Opin Struct Biol* **18**, 236–242 (2008).

85. 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).

86. 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).

87. Erceg, J., Saunders, T. E., Girardot, C., Devos, D. P., Hufnagel, L. & Furlong, E. E. M. Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genet* **10**, e1004060 (2014).

88. Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., *et al.* Low Affinity Binding Site Clusters Confer Hox Specificityand Regulatory Robustness. *Cell* **160**, 191–203 (2015).

89. 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. *Proceedings of the National Academy of Sciences* **113**, 6508–6513 (2016).

90. Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S. & Levine, M. S. Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).

91. Fiore, C. & Cohen, B. A. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res* **26**, 778–786 (2016).

92. Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I. & Ahituv, N. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**, 1021–1028 (2013).

93. Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A. & Segal, E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521–530 (2012).

94. Thanos, D. & Maniatis, T. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

95. Hanes, S. D., Riddihough, G., Ish-Horowicz, D. & Brent, R. Specific DNA Recognition and Intersite Spacing Are Critical for Action of the Bicoid Morphogent. *Mol Cell Biol* **14**, 3364–3375 (1994).

96. 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).

97. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., Mcanany, C., Gagneur, J., Kundaje, A., *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354–366 (2021).

98. Song, B. P., Ragsac, M. F., Tellez, K., Jindal, G. A., Grudzien, J. L., Le, S. H. & Farley, E. K. Diverse logics and grammar encode notochord enhancers. *Cell Rep* **42**, 112052 (2023).

99. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**, 890–898 (2005).

100. Vockley, C. M., McDowell, I. C., D'Ippolito, A. M. & Reddy, T. E. A long-range flexible billboard model of gene activation. *Transcription* **8**, 261–267 (2017).

101. Parker, D. S., White, M. A., Ramos, A. I., Cohen, B. A. & Barolo, S. The cis-regulatory logic of Hedgehog gradient responses: key roles for Gli binding affinity, competition, and cooperativity. *Sci Signal* **4**, ra38 (2011).

102. Jindal, G. A., Bantle, A. T., Solvason, J. J., Grudzien, J. L., D'antonio-Chronowska, A., Lim, F., Le, S. H., Larsen, R. O., Klie, A., Frazer, K. A., *et al.* Affinity-optimizing variants within cardiac enhancers disrupt heart development and contribute to cardiac traits. *bioRxiv* (2022).

103. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J., Mann, R. S. & Zuckerman, M. B. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu. Rev. Cell Dev. Biol* **35**, 357–79 (2019).

104. Delker, R. K., Ranade, V., Loker, R., Voutev, R. & Mann, R. S. Low affinity binding sites in an activating CRM mediate negative autoregulation of the Drosophila Hox gene Ultrabithorax. *PLoS Genet* **15**, e1008444 (2019).

105. 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 (2016).

106. Swanson, C. I., Schwimmer, D. B. & Barolo, S. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Current Biology* **21**, 1186–1196 (2011).

107. Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences* **114**, E1291–E1300 (2017).

108. Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., Samee, M. A. H., He, D., Small, E. M., Mileikovsky, M., *et al.* Complex interdependence regulates heterotypic transcription factor distribution and coordinates cardiogenesis. *Cell* **164**, 999–1014 (2016).

109. Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M. & Levine, M. Immunity Regulatory DNAs Share Common Organizational Features in Drosophila. *Mol Cell* **13**, 19–32 (2004).

110. Passamaneck, Y. J., Katikala, L., Perrone, L., Dunn, M. P., Oda-Ishii, I. & Di Gregorio, A. Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian Ciona intestinalis. *Development* **136**, 3679–3689 (2009).

111. Cave, J. W., Loh, F., Surpris, J. W., Xia, L. & Caudy, M. A. A DNA transcription code for cell-specific gene activation by notch signaling. *Current Biology* **15**, 94–104 (2005).

112. Kazemian, M., Pham, H., Wolfe, S. A., Brodsky, M. H. & Sinha, S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res* **41**, 8237–8252 (2013).

113. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol* **47**, 1–8 (2017).

114. Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. & Taipale, J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

115. Kazemian, M., Pham, H., Wolfe, S. A., Brodsky, M. H. & Sinha, S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res* **41**, 8237–8252 (2013).

