Iterative improvement of deep learning models using synthetic regulatory genomics

André M. Ribeiro-dos-Santos[1] & Matthew T. Maurano[1,2,3]

[1] Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY 10016, USA.

[2] Department of Pathology, NYU Grossman School of Medicine, New York, NY 10016, USA.

[3] Corresponding author: maurano@nyu.edu

ABSTRACT

Generative deep learning models can accurately reconstruct genome-wide epigenetic tracks from the reference genome sequence alone. But it is unclear what predictive power they have on sequence diverging from the reference, such as disease- and trait-associated variants or engineered sequences. Recent work has applied synthetic regulatory genomics to characterized dozens of deletions, inversions, and rearrangements of DNase I hypersensitive sites (DHSs). Here, we use the state-of-the-art model Enformer to predict DNA accessibility across these engineered sequences when delivered at their endogenous loci. At high level, we observe a good correlation between accessibility predicted by Enformer and experimentally measured values. But model performance was best for sequences that more resembled the reference, such as single deletions or combinations of multiple DHSs. Predictive power was poorer for rearrangements affecting DHS order or orientation. We use these data to fine-tune Enformer, yielding significant reduction in prediction error. We show that this fine-tuning retains strong predictive performance for other tracks. Our results show that current deep learning models perform poorly when presented with novel sequence diverging in certain critical features from their training set. Thus an iterative approach incorporating profiling of synthetic constructs can improve model generalizability, and ultimately enable functional classification of regulatory variants identified by population studies.

INTRODUCTION

Most genetic associations with human diseases and traits lie within noncoding regulatory DNA (Maurano et al. 2012). Genome-scale methods to analyze the function of noncoding regulatory elements within their relevant context would dramatically improve our ability for direct assessment of disease-associated variation. Profiling of allelic transcription factor (TF) activity across multiple cell and tissue types has shown that both genomic and tissue context modulates the penetrance of noncoding variants (Halow et al. 2021), but direct biochemical assessment of each variant in every relevant cell type and state remains impractical.

An alternative approach is to train machine learning models on reference genomic sequence and then apply them to genetic variation data. Recent machine learning models including gkmSVM (Lee et al. 2015), DeepSEA (Zhou and Troyanskaya 2015), Enformer (Avsec et al. 2021a), and BPNet (Avsec et al. 2021b) have demonstrated impressive performance in predicting functional genomics data such as expression and DNA accessibility from sequence alone. Those models

were trained on data from large projects such as ENCODE and Roadmap Epigenomics Mapping Project which have generated functional genomics datasets (e.g. DNase-seq) across a wide variety of human and mouse cell and tissue types (Meuleman et al. 2020). Analysis of current models on inter-individual genetic variation data has yielded surprisingly little improvement relative to more parsimonious approaches (Maurano et al. 2015; Dey et al. 2020; Karollus et al. 2023; Toneyan et al. 2022; Sasse et al. 2023; Huang et al. 2023). The infamous opacity of the more complex models reduces their contribution to mechanistic understanding, and could mask overfitting resulting in degraded performance under new circumstances. Furthermore, training sequence is universally drawn from the human or mouse reference genomes, whose genomic features can be highly correlated, particularly for long-range interactions (Xi and Beer 2018; Whalen and Pollard 2019). Such an approach limits the explored sequence space due to constraints imposed by evolution and could reduce model performance on mutations and engineered sequences diverging significantly from reference. Training on variation data (Sasse et al. 2023), genomic sequences from different species (Kelley 2020; Cochran et al. 2022), or random sequence (de Boer and Taipale 2024) could expand the sequence space for training, but the problem remains how to generate sufficient functional data that remains relevant to function at the endogenous genomic locus.

Short reporter constructs and limited genome editing cannot model the effect of hundreds or thousands of regulatory sites which collectively influence genomic context. CRISPR/Cas-based genome engineering strategies provide control over a limited range (100-1000 bp) and are more suited to the deletion of existing sequence than its rearrangement or the addition of novel sequence, and are difficult to use for multi-step engineering of diploid cells. Training on allelic variation in chromatin accessibility (Maurano et al. 2015; Halow et al. 2021) and genome-wide association study variants (Mostafavi et al. 2023) are limited by the standing genetic variation and effects of selection. A synthetic regulatory genomic approach is uniquely able to uncover novel biology, and these new technologies will move the field towards consideration of regulatory function at a locus scale.

Recent work applying synthetic regulatory genomics has highlighted the influence of genomic context on enhancer function (Ribeiro-Dos-Santos et al. 2022; Brosh et al. 2023), and the discrepancy between reporter assays and function at the endogenous locus (Brosh et al. 2023). The Big-IN platform enables targeted integration of large BACs upwards of 160 kb into human and

murine embryonic stem cells, overcoming previous limitations of scalability and portability across genomic loci and cellular contexts (Brosh et al. 2021). Integration of a modular landing pad (LP) enables subsequent recombinase-mediated cassette exchange (RMCE) of large payloads and efficient positive/negative selection for correct clones. Hundreds of different payloads have been delivered to these cell lines at multiple loci (Brosh et al. 2021; Mitchell et al. 2021; Pinglay et al. 2022; Brosh et al. 2023; Ordoñez et al. 2024; Zhang et al. 2023). Compared to editing methods such as saturation genome editing, synthetic regulatory genomics is not limited to derivatization of the reference sequence, but can deliver a nearly unlimited set of sequences.

Here, we explore the application of deep learning models to predict regulatory features at loci engineered through synthetic regulatory genomics. We evaluate the ability of these models to predict the functional behavior of regulatory elements when subject to complex deletions and rearrangements. Further, we evaluate the model prediction of functional genomics for novel synthetic sequences and long-range regulatory elements interaction. Finally, we demonstrate the potential design for iterative improvement of such models by integrating synthetic regulatory dataset. Our work suggests a key role for synthetic regulatory genomics in training future genomic deep learning models.

RESULTS

Enformer performance predicting synthetic payloads

To probe the varied performance of deep learning models across different regimes, we analyzed the Enformer model due to its wide receptive field that allows investigation of large regulatory landscapes (Avsec et al. 2021a). We adapted the published Enformer model to use a flexible input sequence length (**Methods**), and verified the adapted model produces similar results to the published version (**Fig. S1**). We then evaluated the predictive power of Enformer on synthetic sequences related to the reference genome but differing in significant aspects.

We started with a published analysis which dissected the 43-kb locus control region (LCR) required for *Sox2* expression in mouse embryonic stem cells (mESC) (Brosh et al. 2023). The LCR was replaced by 70 synthetic payloads including multiple categories of variation, such as deletions, duplications, rearrangements, inversions, and surgical TF site deletions (**Fig. 1A, Table S1**). We predicted DNA accessibility at 25.6-kb window centered at the LCR for each payload

using Enformer, and summed accessibility at LCR DHSs 19-28 as a proxy for the overall LCR activity (**Table S2**). Predicted LCR accessibility showed a good correlation with experimentally measured *Sox2* expression, confirming it is a good proxy for expression (**Fig. 1B**). To evaluate our accessibility prediction approach on a different locus and cell type, we analyzed a series of enhancer deletions and relocations at the α-globin locus (**Table S1**) characterized in mouse embryonic body-derived erythroid cells (**Fig. 1C**) (Blayney et al. 2023). Starting in mESCs, one α-globin allele was deleted using CRISPR/Cas9, while the other was engineered to support subsequent delivery of engineered *α-globin* payloads ranging up to 85 kb through recombinase-mediated genomic replacement (RMGR). Enhancer deletions and relocations were then engineered with CRISPR/Cas9. Similarly to the predictions on the *Sox2* dataset, the sum of predicted accessibility of all five enhancers (**Table S2**) presented a good correlation to experimentally measured *Hba* expression (**Fig. 1D**).

We then investigated predictive performance for individual TF recognition sequences identified within *Sox2* LCR DHSs 23-24, (Brosh et al. 2023). We used Enformer to perform an in silico scanning deletion analysis of those DHSs replacing 16 bp with Ns (**Fig. 2A**). These predictions correctly identified two regions identified as essential in the experimental analysis. But Enformer predicted no synergy between DHSs 23 and 24 (**Fig. 2B**), inconsistent with the published analysis (Brosh et al. 2023). Furthermore, Enformer completely missed a key cluster of TF recognition sequences (TF sites 23.5–23.8), which experimental deletion analysis showed to be essential (Brosh et al. 2023). Next, we focused on DHS24, which is the strongest single enhancer in the LCR and comprises 4 key TF recognition sequences. We used Enformer to predict accessibility of the payloads harboring combinatorial TF site deletions and observed poor performance predicting activity of these TF recognition sequence deletions (**Fig. 2B**). Thus, while deep learning models can recapitulate gross features of synthetic sequences derived from genomic reference sequence, their performance has gaps, most notably the synergy among neighboring regulatory elements.

Finally, to explore performance of deep learning models on sequences even further diverged from the reference genome, we investigated a synthetic “backwards” sequence in which the *HPRT1* gene sequence was reversed but not complemented (Camellato et al. 2024). This synthetic sequence shares many features with the native *HPRT1* (e.g., base composition, homopolymer runs, and other short repeats) but disrupts higher-order features such as coding sequence and TF

recognition sites (except palindromes). mESCs were engineered to replace the endogenous mouse *Hprt* locus with three synthetic payloads ranging from 95-100 kb (**Fig. 3A**): (i) *HPRT1* containing the native human sequence; (ii) SynHPRT1R containing its reversed but not complemented sequence; and (iii) SynHPRT1R^{noCpG} wherein all CpG sites were removed (**Table S1**). For the *HPRT1* payload in native hESC, Enformer correctly predicted activity at two experimentally observed DHSs (**Fig. 3B**). We then compared Enformer predictions to experimentally measured DNA accessibility (ATAC-seq) in mESC for all three payloads (**Fig. 3C**). The *HPTR1* payload only showed accessibility by ATAC-seq at the gene promoter, and both the reversed payloads (SynHPRT1R and SynHPRT1R^{noCpG}) showed no accessibility at all. In comparison, Enformer predicted false positive peaks in both *HPTR1* and SynHPRT1R^{noCpG} payloads, and it predicted above background accessibility throughout SynHPRT1R (**Fig. 3C**). These SynHPRT1R false positives might be related to sequence CpG content, since they were absent in SynHPRT1R^{noCpG}. These results reinforce model limitations when assessing novel synthetic sequences.

Enformer does not accurately model distal promoter-enhancer regulation

Previous reports have shown that activation of *Sox2* by its LCR decreases with enhancer-gene distance (Zuin et al. 2022). We tested whether Enformer was able to replicate the LCR-*Sox2* distance dependence curve by simulating placement of the *Sox2* LCR at 10-kb intervals both downstream and upstream of the *Sox2* promoter. For each placement, we predicted mESC_CJ7 DNase-seq signal at the gene promoter and LCR (**Fig. 4A**). We found that predicted accessibility remained mostly unchanged and was independent of enhancer-gene distance (**Fig. 4B**, **Fig. S2**). To evaluate the impact of genomic distances in a different sequence context, we flanked the *Sox2* proximal region with two copies of SynHPRT1R^{noCpG} (upstream and downstream) and again evaluated the impact of inserting the whole *Sox2* LCR at 10-kb intervals (**Fig. 4C**). This experiment yielded similar results (**Fig. 4D**). Interestingly, in SynHPRT1R^{noCpG}, the presence of the LCR at any position was predicted to increase *Sox2* promoter accessibility above the ΔLCR background. Thus, Enformer incorrectly predicts no effect of distance between the *Sox2* promoter and LCR, inconsistent with the experimental data. This is consistent with other reports that existing transformer models do not effectively model long-range interactions despite their wide receptive fields (Karollus et al. 2023).

