nature genetics

Article

https://doi.org/10.1038/s41588-024-01923-3

Interpreting *cis*-regulatory interactions from large-scale deep neural networks

Received: 28 July 2023

Accepted: 21 August 2024

Published online: 16 September 2024

Check for updates

Shushan Toneyan & Peter K. Koo

The rise of large-scale, sequence-based deep neural networks (DNNs) for predicting gene expression has introduced challenges in their evaluation and interpretation. Current evaluations align DNN predictions with orthogonal experimental data, providing insights into generalization but offering limited insights into their decision-making process. Existing model explainability tools focus mainly on motif analysis, which becomes complex when interpreting longer sequences. Here we present *cis*-regulatory element model explanations (CREME), an in silico perturbation toolkit that interprets the rules of gene regulation learned by a genomic DNN. Applying CREME to Enformer, a state-of-the-art DNN, we identify *cis*-regulatory elements that enhance or silence gene expression and characterize their complex interactions. CREME can provide interpretations across multiple scales of genomic organization, from *cis*-regulatory elements to fine-mapped functional sequence elements within them, offering high-resolution insights into the regulatory architecture of the genome. CREME provides a powerful toolkit for translating the predictions of genomic DNNs into mechanistic insights of gene regulation.

Recent advances in sequence-based genomic deep neural networks (DNNs) have shown notable success in predicting gene expression by considering substantially larger inputs[1–4], which aim to capture the influence of distal *cis*-regulatory elements (CREs). These DNNs bring promise to decode the *cis*-regulatory codes that drive differential gene expression across cell types, predict the effects of genetic variation and design novel regulatory sequences with desirable properties. However, the extensive sequence size of large-scale DNNs presents a challenge when evaluating their predictions and interpreting learned patterns.

Current methods for evaluating large-scale models have relied on assessing the alignment between predictions and existing experimental perturbation assays[1,5,6]—such as massively parallel reporter assays[7,8] and clustered regularly interspaced short palindromic repeats interference (CRISPRi)[9]—as well as statistical analyses such as expression-quantitative trait loci[6,10,11]. However, these only provide a narrow evaluation of how well DNN predictions agree with the specific biological question being probed within the studied loci. Moreover, the underlying biology can be difficult to assess because different experimental technologies introduce distinct biases and noise sources, which do not generalize across experimental technologies.

Conversely, prevailing post hoc model explainability methods concentrate primarily on the analysis of motifs[12–23], short DNA sequences associated with regulatory functions. As sequence inputs for DNNs grow longer, deciphering the complex coordination of motifs at a scale of hundreds of kilobases (kb) becomes increasingly difficult.

To bridge this gap, we present *cis*-regulatory element model explanations (CREME), an in silico perturbation toolkit designed to examine large-scale DNNs trained on functional genomics data. In contrast to existing model explainability methods, CREME can provide interpretations at various scales, including at a coarse-grained CRE level as well as a fine-grained motif level. CREME is based on the notion that by fitting experimental data, the DNN essentially approximates the underlying 'function' of the experimental assay. Thus, the trained DNN can be treated as a surrogate for the experimental assay, enabling in silico 'measurements' for any sequence, assuming generalization under covariate shifts (that is, predictions are reliable outside the distribution of training sequences).

Drawing inspiration from CRISPRi[24,25], CREME comprises a suite of multiscale in silico perturbation experiments to identify CREs, uncover their interactions with other CREs and quantify the CRE's

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, New York, NY, USA. e-mail: koo@cshl.edu

https://doi.org/10.1038/s41588-024-01923-3

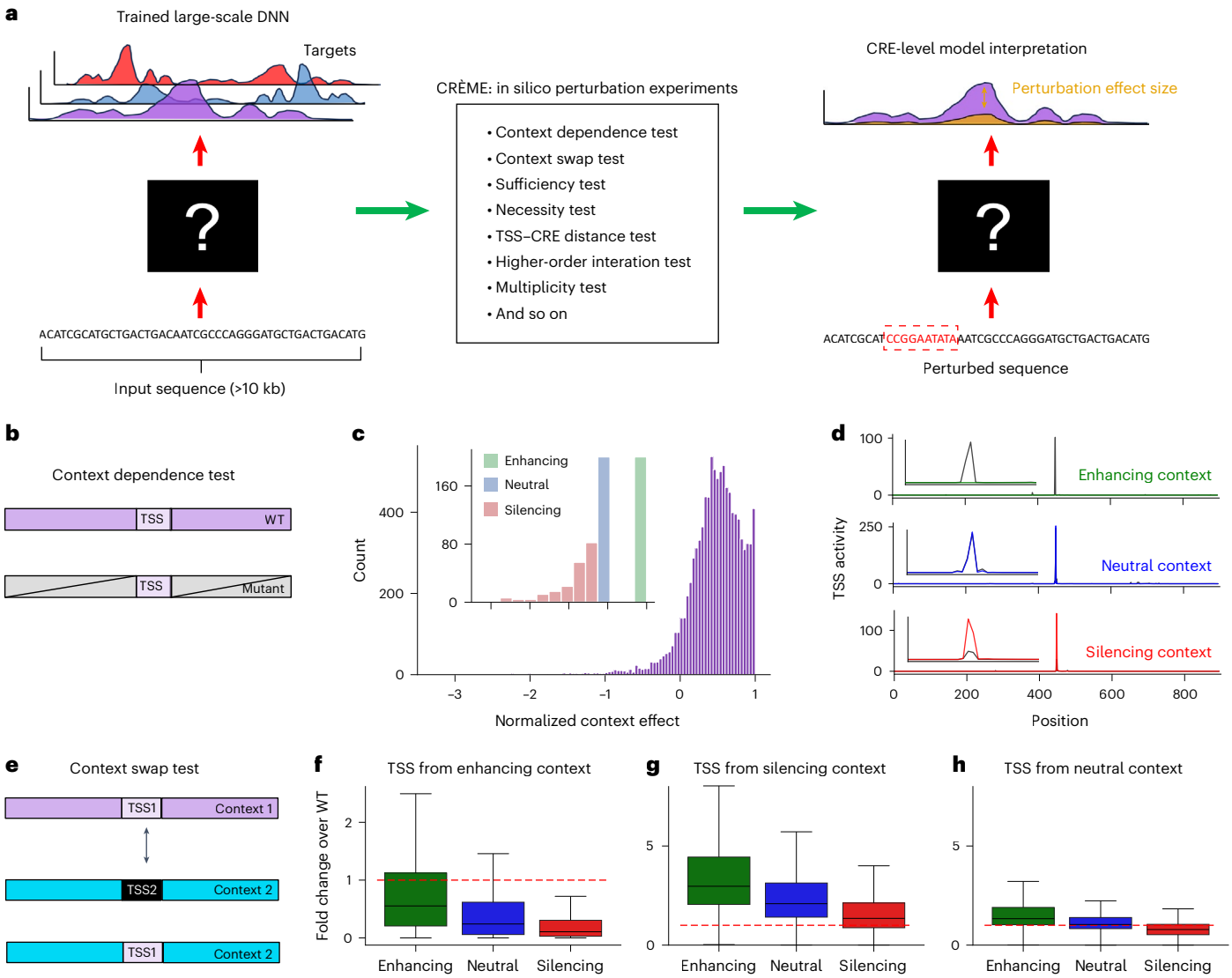


Fig. 1 | CREME overview and results for context perturbations in K562 cells using Enformer. a, CREME offers a suite of in silico perturbation experiments that probe specific biological hypotheses. Perturbations are applied to the input sequence, and the effect size is measured according to the change in model predictions. **b**, A schematic of the context dependence test. The sequence context is perturbed via a dinucleotide shuffle while keeping the central TSS tile (5 kb) intact. **c**, A histogram of the normalized context effect for 10,000 sequences that contain an active, annotated gene in K562 cells. Inset: the subset of sequences categorized as enhancing, silencing and neutral contexts ($N = 200$ randomly selected sequences for each context). **d**, Representative sequences from the three context categories showing Enformer's predictions before and after a context perturbation, with a zoomed in version shown in the inset. **e**, A schematic of the context swap test. **f**–**h**, Box plots of normalized fold change over WT–TSS activity predictions of chimeric divided by WT sequences, organized according to the original context categories: enhancing (**f**), silencing (**g**) and neutral (**h**). The number of data points in each box plot represents an all-versus-all comparison of each respective TSS in each possible context (40,000 data points in each box). Box plots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers). The red dashed line represents a fold change over WT of 1, which indicates no change in effect.

effect on target genes (Fig. 1a). Unlike wet laboratory experiments, the perturbations by CREME are performed in silico, so there are minimal restrictions on the scale of perturbations that can be applied. However, the challenge is that as the number of perturbations becomes larger, the space of possible combinatoric sets of perturbations grows exponentially. Thus, CREME proposes a suite of thoughtfully designed perturbation experiments that enable calibrated claims of gene regulation through the lens of genomic DNNs.

To demonstrate the utility of CREME, we interpret Enformer[1], a DNN that takes ~200 kb DNA sequences as input and predicts the corresponding read coverage profiles for 5,313 experiments that include chromatin accessibility, transcription factor binding, histone marks and gene expression across various human cell lines and tissues. In this study, we investigate the regulation of gene expression in K562, GM12878 and PC-3 cells. Using a curated list of sequences centered on transcription start site (TSS) annotations from GENCODE[26], we examine how specific sequence perturbations around a gene's TSS affect Enformer's predictions of a cap analysis gene expression sequencing (CAGE-seq track for the cell type under investigation. The results are organized according to specific biological questions.

Results

How much does distal context affect TSS activity?

To quantify the extent that Enformer relies on complex interactions of distal regulators[27,28], we use the context dependence test, which measures the effect of shuffling the sequence context beyond the proximal regions, ~5 kb centered on the TSS (Fig. 1b). We performed 100 independent dinucleotide shuffles and averaged predictions, following

a causal explanation method called the global importance analysis (GIA[15]). Similar to previous studies using Enformer[6,29], this assumes that dinucleotide-shuffled sequences are neutral and not out of distribution for the model.

For each cell line, we curated 10,000 genes with high predicted TSS activity, calculating the normalized context effect, which is the predicted difference for wild-type (WT) and mutant sequences, normalized by WT. Positive values (up to a maximum of 1) indicate decreased activity from context shuffling, while negative values indicate increased activity.

Interestingly, Enformer exhibited a range of responses to the context dependence test. Most context shuffles had positive effects, probably due to the enhancer disruption (Fig. 1c,d). Some cases showed no change in activity, possibly due to neutral context (that is, depletion of CREs), a balanced set of CREs with net neutral effect, or TSS activity is independent of context. In rarer cases, TSS activity increased, suggesting disruption of active silencing elements[30–33].

We replicated the analysis in GM12878 and PC-3 cell lines, finding similar context effect distributions (Extended Data Fig. 1a,b). While most genes showed consistent effects across cell lines, some distal context exhibited opposite effects—positive effect in one cell line and negative effect in another cell line (Extended Data Fig. 1c). However, the coarse resolution of this test cannot resolve whether the differences stem from the same or different CREs in the context.

We also applied the context dependence test to Borzoi, which comprises an ensemble of DNNs that consider longer input sequences (524 kb), observing similar trends across cell lines but with weaker effect sizes (Extended Data Fig. 2). This may be due to Enformer's stronger emphasis on fewer CREs within its smaller receptive field. Due to Borzoi's computational cost, we focus on Enformer for further analysis, using 200 randomly selected genes from each context category (Fig. 1c, inset), namely, enhancing, silencing and neutral.

How compatible are contexts for different genes?

Next, we used CREME's tile swap test to examine how gene expression changes when a TSS and its proximal context (5 kb) are inserted into nonnative genomic contexts (Fig. 1e). We quantified this using fold change over WT—the ratio of predicted activity for mutant versus WT sequences. We tested all TSS–context pair combinations (N = 360,000 sequences in K562) and stratified results by the source TSS tile's context category.

Interestingly, swapping a TSS tile from an enhancing context into other enhancing contexts resulted in ~50% decrease in activity, on average (Fig. 1f, middle). This suggests that, according to Enformer, enhancers can somewhat enhance any gene, but are better tuned for their native gene, possibly through compatibility rules[34–36]. Placing the TSS tile into nonenhancing contexts caused a larger activity drop, indicating these contexts lack enhancers or have weaker CREs.

On the other hand, silencing contexts acted more generically, effectively silencing any gene from silencing contexts (Fig. 1g). Inserting the TSS tile from silencing to nonsilencing contexts significantly increased activity, suggesting a depletion of silencing elements. Swapping the TSS tile from neutral contexts into other contexts showed modest changes (Fig. 1h). Neutral context genes were enriched for housekeeping genes (22% overlap) compared with enhancing and silencing contexts (5% overlap). Similar trends were observed across other cell types (Extended Data Fig. 1d,e).

Which CREs are necessary for gene expression?

The previous analyses provided a coarse-grained view of how distal sequence context (beyond 5 kb) influences TSS activity. CREME's necessity test provides a higher-resolution view of putative enhancers and silencers and quantifies their effect on TSS activity (Fig. 2a). We tiled the input sequence into nonoverlapping 5 kb bins and observed how shuffling each tile independently affects TSS activity. We chose 5 kb tiles to match average experimental perturbation lengths (for example, CRISPRi[37]), while remaining computationally tractable for perturbation experiments at scale. We average predictions over ten shuffles per tile perturbation.

We applied the necessity test to every context tile for each sequence (N = 22,800 tile perturbations for K562). The effect size was quantified by calculating the normalized shuffle effect, which is the difference in predicted activity between WT and mutant sequences divided by WT activity. Positive values indicate enhancing effects, negative values indicate silencing effects and zero indicates no effect. According to Enformer, all contexts contain a mix of weak enhancing and silencing tiles, with tiles from enhancing and silencing contexts tending toward positive and negative effects, respectively (Fig. 2a and Supplementary Data 1 for the list of CREs). Similar trends were observed across cell types (Extended Data Fig. 3a,b). Together, Enformer learns that 5 kb sequence elements exhibit a continuous range of mostly weak enhancing and silencing effects on gene expression.

Are individual CREs sufficient to drive TSS activity?

To examine whether individual tiles can explain the observed TSS activity levels[38], we employed CREME's sufficiency test. This involves embedding a tile of interest and the TSS tile into dinucleotide-shuffled sequences at their original positions and predicting TSS activity (Fig. 2b; N = 22,800 tile perturbations per cell type). This GIA experiment reveals the combined importance of a TSS–CRE pair while removing contributions from background context[15].

We used context-specific metrics owing to varying intrinsic TSS activity levels. For enhancing context, we calculated the tile effect over WT, which is the difference in predicted activity between context-shuffled sequences with both TSS and tile embedded and those with TSS alone, divided by WT activity. A value of 1 indicates the recovery of WT activity. For silencing and neutral contexts, we used the tile effect over control, calculated similarly but divided by the intrinsic TSS activity levels. A value of 0 indicates no change, 1 reflects doubled activity and negative values show decreased activity.

In all three cell lines, individual tiles in enhancing contexts act as enhancers or have no effect, with most showing minor positive effects and few driving substantial TSS activity (Fig. 2b and Extended Data Fig. 3c,d). In silencing contexts, tiles contain both strong enhancers and silencers relative to intrinsic TSS activity levels (Fig. 2c and Extended Data Fig. 3c,d), suggesting either an enrichment of silencers or synergistic interactions overpowering enhancers. Neutral contexts contain a balanced set of weak enhancers and silencers, which presumably cancel each other's effects (Fig. 2c and Extended Data Fig. 3c,d). Notably, a tile's necessity does not imply its sufficiency (Extended Data Fig. 3e). We observe similar trends when performing the sufficiency test using Borzoi (Supplementary Fig. 1), albeit with a narrower range of effect sizes compared with Enformer.

Characterization of sufficient CREs

Next, we analyzed the properties of sufficient tiles that act as strong enhancers or silencers for each cell line (Supplementary Data 1 for selected CREs). First we explored their positional distribution with respect to their target TSS. We found that enhancer tiles are concentrated near their target TSS, with density decreasing as distance increases (Fig. 2d and Extended Data Fig. 4a), in agreement with previous studies[6]. Silencer tiles exhibit a broader distribution but are still predominantly located close to the TSS (Fig. 2d and Extended Data Fig. 4a).