116. Hutchins, A. P., Choo, S. H., Mistri, T. K., Rahmani, M., Woon, C. T., Ng, C. K. L., Jauch, R. & Robson, P. Co-motif discovery identifies an esrrb-Sox2-DNA ternary complex as a mediator of transcriptional differences between mouse embryonic and epiblast stem cells. *Stem Cells* **31**, 269–281 (2013).

117. Cai, H. N., Arnosti, D. N. & Levine, M. Long-range repression in the Drosophila embryo. *Proceedings of the National Academy of Sciences* **93**, 9309–9314 (1996).

118. Kulkarni, M. M. & Arnosti, D. N. Logic of Short-Range Transcriptional Repression in Drosophila Melanogaster. *Mol Cell Biol* **25**, 3411–3420 (2005).

119. Merika, M. & Thanos, D. Enhanceosomes. *Curr Opin Genet Dev* **11**, 205–208 (2001).

120. Markstein, M. & Levine, M. Decoding cis-regulatory DNAs in the Drosophila genome. *Curr Opin Genet Dev* **12**, 601–606 (2002).

121. Crocker, J., Tamori, Y. & Erives, A. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6**, 2576–2587 (2008).

122. Scully, K. H., Jacobson, E. M., Jepsen, K., Lunyak, V., Viadiu, H., Carriere, C., Rose, D. W., Hooshmand, F., Aggarwal, A. K. & Rosenfeld, M. G. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* **290**, 1127–1131 (2000).

123. Rastegar, S., Hess, I., Dickmeis, T., Nicod, J. C., Ertzer, R., Hadzhiev, Y., Thies, W. G., Scherer, G. & Strähle, U. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* **318**, 366–377 (2008).

124. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**, (1998).

125. Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S. & Ovcharenko, I. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* **21**, 1139–1149 (2011).

126. Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. & McCallion, A. S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276–279 (2006).

127. He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A. & Zeitlinger, J. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat Genet* **43**, 414–421 (2011).

128. Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* **370**, (2020).

129. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics* **26**, 66–74 (2010).

130. Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., *et al.* ChIP-seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806–812 (2010).

131. Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., *et al.* Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**, 1036–1040 (2010).

132. May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., *et al.* Large-scale discovery of enhancers from human heart tissue. *Nat Genet* **44**, 89–93 (2012).

133. Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A. & Regev, A. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022).

134. 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).

135. Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B. & Celniker, S. E. 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).

136. 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).

137. 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).

138. 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).

139. Beer, M. A. & Tavazoie, S. Predicting Gene Expression from Sequence. *Cell* **117**, 185–198 (2004).

140. Zinzen, R. P. & Papatsenko, D. Enhancer responses to similarly distributed antagonistic gradients in development. *PLoS Comput Biol* **3**, 0826–0835 (2007).

141. 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).

142. 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).

143. 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**, 1–28 (2015).

144. Dibaeinia, P. & Sinha, S. Deciphering enhancer sequence using thermodynamics-based models and convolutional neural networks. *Nucleic Acids Res* **49**, 10309–10327 (2021).

145. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: A quantitative approach. *Nat Rev Genet* **10**, 443–456 (2009).

146. Janssens, H., Hou, S., Jaeger, J., Kim, A. R., Myasnikova, E., Sharp, D. & Reinitz, J. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nat Genet* **38**, 1159–1165 (2006).

147. Reinitz, J., Hou, S. & Sharp, D. H. Transcriptional Control in Drosophila. *Complexus* **1**, 54–64 (2003).

148. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (The MIT Press, 2016).

149. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (2012).

150. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 580–587 (2014).

151. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3431–3440 (2015).

152. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., *et al.* Deep Speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567* (2014).

153. Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144* (2016).

154. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., *et al.* Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).

155. Ramos, S., Gehrig, S., Pinggera, P., Franke, U. & Rother, C. Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling. in *2017 IEEE Intelligent Vehicles Symposium (IV)* 1025–1032 (2017).

156. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* **30**, 595–608 (2016).

157. Albrecht, T., Slabaugh, G., Alonso, E. & Al-Arif, S. M. M. R. Deep learning for single-molecule science. *Nanotechnology* **28**, 423001 (2017).

158. Agrawal, A. & Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Commun* **9**, 779–792 (2019).

159. Erdmann, M., Glombitza, J., Kasieczka, G. & Klemradt, U. *Deep learning for physics research*. (World Scientific, 2021).

160. Ge, M., Su, F., Zhao, Z. & Su, D. Deep learning analysis on microscopic imaging in materials science. *Mater Today Nano* **11**, 100087 (2020).