Enformer predictions depend on background sequence properties

This lack of distance dependence contrasts with a prior report that Enformer predicts a key role for enhancer-promoter distance (Toneyan and Koo 2024). To explore this apparent inconsistency, we evaluated the effect of the position of their 74 candidate enhancers on CAGE signal reflecting transcription initiation at their target promoter in K562 cells. We built a virtual sequence for each enhancer-promoter pair by extracting a 196-kb window centered at the targeted gene transcription start site (TSS) from the human reference genome. We randomly shuffled dinucleotides within the targeted region sequence and restored the gene TSS block (central 5-kb region). We then generated a series of virtual sequences by inserting the candidate enhancer in 10-kb intervals from the TSS. The TSS CAGE signal was measured for each TSS-enhancer position by averaging the signal of ten random dinucleotide shuffles. Consistent with the published analysis, all promoter-enhancer pairs showed a steep reduction of CAGE signal with increasing TSS-enhancer distance (**Fig. S3A**). We reasoned that the differing results might be related to the choice of background sequences. Thus, we repeated the experiment using 10 randomly generated background sequences based on the dinucleotide content of SynHPRT1R^{noCpG}. In this case, we observe a selective distance response on TSS CAGE signal, with many DHS relocations producing little to no impact on TSS CAGE signal (**Fig. S3B**). Thus, we observed completely different responses depending on the background sequence employed. Interestingly, putative enhancers from high GC-content locus presented the steepest reduction of CAGE signal using our neutral background (**Fig. S3C**). This highlights the importance of background generation strategy and the impact of key genome sequence parameters such as G+C content on evaluation of model predictions.

Deep learning models struggle to predict cell-type selective DHSs

Current deep learning models are trained to predict multiple regulatory tracks (i.e. cell types) at the same time (Kelley et al. 2016, 2018; Avsec et al. 2021b). We reasoned that this multi-task strategy could favor detection of pervasive regulatory sites present across multiple cell types to the detriment of more cell-selective features (Schreiber et al. 2020; Kathail et al. 2024). We compared peaks called from Enformer predicted accessibility against matching experimental DNase-seq peaks. Peaks were partitioned according to their activity across 141 cell and tissue types from reference DNase-seq datasets which included including 105 human tracks and 36 mouse tracks (**Table S3**). This analysis showed that tissue-specific peaks were highly enriched for false positives, while true positives tended to be constitutively active sites (**Fig. 5A-B**). Enformer presented

increasingly higher positive predictive value (PPV) with decreasing tissue-specificity (**Fig. 5C**). Similar trends were observed when investigating mESC promoter, CTCF and non-CTCF sites (**Fig. S4**), suggesting the lack of sensitivity to cell-type selective sites is not limited to particular classes of genomic regulatory elements. The relatively low proportion of false positive sites among the more cell-type specific sites suggests that some of them can be explained by features bleeding through from other training tracks.

Fine-tuning on synthetic regulatory genomics datasets improves predictive performance

We then developed a fine-tuning strategy to improve model performance through incorporation of synthetic regulatory genomics datasets. We added a new independent output layer that uses the baseline Enformer feature extraction trunk to predict *Sox2* expression (**Fig. 6A**). The new output layer is composed of a self-attention layer to capture relevant features independently of position and a dense layer to combine the resulting signal into a single expression prediction value. This allowed us to fine-tune Enformer internal feature extraction without affecting mouse or human regulatory track heads. We evaluated three configurations of the new output self-attention layer: SingleHead 64/64, SingleHead 64/128, and MultiHead 64/64 (detailed in **Methods**). They all presented similar performance and any can be considered adequate architectures.

For training, we expanded the experimentally assessed *Sox2* LCR payload sequences by generating 1,280 1-bp shifts of a 25.6-kb focal window centered at the payload after replacing the endogenous *Sox2* LCR on the mouse reference genome (**Fig. S5A**). Among those shifts, 48 were randomly selected to fine-tune models and another 48 to evaluate the results (**Fig. 6AB; Fig. S5B**). To assess whether our strategy was at risk of overfitting due to training and evaluating our model on the same collection of payloads, we also trained ten batches of models with payloads randomly partitioned into independent training and validation datasets (**Fig. S5C**). Partitioned training produced similar results to the full dataset training, but exhibited higher variance due to the reduced numbers of payloads per category (**Fig. S6**). Thus, we employed the model trained with the full dataset for all other analysis.

Across payload categories, fine-tuned models outperformed baseline Enformer predictions, especially for complex sequence modification such as DHS ordering and inversion (**Fig. 6C**). When comparing baseline and fine-tuned genome-wide accessibility predictions, fine-tuning eliminated

several baseline false positives (**Fig. 6D**). Genome-wide assessment of mESC accessibility predictions demonstrate that fine-tuned models presented higher peak correlation to source DNase-seq and higher peak calling accuracy (**Fig. 6E-F**). Predictions from the fine-tuned Enformer retained high predictive performance across multiple assay categories (**Fig. S7**). The final tuning step for Enformer was towards the human dataset (Avsec et al. 2021a), which may have slightly favored detection of human regulatory features. Fine-tuned predictions outperformed baseline for all mouse assays and showed only minimal performance decay against human assays (**Fig. S8**). Our results demonstrate a unique approach to iteratively improve the predictive performance of pre-trained deep learning models through incorporation of high-yield synthetic regulatory assay data.

DISCUSSION

We showed that Enformer has good performance modeling basic aspects of synthetic regulatory data at both *Sox2* and *α-globin* loci, but it fails to capture DHS synergy and was only partially sensitive to TF site deletions. For sequences (i.e. SynHPRT1R) that were more diverged from the reference used to train Enformer, the model performed poorly. We found Enformer unable to effectively model long-range interactions, consistent with previous reports (Karollus et al. 2023; Toneyan and Koo 2024). Enformer's architecture penalizes feature interaction proportionally to the distance between elements, which may contribute to this shortcoming. Further, long-range interactions are relatively uncommon thus will not be favored during training (Karollus et al. 2023). This highlights the limitations imposed by a limited training sequence space and poses caveats to future analysis of novel sequences.

Multitask training to predict multiple regulatory genomic assays across different cell and tissue types enables information sharing between the different contexts (Kelley et al. 2018). However, performance remains limited by the training sequence space. Here we demonstrated that predictive power is related to the degree of DHS activity across multiple cell and tissue types, consistent with previous reports (Kathail et al. 2024). Broadly active sites across regulatory assays and cell types are favored during training, since prediction errors for such features are weighed more strongly. Thus correlation structures among genomic features can impede model training (Whalen et al. 2022). Simply expanding the number of training tasks has limited benefit as the sequence space explored remains the same (Schreiber et al. 2020; Kathail et al. 2024).

Deep learning models are promising tools to expand our understanding of regulatory genomics and to interpret variant functional impact (Avsec et al. 2021a, 2021b; Zhou and Troyanskaya 2015). However, our results show they are currently limited by the limited scope of their training sequences, which largely reflect the reference genome. We argue that deep learning models would benefit from training on functional characterization of regulatory elements that have been systematically deleted, reordered, and inverted, and perturbed by disease- and trait-associated variation. An iterative training strategy in which these results feed into model development, which in turn directs future experimental exploration would allow systematic exploration of disease- and trait-relevant sequence space and improve model performance across contexts. As a proof of concept of this strategy, our results showed that fine-tuning improved prediction performance across all payload categories and improved genome-wide prediction performance, cleaning many false positives signals and producing higher peak calling accuracy. Thus, synthetic regulatory genomics and machine learning are highly complementary in genomics' dual status as both a "big data" and experimental science to uniquely permit pushing the limits of model knowledge while maintaining generality through iterative large-scale testing of model predictions.

COMPETING INTEREST STATEMENT

M.T.M. is listed as an inventor on a patent application describing Big-IN.

ACKNOWLEDGEMENTS

We thank Raquel Ordoñez for helpful comments, and Ran Brosh and Brendan Camellato for sharing data. This work was partially funded by National Institutes of Health (NIH) grants RM1HG009491 and R01MH136353.

METHODS

Enformer model

The published Enformer model is hosted in the kaggle repository (https://www.kaggle.com/models/deepmind/enformer) cited in (Avsec et al. 2021a). This model (referred to here as Avsec2021) requires a 393,216-bp input sequence but only outputs predictions for the central 114,688-bp region (**Fig. S1A**).

For all analyses, we employed an adapted version of Enformer due to computational constraints. This model was based on an equivalent model hosted at the DeepMind github repository (https://github.com/google-deepmind/deepmind-research/tree/master/enformer) and pre-trained weights from Google storage (gs://dm-enformer/models/enformer/sonnet_weights/). The adapted version removed an internal cropping layer to allow a flexible input sequence size instead of the fixed 196,608 bp, and to generate predictions for the whole input (**Fig. S1B**). We compared the Avsec2021 version to our adapted model with decreasing input sizes (196,608 bp, 114,688 bp, and 25,600 bp). We selected ten 114,688 bp sites containing at least one mESC_CJ7 hotspot V1 (FDR 0.01) and predicted their mESC_CJ7 DNase-seq density using all four implementations. To accommodate larger input sequence requirements, we expanded the targeted sites to the required size. For smaller input sequence requirements, the targeted regions were broken down into non-overlapping windows. Deviations from the Avsec2021 predictions were measured by Root Mean Squared Error (RMSE) and internal prediction deviations between ten 1-bp strides by standard deviation (**Fig. S6C-D**). We found no significant prediction deviations (**Fig. S6C-E**) between Avsec2021 and our adapted versions.

Enformer predictions throughout the manuscript were measured as the average of ten 1-bp strides (5 bp downstream and upstream) unless otherwise noted. Predicted accessibility was measured as the maximum signal across all bins at each site.

Synthetic regulatory biology datasets

Sox2

Expression data was previously published: Fig4A and TableS4 (Brosh et al. 2023). Briefly, engineered allele *Sox2* expression was measured by qRT-PCR using allele-specific primers. Fold-change was calculated as $2^{\Delta CT[CAST-B6]}$ and scaled to yield expression values ranging from 0

(ΔSox2) and 1 (WT). Previously published DNase-seq data was taken from ENCODE: DS13320 (mESC_CJ7) and DS21450 (mESC_E14) (Vierstra et al. 2014).

When employing Enformer to predict mESC_CJ7 DNase-seq signal of these payloads, a virtual sequence was built by replacing *Sox2* LCR from the mouse reference genome (mm10) with each synthetic payload sequence. Predictions were generated for a 25.6-kb window centered at payload location.

α-globin

Expression data was previously published: Fig5A and Fig6B-D (Blayney et al. 2023). Engineering was performed in mESC (mDist cells derived from mESC_E14), and expression values from embryonic body-derived erythroid cells were analyzed. *Hba* expression in engineered cells was profiled by qRT-PCR, normalized to *Hbb* and scaled to a proportion of WT expression. Previously published DNase-seq data was taken from ENCODE: DS20937 (mfLiver-F3) (Vierstra et al. 2014).

We employed Enformer to predict accessibility of those five enhancers (R1-R4; **Table S1**) in mouse embryonic liver (mfLiver-F3) as a proxy for *Hba* expression. Different enhancer configurations were generated by deleting and relocating R1-R4 enhancers on the mouse reference genome (mm10). Predictions were generated for a 25.6-kb window centered around the R1-R4 enhancers.

HPRT1R

ATAC-seq data was previously published: BS15734A, BS15738A, BS21951A (Camellato et al. 2024). Previously published DNase-seq data was taken from ENCODE: DS13320 (mESC_CJ7), and DS18873 (H7_hESC) (Vierstra et al. 2014; Meuleman et al. 2020).

A virtual sequence was built by replacing the mouse *Hprt* locus (Table S1) from the mouse reference genome (mm10) with each of the three payloads. Accessibility in mESC_CJ7 was predicted in a window of 393,216 bp centered at the *Hprt* locus (**Fig. 3C**). We similarly predicted human *HPRT1* locus accessibility in human H7_hESC.

Enformer fine-tuning

To fine-tune Enformer using the *Sox2* LCR synthetic payload dataset, we added a new output head to Enformer to predict *Sox2* expression based on *Sox2* LCR sequence. The new head employs a self-attention layer to capture relevant features from the Enformer prediction trunk independently of position. The resulting signal is combined into a scalar prediction of *Sox2* expression by a fully-connected layer using a softplus activation function (**Fig. 6A**).

Here we evaluated three configurations of the self-attention layer by varying the internal key/value matrix sizes, and number of independent attention projections (or heads). SingleHead 64/64 applies a single projection of 64 key and value matrices, SingleHead 64/128 applies a single projection of 64 key and 128 value matrices, and MultiHead 64/64 applies four independent projections of 64 key and value matrices.

Sox2 LCR synthetic payload sequences were expanded to 1,280 entries by striding 1 bp at a time up to ±640 bp upstream and downstream through a 25.6-kb focal region centered at the LCR (**Fig. S5A**). 48 strides were sampled to train three fine-tuned model configurations and another 48 strides were sampled for evaluation (**Fig. 6B, Fig. S5B**). We included all payloads (n=70) during both training and evaluation.