Next, we compared epigenetic features for enhancer and silencer tiles using ENCODE tracks[39] (Supplementary Data 1 for epigentic track list). Enhancer tiles were enriched for well-recognized features of enhancers[40], such as chromatin accessibility, H3K27ac, H3K9ac, and other known enhancer-associated histone marks (Fig. 2e,f). By contrast, silencer tiles were enriched for H3K9me3, which are associated with repressing gene repression[41]. These patterns were consistent

https://doi.org/10.1038/s41588-024-01923-3

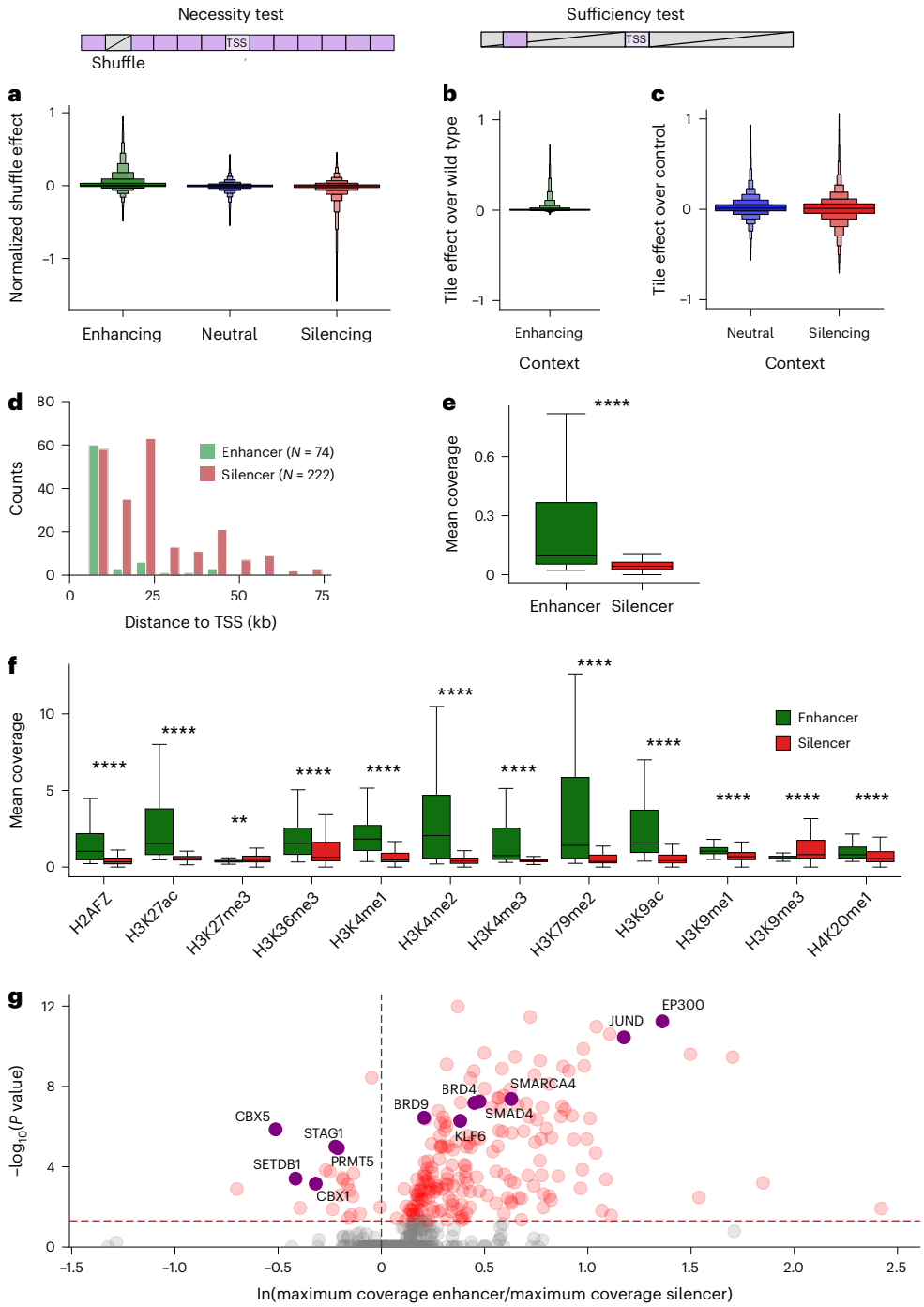


Fig. 2 | CRE-level analysis in K562 using Enformer. a, A box plot of the normalized shuffle effect for each tile in sequences from enhancing, neutral and silencing context categories (necessity test). The schematic of the necessity test above the panel shows a TSS-centered sequence with a second shuffled tile indicated by a gray slashed box. **b**,**c**, Box plots of tile effects normalized by WT (**b**) and control (**c**), that is the intrinsic TSS activity, for each tile in sequences from enhancing, neutral and silencing context categories (sufficiency test). The schematic of the sufficiency test above the panel shows a TSS tile and a tile of interest embedded in dinucleotide-shuffled sequences at their original positions. In **a**–**c**, the box plots have 7,600 context-derived tiles in each sequence. The context corresponds to 38 tile shuffles in each of the 200 sequences. **d**, A histogram of the distance between CRE tiles from TSS for 74 sufficient enhancer and 222 sufficient silencer tiles, defined by sufficiency thresholds (enhancers >0.3 from **b** and silencers <−0.3 from **c**). **e**,**f**, Box plots of the mean DNase-seq coverage (**e**) and mean histone mark coverage (**f**) of sufficient enhancer and silencer tiles. The number of data points in green and red boxes is 76 and 222, respectively. Statistical significance is given by the two-sided Mann–Whitney *U* test (*$P < 0.05$; ***$P < 0.001$; ****$P < 0.0001$). Box plots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers). **g**, A scatter plot of the TF enrichment analysis given by the significance level according to the two-sided Mann–Whitney *U* test (adjusted for multiple testing using Bonferroni correction) versus the log-fold enrichment of the maximum TF ChIP–seq coverage in sufficient enhancers versus silencers. Some activating and repressive TFs are annotated. The red dashed line indicates $P = 0.05$.

https://doi.org/10.1038/s41588-024-01923-3

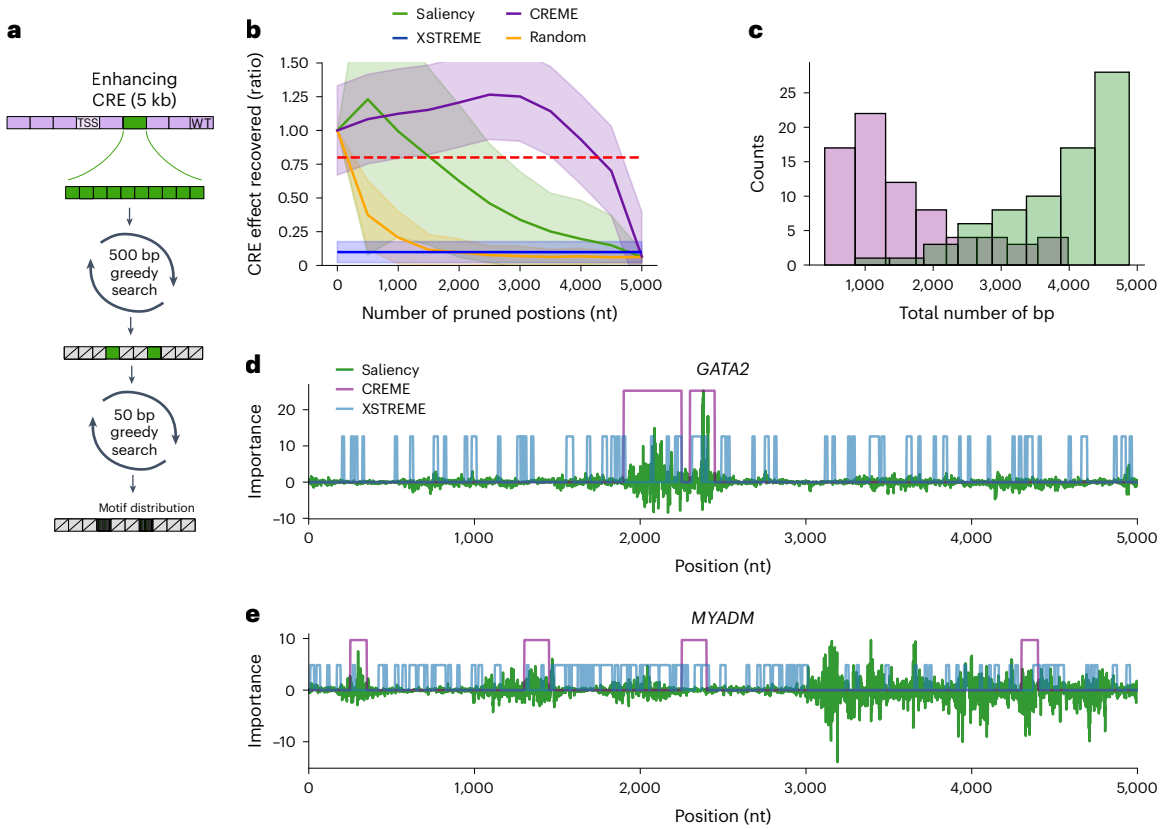


Fig. 3 | Fine-tile search results for enhancing tiles in K562. a, A schematic of the fine-tile search procedure. The green 5 kb tile represents a previously identified sufficient enhancer tile. The greedy search aims to prune the least enhancing subtiles, first at a 500 bp subtile resolution, followed by a second greedy search at a 50 bp sub-subtile resolution among the surviving 500 bp subtiles. The final output is a sequence with the core 50 bp elements among randomized sequences that recapitulates the sufficiency of the whole tile. **b**, A plot of the average CRE effect recovered ratio (which is the TSS activity when only sub-tiles are embedded in shuffled context divided by the the activity when the full tile is embedded) versus the number of pruned positions for 74 enhancer tiles using different methods. The dotted red line shows an arbitrary threshold of 80% explained. The shaded regions represent the standard deviation of the mean. **c**, A histogram of the number of remaining positions at a score of 0.8 for CREME (purple) versus saliency analysis. (green) **d**,**e**, Example annotations of important sequence elements for a putative enhancer for different genes: GATA2 (**d**) and MYADM (**e**). Note that the importance scale (*y* axis) is set according to saliency maps and the heights of binary annotations given by XSTREME and CREME are set arbitrarily for comparison. nt, nucleotide.

across cell types, except for H3K36me3 and H3K4me3, which showed cell type-specific enrichment (Extended Data Fig. 4b–d). Neutral background tiles showed intermediate activity or aligned more closely with silencing tiles (Supplementary Figs. 2–4).

Using RepeatMasker[42], we found silencer tiles had higher percentages of long interspersed nuclear elements (16.6% versus 9.7%) and long terminal repeats (7.4% versus 2.6%) compared with enhancer tiles (Supplementary Table 1). Most putative CREs mostly do not overlap with existing CRE annotations; only 27% of enhancer tiles and 24% of silencer tiles overlap with promoter annotations[26] or enhancer annotations[43].

Using transcription factor (TF) chromatin immunoprecipitation sequencing (ChIP–seq) data (Supplementary Data 1 for TF track list), we found putative silencers were enriched for known repressor TFs—including CBX5[44], PRMT5[45] and SETDB1[46]—while enhancers were enriched for activator TFs and remodelers, including EP300[47], BRD9[48] and SMARCA4[49].

These results support that CREME-derived putative CREs align with characteristic properties of enhancers and silencers. Unlike traditional approaches that define CREs through observational statistics of biochemical features[50], CREME identifies CREs that directly impact gene expression in an unbiased manner through a DNN's lens.

Fine-mapping sufficient sequence elements within CREs

While CREME can provide insights at any length scale, 5 kb was chosen to balance the scale of perturbations with the computational costs of Enformer. To achieve a higher-resolution interpretation within 5 kb tiles, we use CREME's fine-tile search (Fig. 3a). This method is a nested greedy search algorithm that systematically prunes subtiles via shuffle replacement at two resolutions to efficiently identify a sufficient set of subtiles that explains the whole tile's behavior. We monitor the CRE effect recovered—the ratio of subtiles' sufficiency test to the whole tile's sufficiency test—to ensure subtiles fully recapitulate the 5 kb tile's effect.

For comparison, we benchmarked CREME's fine-tile search with a motif enrichment tool called XSTREME[51] and a post hoc interpretability method called Saliency Maps[12], which quantifies the sensitivity of each nucleotide of a given sequence on model predictions. Focusing on sufficient enhancer tiles ($N = 74$), fine-tile search identified more compact subsets of sequences that better explained the full enhancer tile compared with saliency-based search and motif analysis (Fig. 3b,c). This suggests that motif analysis poorly characterizes sufficient sequence elements, while saliency maps (applied to Enformer) identify some essential positions but also attribute importance to random nucleotides.

A visual comparison of the GATA2 and MYADM enhancer tiles showed occasional agreement between saliency maps and CREME (Fig. 3d) but also significant differences (Fig. 3e and Supplementary Fig. 5). Motif hits were abundant throughout each 5 kb locus, with most not driving Enformer's predicted TSS activity. Similar trends were observed across different cell types (Supplementary Fig. 6).

https://doi.org/10.1038/s41588-024-01923-3

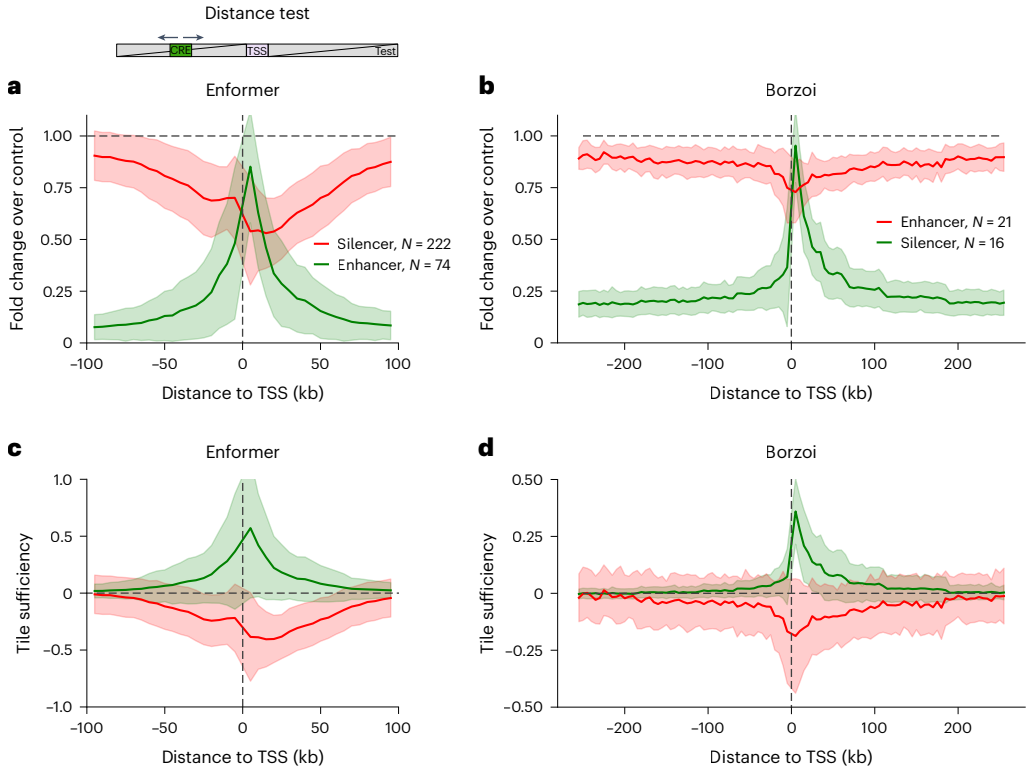


Fig. 4 | TSS–CRE distance test schematic and results. a,**b**, Average plots of the fold change over the maximum of moving a CRE tile along different positions in a sequence (5 kb steps) about a fixed TSS tile in shuffled sequences; the maximum represents the maximum TSS activity across all embedded positions for Enformer (**a**) and Borzoi (**b**). Shaded regions represent the standard deviation of the mean. The schematic above **a** illustrates the distance test. **c**,**d**, Average plots of the tile sufficiency versus distance to TSS for Enformer (**c**) and Borzoi (**d**). Tile sufficiency is calculated according to the predicted TSS activity with a TSS–CRE pair at a given distance minus the control sequence (shuffled context with just the TSS) divided by the WT sequence for enhancers and by the control sequence for silencers. Shaded regions represents the standard deviation of the mean. **a**–**d**, Vertical dashed line is a guide-to-the-eye for the position directly to the right of the TSS tile. Horizontal dashed line serves as a guide-to-the-eye for a fold change over control of 1.0 (**a**,**b**) and a tile sufficiency of 0.0 (**c**,**d**).