161. Agrawal, A., Gopalakrishnan, K. & Choudhary, A. *In Handbook on Big Data and Machine Learning in the Physical Sciences: Volume 1. Big Data Methods in Experimental Materials Discovery*. (World Scientific, 2020).

162. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

163. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

164. 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).

165. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931–934 (2015).

166. 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).

167. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389–403 (2019).

168. Elman, J. L. Finding structure in time. *Cogn Sci* **14**, 179–211 (1990).

169. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **9**, 1735–1780 (1997).

170. Quang, D. & Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44**, (2016).

171. Lanchantin, J., Singh, R., Wang, B. & Qi, Y. Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. *Pac Symp Biocomput* **22**, 254–265 (2017).

172. Minnoye, L., Taskiran, I. I., Mauduit, D., Fazio, M., Van Aerschot, L., Hulsemans, G., Christiaens, V., Makhzami, S., Seltenhammer, M., Karras, P., *et al.* Cross-species analysis of enhancer logic using deep learning. *Genome Res* **30**, 1815–34 (2020).

173. Janssens, J., Aibar, S., Taskiran, I. I., Ismail, J. N., Spanier, K. I., Gonzalez-Blas, C. B., Quan, X. J., Papasokrati, D., Hulselmans, G., Makhzami, S., *et al.* Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022).

174. Agarwal, V. & Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol* **23**, 245 (2022).

175. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y. & Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739–750 (2018).

176. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**, 1–27 (2020).

177. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P. & Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).

178. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* **24**, 125–137 (2023).

179. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, 878 (2016).

180. Wang, M., Tai, C., E, W. & Wei, L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res* **46**, e69–e69 (2018).

181. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* **54**, 940–949 (2022).

182. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res* **45**, e99–e99 (2017).

183. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18**, 67 (2017).

184. Ameen, M., Sundaram, L., Shen, M., Banerjee, A., Kundu, S., Nair, S., Shcherbina, A., Gu, M., Wilson, K. D., Varadarajan, A., *et al.* Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. *Cell* **185**, 4937-4953.e23 (2022).

185. Atak, Z. K., Taskiran, I. I., Demeulemeester, J., Flerin, C., Mauduit, D., Minnoye, L., Hulselmans, G., Christiaens, V., Ghanem, G. E., Wouters, J., *et al.* Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning. *Genome Res* **31**, 1082–1096 (2021).

186. Maslova, A., Ramirez, R. N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., Mostafavi, S. & null, null. Deep learning of immune cell differentiation. *Proceedings of the National Academy of Sciences* **117**, 25655–25666 (2020).

187. Trevino, A. E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H. Y., Pașca, A. M., Kundaje, A., *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053-5069.e23 (2021).

188. Kim, D., Risca, V., Reynolds, D., Chappell, J., Rubin, A., Jung, N., Donohue, L., Kathiria, A., Shi, M., Zhao, Z., *et al.* The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nat Genet* **53**, 1564–1576 (2021).

189. Schreiber, J., Libbrecht, M., Bilmes, J. & Noble, W. S. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv 103614* (2018).

190. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* **17**, 1111–1117 (2020).

191. Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., Thandapani, P., Skok, J., Aifantis, I., Fenyö, D., *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol* (2023).

192. Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., Teh, Y. W., Lunter, G. & Hughes, J. R. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* **17**, 1118–1124 (2020).

193. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* **54**, 725–734 (2022).

194. Cheng, J., Çelik, M. H., Kundaje, A. & Gagneur, J. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol* **22**, 94 (2021).

195. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

196. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K. & Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171–1179 (2018).

197. Kowalski, M. H., Wessels, H.-H., Linder, J., Choudhary, S., Hartman, A., Hao, Y., Mascio, I., Dalgarno, C., Kundaje, A. & Satija, R. CPA-Perturb-seq: Multiplexed single-cell characterization of alternative polyadenylation regulators. *bioRxiv* (2023).

198. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23**, (2022).

199. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91–106 (2019).

200. Slutskin, I. V., Weinberger, A. & Segal, E. Sequence determinants of polyadenylation-mediated regulation. *Genome Res* **29**, 1635–1647 (2019).

201. Pan, X. & Shen, H.-B. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics* **18**, 136 (2017).

202. Avsec, Ž., Barekatain, M., Cheng, J. & Gagneur, J. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics* **34**, 1261–1269 (2018).