In parallel, we trained models with payloads partitioned into training and evaluation datasets to avoid data leakage. Payloads were randomly assigned into a training or validation dataset in a 3:1 ratio in ten independent batches. Models were trained and their performance evaluated using the training and validation datasets for each batch (**Fig. S5C**). Performance was average across training batches. Retrained models presented similar results to full dataset training, but higher variance likely attributable to the reduced size of the training sets.

For both strategies, fine-tuning was conducted using only the baseline Enformer trunk and the new output head. This strategy allows us to adjust Enformer internal feature extraction based on our synthetic dataset, while preserving mouse and human prediction heads. We employed the Adam optimization algorithm with a learning rate of 10^{-5} to fine-tune the models for 10 epochs,

each visiting 400 sequences at time in batches of 4. Training error was measured by Mean-Squared Error (MSE).

Expression was estimated directly from the new output head and indirectly by summing predicted accessibility of all LCR DHSs in mESC_CJ7. For comparison, accessibility estimates were scaled to match measured expression using a linear regression fitted with the training dataset. Predictions were averaged across all payload strides and error measured by Mean Squared-Error (MSE) across synthetic payloads (**Table S1**).

Genome-wide prediction performance

Genome-wide evaluations were conducted by comparing experimental DNase-seq results to baseline and fine-tuned models for 8 selected tracks (**Table S4**). For each selected track, we evaluated predictions at 2,000 randomly selected 25.6-kb regions containing at least one hotspot V1 (FDR 0.01) for the selected track. Within those regions, peaks were identified as continuous runs of DNase-seq density or prediction values above the track cutoff and larger than 128 bp. The track signal cutoff was established by calculating the signal 95th percentile at 2,000 random 128-bp sites within the targeted regions. A final collection of evaluated sites was generated by merging peaks called for all DNase-seq assays (selected track, technical replicate and others), baseline and fine-tune predictions and combining it to non-overlapping random sites used to the signal cutoff. Activity across cell-types and tissues was estimated by annotating the number of tracks with overlapping hotspots V1 (FDR 0.01) among a collection of reference mouse (N=36) and human (N=105) ENCODE DNase-seq tracks (**Table S3**).

REFERENCES

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. 2021a. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**: 1196–1203.

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

Blayney JW, Francis H, Rampasekova A, Camellato B, Mitchell L, Stolper R, Cornell L, Babbs C, Boeke JD, Higgs DR, et al. 2023. Super-enhancers include classical enhancers and facilitators to fully activate gene expression. *Cell* **186**: 5826-5839.e18.

Brosh R, Coelho C, Ribeiro-Dos-Santos AM, Ellis G, Hogan MS, Ashe HJ, Somogyi N, Ordoñez R, Luther RD, Huang E, et al. 2023. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *Mol Cell* **83**: 1140-1152.e7.

Brosh R, Laurent JM, Ordoñez R, Huang E, Hogan MS, Hitchcock AM, Mitchell LA, Pinglay S, Cadley JA, Luther RD, et al. 2021. A versatile platform for locus-scale genome rewriting and verification. *Proc Natl Acad Sci U S A* **118**: e2023952118.

Camellato BR, Brosh R, Ashe HJ, Maurano MT, Boeke JD. 2024. Synthetic reversed sequences reveal default genomic states. *Nature* **628**: 373–380.

Cochran K, Srivastava D, Shrikumar A, Balsubramani A, Hardison RC, Kundaje A, Mahony S. 2022. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Res* **32**: 512–523.

de Boer CG, Taipale J. 2024. Hold out the genome: a roadmap to solving the cis-regulatory code. *Nature* **625**: 41–50.

Dey KK, van de Geijn B, Kim SS, Hormozdiari F, Kelley DR, Price AL. 2020. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat Commun* **11**: 4703.

Halow JM, Byron R, Hogan MS, Ordoñez R, Groudine M, Bender MA, Stamatoyannopoulos JA, Maurano MT. 2021. Tissue context determines the penetrance of regulatory DNA variation. *Nat Commun* **12**: 2850.

Huang C, Shuai RW, Baokar P, Chung R, Rastogi R, Kathail P, Ioannidis NM. 2023. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat Genet* **55**: 2056–2059.

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

Kathail P, Shuai RW, Chung R, Ye CJ, Loeb GB, Ioannidis NM. 2024. Current genomic deep learning models display decreased performance in cell type-specific accessible regions. *Genome Biol* **25**: 202.

Kelley DR. 2020. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**: e1008050.

Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**: 739–750.

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**: 990–999.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics* **47**: 955–961.

Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA. 2015. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature Genetics* **47**: 1393–1401.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.

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

Mitchell LA, McCulloch LH, Pinglay S, Berger H, Bosco N, Brosh R, Bulajić M, Huang E, Hogan MS, Martin JA, et al. 2021. De novo assembly and delivery to mouse cells of a 101 kb functional human gene. *Genetics* **218**: iyab038.

Mostafavi H, Spence JP, Naqvi S, Pritchard JK. 2023. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**: 1866–1875.

Ordoñez R, Zhang W, Ellis G, Zhu Y, Ashe HJ, Ribeiro-Dos-Santos AM, Brosh R, Huang E, Hogan MS, Boeke JD, et al. 2024. Genomic context sensitizes regulatory elements to genetic disruption. *Mol Cell* **84**: 1842-1854.e7.

Pinglay S, Bulajić M, Rahe DP, Huang E, Brosh R, Mamrak NE, King BR, German S, Cadley JA, Rieber L, et al. 2022. Synthetic regulatory reconstitution reveals principles of mammalian Hox cluster regulation. *Science* **377**: eabk2820.

Ribeiro-Dos-Santos AM, Hogan MS, Luther RD, Brosh R, Maurano MT. 2022. Genomic context sensitivity of insulator function. *Genome Res* **32**: 425–436.

Sasse A, Ng B, Spiro AE, Tasaki S, Bennett DA, Gaiteri C, De Jager PL, Chikina M, Mostafavi S. 2023. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat Genet* **55**: 2060–2064.

Schreiber J, Nair S, Balsubramani A, Kundaje A. 2022. Accelerating in silico saturation mutagenesis using compressed sensing. *Bioinformatics* **38**: 3557–3564.

Schreiber J, Singh R, Bilmes J, Noble WS. 2020. A pitfall for machine learning methods aiming to predict across cell types. *Genome Biol* **21**: 282.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Toneyan S, Koo PK. 2024. Interpreting cis-regulatory interactions from large-scale deep neural networks. *Nat Genet* **56**: 2517–2527.

Toneyan S, Tang Z, Koo PK. 2022. Evaluating deep learning for predicting epigenomic profiles. *Nat Mach Intell* **4**: 1088–1100.

Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**: 1007–1012.

Whalen S, Pollard KS. 2019. Reply to “Inflated performance measures in enhancer-promoter interaction-prediction methods.” *Nat Genet* **51**: 1198–1200.

Whalen S, Schreiber J, Noble WS, Pollard KS. 2022. Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* **23**: 169–181.

Xi W, Beer MA. 2018. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLoS Comput Biol* **14**: e1006625.

Zhang W, Golynker I, Brosh R, Fajardo A, Zhu Y, Wudzinska AM, Ordoñez R, Ribeiro-Dos-Santos AM, Carrau L, Damani-Yokota P, et al. 2023. Mouse genome rewriting and tailoring of three important disease loci. *Nature* **623**: 423–431.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931–934.

Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, Mach P, Kryzhanovska M, Tihanyi G, Kohler H, et al. 2022. Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**: 571–577.

FIGURE LEGENDS

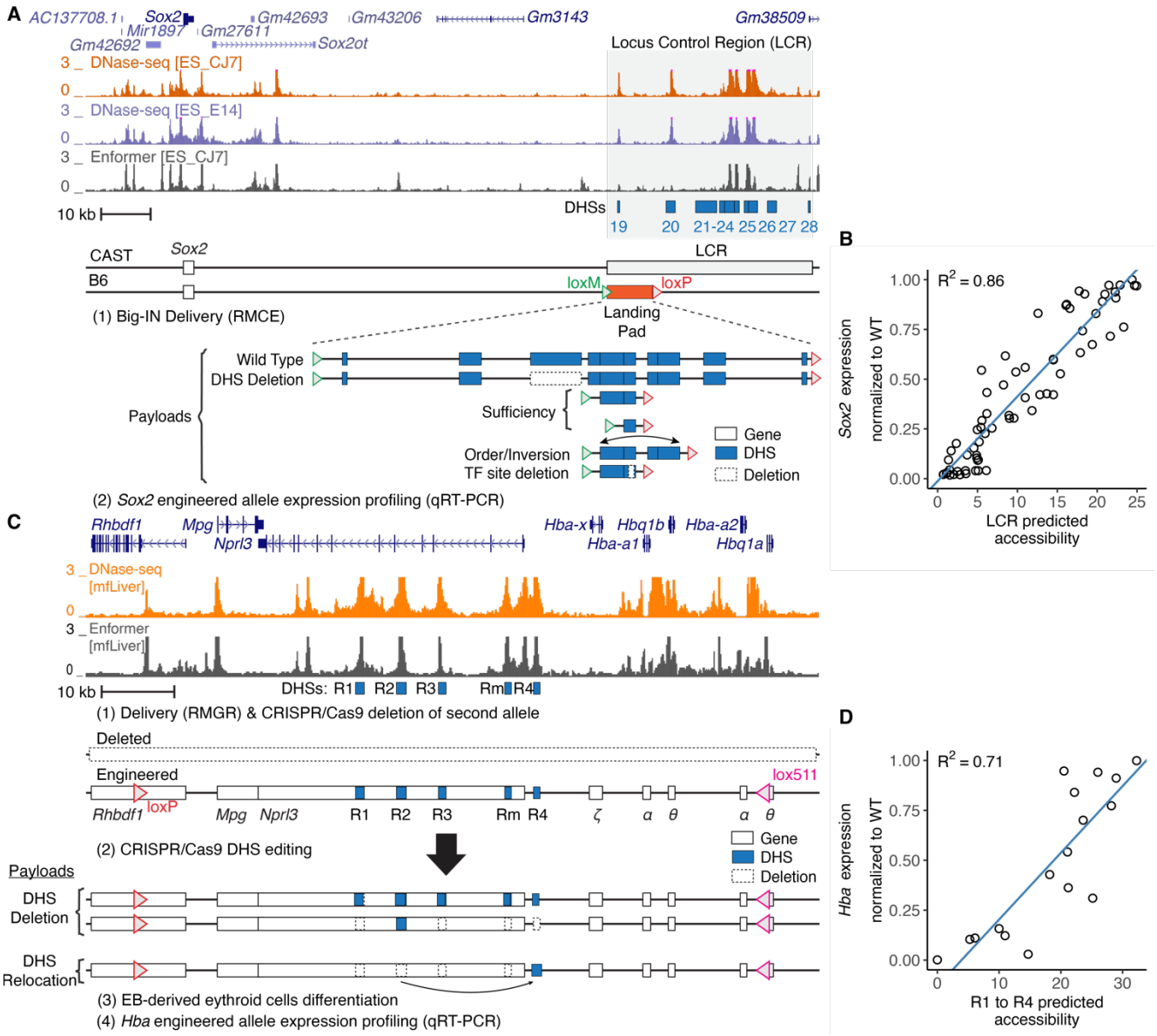

Fig. 1. Assessment of Enformer performance on synthetic sequences.

(A) Schematic of *Sox2* locus showing DNase-seq in mouse embryonic stem cells (mESC) and Enformer prediction Engineered CAST/B6 mESC cells harbor a Big-IN landing pad replacing one allele of the *Sox2* LCR for delivery of multiple synthetic payloads (Brosh et al. 2023). Locus control region (LCR) is highlighted in gray. DHSs are indicated by blue boxes. DHS deletions are indicated by dashed boxes. (B) *Sox2* expression was characterized by allele-specific qRT-PCR. Each point represents experimentally measured *Sox2* expression and Enformer prediction of a synthetic payload (N=70). Engineered *Sox2* allele expression was scaled between 0 (ΔSox2) and 1 (WT). LCR activity was measured as the summed accessibility over all LCR (DHSs 19-28) in mESC_CJ7. (C) Schematic of *α-globin* locus showing DNase-seq of mouse fetal liver (mfLiver), Enformer predictions, and *α-globin* locus engineering strategy by CRISPR/Cas9 editing and characterization using qRT-PCR (Blayney et al. 2023). Enhancers R1-R4 are indicated by blue boxes. Deleted regions are represented by dashed boxes. (D) *Hba* expression of the resulting cells was profiled by qRT-PCR, normalized to *Hbb* and scaled as a proportion of WT expression. Each point represents experimentally measured *Hba* expression and Enformer prediction of an enhancer configuration (N=17). *Hba* expression was measured by qRT-PCR, normalized to *Hbb* and scaled as a proportion of WT expression. Predicted enhancers (R1-R4) activity was measured as the sum of maximum predicted accessibility at enhancers R1 to R4.