The multiscale fine-tile search effectively fine-maps functional sequence elements within putative CREs. Unlike attribution methods that identify important positions through gradients or additive approximations[13,14], CREME identifies a compact set of nucleotides that are both necessary and sufficient for model predictions. Due to computational costs, we continue our study defining putative CRE elements at a 5 kb resolution.

How does a CRE's effect on a target gene depend on distance?

To map a CRE's distance-dependent effect on gene expression, we developed the TSS–CRE distance test. This GIA experiment monitors TSS activity while systematically varying the distance of an enhancer or silencer tile from the target TSS (Fig. 4a). To focus on general trends across genes, we normalized by the maximum value for each gene; a value close to 1 indicates the location of the highest and lowest activity for enhancers and silencers, respectively.

For TSS–enhancer pairs (N = 2,812 GIA experiments), Enformer learned a strong asymmetric distance-dependent relationship (Fig. 4a). TSS activity is highest when the enhancer is juxtaposed next to the TSS tile, preferring the gene body, with effect decaying with distance, similar to experimental studies[52–54]. This suggests that, according to Enformer, a weak enhancer can increase its effect on gene expression by moving closer to the TSS in a preferred direction.

TSS–silencer pairs (N = 8,436 experiments) show a similar asymmetric effect decreasing with distance, but dropping off less rapidly than enhancers (Fig. 4a). Similar trends were observed across cell types (Extended Data Fig. 5).

Similarly, Borzoi-derived enhancers (N = 2,142 experiments) exhibit a strong asymmetric distance-dependent decay from the TSS, albeit Borzoi appears to consider distal enhancers up to 100 kb (Fig. 4b). By contrast, Borzoi-derived silencers (N = 1,632) yielded weaker effects and decayed slower with distance relative to Enformer (Fig. 4b).

We repeated the TSS–CRE distance test for Borzoi[4], using putative CREs identified in Borzoi's sufficiency test; 102 positions were explored for 21 enhancers and 16 silencers, yielding 2,142 and 1,632 GIA experiments, respectively. Similarly, Borzoi-derived enhancers exhibit a strong asymmetric distance-dependent decay from the TSS, albeit Borzoi appears to consider distal enhancers up to 100 kb (Fig. 4c). By contrast, Borzoi-derived silencers yielded weaker effects and decayed slower with distance relative to Enformer (Fig. 4d).

Identifying minimal sets of CREs that drive TSS activity

As individual tiles were not sufficient to recapitulate WT TSS activity levels, we show how CREME's higher-order interaction test can be used to identify minimal CRE sets that maximally alter TSS activity via coordinated, multitile perturbations (Fig. 5a). In anticipation of complex CRE interactions[27,55], we searched for necessary tile sets via an iterative greedy search, where, in each round, a new, necessary tile is identified (via a shuffle perturbation) that yields the largest effect size. The tile will then be fixed with random sequences and the next round will search among the remaining tiles. The objective is to optimize for lower or higher activity for enhancer or silencer sets, respectively; both tests converge to activity in fully random contexts, mirroring the context dependence test.

For enhancer sets, across all cell lines, five enhancers on average drive over 80% of TSS activity in enhancing contexts (Fig. 5b and Extended Data Fig. 6a). These enhancers are mostly centered around the TSS, but can span Enformer's entire receptive field (Fig. 5d and

https://doi.org/10.1038/s41588-024-01923-3

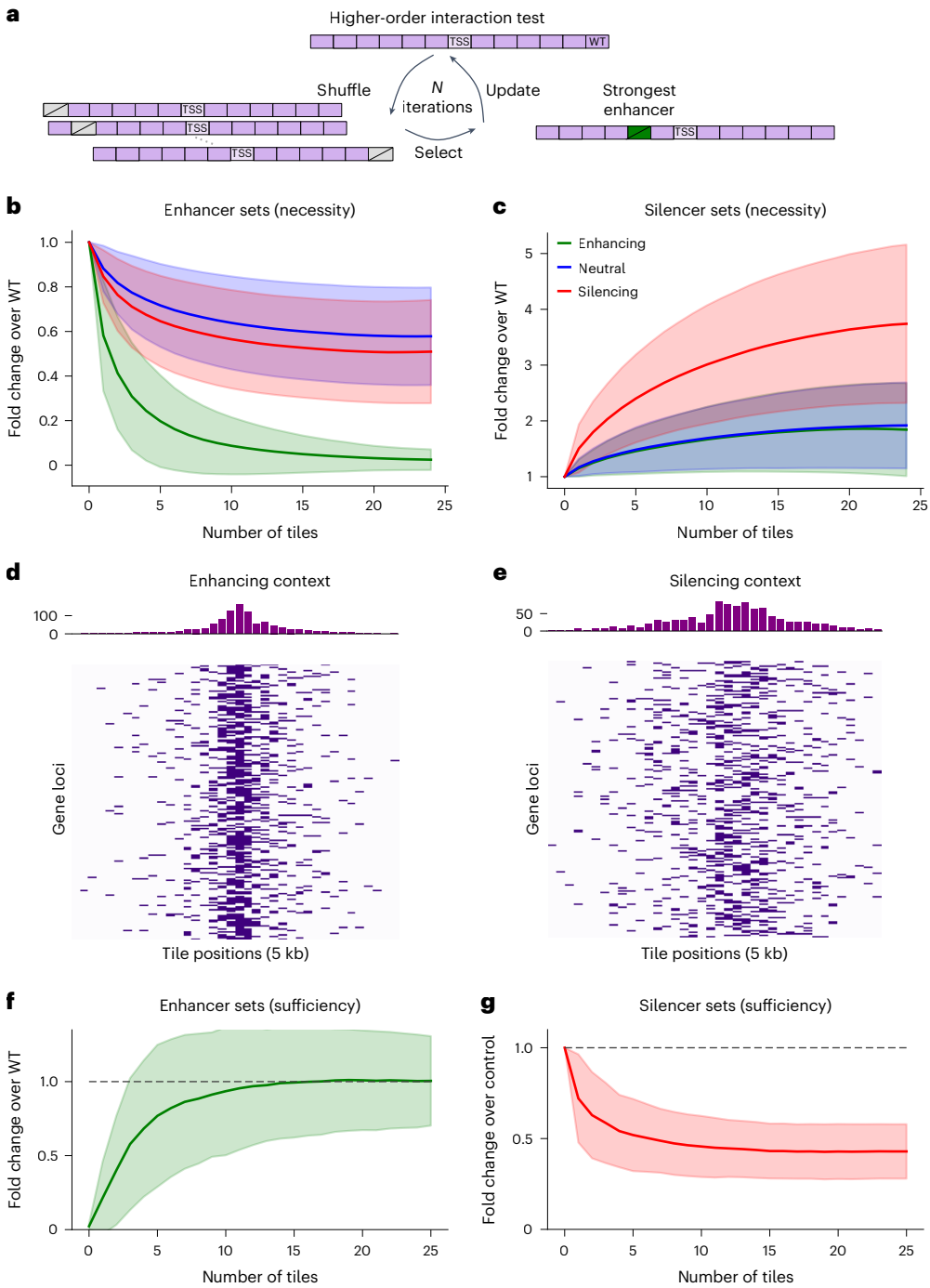


Fig. 5 | Optimal CRE sets reveal complex interactions for K562 using Enformer. a, A schematic of the higher-order interaction test for enhancer CRE sets (gray slashed boxes represent a shuffled tile and green box indicates an enhancer tile). **b**,**c**, Average plots of the greedy search results for enhancer tile sets (**b**) and silencer tile sets (**c**) for sequences from different context categories ($N = 200$ sequences for each context category). The fold change over WT is the predicted TSS activity of the shuffled CRE tiles in each round of the greedy search (indicated by the number of tiles). **d**,**e**, Heatmaps of the locations of the first five tiles identified in the enhancer or silencer greedy search within sequences from an enhancing (**d**) or silencing (**e**) context, respectively. The histograms on top show the distribution of tile positions. **f**,**g**, The sufficiency of the tile sets identified in each round of greedy search: the average fold change over WT (**f**) and control (**g**), which represents shuffled sequences with just the TSS tile. Sufficiency places the tile sets along with the TSS tile into shuffled sequences. The shaded region represents the standard deviation of the mean. **f**,**g**, Horizontal dashed line serves a a guide-to-the-eye for a fold change of 1.0.

Supplementary Fig. 7). Neutral and silencing contexts also contain enhancers, but with smaller effects, never driving activity to zero.

For silencer sets, silencing contexts are enriched with more silencers, each with smaller effect sizes (Fig. 5c and Extended Data Fig. 6b). Silencers are more broadly distributed than enhancers (Fig. 5e and Supplementary Fig. 8).

By testing the sufficiency of the set of necessary tiles, we found that while five enhancers were sufficient to explain 80% of WT activity, 11 tiles were needed to fully explain it (Fig. 5f and Extended Data Fig. 6c). On the other hand, three silencer tiles were sufficient to suppress 50% of intrinsic TSS activity, with saturation reached by round 6 (Fig. 5g and Extended Data Fig. 6d). The discrepancy between necessity

https://doi.org/10.1038/s41588-024-01923-3

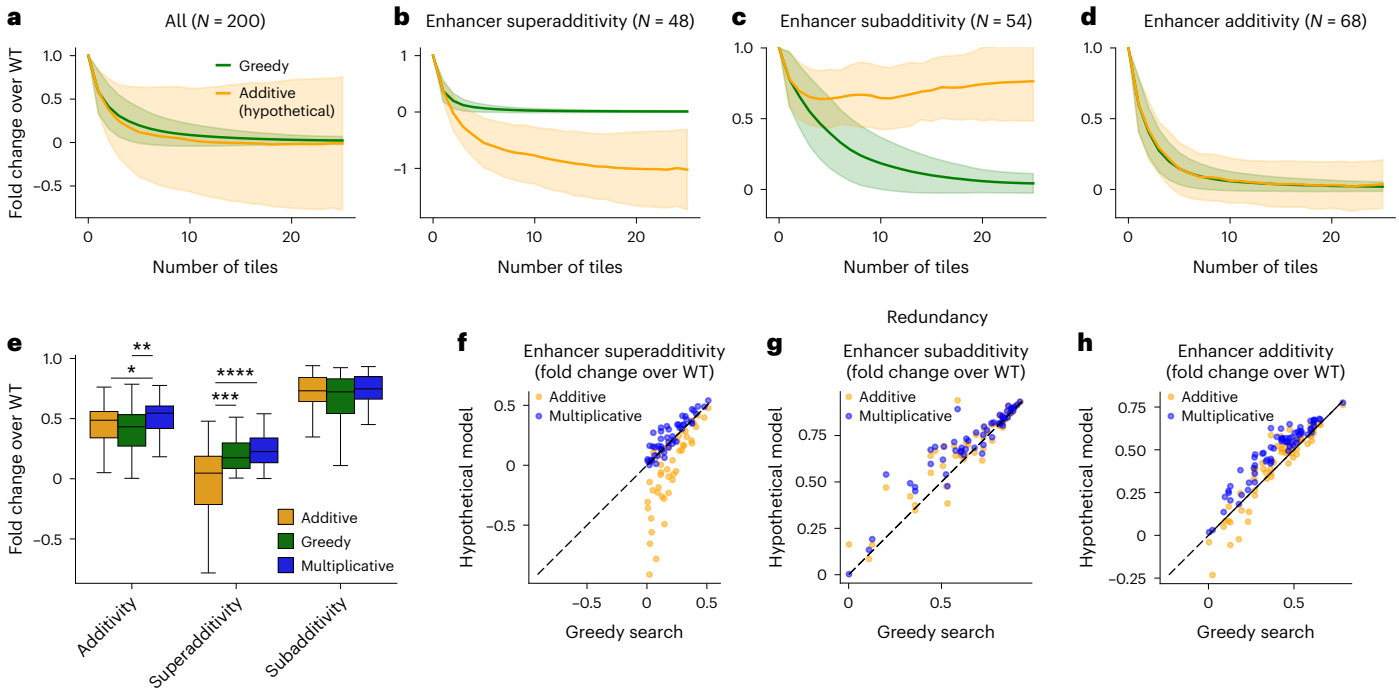


Fig. 6 | Investigation of CRE interactions. a, A comparison of the average fold change over control for enhancer sets for 200 sequences categorized as enhancing context versus a hypothetical additive effects model. **b**–**d**, The 200 sequences from enhancing contexts are stratified according to interaction type, superadditivity (**b**), subadditivity (**c**) and additivity (**d**), using mean squared error-based thresholds of 0.1 for superadditivity and subadditivity and 0.05 for additivity definition (with some ambiguous cases left out of the classification). The shaded region represents the standard deviation of the mean. **e**, A comparison of the hypothetical additive model and hypothetical multiplicative model versus greedy search outcomes at iteration 2 of the higher-order interaction test. The number of points in each box is 68, 48 and 54 for additivity, superadditivity and subadditivity, respectively. Box plots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers). Statistical significance was given according to the two-sided Mann–Whitney *U* test (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$). **f**–**h**, Greedy search versus hypothetical additive or multiplicative models for sequences categorized as enhancer superadditivity (**f**), enhancer subaddtivitiy (**g**) and enhancer additivity (**h**) from sequences categorized in **b**–**d**, respectively. Scatter plots show a more detailed view of the data in **e** with the *x* axis showing the higher-order interaction test outcomes and the *y* axis showing the hypothetical model outputs (additive or multiplicative). Dashed line indicates a guide-to-the-eye for a perfect correlation.

and sufficiency for silencers could be due the presence of enhancers within the context, which are present in the necessity test but not in the sufficiency test.

Together, this suggests that, according to Enformer, sequence context contains numerous enhancers and silencers and their overall net effect drives the observed TSS activity.

How do sets of CREs interact?

While the higher-order interaction test can identify sets of CREs, it does not inform their interactions. To gain insights into CRE interactions, we compared higher-order interaction test results with a hypothetical additive effects model to gain insights into CRE interactions. According to Enformer, sets of enhancer tiles are largely additive on average (Fig. 6a and Extended Data Fig. 7a,b), in agreement with previous studies[56]. However, when stratified, enhancers exhibit complex nonadditive behaviors (Fig. 6b–d and Extended Data Fig. 7a,b), including superadditivity and subadditivity. Subadditivity occurs when multiple enhancers have small individual effects due redundancy[57–60]. Superadditivity suggests CRE interdependence, possibly from synergistic[27,55] or multiplicative effects[6,35,61,62]. The non-monotonic change in individual enhancer effect sizes during the greedy search—a perturbation to one enhancer modifies the effect size of other enhancers—suggests strong dependencies between enhancers (Extended Data Fig. 8).

Next, we tested whether superadditivity could be explained by a multiplicative effects model by analyzing the natural log of predicted TSS activity. For enhancer pairs, the multiplicative model better explained superadditive cases, while the additive model explained additive cases well (Fig. 6e–h and Extended Data Fig. 7c–f). However, for larger enhancer sets, the multiplicative model diverged from observed effects (Extended Data Fig. 9 and Supplementary Fig. 9), suggesting synergistic interactions[63].

We also compared multitile perturbations with a hypothetical additive model for silencer sets and found that silencers predominantly exhibit superadditivity (Extended Data Fig. 9 and Supplementary Fig. 10). Notably, we did not observe any cases of silencer subadditivity, which may be due to our biased selection of active genes. A large subset did, however, align better with a multiplicative model. Together, these results suggests that, according to Enformer, CREs interact in complex ways and their behavior varies in a locus-specific manner across genes.

How does Enformer respond to dosage of CREs?

To test whether CRE subadditivity arises from TSS activity following a sigmoidal function[64–69], we used CREME's multiplicity test. This measures the effect of greedily inserting multiple enhancers (or silencers) within dinucleotide-shuffled sequences to maximize (or minimize) TSS activity. We observed that Enformer's predictions saturate as the dosage of enhancers (or silencers) increases, with different genes plateauing at different TSS activity levels (Extended Data Fig. 10). This suggests that subadditivity cases for enhancer sets probably reflect reaching the gene's saturation levels. Further research is needed to understand factors regulating saturation properties.