203. Budach, S. & Marsico, A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics* **34**, 3035–3037 (2018).

204. Cheng, S., Guo, M., Wang, C., Liu, X., Liu, Y. & Wu, X. MiRTDL: A Deep Learning Approach for miRNA Target Prediction. *IEEE/ACM Trans Comput Biol Bioinform* **13**, 1161–1169 (2016).

205. Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., Fields, S. & Seelig, G. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* **27**, 2015–2024 (2017).

206. Zheng, A., Lamkin, M., Zhao, H., Wu, C., Su, H. & Gymrek, M. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat Mach Intell* **3**, 172–180 (2021).

207. 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).

208. 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).

209. Shrikumar, A., Tian, K., Shcherbina, A., Avsec, Ž., Banerjee, A., Sharmin, M., Nair, S. & Kundaje, A. TF-MoDISco v0.4.4.2-alpha: Technical Note. *arXiv 1811.00416* (2018).

210. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *arXiv 1704.02685* (2017).

211. de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54**, 613–624 (2022).

212. Reiter, F., de Almeida, B. P. & Stark, A. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. *Genome Res* **33**, 346–358 (2023).

213. Galupa, R., Alvarez-Canales, G., Borst, N. O., Fuqua, T., Gandara, L., Misunou, N., Richter, K., Alves, M. R. P., Karumbi, E., Perkins, M. L., *et al.* Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. *Dev Cell* **58**, 51-62.e4 (2023).

214. Fuqua, T., Jordan, J., Breugel, M. E. Van, Halavatyi, A., Tischer, C., Polidoro, P., Abe, N., Tsai, A., Mann, R. S., Stern, D. L., *et al.* Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).

215. Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, W., *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

216. Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C. A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

217. Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R. M., Martin, B., Dulja, A., Schaub, C., Trapnell, C., *et al.* The continuum of Drosophila embryonic development at single-cell resolution. *Science* **377**, eabn5800 (2022).

218. Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

219. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

220. Lalanne, J.-B., Regalado, S. G., Domcke, S., Calderon, D., Martin, B., Li, T., Suiter, C. C., Lee, C., Trapnell, C. & Shendure, J. Multiplex profiling of developmental enhancers with quantitative, single-cell expression reporters. *bioRxiv* (2023).

221. Zhao, S., Hong, C. K. Y., Myers, C. A., Granas, D. M., White, M. A., Corbo, J. C. & Cohen, B. A. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat Genet* **55**, 346–354 (2023).

222. Agarwal, V., Inoue, F., Schubach, M., Martin, B. K., Dash, P. M., Zhang, Z., Sohota, A., Noble, W. S., Yardimci, G. G., Kircher, M., *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv* (2023).

223. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. & Rinn, J. L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

224. Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K. W., Le, N.-T., Abe, J.-I., Cooke, J. P. & Wang, G. A relay velocity model infers cell-dependent RNA velocity. *Nat Biotechnol* (2023).

225. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat Methods* **19**, 1088–1096 (2022).

226. Taskiran, I. I., Spanier, K. I., Christiaens, V., Mauduit, D. & Aerts, S. Cell type directed design of synthetic enhancers. *bioRxiv* (2022).

227. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv 1810.04805* (2019).

228. OpenAI. GPT-4 Technical Report. *arXiv 2303.08774* (2023).

229. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Lawrence Zitnick, C., Ma, J., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

230. Brandes, N., Goldman, G., Wang, C. H., Jimmie Ye, C., Ntranos, V., Zuckerberg Biohub, C. & Francisco, S. Genome-wide prediction of disease variants with a deep protein language model. *bioRxiv* (2022).

231. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

232. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127 (2022).

233. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

234. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Henryk Grzywaczewski, A., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv* (2023).

235. Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S. & Aerts, S. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *bioRxiv* (2023).

236. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol* **24**, 56 (2023).

237. Child, R., Gray, S., Radford, A. & Sutskever, I. Generating Long Sequences with Sparse Transformers. *arXiv 1904.10509* Preprint at (2019).

238. Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., *et al.* The Human Cell Atlas White Paper. *arXiv 1810.05192* (2018).

239. Sasse, A., Ng, B., Spiro, A., Tasaki, S., Bennett, D. A., Gaiteri, C., De Jager, P. L., Chikina, M., Mostafavi, S. & Allen, P. G. How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks? *bioRxiv* (2023).