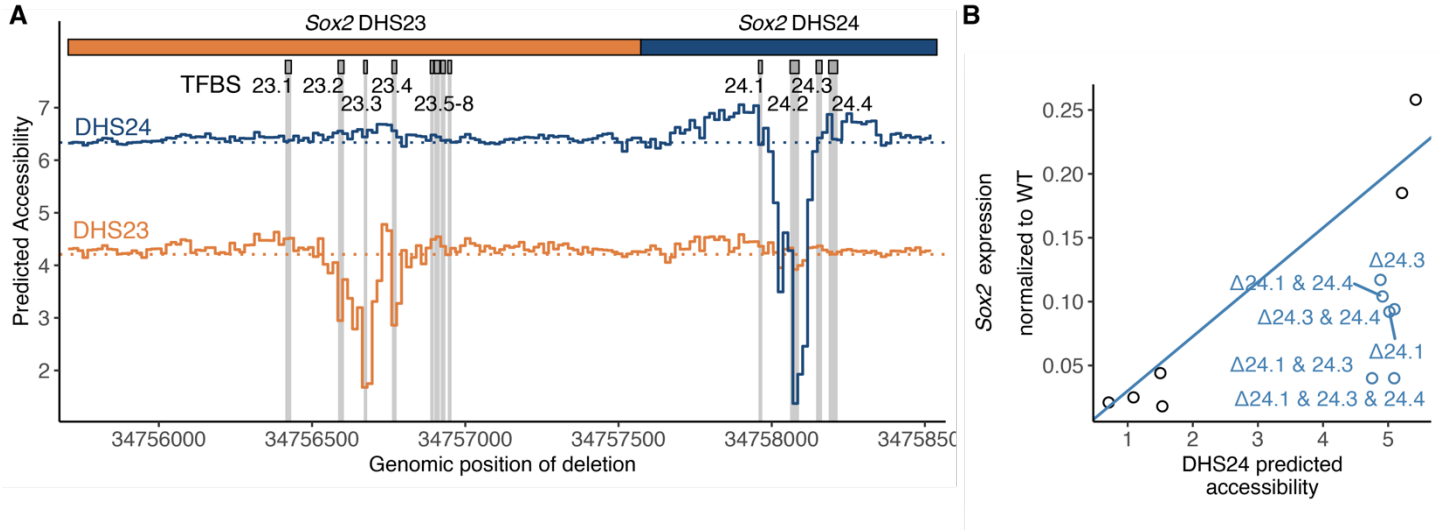

Fig. 2. Enformer poorly models individual TF recognition site deletions.

(A) In silico scanning deletion analysis of Sox2 DHS23 and DHS24 accessibility. Enformer was used to predict accessibility (mESC mESC_CJ7 DNase-seq) for a series of DHS23-24 virtual payloads delivered replacing the *Sox2* LCR (Brosh et al. 2023). 16-bp deletions were encoded by replacing the reference sequence with Ns sliding across the length of DHS23 and DHS24. Colored lines indicate the DHS accessibility for each deletion location (x-axis). Shown are DHS23 and DHS24 accessibility according to deletion position are represented by the orange and blue lines respectively. Horizontal dotted lines indicate baseline accessibility. Boxes above indicate relevant TF recognition sequences. (B) Each point represents a synthetic DHS24 payload (N=12) with TF recognition sequences deleted or mutated and delivered in place of the *Sox2* LCR and profiled for expression (Brosh et al. 2023). Enformer was used to predict accessibility of each experimentally profiled payload. Maximum accessibility at DHS24 (x-axis) was compared to experimentally determined *Sox2* expression (y-axis, 1 represents WT LCR and 0 represents ΔLCR). Selected points off-diagonal are highlighted in blue and labeled.

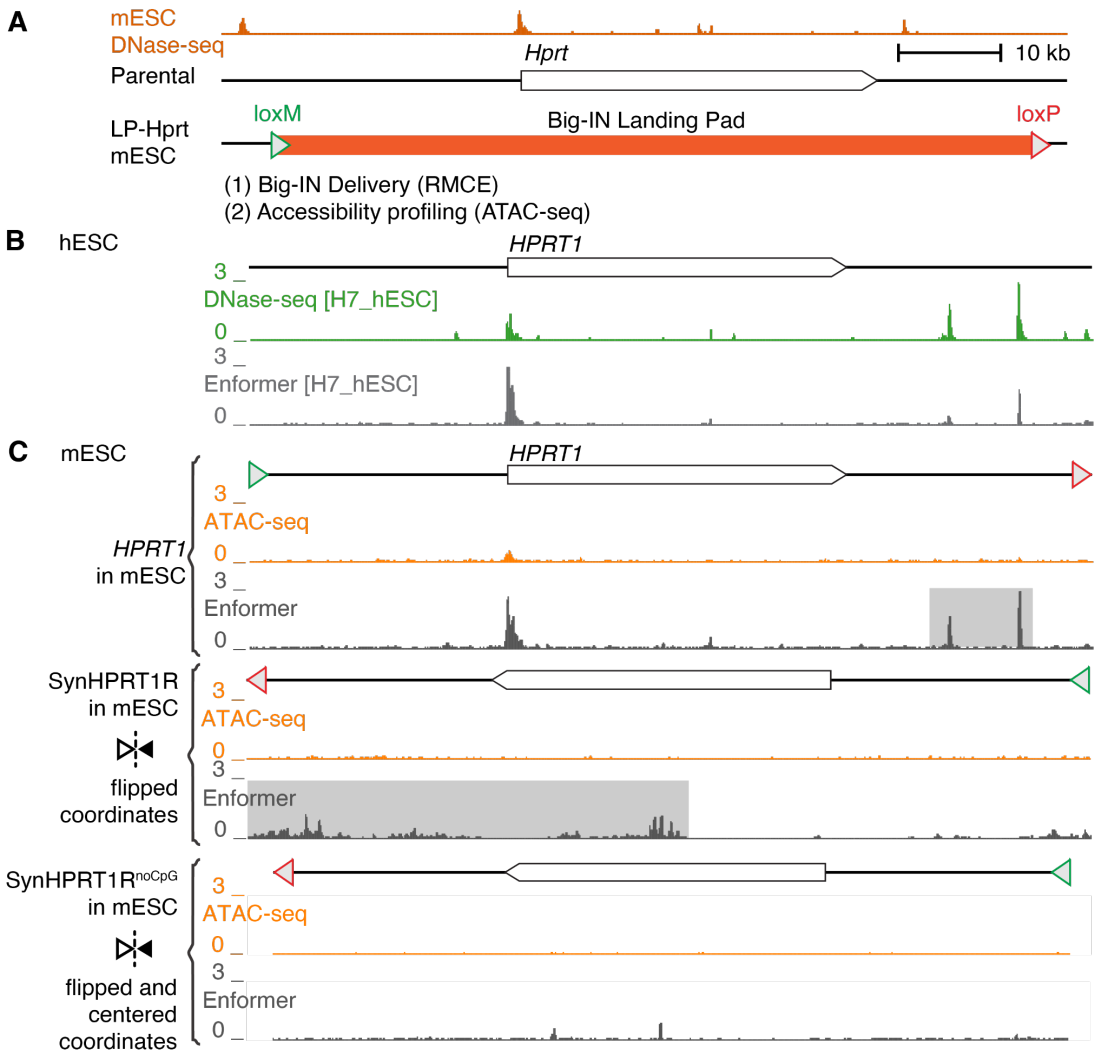

Fig. 3. Limited predictive performance on novel synthetic sequences.

(A) Schematic of mouse *Hprt* locus and synthetic payload profiling strategy (Camellato et al. 2024). mESC DNase-seq is shown at top. mESCs were engineered to replace their *Hprt* locus with Big-IN landing pad, which was used to deliver three synthetic payloads: human *HPRT1* locus, SynHPRT1R (reversed but not complemented human *HPRT1* locus) and SynHPRT1R^{noCpG} (SynHPRT1R with all CpG removed) payloads. Engineered cells were characterized by ATAC-seq. (C) Schematic of human *HPRT1* locus. Shown is DNase-seq of human H7_hESC and corresponding Enformer prediction. (B) ATAC-seq and Enformer accessibility predictions for engineered mESC. Shown synthetic payloads (*HPRT1*, SynHPRT1R, and SynHPRT1R^{noCpG}) profiled and predicted accessibility. Schematics of each synthetic payload are shown above accessibility tracks. Reversed payloads (SynHPRT1R, and SynHPRT1R^{noCpG}) are flipped horizontally and SynHPRT1R^{noCpG} was centralized to match *HPRT1* coordinates. Enformer false positive predictions are highlighted in gray.

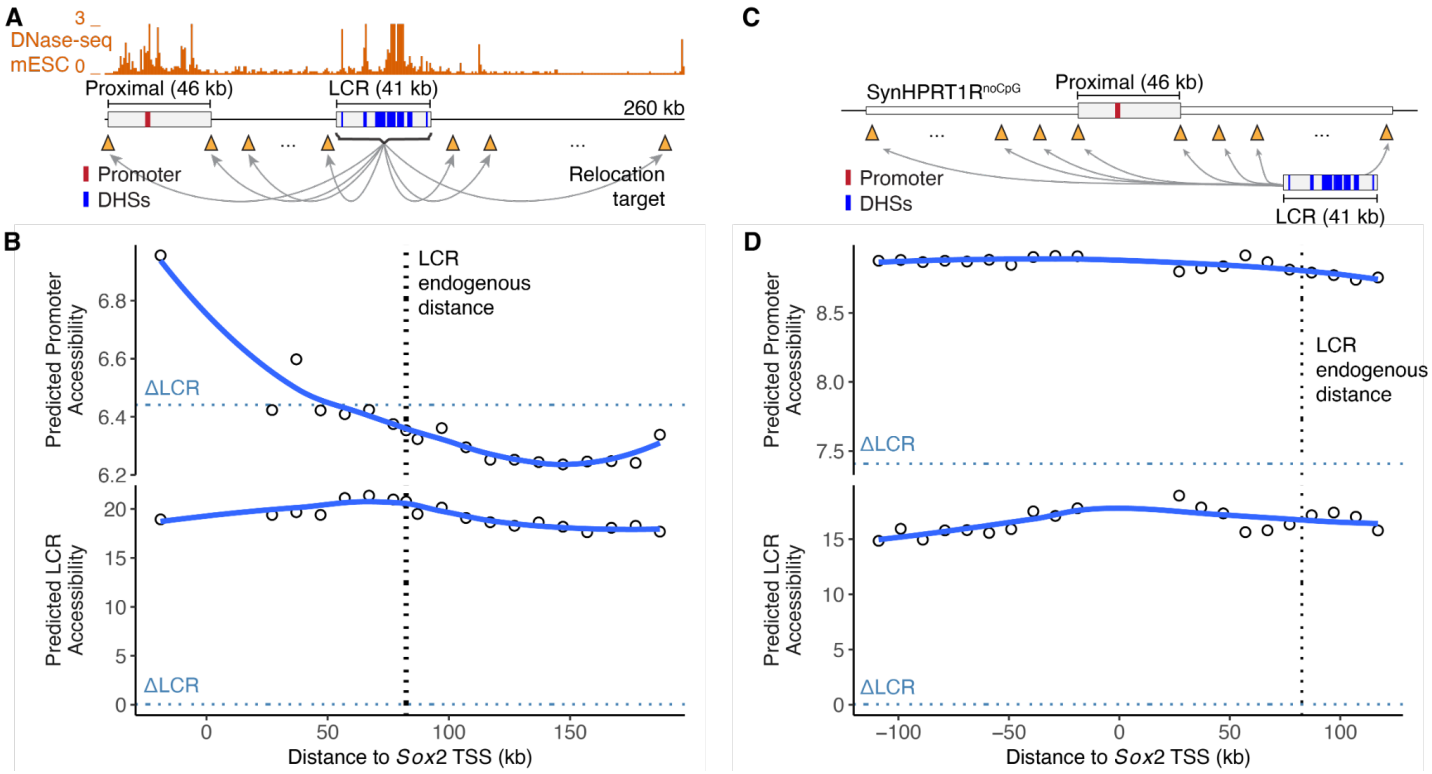

Fig. 4. Enformer does not predict effect of LCR relocation on *Sox2* accessibility.