Discussion

CREME provides a suite of in silico experiments for unbiased interpretations of large-scale sequence-based DNNs, enabling CRE-level analysis similar to CRISPRi perturbations. To help navigate perturbation design space, CREME proposes systematic multiscale perturbation experiments that enable drawing precise claims of gene regulation. By interpreting Enformer using various necessity and sufficiency tests, we found that each gene's sequence context contains numerous enhancers and silencers. These CREs exhibit a continuum of positive and negative effect sizes, which highlights the difficulty in categorical definitions of enhancers and silencers through arbitrary activity cutoffs. Moreover, sets of enhancing and silencing CREs interact in complex ways, including additivity, superadditivity (multiplicative and synergistic) and subadditivity. These interactions vary in a locus-specific manner across genes.

The complex CRE interactions suggest that single or paired CRISPRi perturbations would be insufficient to fully characterize the key regulatory factors for a given gene. Thus, CREME provides a roadmap to improving perturbation experiments to better characterize *cis*-regulatory mechanisms. However, insights gained through DNN interpretation should be treated as hypotheses and validated by laboratory experiments.

One major limitation is a potential misalignment between DNN understanding and biological reality. In this case, CREME results might reflect artifacts learned by the model. For instance, Enformer's underestimation of distal enhancer effects, which was previously identified through a comparison with CRISPRi data[6], was directly revealed through CREME's TSS–CRE distance test. In addition, out-of-distribution shifts in perturbed sequences may reduce the reliability of DNN predictions[70]. In this study, we sought to limit the negative impacts of distributional shifts by staying close to genomic sequences, performing multiple trials and considering control experiments. Moreover, CREME is currently limited to sequence-to-function models.

CREME's key hyperparameters include tile size and number of perturbations. In this study, we opted for a 5 kb tile size as it balances potential insights to be gained with scaling perturbation experiments. Nevertheless, CREME's fine-tile search can provide a higher-resolution understanding within 5 kb tiles. Moreover, choice of metrics to interpret results can probably lead to variable takeaways. While we provide our rationale for the metric used for each test, alternative choices could be more insightful; this depends on the question being asked. Additionally, further stratification could help to further dissect the rules of gene regulation, especially for highly expressed versus lowly expressed genes.

While CREME comes equipped with many useful perturbation experiments, the toolkit is extensible; different perturbation experiments can be constructed to address different biological questions.

CREME advances our ability to interpret sequence elements at multiple scales using a DNN's predictions, instead of surrogates or additive approximations. CREME elucidates biological factors directly driving DNN predictions, as opposed to observational analysis of functional activity or motif analysis. While this study focused on CRE-level perturbations, CREME can also interpret modestly sized DNNs using smaller tile sizes, providing insights into motifs and their interactions. With further advances in machine learning that drive the alignment between genomic DNNs and biological reality to narrow, methods like CREME could facilitate better-informed hypotheses that can help guide more efficient laboratory-based perturbation experiments. Moving forward, we envision CREME will help to study challenging questions of gene regulation, such as enhancer–promoter compatibility rules and the sequence basis of interactions among weak enhancers or silencers, opening the doors to synthetically designing or rewiring *cis*-regulatory networks that control gene expression.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-024-01923-3.

References

1. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
2. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
3. Karbalayghareh, A., Sahin, M. & Leslie, C. S. Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Res.* **32**, 930–944 (2022).
4. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. Preprint at *bioRxiv* https://doi.org/10.1101/2023.08.30.555582 (2023).
5. Toneyan, S., Tang, Z. & Koo, P. K. Evaluating deep learning for predicting epigenomic profiles. *Nat. Mach. Intell.* **4**, 1–13 (2022).
6. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 1–29 (2023).
7. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
8. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science* **339**, 1074–1077 (2013).
9. Qi, L. S. et al. Repurposing crispr as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
10. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
11. Huang, C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat Genet.* **55**, 2056–2059 (2023).
12. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proc. of the International Conference on Learning Representations* (ICLR, 2014).
13. Scott, M., and Lee Su-In. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* **30**, 4765–4774 (2017).
14. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning* 3145–3153 (2017).
15. 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).
16. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
17. Koo, P. K. & Ploenzke, M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat. Mach. Intell.* **3**, 258–266 (2021).
18. Hammelman, J. & Gifford, D. K. Discovering differential genome sequence activity with interpretable and efficient deep learning. *PLoS Comput. Biol.* **17**, e1009282 (2021).
19. Liu, G., Zeng, H. & Gifford, D. K. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC Bioinform.* **20**, 401 (2019).

20. 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).
21. Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D. & Barash, Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biol.* **21**, 149 (2020).
22. Linder, J. et al. Interpreting neural networks for biological sequences by learning stochastic masks. *Nat. Mach. Intell.* **4**, 41–54 (2022).
23. Seitz, E. E., McCandlish, D. M., Kinney, J. B. & Koo, P. K. Interpreting *cis*-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nat. Mach. Intell.* **6**, 701–713 (2024).
24. Fulco, C. P. et al. Systematic mapping of functional enhancer–promoter connections with crispr interference. *Science* **354**, 769–773 (2016).
25. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
26. Frankish, A. et al. Gencode 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
27. Lin, X. et al. Nested epistasis enhancer networks for robust genome regulation. *Science* **377**, 1077–1085 (2022).
28. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region capture micro-c reveals coalescence of enhancers and promoters into nested microcompartments. *Nat. Genet.* **6**, 1048–1056 (2023).
29. Luthra, I. et al. Regulatory activity is the default dna state in eukaryotes. *Nat. Struct. Mol. Biol.* **3**, 559–567 (2024).
30. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Geneti.* **52**, 254–263 (2020).
31. Stampfel, G. et al. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).
32. Kulkarni, M. M. & Arnosti, D. N. *cis*-regulatory logic of short-range transcriptional repression in drosophila melanogaster. *Mol. Cell. Biol.* **25**, 3411–3420 (2005).
33. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).
34. Martinez-Ara, M., Comoglio, F., van Arensbergen, J. & van Steensel, B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol. Cell* **82**, 2519–2531 (2022).
35. Bergman, D. T. et al. Compatibility rules of human enhancer and promoter sequences. *Nature* **607**, 176–184 (2022).
36. Narita, T. et al. The logic of native enhancer-promoter compatibility and cell-type-specific gene expression variation. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.18.500456 (2022).
37. Armendariz, D. A., Sundarrajan, A. & Hon, G. C. Breaking enhancers to gain insights into developmental defects. *eLife* **12**, e88187 (2023).
38. 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).
39. Luo, Y. et al. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
40. Igolkina, A. A. et al. H3k4me3, h3k9ac, h3k27ac, h3k27me3 and h3k9me3 histone tags suggest distinct regulatory evolution of open and condensed chromatin landmarks. *Cells* **8**, 1034 (2019).
41. Monaghan, L. et al. The emerging role of h3k9me3 as a potential therapeutic target in acute myeloid leukemia. *Front. Oncol.* **9**, 705 (2019).
42. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
43. Gao, T. & Qian, J. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2020).
44. Zhang, Y., See, Y. X., Tergaonkar, V. & Fullwood, M. J. Long-distance repression by human silencers: chromatin interactions and phase separation in silencers. *Cells* **11**, 1560 (2022).
45. Jin, Y. et al. Targeting methyltransferase prmt5 eliminates leukemia stem cells in chronic myelogenous leukemia. *J Clin Invest.* **126**, 3961–3980 (2016).
46. Griffin, G. K. et al. Epigenetic silencing by setdb1 suppresses tumour intrinsic immunogenicity. *Nature* **595**, 309–314 (2021).
47. Garcia-Carpizo, V. et al. CREBBP/EP300 bromodomains are critical to sustain the GATA1/MYC regulatory axis in proliferation. *Epigenetics Chromatin* **11**, 30 (2018).
48. Del Gaudio, N. et al. BRD9 binds cell type-specific chromatin regions regulating leukemic cell survival via STAT5 inhibition. *Cell Death Dis.* **10**, 338 (2019).
49. Lazar, J. E. et al. Global regulatory DNA potentiation by SMARCA4 propagates to selective gene expression programs via domain-level remodeling. *Cell Rep.* **31**, 107676 (2020).
50. Benton, M. L., Talipineni, S. C., Kostka, D. & Capra, J. A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics* **20**, 511 (2019).
51. Grant, C. E. & Bailey, T. L. XSTREME: comprehensive motif analysis of biological sequence datasets. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.02.458722 (2021).
52. Zuin, J. et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).
53. Zhan, Y. et al. Reciprocal insulation analysis of Hi-C data shows that tads represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* **27**, 479–490 (2017).
54. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
55. Choi, J. et al. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *eLife* **10**, e65381 (2021).
56. Martinez-Ara, M., Comoglio, F. & van Steensel, B. Large-scale analysis of the integration of enhancer-enhancer signals by promoters. Preprint at *bioRxiv* https://doi.org/10.1101/2023.08.11.552995 (2023).
57. Kvon, E. Z., Waymack, R., Gad, M. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* **22**, 324–336 (2021).
58. Frankel, N. et al. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493 (2010).
59. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
60. Perry, M. W., Boettiger, A. N. & Levine, M. Multiple enhancers ensure precision of gap gene-expression patterns in the drosophila embryo. *Pro. Natl Acad. Sci. USA* **108**, 13570–13575 (2011).
61. Hong, C. K. Y. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters. *Genome Res.* **32**, 85–96 (2022).
62. Zhou, J. L., Guruvayurappan, K., Chen, H. V., Chen, A. R. & McVicker, G. P. Genome-wide analysis of crispr perturbations indicates that enhancers act multiplicatively and without epistatic-like interactions. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.26.538501 (2023).

63. Sanford, E. M., Emert, B. L., Coté, A. & Raj, A. Gene regulation gravitates toward either addition or multiplication when combining the effects of two signals. *eLife* **9**, e59388 (2020).
64. 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).
65. Melen, G. J., Levy, S., Barkai, N. & Shilo, B.-Z. Threshold responses to morphogen gradients by zero-order ultrasensitivity. *Mol. Syst. Biol.* **1**, 2005–0028 (2005).
66. Burz, D. S., Rivera-Pomar, R., Jäckle, H. & Hanes, S. D. Cooperative DNA-binding by bicoid provides a mechanism for threshold-dependent gene activation in the drosophila embryo. *EMBO J.* **17**, 5998–6009 (1998).
67. Doughty, B. R. et al. Single-molecule chromatin configurations link transcription factor binding to expression in human cells. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.02.578660 (2024).
68. Bothma, J. P. et al. Enhancer additivity and non-additivity are determined by enhancer strength in the drosophila embryo. *eLife* **4**, e07956 (2015).
69. Scholes, C., Biette, K. M., Harden, T. T. & DePace, A. H. Signal integration by shadow enhancers and enhancer duplications varies across the drosophila embryo. *Cell Rep.* **26**, 2407–2418 (2019).
70. Ovadia, Y. et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Adv. Neural Inf. Process. Syst.* https://papers.nips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf (2019).

Methods

Inclusion and ethics

This study relied solely on computational models that were trained using public resources. It did not make use of individual-level data, and no specific ethics approval was required. Some of the data sources might present biases toward European ancestries.

Enformer

Enformer is a previously established DNN that takes genomic sequences of the length 196,608 bp as input and predicts 5,313 epigentic tracks for human and 1,643 epigenetic tracks for mouse biosamples through two output heads[1]. For each track, Enformer's predictions cover 896 binned positions, with each bin representing 128 bp. This represents the central 114,688 bp of the input sequence. The extended input sequence, provides context for the edge cases, that is the start and end of the predictions. The epigenetic tracks consist of processed coverage values of expression (CAGE), DNA accessibility (DNase-seq), TF binding and histone modification (ChIP–seq). Enformer is composed of convolutional layers that initially summarize the input sequence into representations of 128 bp bins. This is followed by 11 transformer blocks that use multihead self-attention[71]. We acquired code for the Enformer model along with trained weights from https://tfhub.dev/deepmind/enformer/1as per instructions in the 'Methods' section of ref. 1.

Borzoi

Borzoi is a DNN similar to Enformer in that it predicts a range of epigenetic tracks from input DNA sequences. However, it considers a larger input size of 524,288 bp and makes predictions for strand specific outputs of CAGE (and other tracks) at a 32 bp resolution. Borzoi is an ensemble of four models trained on different data splits. Here, we adopted the same strategy as the authors of Borzoi of running inference using all four models and averaging the results. Each model in the ensemble is composed of convolutional layers that summarize the input into representations of 128 bp bins, eight transformer layers that use multihead self-attention and deconvolution layers that upsample the representations to 32 bp resolution. We acquired code for the Borzoi model along with trained weights from https://github.com/calico/borzoi/tree/main as per instructions in the 'Methods' section of ref. 4.

Transcription start site selection

We acquired human annotations from GENCODE[26] (https://www.gencodegenes.org/human/) and filtered for 'transcript' annotations and 'protein coding' genes. We then extracted sequences of the length 196,608 bp (or 524,288 bp for Borzoi) from the GRCh38 reference genome centered at each filtered TSS. We converted the sequences to a one-hot representation, treating *N* characters as a uniform probability (that is 0.25). We calculated Enformer's prediction for these sequences and considered the mean at positions 447 and 448 (of the 896 binned predictions), which corresponds to the central TSS. We used tracks 4,824, 5,110 and 5,111 of the human output head (corresponding to PC-3, GM12878 and K562 CAGE predictions, respectively). We refer to this scalar coverage value per cell line as the TSS activity. To focus our study on genes that yield high TSS activity, we considered the top 10,000 unique genes per cell line with the highest predicted read coverage.

CREME: *cis*-regulatory element model explanations

CREME is an in silico perturbation toolkit that can uncover rules of gene regulation learned by a large-scale DNN. The rationale behind CREME stems from the concept that DNNs are function approximators. Thus, by fitting experimental data, the DNN is effectively approximating the underlying 'function of the experimental assay'. By treating the DNN as a surrogate for the experimental assay, CREME can be queried with new sequences and provide in silico 'measurements', albeit through the lens of the DNN. Inspired by wet laboratory experiments, such as CRISPRi[24,25,72], that perturb genomic loci to uncover how CREs influence gene expression, we devised a suite of in silico perturbation experiments that interrogate a DNN's understanding of long-standing questions of gene regulation, including the context dependence of gene expression[31,73], identification of enhancing and silencing CREs and their target genes[24,30], distance dependence of CREs to target genes on gene expression, the complex higher-order interactions of CREs and the effect of finer-scale elements on gene expression[34,35,58–60,64]. Since DNN predictions may not fully capture the underlying biology when fitting experimental data, CREME is strictly a model interpretability tool. Below, we detail the different in silico perturbation tests explored in this paper.

Regaring the CREME investigation of Enformer, for the vast majority of the experiments, we only considered TSS activity, which we define as the central 5 kb tile centered on the input sequence. Enformer's receptive field for this tile covers roughly 200 kb sequences, so the 200 kb region centered on the sequence is what is probed in our experiments. We split the central 200 kb sequences into 38 nonoverlapping 5 kb tiles (such that the tiles are fully within the input sequence), with the central tile corresponding to the TSS of an annotated gene. We define the TSS activity as the mean of Enformer's prediction for bins 447 and 448, which are the central bins. Predictions for tracks 4,824, 5,110 and 5,111 of the human output head correspond to PC-3, GM12878 and K562 cell lines.

Context dependence test. The context dependence test aims to measure the effect size of TSS activity in random contexts (derived from dinucleotide-shuffled versions of the WT sequence). This test measures the extent to which a prediction of a given TSS activity is influenced by its context which may contain enhancers and silencers. To do this, we computed the difference between WT TSS activity and that of a dinucleotide-shuffled context case. We normalized by WT TSS activity. Detailed steps can be found in Supplementary Note 1.