Simulated relocation of the *Sox2* LCR in 10-kb intervals surrounding the promoter in its endogenous locus (A-B) or neutral synthetic sequence (SynHPRT1R^{noCpG}) (C-D). Enformer predicted accessibility for a 260-kb window encompassing the whole virtual locus and any remaining space was filled with Ns. Schematic of relocation assay within the endogenous genomic sequence (A) and SynHPRT1RnoCpG (C); *Sox2* locus DNase-seq of *mESC* is shown on top. Red rectangles indicate *Sox2* promoter and blue rectangles LCR DHSs 19-28. (B, D) Each point represents predicted accessibility of *Sox2* promoter (top panel) and LCR (bottom panel) for a given LCR distance to *Sox2* TSS. Promoter accessibility was measured as the site maximum predicted signal, and LCR accessibility as the sum of DHSs 19-28 maximum predicted signal. Vertical black line indicates endogenous LCR distance to TSS. Horizontal blue line indicates predicted accessibility when LCR was replaced by SynHPRT1R^{noCpG} fragment of the same size. Blue line represents a LOESS curve fitted to the data.

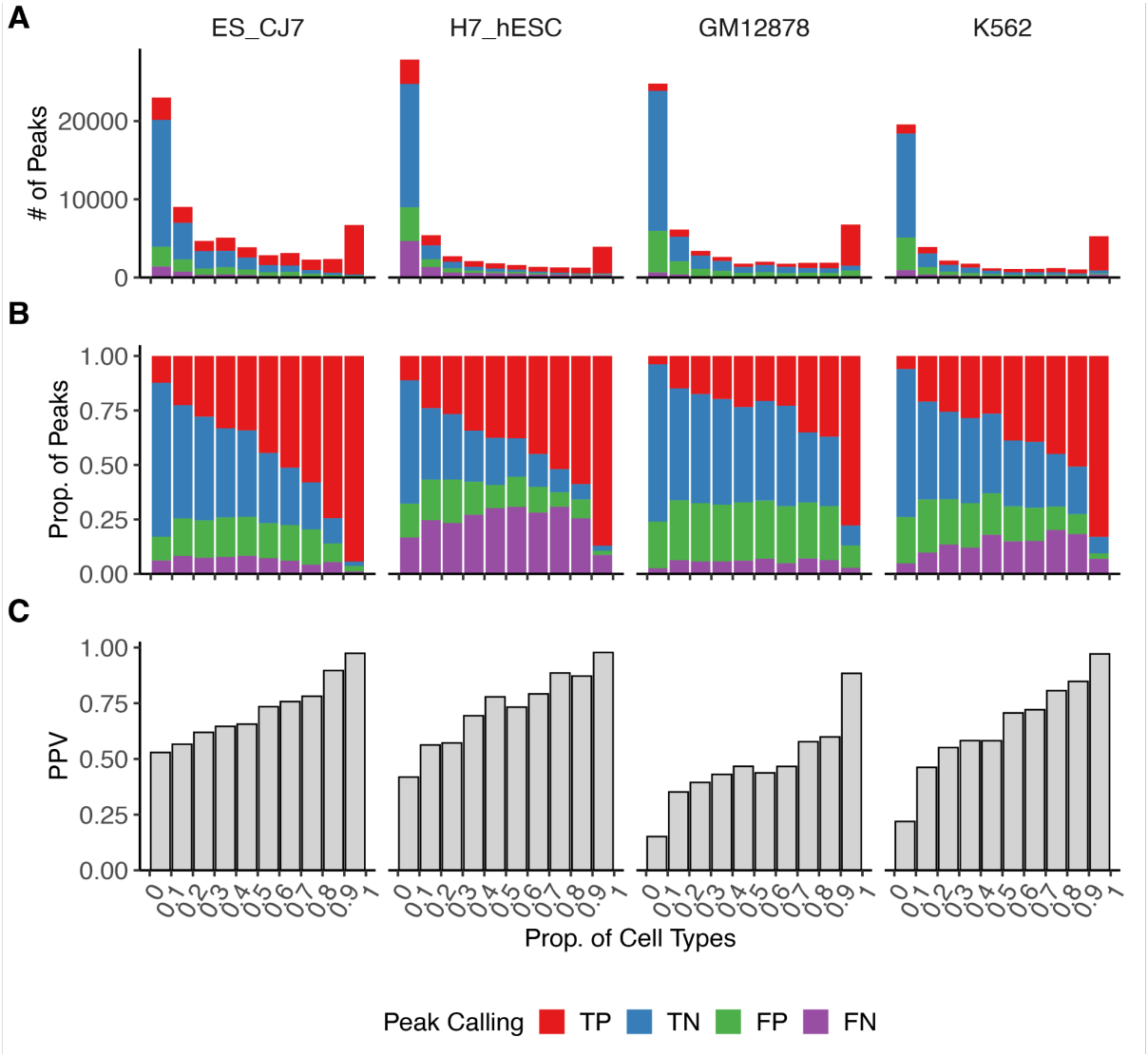

Fig. 5. Enformer favors detection of shared regulatory sites.

(A-C) Peak calling comparison between mESC_CJ7, H7_hESC, GM12878, and K562 DNase-seq and their respective Enformer predictions (detailed in **Methods**). Peaks were binned according to the proportion of cell types represented in a collection of mouse (N=36) and human (N=105) reference DNase-seq datasets (**Table S2**). Peaks are classified as true positives (TP), when found in both reference DNase-seq and Enformer predictions; true negative (TN), when not found in either; false positive (FP), when found only in Enformer predictions; and false negative (FN), when found only in reference DNase-seq. (A) Total number of peaks colored by category. (B) Relative proportion of categories per bin. (C) Positive predictive value (PPV) per bin.

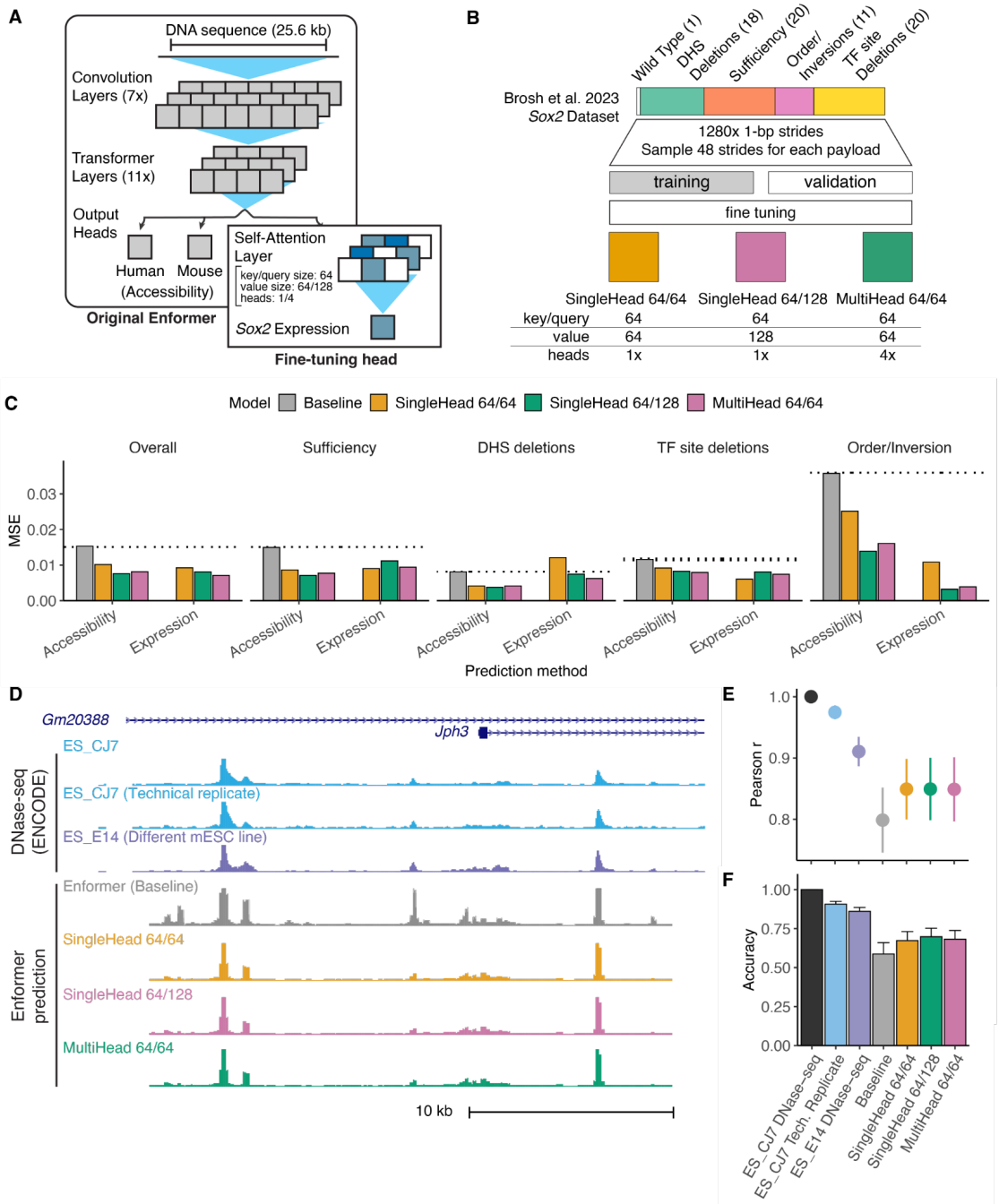

Fig. 6. Improved model prediction by incorporating synthetic regulatory genomics data.

(A) Enformer fine-tuning architecture. (B) Fine-tuning dataset augmentation and training strategy. (C) Predictive performance of the baseline and fine-tuned Enformer models for *Sox2* expression. Expression was estimated either by summing the maximum predicted accessibility of all LCR DHS (Accessibility) or directly from the new output head (Expression). Performance was measured by mean squared error (MSE). Results presented for overall dataset and payload categories with more than two payloads. Horizontal dotted lines indicate baseline performance. (D) Example of experimental and predicted mESC DNase-seq at *Jph3* locus. Shown is DNase-seq of the prediction target (mESC_CJ7), a technical replicate, an independent cell line (mESC_E14), and accessibility predictions made with baseline and fine-tuned Enformer models. (E-F) Peak calling comparison between the prediction target (mESC_CJ7) DNase-seq, a technical replicate, an independent cell line (mESC_E14), and Enformer predictions. Error bars indicate ±1 s.d.. (E) Peak signal correlation score (*r*) with the prediction target (mESC_CJ7). Peak signal was measured as the maximum DNase-seq or Enformer accessibility prediction at the site. Points indicate average *r* among autosomal chromosomes. (F) Average peak calling accuracy among autosomal chromosomes.