In an example case, assuming WT prediction is 10 and mutant is 5 (that is, shuffling leads to a drop in activity), the normalized effect would equal

$$\frac{\mathrm{WT}-\mathrm{MUTANT}}{\mathrm{WT}}=\frac{10-5}{10}=0.5.$$

For the interpretation, the effect size of 0 means that the context is neutral and has no effect on the TSS predictions (that is, WT and mutant yield the same prediction). Positive effect size means that the central TSS prediction for the mutated sequence is lower than WT, which indicates that we have perturbed an enhancing context. Negative effect size means that the central TSS prediction for the mutated sequence is higher than WT, which suggests that we have perturbed a silencing context.

For analysis, we categorized the sequences into silencing, neutral and enhancing contexts based on their context effect on TSS. We identified three regions: (1) enhancing context were chosen based on an effect size of more than 0.95, (2) neutral context was chosen if the absolute effect size was less than 0.05 and (3) silencing context was chosen based on an effect size of less than −0.3. If the number of data points in a category was above 200, we randomly sampled 200 sequences to cap the number for further experiments. We used these groups per cell line throughout the experiments.

For each cell line, the breakdown of contexts in each category is given in Supplementary Table 2.

For Borzoi, we chose to proceed with less strict thresholds: −0.9, 0.05 and −0.2 for enhancing, neutral and silencing contexts, respectively, to address the weaker context effect sizes given by Borzoi. The number of detected sequences in each category is given in Supplementary Table 3.

Context swap test. The context swap test aims to measure the extent that TSS activity depends on a specific genomic context or measure compatibility with other contexts. For this, we embedded the TSS of one sequence into another, to replace the existing TSS. We then obtained TSS activity prediction and normalized it by the WT prediction of the first sequence. Detailed steps can be found in Supplementary Note 1.

For an example case, assuming WT prediction for a given sequence from which the TSS is taken is 20 and mutant, that is, the activity of the same TSS in a new background, is 10 (that is, moving the TSS to the new background leads to a lower activity), the normalized effect would equal

$$\frac{\text{MUTANT}}{\text{WT}} = \frac{10}{20} = 0.5.$$

Assuming both sequences are from enhancing backgrounds, this would indicate that the new context is less enhancing compared with its native TSS (classified based on context dependence test).

For the interpretation, if the fold change over the control is 1, the TSS activity in the new context is the same as in its native context. If the value is below 1, the TSS is less active in the new context and vice versa for values above 1.

For analysis, we performed the context swap test on the sequences filtered by the context dependence test per cell line. Specifically, we placed the TSSs in each context category across all other context categories, separately keeping track of the source TSS and the context category.

Necessity test. The necessity test measures the importance of a putative CRE on the central TSS activity for a given sequence context while the other tiles remain intact. To perform this, we independently shuffled each 5 kb tile and obtained mutant TSS predictions. We then normalized these by subtracting the WT activity and divided the difference by WT. Detailed steps can be found in Supplementary Note 1.

For example, assuming WT activity is 10 and it drops to 5 after shuffling 1 CRE (indicating that the shuffled CRE is enhancing), the normalized shuffle effect would equal

$$\frac{\text{WT} - \textit{MUTANT}}{\text{WT}} = \frac{10-5}{10} \approx 0.5.$$

For the interpretation, an effect size of 0 means that the WT value is similar to the case when the tile is shuffled, and therefore, the tile shuffle has no effect on the TSS. A large positive value means that the shuffled case yields a lower TSS signal than the WT, that is, the tile shuffling leads to a drop in TSS activity, and a large negative value means the shuffling leads to a higher TSS activity compared with WT.

For analysis, we performed the necessity test on all tiles within the sequences from the context dependence test that had enhancing, silencing or neutral backgrounds (as classified by selected thresholds).

Sufficiency test. The sufficiency test measures the effect of a given CRE on its TSS in otherwise random contexts, that is, in isolation from the rest of the tiles from the original WT sequence. This essentially measures whether the CRE by itself is enough to up- or downregulate the TSS. For this, we individually inserted 5 kb tiles into context-shuffled sequences and obtained mutant predictions. We also computed TSS activity in context-shuffled case as control. We then normalized by subtracting control from mutant. For enhancing context sequences we divided this by the WT activity, in others by control. Detailed steps can be found in Supplementary Note 1.

For example, assume an enhancing sequence with a WT prediction of 10 and control (only TSS) sequence with predicted activity of 2. Given a strong enhancer CRE that recovers the activity in the mutant (TSS and CRE) to 12, the normalized activity would equal

$$\frac{\text{MUTANT} - \textit{CONTROL}}{\text{WT}} = \frac{12-2}{10} = 1.$$

Given a silencing context sequence with WT activity of 10, control activity of 20 (note that silencing context sequences by definition have higher activity after context shuffle) and a strong silencer that represses the signal to 10, the normalized effect would be computed as

$$\frac{\text{MUTANT} - \textit{CONTROL}}{\text{CONTROL}} = \frac{10-20}{20} = -0.5.$$

For the interpretation, in case of enhancing context sequences, the TSS activity drops to a small value after the whole context shuffling (control). Therefore, here we compare the value to the original WT sequence, subtracting the small activity of the TSS on its own (control). This can be interpreted as the extent to which a given tile individually restores TSS activity to the original WT value. A value of 0 means that the tile has no positive effect on the TSS (mutant equals control), positive values indicate an enhancing effect (mutant > control) and vica versa for negative values.

In case of silencing and neutral context sequences, the TSS activity in the CONTROL condition is not always a small value (by definition the silencing contexts are the ones where shuffling the context leads to a high TSS activity). This leads to ambiguities in cases when the WT and control contribute to the normalized values. Therefore, we simply computed the tile effect as fraction change observed when adding the tile compared with CONTROL. A value of 0 means that tile addition has no effect on TSS, a positive value indicates that tile addition activates the TSS and a negative value means that the tile lowers TSS activity.

For analysis, we performed the sufficiency test on the same subset of sequences as necessity test that had enhancing, silencing or neutral backgrounds (as classified by selected thresholds). Based on sufficiency test results, we denote CREs from enhancing contexts with effect size larger than 0.3 as enhancers. Similarly, we define silencers as tiles from silencing contexts with effect size smaller than −0.3. In case of Borzoi, we used −0.15, a higher threshold, for silencers to include more elements and considered those in both silencing and neutral background.

TSS–CRE distance test. The TSS–CRE distance test is a GIA experiment where we systematically shift the position of a tile in shuffled sequences and measure its effect on TSS activity. For this, we inserted a given CRE in various positions across the context-shuffled sequence. We normalized this by dividing the activity at each position by that of the maximum activity across all positions in the sequence. Detailed steps can be found in Supplementary Note 1.

For example, assuming we consider only five possible test positions with mutant sequence activity values of 2, 5, 10, 5 and 2, the CONTROL would be set to 10, and the normalized activities would equal 0.2, 0.5, 1, 0.5 and 0.2, respectively.

For analysis, we used the definition of enhancers and silencers based on sufficiency test results. We performed the TSS–CRE distance test on CREs defined as enhancers within enhancing contexts and silencers in silencing contexts for each cell line.

Higher-order tile interaction test. The aim of the higher-order tile interaction test is to dissect CRE networks. Specifically, we compute the combined effect of multiple tile shuffles that have large effects through a greedy search. For enhancers, the iterative greedy search systematically identifies tiles that lead to a lower TSS activity when shuffled. We followed the same steps for silencer search but instead of choosing the minimum predicted value we chose the maximum predicted value in each iteration. Detailed steps can be found in Supplementary Note 1.

For example, we assume a sequence with two enhancing tiles E1 and E2 and WT activity of 100. In the first iteration, after shuffling E1,

the prediction drops to 50, and after shuffling E2, it yields a prediction of 70, while shuffling the other tiles maintains activity at WT levels. The normalized TSS activity in case of shuffling E1 would be computed as

$$\frac{\text{MUTANT}}{\text{WT}} = \frac{50}{100} = 0.5.$$

Given that E1 elicits the sharpest drop in TSS activity, if searching for enhancers, the E1 tile would be fixed with shuffled sequences and a subsequent round would search among the remaining tiles.

For analysis, we performed higher-order tile interaction test for maximally enhancing TSS activity and maximally silencing TSS activity for all sequences from different context categories.

Comparison with additive effects. To help understand the trajectories from the higher-order tile interaction test, we calculated the hypothetical effects of an additive model. In brief, the additive effects are calculated on the basis of combining the effects on TSS activity from the individual effects of each CRE (that is, calculated in the first round of the greedy search), following the CRE tile order found by the greedy search. This does not take into account cooperative or redundant relationships within sets of CREs, which would be captured in the greedy search.

For example, assuming a sequence with two strong enhancers E1 and E2 with fully additive effects, we would expect their effect sizes to add up to the prediction from when both are shuffled simultaneously. For instance, if the WT prediction is 100, and in iteration 1 shuffling just E1 leads to a prediction of 60 and E2 of 70, then the effect sizes are −40 and −30, respectively. The hypothetical additive model would then predict that the simultaneous shuffling of E1 and E2 would lead to a normalized effect size of

$$\frac{\text{WT} + \text{effect(E1)} + \text{effect(E2)}}{\text{WT}} = \frac{100 - 40 - 30}{100} = 0.3.$$

For the interpretation, if the hypothesis of additive effect holds, we would expect the greedy search trace to be the same as the additive or hypothetical trace for each sequence. If the results are different we can categorize these as superadditive or subadditive for cases when the additive model overestimates the shuffle effect or underestimates it, respectively. To illustrate an example scenario leading to a superadditive case let us assume two enhancers are cooperating, that is, their combined effect is larger than individual effects (a nonadditive case). We would expect their individual shuffle effects to also be larger than shuffling them simultaneously (because disabling one leads to a large effect size already). In contrast, subadditivity can arise if two enhancers are redundant, that is, their roles are overlapping, and the effect size will be small when only a single tile is shuffled (because the other enhancer tile can compensate). Therefore, the estimated additive effect (based on single tile shuffles of iteration 1) will underestimate the effect of shuffling both of the enhancers. This will thus lead to the additive hypothetical trace being higher than the one based on the greedy search.

For analysis, to characterize deviations, we computed the mean squared error (m.s.e.) of the greedy and hypothetical additive outputs for each sequence. We classified the cases where the m.s.e. value is above 0.1 (arbitrary threshold) and the greedy search results on average is greater than the average of additive as superadditivity cases. Similarly, we classified the cases where the m.s.e. value is above 0.1 (arbitrary threshold) and the greedy search result on average is lower than the average of additive as subadditivity cases. We used a stricter threshold of 0.05 for the m.s.e. value to classify sequences as additive. We applied these thresholds to both the enhancer and silencer search cases after assessing visually that the traces overlap substantially.

Comparison with multiplicative effects. Similarly, we compared greedy search outcomes to a multiplicative model. For this we computed the natural logarithm of all predicted activities and otherwise proceeded with the same steps as in 'Comparison with additive effects' section. We interpreted the results using the same logic as in the additive effect case, but we used different thresholds for classifying sequences into multiplicative and nonmultiplicative categories. We mostly observed cases of the hypothetical multiplicative model predicting higher values than the greedy search yielded and some where they aligned (hence the binary classification). We used the MSE thresholds 0.04 for enhancers (cases of m.s.e. above 0.04 were classified as nonmultiplicative) and 0.01 for silencers. Both were selected after visual inspection of the traces.

Sufficiency of CRE sets from higher-order test. In a complementary set of experiments to higher-order interaction test, we tested whether the identified CREs are sufficient to recover or suppress the TSS activity. For this, we inserted tiles into shuffled context sequences progressively adding more tiles, following the order of shuffling from the higher-order interaction test.

Multiplicity test. The multiplicity test measures how TSS activity scales upon repeated addition of an enhancing or silencing tile. With this GIA experiment, we aim to test the model's extrapolation behavior. Specifically, we probed whether TSS activity reaches saturation upon a high dosage of a CRE; saturation is when the predictions reach a plateau when we enrich for enhancers or silencers. The multiplicity test is similar to the greedy search used in the higher-order tile interaction test, with the exception that we are systematically adding the same CRE of interest into optimal positions in each round of dinucleotide-shuffled sequences. Detailed steps can be found in Supplementary Note 1.

For example, assuming a control activity of 10, if inserting CRE at a new position yields a prediction of 20, the normalized TSS activity would be computed as

$$\frac{\text{MUTANT}}{\text{CONTROL}} = \frac{20}{10} = 2.$$

For analysis, using the enhancers defined in 'Sufficiency test' section, we performed 15 iterations of the multiplicity test for each CRE.

Fine-tile search. This analysis aims to identify subsets of sequences that recover the majority of a given 5 kb enhancer CRE effect. To formalize the problem, we aimed to identify the smallest set of 50 bp sequences that, when embedded in otherwise shuffled context, recovers at least 80% of the enhancer CRE effect on the TSS. We did this using a two-stage nested greedy search, reducing the search space in each stage: first at a 500 bp resolution followed by a 50 bp resolution among the surviving 500 bp tiles. Detailed steps can be found in Supplementary Note 1.

We compared this approach with XSTREME[51] (using both de novo motifs and known motifs from the JASPAR database) and saliency-based motif embedding. For XSTREME analysis, we scanned the WT sequence for motif hits with FIMO[74] and embedded those subsequences into otherwise shuffled background sequences. We evaluated how much of the CRE effect is recovered by XSTREME motifs using the same normalized value (that is, by dividing with the activity when the whole CRE is embedded).

For this analysis, we used all the enhancing context sequences and the enhancer CREs identified within those in each cell line (74 in K562, 41 in GM12878 and 35 in PC-3). We used the same set of ten backgrounds per sequences for comparing CREME, XSTREME and saliency-based analyses.

Biochemical characterization of CREs

To characterize and contrast the chromatin state of the putative enhancing and silencing CREs (at a 5 kb window size) we downloaded (default) bigwig files from ENCODE and summarized the values at coordinates of the CREs according to the mean read coverage across

the 5 kb read coverage for epigenetic marks and the max read coverage for TF tracks (Supplementary Data 1).

For Supplementary Figs. 8–10, we sampled neutral tiles as a background for comparison. We selected tiles from neutral contexts, filtered only those with lower than 0.1 absolute sufficiency values, then randomly selected one tile per gene. We additionally subsampled the set to match the the average number of enhancing and silencing tiles.

We computed the percentage of enhancer and silencer sequences containing repeating sequences using RepeatMasker using the default parameters. The detailed outputs per repeating sequence category and cell line are given in the Supplementary Table 1.

Statistics and reproducibility

The main general design consideration in CREME experiments includes the choice of the number of shuffles to perform. This was chosen based on the relatively small standard deviation of the predictions and the limitations on the compute time based on how long each inference cycle took. Therefore, no statistical method was used to predetermine sample size. We did not exclude any samples; however, we did subsample sequences randomly, to cap the compute time in case of selecting enhancing, silencing and neutral sequences.

Reporting summary

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

Data availability

Final and intermediate results for paper reproducibility are available via Zenodo at https://doi.org/10.5281/zenodo.12584210 (ref. 75).

Code availability

Static code for reproducing the analyses in the manuscript is available via Zenodo at https://zenodo.org/records/12594513 (ref. 76). A bleeding-edge version of CREME is available via GitHub at https://github.com/p-koo/creme-nn and https://github.com/p-koo/CREME_paper_reproducibility. A stable version of CREME is installable via pip (PyPI at https://pypi.org/project/creme-nn/). Comprehensive documentation is provided on ReadTheDocs.org (API at https://creme-nn.readthedocs.io/en/latest/index.html and tutorials at https://creme-nn.readthedocs.io/en/latest/tutorials.html).

References

71. Vaswani, A. et al. Attention is all you need. In *Adv. Neural Inf. Process. Syst.* https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (2017).
72. Chen, P. B. et al. Systematic discovery and functional dissection of enhancers needed for cancer cell fitness and proliferation. *Cell Rep.* **41**, 111630 (2022).
73. Crocker, J. et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191–203 (2015).
74. Grant, C. E., Bailey, T. L. & Noble, W. S. Fimo: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
75. Toneyan, S. & Koo, P. Creme-nn data and results. *Zenodo* https://doi.org/10.5281/zenodo.12584210 (2024).
76. Toneyan, S. & Koo, P. Creme-nn code. *Zenodo* https://zenodo.org/records/12594513 (2023).

Acknowledgements

We thank S. Navlakha, J. Desmarais, J. Kinney and members of the Koo Lab for helpful comments on the manuscript. Research reported in this publication was supported in part by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG012131 (P.K.K.), the National Institute Of General Medical Sciences of the National Institutes of Health under award number R01GM149921 (S.T. and P.K.K.) and the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. This work was performed with assistance from the US National Institutes of Health Grant S10OD028632-01. We also thank the NVIDIA GPU Grant Program for support.

Author contributions

S.T. and P.K.K. conceived of the method and designed the experiments. S.T. developed code, ran the experiments and analyzed the results. S.T. and P.K.K. interpreted the results and contributed to writing the paper.

Competing interests

The authors declare no competing interests.

Additional information

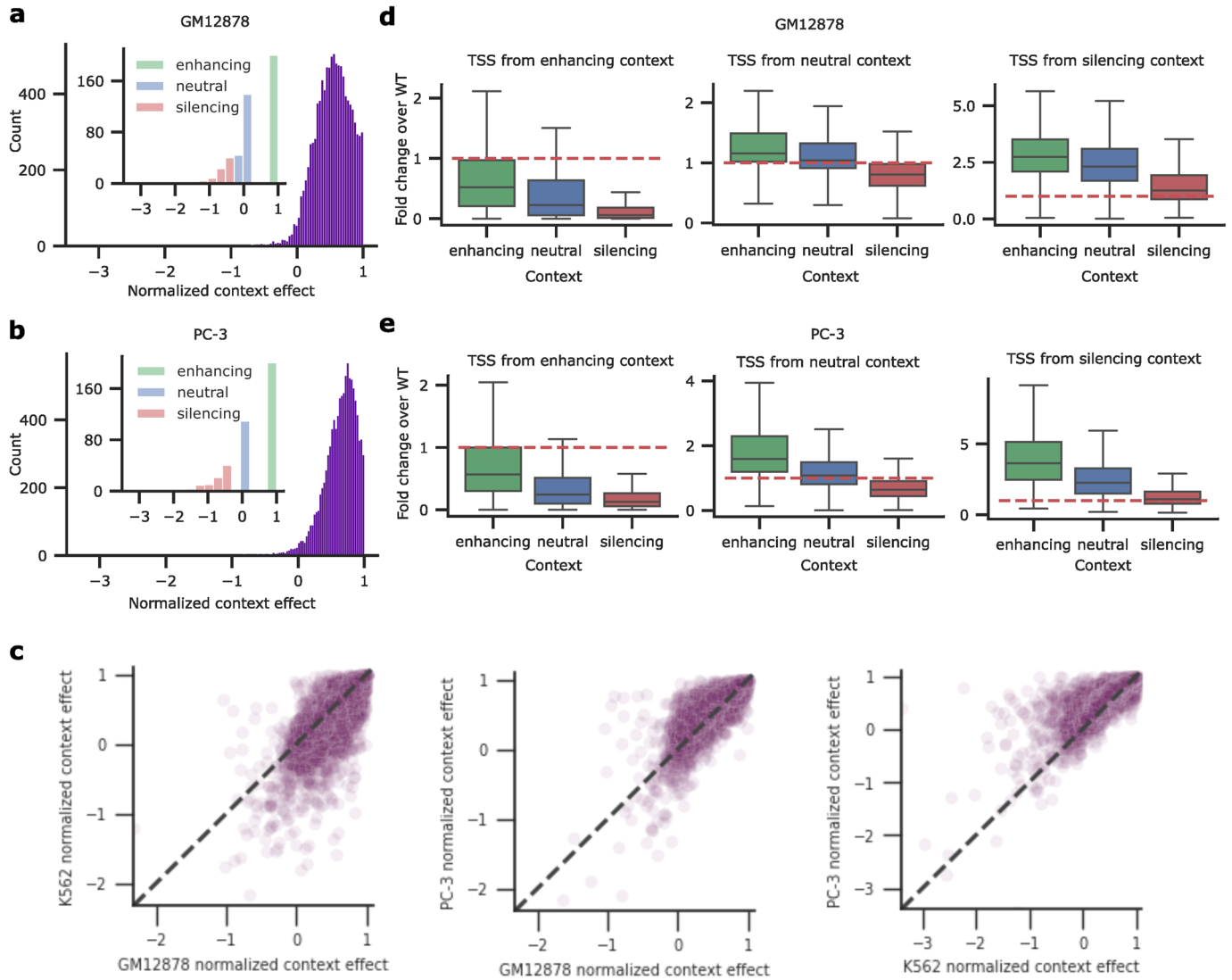
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01923-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01923-3.

Correspondence and requests for materials should be addressed to Peter K. Koo.

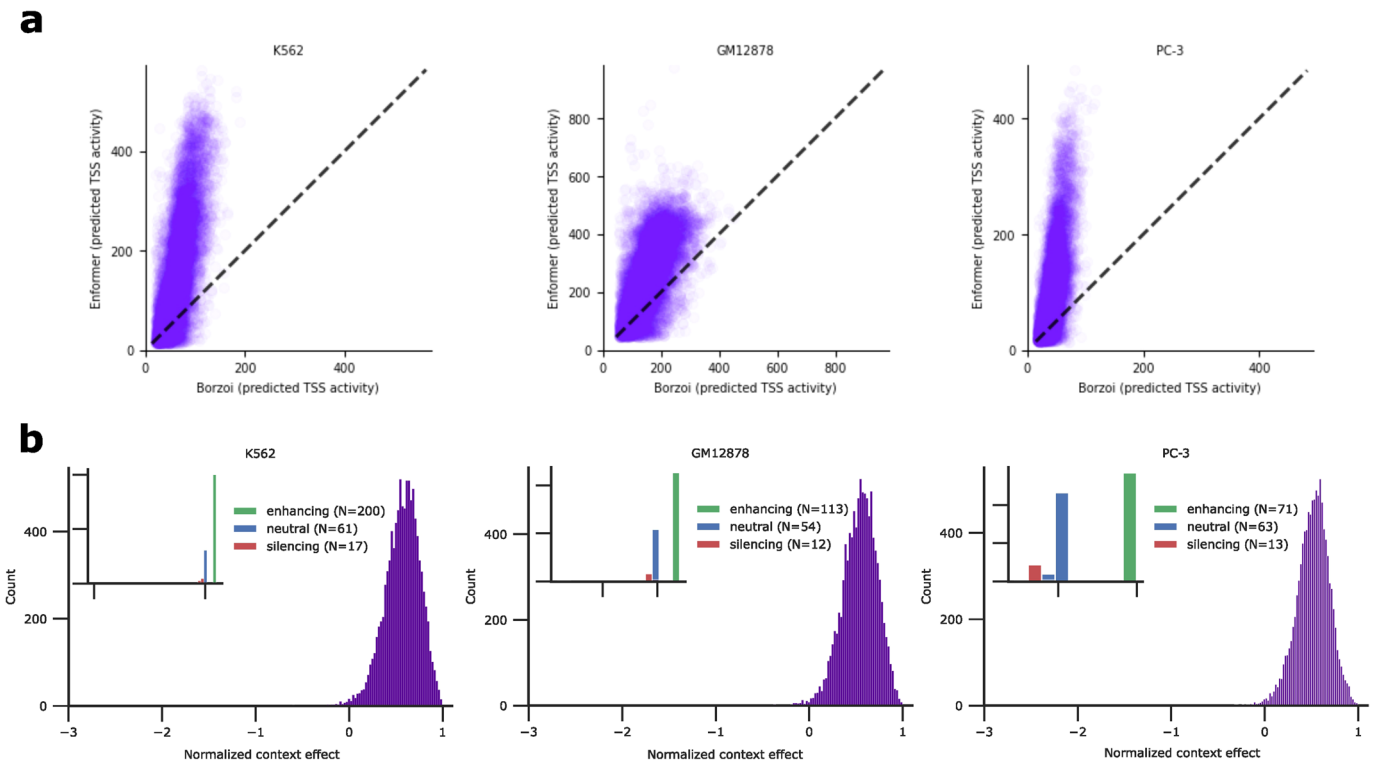
Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01923-3



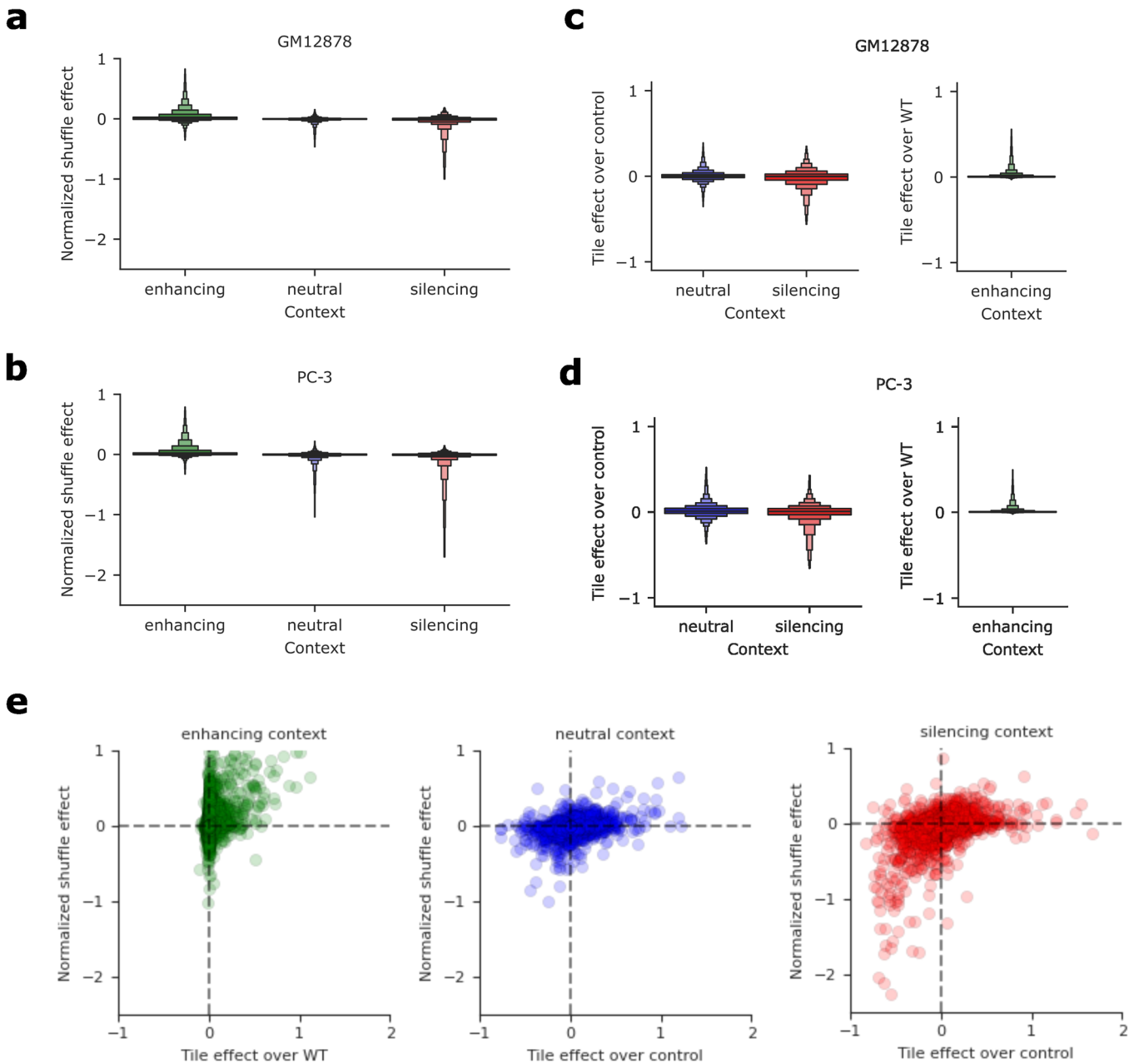
Extended Data Fig. 1 | Results of the Context Dependence Test and Context Swap Test for GM12878 and PC-3. a,b Histogram of normalized context effect from the Context Dependence Test for 10,000 sequences that contain an active, annotated gene in GM12878 and PC-3 cells. Inset shows the subset of sequences for enhancing, silencing and neutral contexts. **a** inset contains 200, 78 and 183 data points in enhancing, silencing and neutral context respectively. **b** inset contains 200, 90 and 110 data points in enhancing, silencing and neutral context respectively. **c**, Pairwise comparison of normalized context effects between cell lines for matched genes. The number of data points is 7688, 6946, 7492 from left to right. **d,e**, Context Swap Test results. Boxplots of normalized context effect on TSS for sequences with context perturbations given by insertion of the original TSS in different context categories. Results are organized according to the original TSS category: enhancing (*left*), neutral (*middle*), and silencing (*right*). The number of data points in each boxplot represent an all-vs-all comparison of each respective TSS in each possible context. The number of data points in **d** is 40,000, 36,600, 15,600 in boxplots for TSS from enhancing context, 36,600, 33,489, 14,274 in TSS from neutral context and 15600, 14274, 6084 in TSS from silencing context. The number of data points in e is 40,000, 22,000, 18,000 in boxplots for TSS from enhancing, 22,000, 12,100, 9,900 in TSS from neutral context and 18,000, 9,900, 8,100 in TSS from silencing context. Boxplots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers).

https://doi.org/10.1038/s41588-024-01923-3



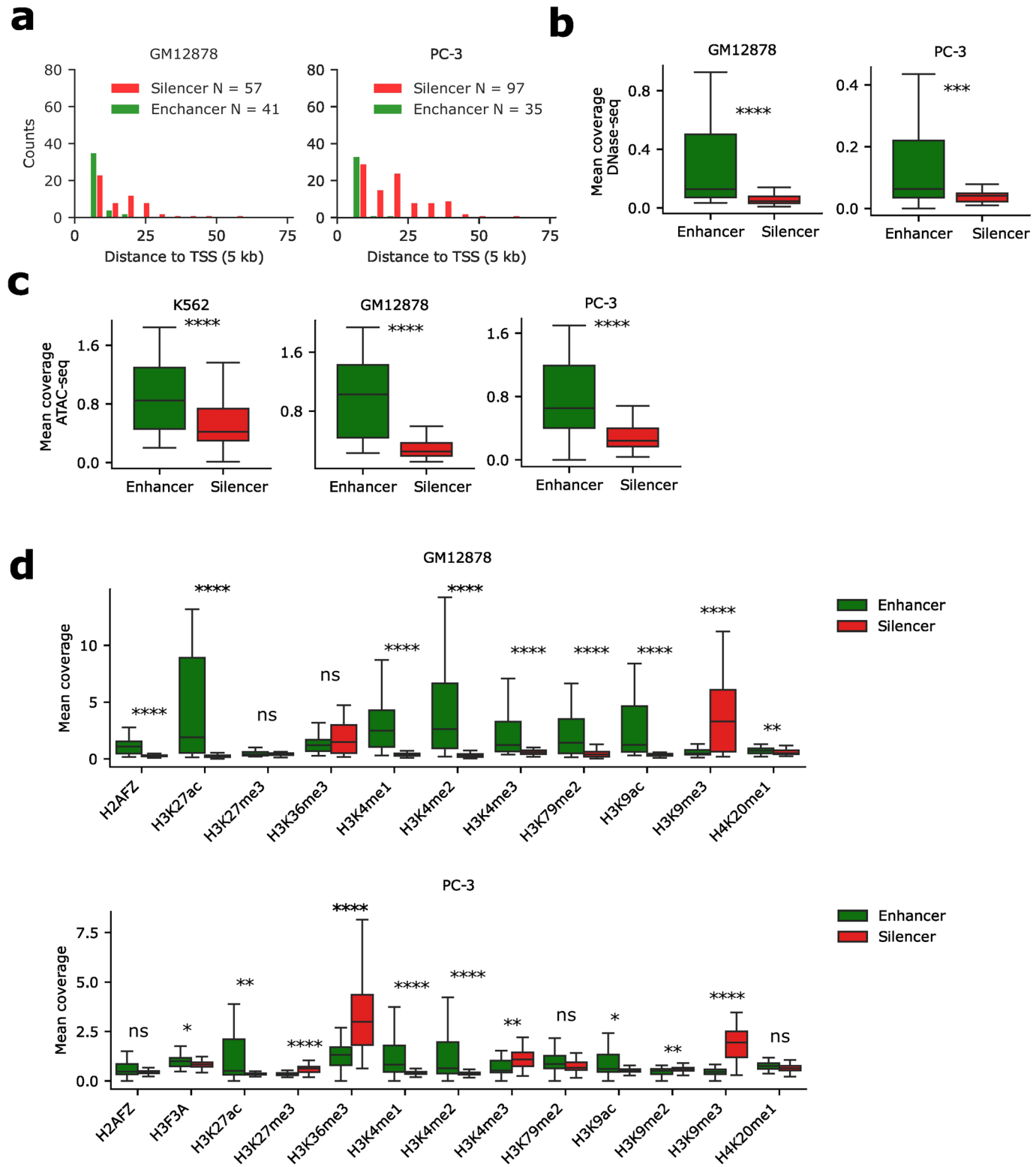
Extended Data Fig. 2 | Borzoi Context Dependence Test results. a, Scatter plot comparing the wild-type activity predicted by Enformer versus Borzoi for the matched cell types and for matched genes. **b**, Histogram of normalized context effect for the 10,000 highest activity, annotated genes (according to Borzoi's predictions) for K562, GM12878 and PC-3 cells. Inset shows the subset of sequences for enhancing, silencing and neutral contexts. The number of data points is shown in inset legend.