Supplementary Materials

TABLE OF CONTENTS

SUPPLEMENTARY FIGURES **25**

Fig. S1. Baseline and adapted Enformer prediction comparison. 26
Fig. S2. Enformer predictions of in locus LCR relocation. 27
Fig. S3. Impact of background sequence on Enformer predictions. 28
Fig. S4. Cell-type activity analysis at promoter, CTCF, and non-CTCF sites. 29
Fig. S5. Schematic of fine-tuning training and dataset augmentation. 30
Fig. S6. Partitioned fine-tuning training performance. 31
Fig. S7. mESC_CJ7 peak analysis. 32
Fig. S8. Peak calling performance across multiple cell/tissues. 33

SUPPLEMENTARY TABLES **34**

Table S1. Genomic coordinates. 34
Table S2. *Sox2* and *α-globin* accessibility predictions. 34
Table S3. ENCODE data used for cross-cell type activity analysis. 34
Table S4. Investigated ENCODE tracks. 35

SUPPLEMENTARY FIGURES

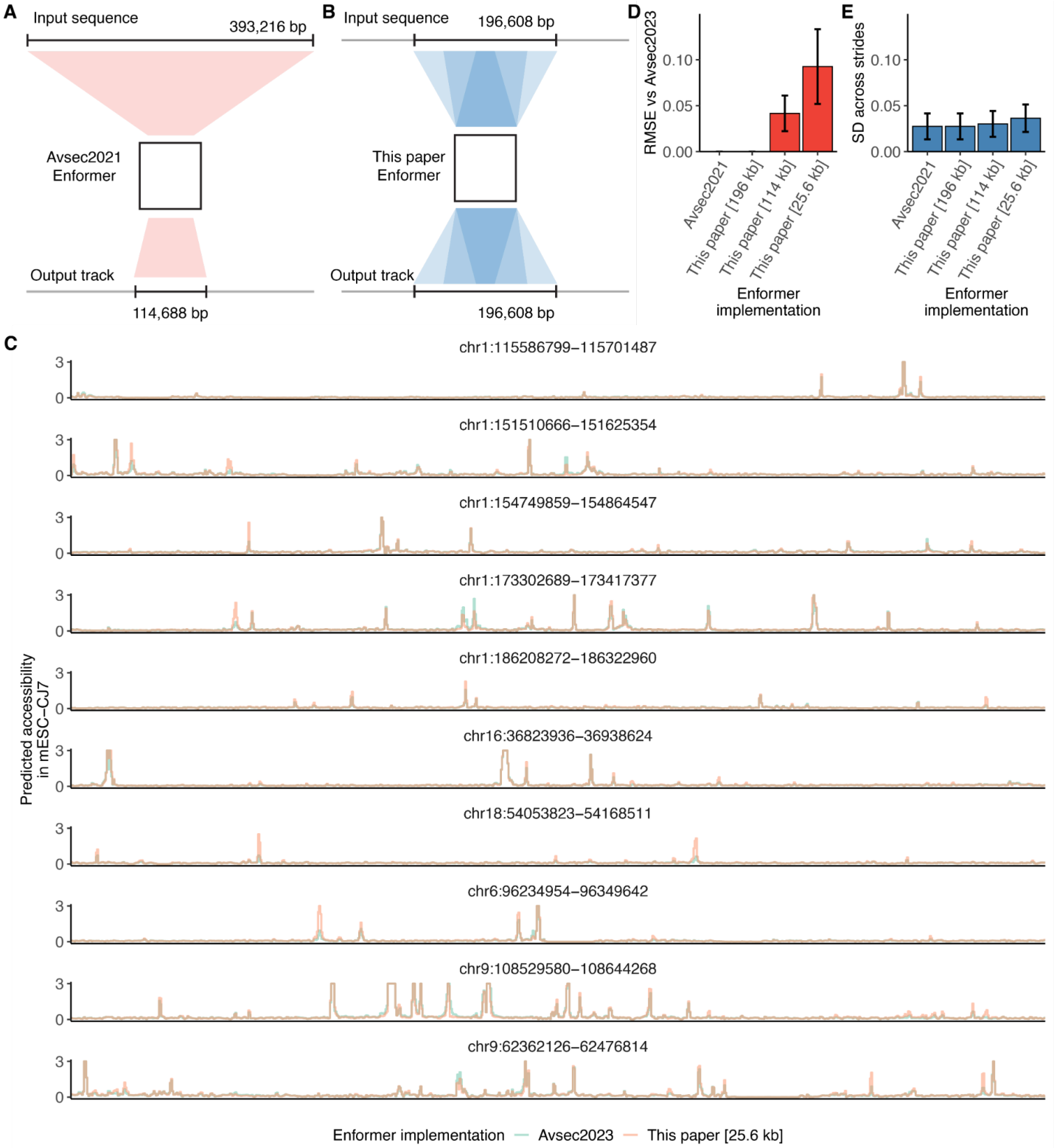

Fig. S1. Baseline and adapted Enformer prediction comparison.

Comparison of Avsec2021 and adapted Enformer implementation in this paper to ensure no major differences were introduced. Here, we evaluated mESC_CJ7 accessibility prediction for these implementations at ten randomly selected 114,688 bp sites containing at least one mESC_CJ7 hotspot V1 (FDR 0.01). (A) Schematic of Avsec2021 baseline model (Avsec et al. 2021a). (B) Schematic of model in this paper. Enformer model code from DeepMind GitHub page with pre-trained weights (see Methods) was adapted to remove internal cropping and allow a flexible input size and output prediction for the whole input sequence. (C) Accessibility predictions for all evaluated sites for both Avsec2021 and this paper [25.6 kb]. (D) Deviation from Avsec2023 baseline measured by Root Mean Squared Error (RMSE). (E) Prediction internal standard deviation across ten 1-bp strides used to estimate accessibility.

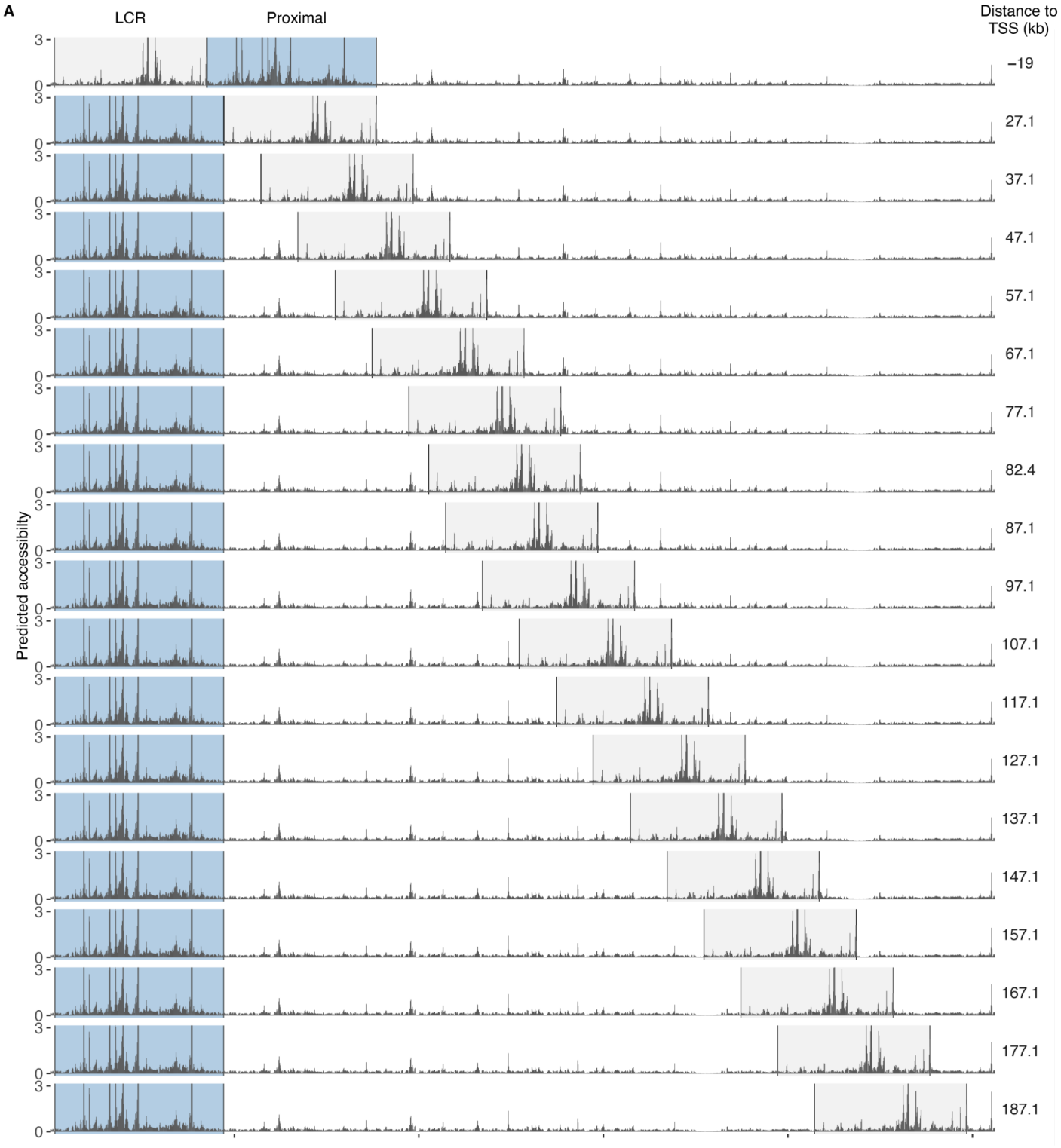

Fig. S2. Enformer predictions of in locus LCR relocation.

(A) Enformer prediction of in locus LCR relocation tracks. *Sox2* proximal region is highlighted in blue and LCR highlighted in gray. LCR distance to *Sox2* TSS is indicated to the right.

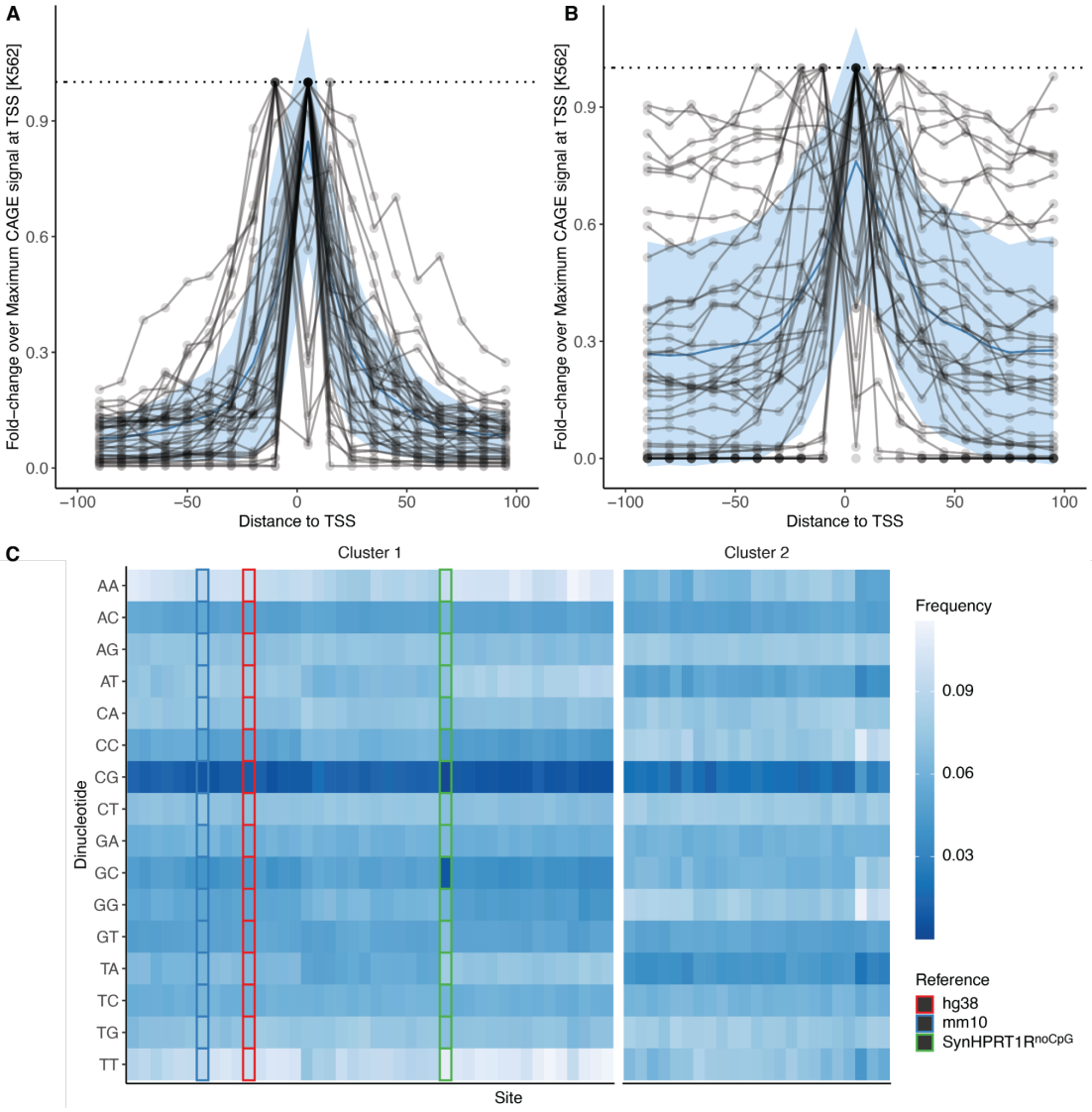

Fig. S3. Impact of background sequence on Enformer predictions.