https://doi.org/10.1038/s41588-024-01923-3



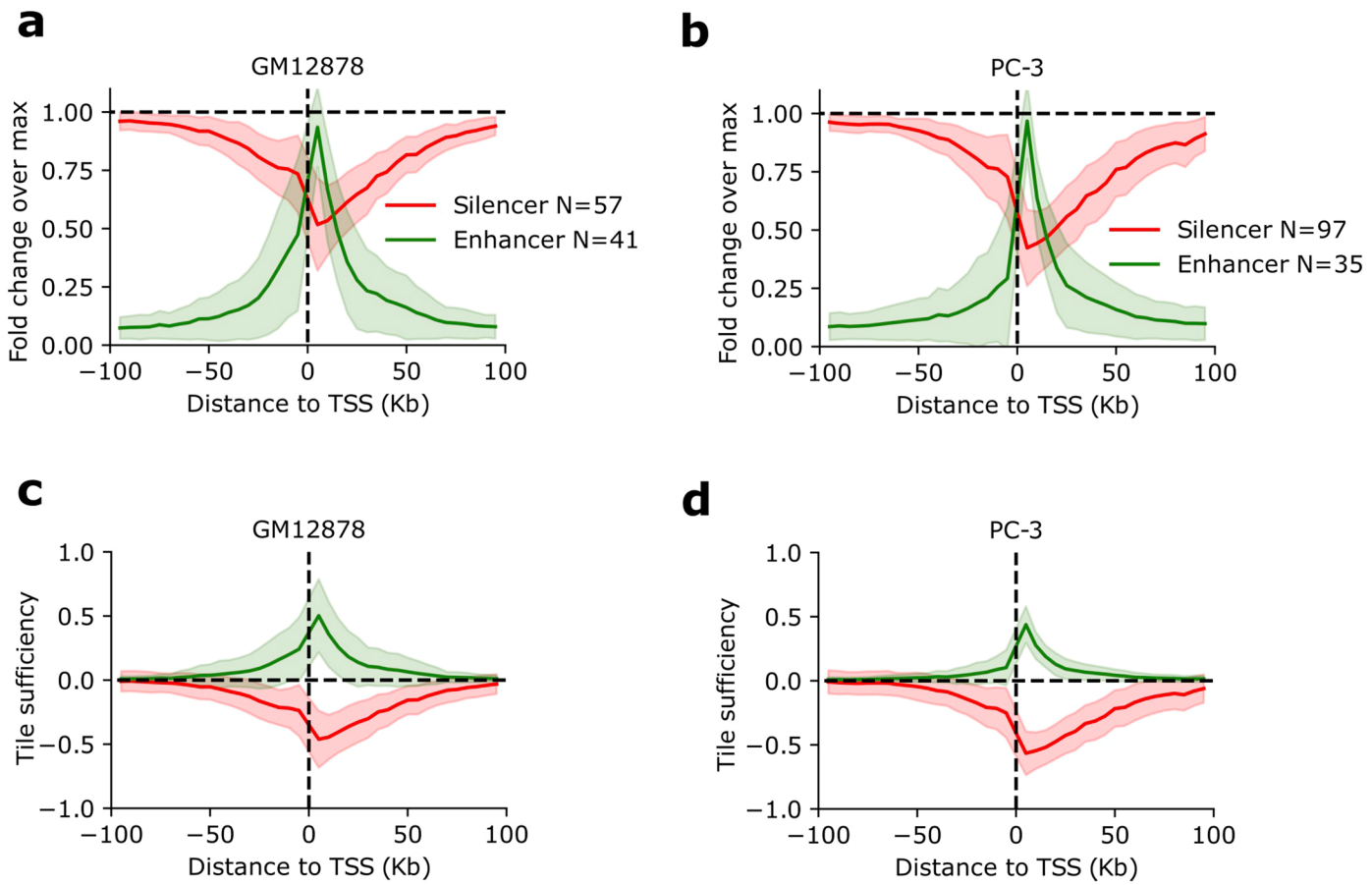
Extended Data Fig. 3 | CRE effects on TSS activity in GM12878 and PC-3 cell lines. a,b, Boxen plot of the normalized shuffle effect for each tile in sequences from enhancing, neutral and silencing context categories (Necessity Test) for GM12878 (**a**) and PC-3 (**b**). The number of data points in **a** is 7600, 6954, 2964 and in **b** is 7600, 4180, 3420 in enhancing, neutral and silencing contexts respectively. **c, d**, Boxen plot of tile effects for each tile in sequences from enhancing, neutral and silencing context categories (Sufficiency Test) for GM12878 (**c**) and PC-3 (**d**). Normalization is with predicted TSS activity for wild-type (enhancing context) and control, that is the intrinsic TSS activity (neutral and silencing context). Boxen-plots have the same number of data points as in **a** and **b**. In panels **a** – **d** center lines of boxenplots show the median and boxes in both directions always indicate half of the remaining data. **e**, Scatter plot between the results from the Necessity Test (y-axis) versus the results from the Sufficiency Test (x-axis) in K562 cell line (N = 7,600 in each plot corresponding to 200 sequences with 38 tiles in each).

https://doi.org/10.1038/s41588-024-01923-3



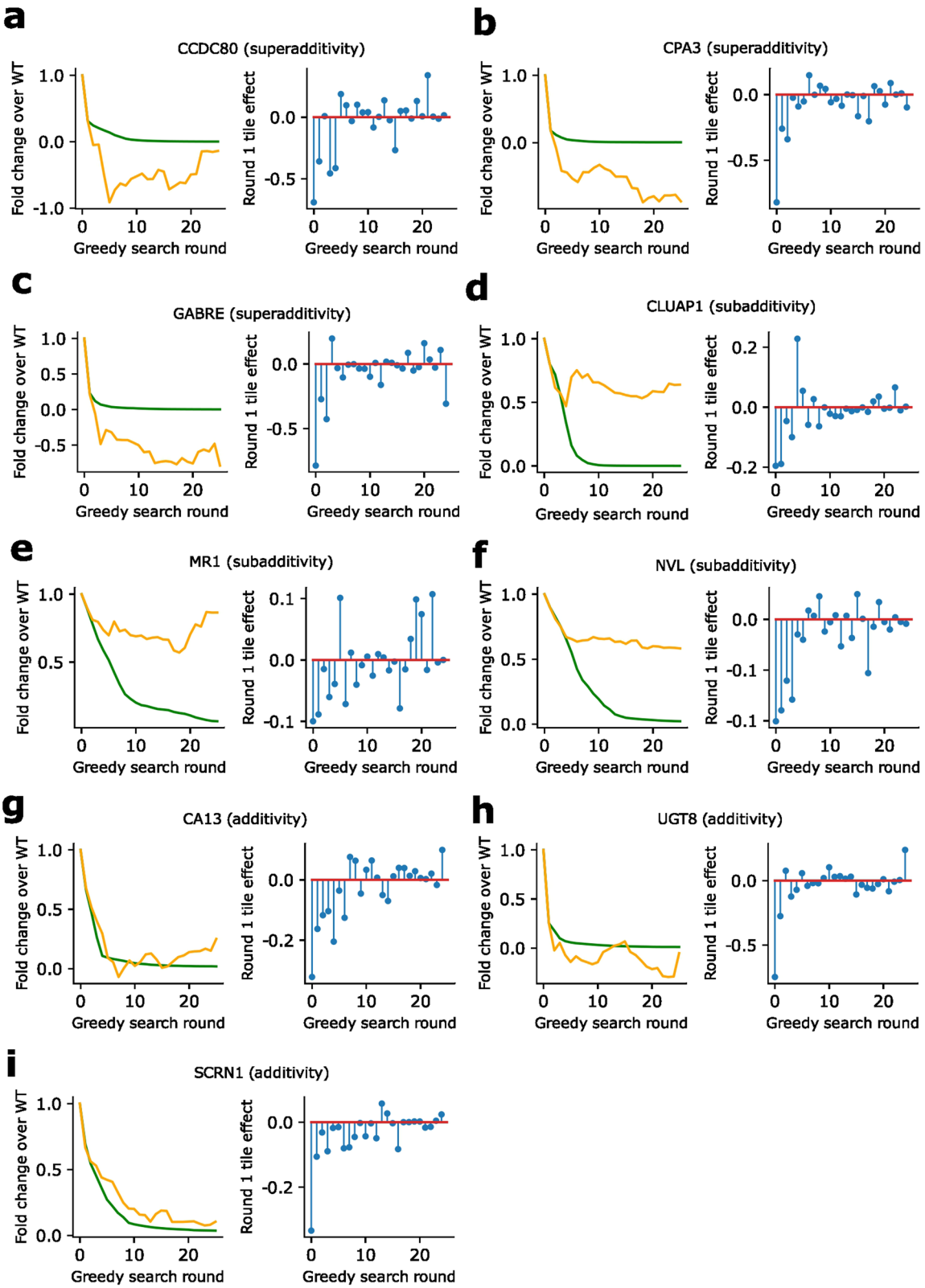
Extended Data Fig. 4 | Characterization of sufficient CREs in GM12878 and PC-3. a, Histogram of the distance between CRE tiles from TSS for sufficient enhancers and silencers in GM12878 and PC-3. **b**–**d**, Boxplots of mean DNase-seq coverage (**b**), mean ATAC-seq coverage (**c**), and mean histone mark coverage (**d**) of sufficient enhancer and silencer tiles in various cell types. The number of points in green and red boxes is 76 and 222 in K562, 41 and 57 for GM12878 and 35 and 97 for PC-3. Significance is given by the two-sided Mann-Whitney U test (*: $p < 0.05$; **: $p < 0.05$; ***: $p < 0.001$; ****: $p < 0.0001$). Boxplots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers).

https://doi.org/10.1038/s41588-024-01923-3



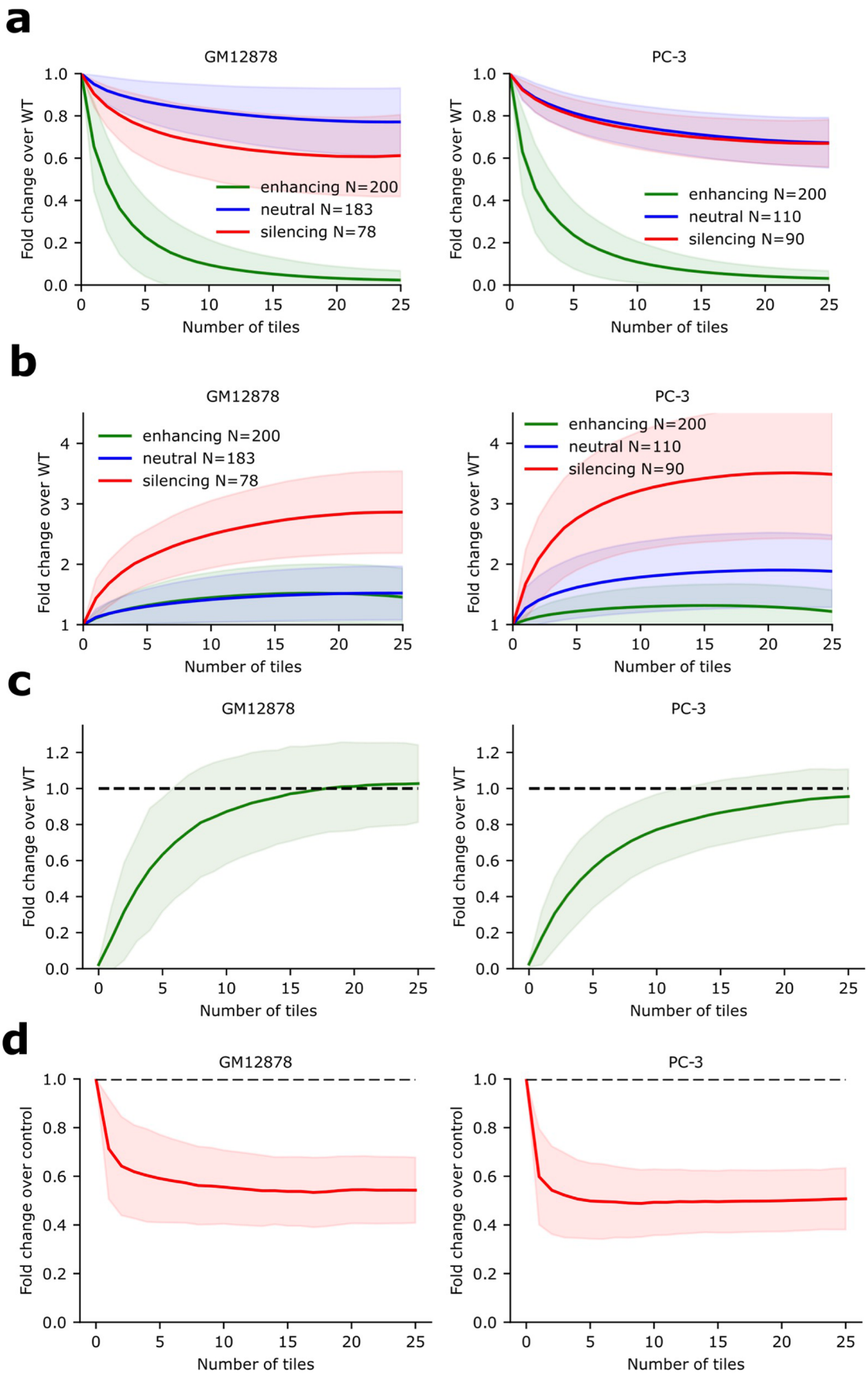
Extended Data Fig. 5 | TSS-CRE Distance Test results across cell lines. **a**–**c**, Average plot of the fold change over max versus distance to TSS for GM12878 (**a**) and PC-3 (**b**). Max represents the maximum TSS activity across all embedded positions within each sequence using Enformer. **c, d**, Plot of the tile sufficiency versus distance to TSS for GM12878 (**c**) and PC-3 (**d**), respectively. Tile sufficiency is calculated according to the predicted TSS activity with a TSS-CRE pair at a given distance minus the control sequence (shuffled context with just the TSS) divided by the WT sequence for enhancers and by the control sequence for silencers. In panels **a** – **d** shaded regions represent standard deviation of the mean.

https://doi.org/10.1038/s41588-024-01923-3



Extended Data Fig. 6 | Example sequences showing individual tile effect sizes from the Higher-Order Interaction Test results. a–i, the left panels show results of the greedy search (green) and the additive model (orange) for a particular gene; the right panel shows the independent tile effect size (calculated from the first iteration) sorted according to greedy search tile order. **a–c** shows example sequences classified as superadditivity; **d–f** shows sequences classified as subadditivity; **g–i** shows example sequences classified as additivity.

https://doi.org/10.1038/s41588-024-01923-3

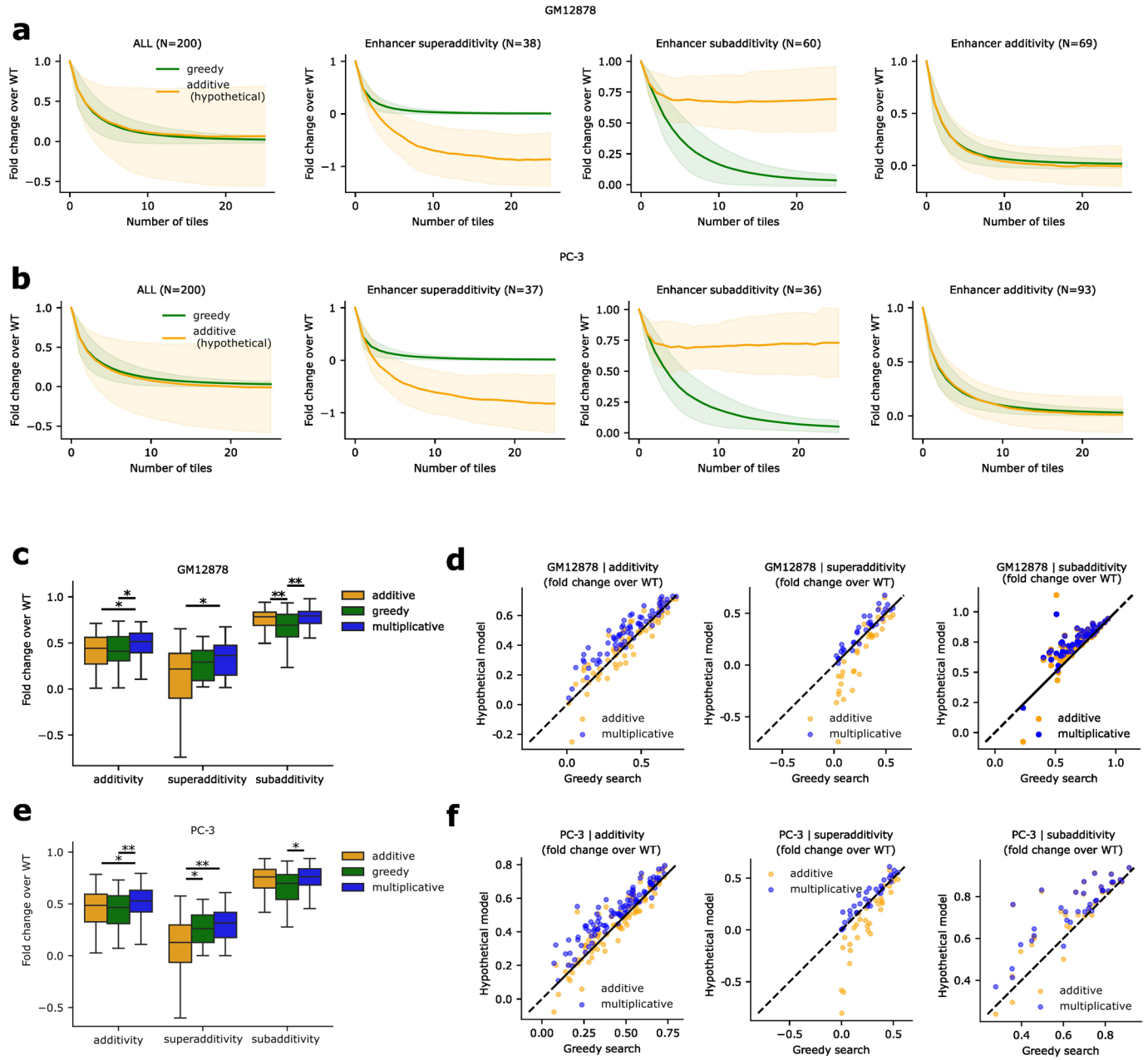


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41588-024-01923-3

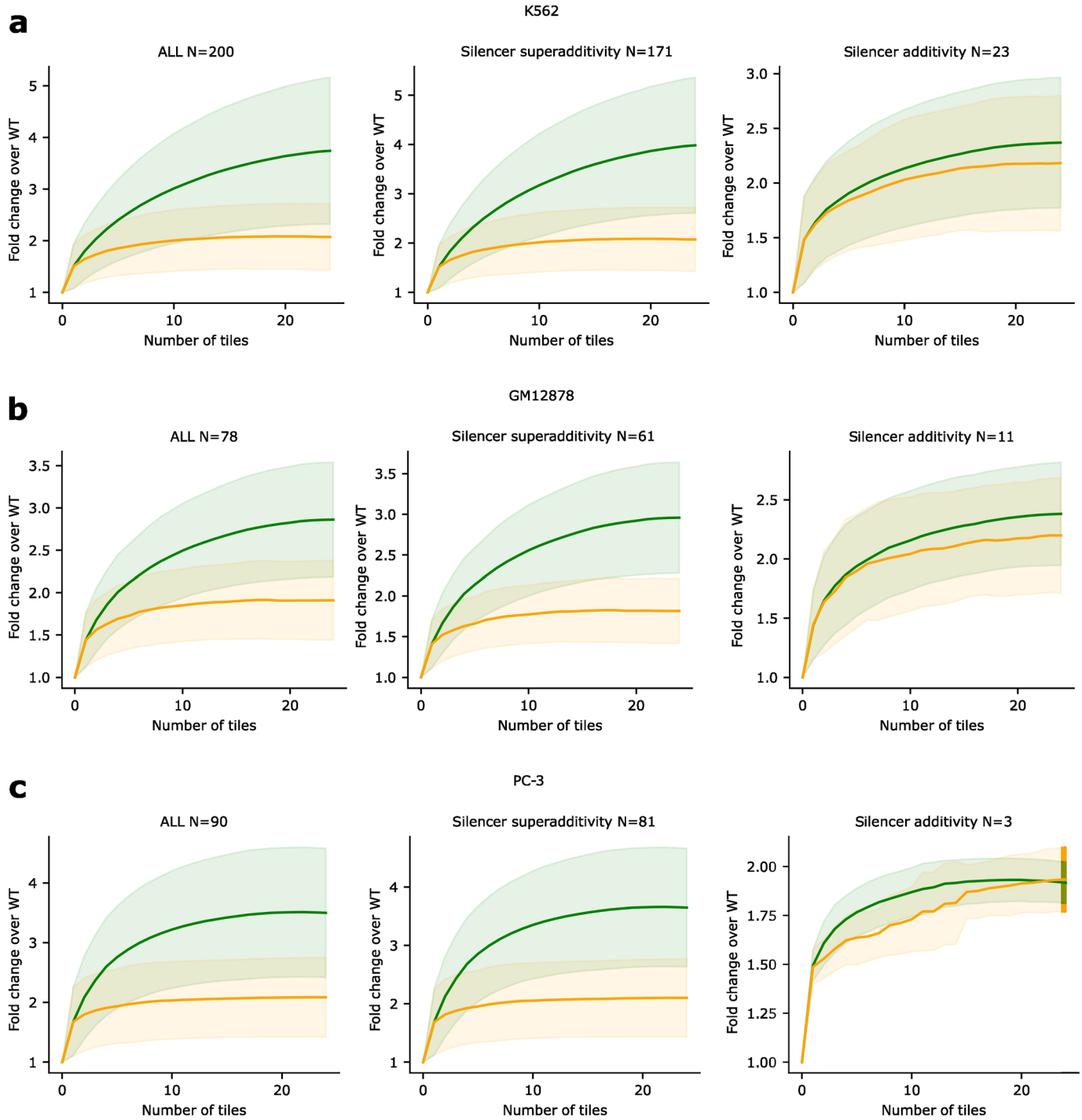
Extended Data Fig. 7 | Optimal CRE sets reveal complex interactions in GM12878 and PC-3. a, b, Average plot of the greedy search results for enhancer tile sets (**a**) and silencer tile sets (**b**) for sequences from different context categories for various cell lines. The fold change over wild-type (WT) is the predicted TSS activity of the shuffled CRE tiles in each round of the greedy search (indicated by the number of tiles). **c, d**, Sufficiency of the tile sets identified in each round of greedy search. Average fold change over wild-type (**c**) and control (**d**), which represents shuffled sequences with just the TSS tile. Sufficiency places the tile sets along with the TSS tile into shuffled sequences, averaging over 10 total shuffles. Shaded region represents the standard deviation of the mean.

https://doi.org/10.1038/s41588-024-01923-3



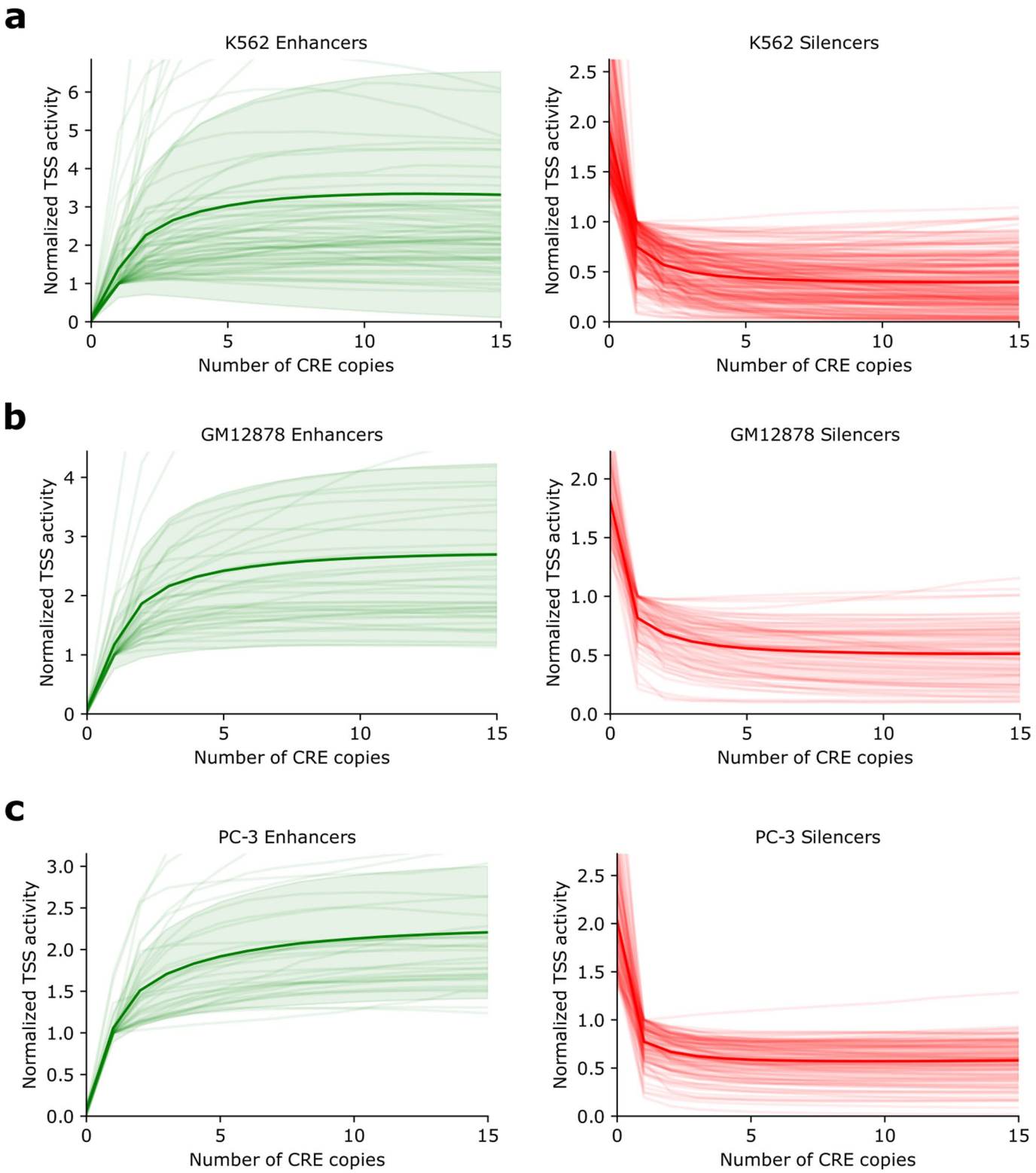
Extended Data Fig. 8 | Comparison of enhancer sets identified by the Higher-Order Interaction Test and a hypothetical additive model for GM12878 and PC-3. a, b, Comparison of the average fold change over wild-type (WT) for enhancer sets for sequences categorized as enhancing context versus a hypothetical additive effects model. The sequences from enhancing contexts are stratified according to interaction type, superadditivity, subadditivity, and additivity. Sequences were classified using mean squared error based thresholds of 0.1 for superadditivity and subadditivity and 0.05 for additivity definition (with some ambiguous cases left out of classification). Shaded region represents standard deviation of the mean. **c, e**, Comparison of hypothetical additive model and hypothetical multiplicative model versus greedy search outcomes at iteration 2 of the higher-order interaction test. The number of points in each box is 69, 38 and 60 in GM12878 and 93, 37, 36 in PC-3 for additive, superadditivity and subadditivity cases. Note, that some ambiguous cases were left out of the classification if they were outside of the selected thresholds. Statistical significance was given according to the two-sided Mann-Whitney U test (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$). Boxplots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers). **d, f**, Greedy search versus hypothetical additive or multiplicative models. Scatter plots show a more detailed view of the data in **c, e** with x-axis showing the higher-order interaction test outcomes and the y-axis showing the hypothetical model outputs (additive or multiplicative).

https://doi.org/10.1038/s41588-024-01923-3



Extended Data Fig. 9 | Comparison of silencer sets identified by the Higher-Order Interaction Test and a hypothetical additive model for K562, GM12878 and PC-3. a–**c**, Comparison of the average fold change over wild-type (WT) for silencer sets for sequences categorized as silencing context versus a hypothetical additive effects model for K562 (**a**), GM12878 (**b**), PC-3 (**c**). The sequences from silencing contexts are stratified according to interaction type, superadditivity and additivity. Shaded region represents standard deviation of the mean. Notably, we did not identify any subadditivity cases.

https://doi.org/10.1038/s41588-024-01923-3



Extended Data Fig. 10 | Saturation behavior of TSS activity predictions by Enformer in various cell lines. The results from a CRE Multiplicity Test applied to sequences from enhancing context (left) and silencing context (right) in **a**–**c**. Each line represents a particular enhancer or silencer CRE embedded into shuffled sequences at optimal positions (according to a Greedy Search) versus the copy number of the CRE in the sequence. The number of enhancers in each plot in **a**–**c** is 200, the number of silencers is 200, 78, 90 in **a**–**c**, respectively. The normalized TSS effect represents the predicted TSS activity of the mutated sequence divided by the control, which is the shuffled sequence with the TSS tile and the CRE in their original positions. The average across all CREs is shown with a thicker line and the shaded region represents the standard deviation of the mean.

nature portfolio

Corresponding author(s): Peter Koo

Last updated by author(s): Jul 31, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	RepeatMasker v4.1.5, BedTools v2.30.0, XSTREME at https://meme-suite.org/meme/tools/xstreme, kipoiseq v0.5.2, pyfaidx v0.6.4, pyranges v0.0.117
Data analysis	Python v3.9.16, numpy v1.22.3, pandas v2.0.1, seaborn v0.13.2 tensorflow-gpu v2.11.1, tensorflow-hub v0.13.0, matplotlib v3.7.5, tqdm v4.65.0, natsort v8.3.1, logomaker v0.8, pymemesuite v0.1.0a2, biopython v1.81 Static code for reproducing the analyses in the manuscript is available on Zenodo: https://zenodo.org/records/12594513. Bleeding-edge version of CREME is available at GitHub: https://github.com/p-koo/creme-nn and https://github.com/p-koo/CREME_paper_reproducibility.

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

Data

Policy information about availability of data

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

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

We obtained data for Biochemical analysis from www.encodeproject.org and deposited final and intermediate results for paper reproducibility on Zenodo at: https://doi.org/10.5281/zenodo.12584210.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No sample size calculation was performed. Sample sizes were determined based on available resources (i.e. compute time) and were acceptable based on the standard deviation values between repeated shuffles.
Data exclusions	For TSS analysis the subset of sequences was selected based on predetermined threshold of prediction values at TSS from the model (sequences yielding lower prediction and those where the TSS was not aligned with maximum prediction were excluded to avoid cases with multiple TSSs at the center of the sequence).
Replication	Replication is not relevant for this study because no experimental data was collected. The output of a DNN model does not change if the input is the same and therefore we have not performed replication of inference.
Randomization	All perturbations of background or tiles involved dinucleotide shuffling of sequences repeated multiple times. All the shuffled sequences were allocated into 'test' experimental group and the wild type sequence was selected as the control.
Blinding	Blinding is not relevant for this study because this would not affect the outputs of the DNNs because group assignment does not affect outputs from the models.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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