(A-B) Enformer estimates of transcription driven by 74 K562 candidate CREs (Toneyan and Koo 2024) as a function of their distance to targeted gene TSS. Simulations were conducted by moving CRE regions 10 kb downstream and upstream from their targeted TSS and measuring K562 CAGE signal at the TSS. A 5-kb window centered at both CRE and TSS were preserved while the remaining background sequence was generated by shuffling dinucleotides from the CRE endogenous loci (A) or from SynHPRT1R^{noCpG} (B). CAGE signal was estimated as the average from 10 random dinucleotides shuffles and normalized by the maximum estimated signal across all positions. Each line represents a CRE candidate, and each point a CAGE signal estimate. Solid blue line represents the average signal across all CREs and the blue shaded area ±1 s.d.. (C) Heatmap of dinucleotide frequency for mouse (mm10) and human (hg38) reference genomes, neutral synthetic sequence (SynHPRT1R^{noCpG}), and 74 K562 candidate CRE loci. Candidate CREs loci correspond to a 196,608-bp region centered at the targeted gene TSS. Each column represents a candidate CRE or reference, each row a dinucleotide, and their frequency represented by a color gradient from blue to white. Sites were ordered according to a hierarchical clustering and divided into two clusters indicated above. Dinucleotide sequences for hg38, mm10, and SynHPRT1R^{noCpG} are highlighted in red, blue, and green, respectively.

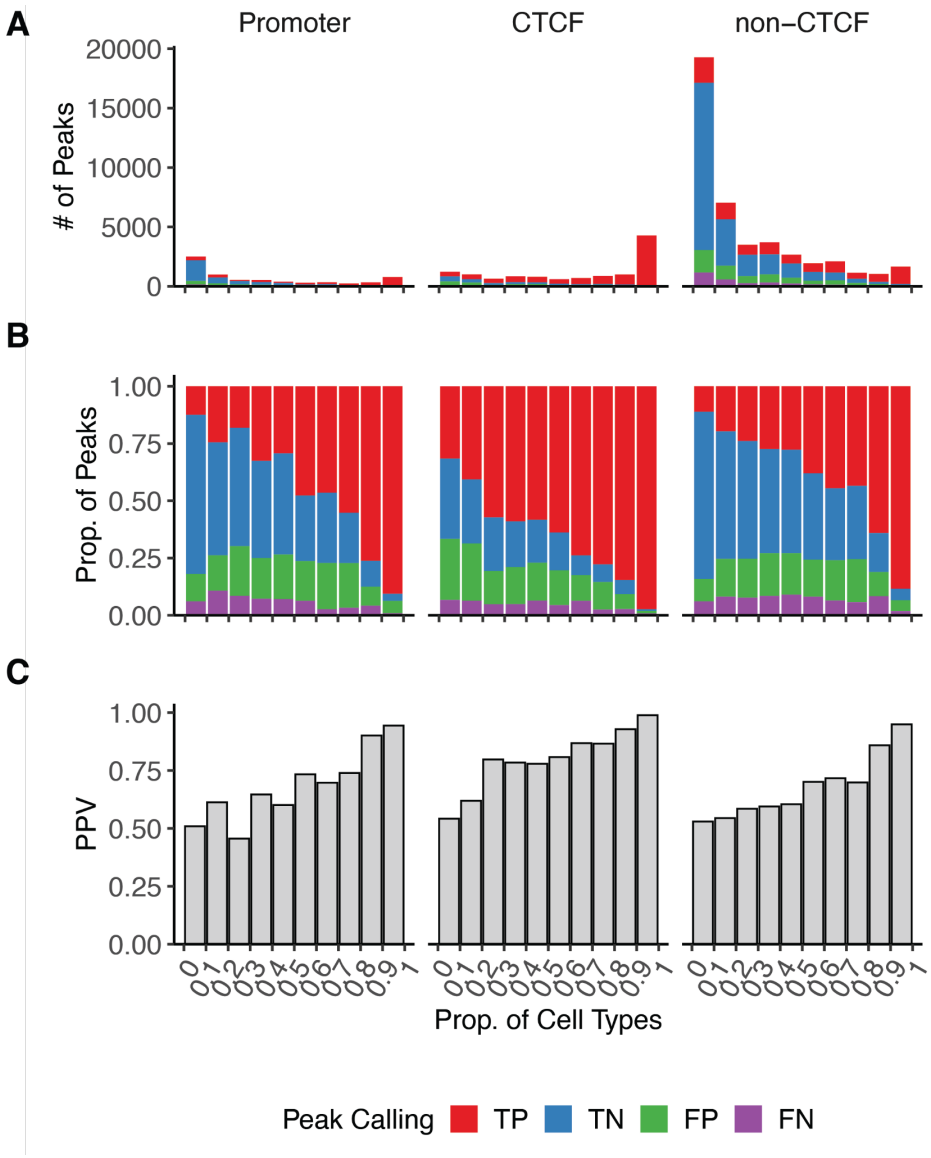

Fig. S4. Cell-type activity analysis at promoter, CTCF, and, non-CTCF sites.

(A-C) Peak calling comparison of mESC_CJ7 DNase-seq track and Enformer prediction (detailed in **Methods**) of Promoter (within 2.5 kb of TSS), CTCF (overlapping CTCF hotspots) and non-CTCF sites. Peaks were binned according to the proportion of cell types represented in a collection of mouse (N=36) reference DNase-seq datasets (**Table S2**). Peaks are classified as true positives (TP), when found in both reference DNase-seq and Enformer predictions; true negative (TN), when not found in either; false positive (FP), when found only in Enformer predictions; and false negative (FN), when found only in reference DNase-seq. (A) Total number of peaks colored by category. (B) Relative proportion of categories per bin. (C) Positive predictive value (PPV) per bin.

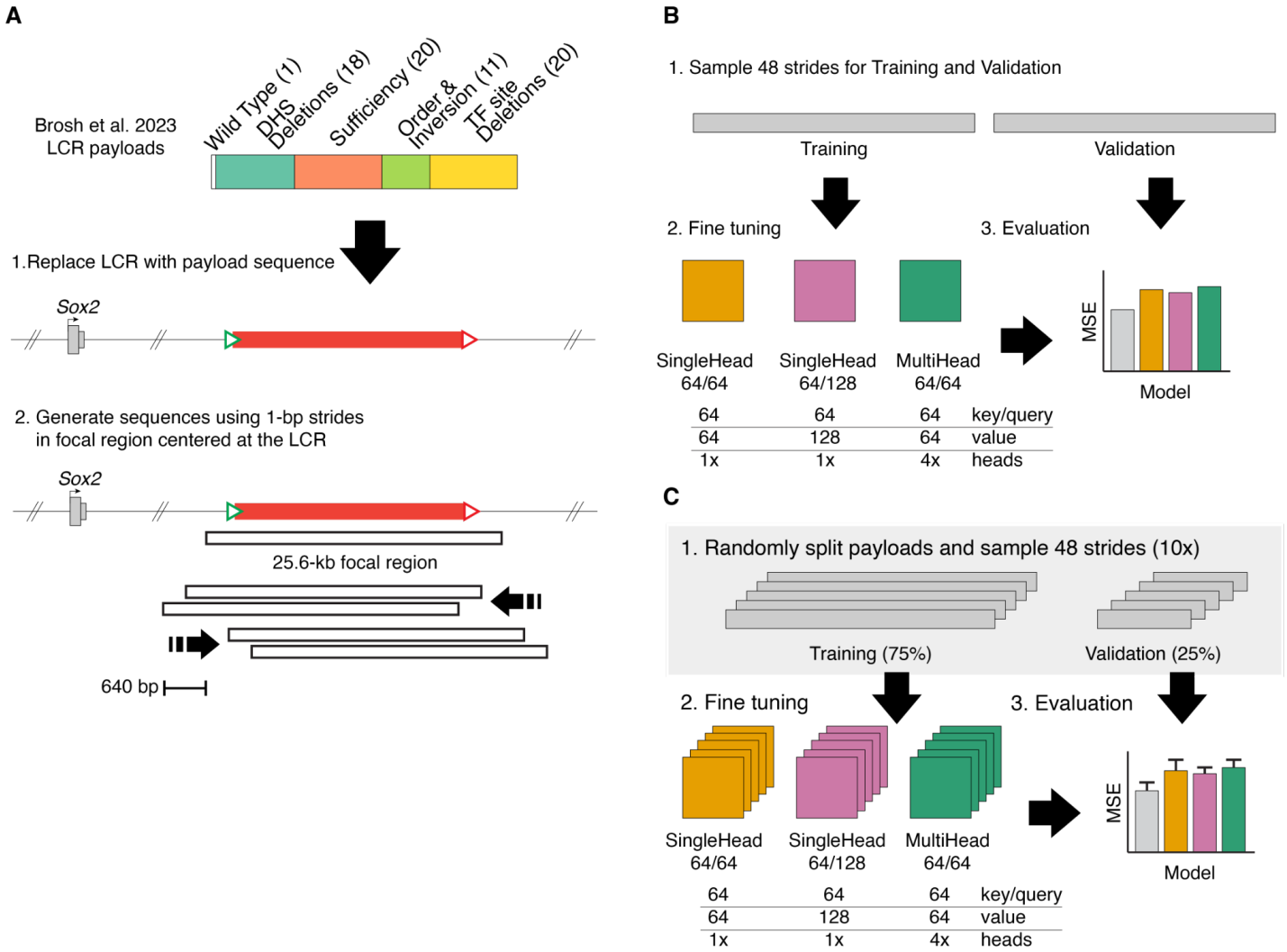

Fig. S5. Schematic of fine-tuning training and dataset augmentation.

(A) *Sox2* LCR dataset augmentation strategy. Virtual input sequences were built by replacing *Sox2* LCR by each synthetic payload sequence. Dataset was augmented by striding a 25.6-kb focal region centered at the LCR 1 bp at a time (upstream and downstream) into a total of 1,280 sequences. (B-C) Fine-tuning strategy using all payloads (B) and partitioning payloads into training and validation datasets (C). (B) Fine-tuning strategy including all payloads. For each payload, 48 strides were randomly selected for training and another 48 strides were selected to evaluate the trained models. (C) Fine-tuning strategy on partitioned payloads. Payloads were randomly partitioned into training and validation datasets in ten independent batches and 48 strides of each payload selected. For each batch, models were trained using the training dataset and evaluated with their respective validation dataset.

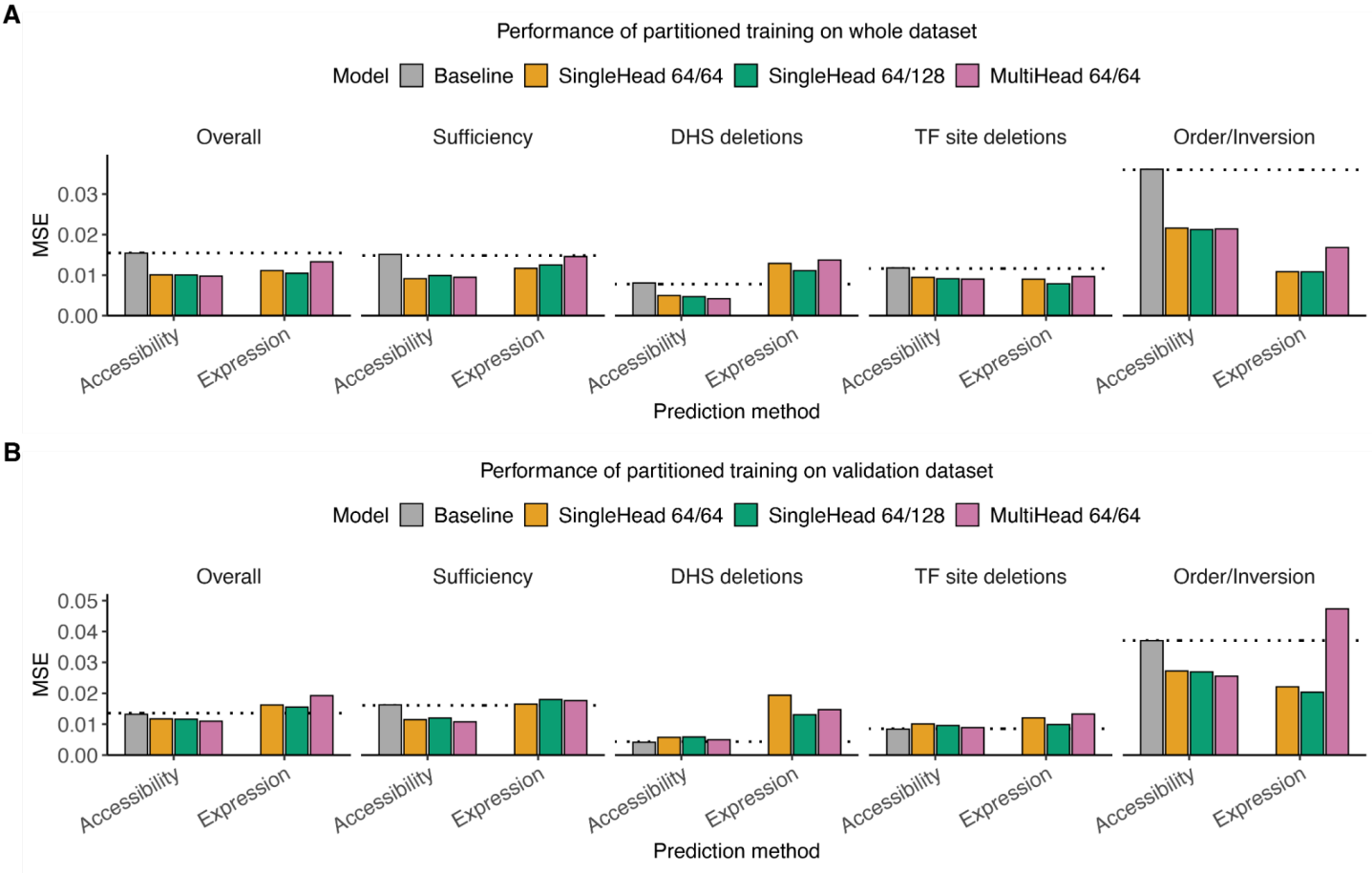

Fig. S6. Partitioned fine-tuning training performance.

(A-B) Prediction performance of baseline and fine-tuned models trained using payload partitioning strategy. Expression was estimated either by summing the maximum predicted accessibility of all LCR DHS (Accessibility) or directly from the new output head (Expression). Models performance was measured as mean squared error (MSE) and average between ten independent partition batches and error bars ±1 s.d.. Presented the results for the overall dataset and payload categories with more than two payloads. Horizontal dotted lines indicate baseline performance. (A) Shows prediction performance for the whole training and validation dataset, and (B) performance for the validation dataset.

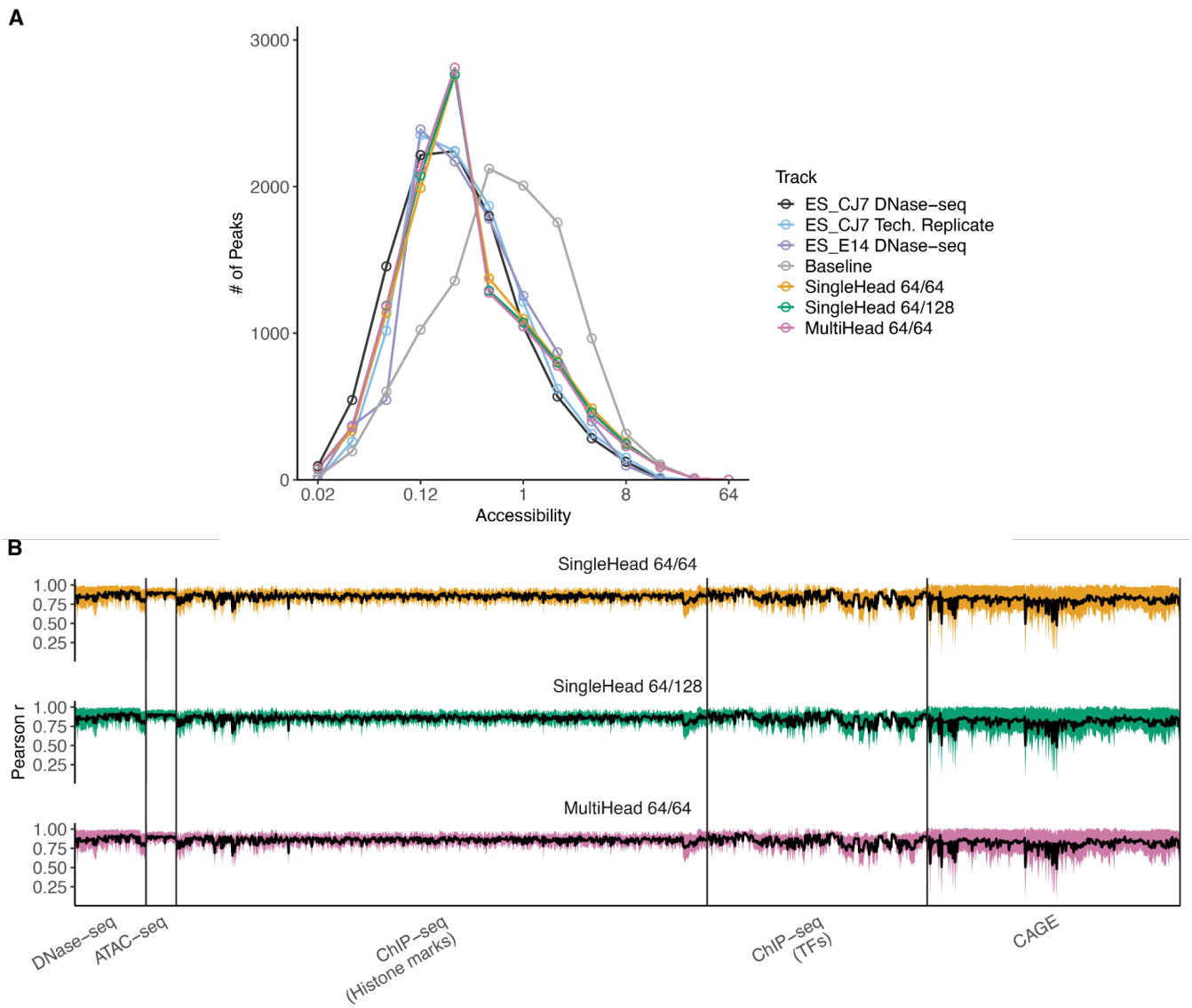

Fig. S7. mESC_CJ7 peak analysis.

(A-B) DNase-seq density and Enformer prediction comparison at 2,000 randomly selected 25.6-kb regions containing at least one mESC_CJ7 hotspot V1 (FDR 0.01). (A) Peak signal distribution at randomly selected regions. Peak calling detailed in **Methods**. Peak signal was measured as maximum DNase-seq density or mESC_CJ7 predicted accessibility. Shown the signal distribution for DNase-seq data for mESC_CJ7, a technical replicate, and an independent cell line (mESC_E14), and Enformer predictions. (B) Fine-tuned models prediction correlation to baseline Enformer at randomly selected regions. Shown Pearson correlation score (*r*) of fine-tuned models to baseline model prediction for all mouse output tracks (single-cell ATAC-seq were excluded). Black line indicates average *r* across all evaluated regions and shaded area represents ±1 s.d.. Tracks are grouped according to their type indicated in x-axis.

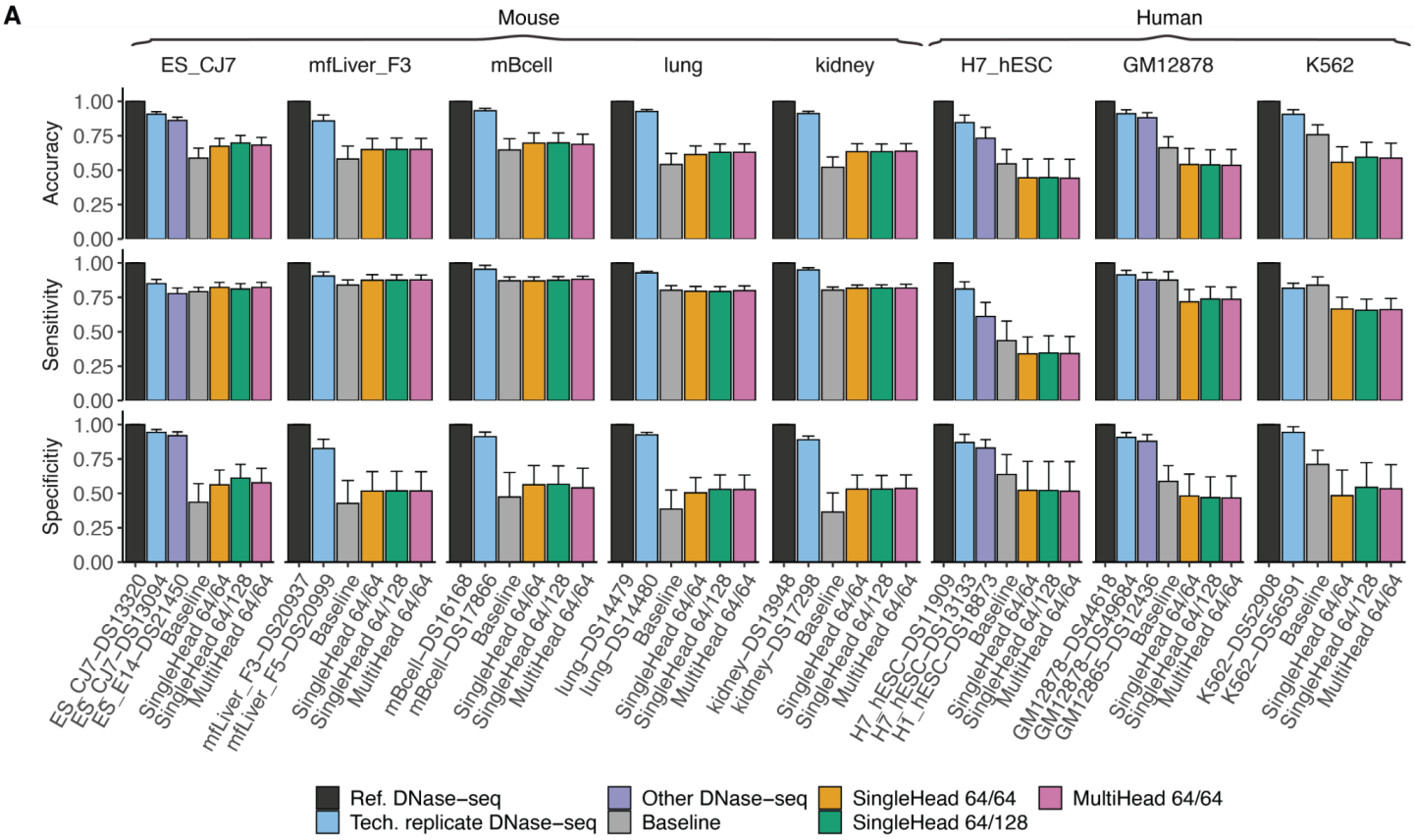

Fig. S8. Peak calling performance across multiple cell/tissues.

(A) Peak calling accuracy, sensitivity (true positive ratio) and specificity (true negative ratio) for various targeted ENCODE cell/tissues DNase-seq tracks. Peak calling is detailed in **Methods**. Results shown for the targeted DNase-seq track, a technical replicate, an independent similar track, and predictions made by baseline and fine-tuned Enformer. Error bar indicates ±1 s.d..

SUPPLEMENTARY TABLES

Table S1. Genomic coordinates.

Genomic coordinates of DHSs and other relevant regions.

Table S2. *Sox2* and *α-globin* accessibility predictions.

Enformer accessibility predictions of synthetic payloads delivered in place of *Sox2* LCR (Brosh et al. 2023) and CRISPR/Cas9 editing of *α-globin* enhancers (Blayney et al. 2023), and their experimentally characterized engineered gene expression (qRT-PCR). Payload.ID, unique descriptor of payload or CRISPR/Cas9 edit; Group, payload or edit category; Locus, indicator of *Sox2* or *α-globin* locus; Activity, normalized observed gene expression; AccessibilityPrediction, sum of DHSs maximum accessibility for all *Sox2* LCR DHSs 19-28 or *α-globin* DHSs (R1-R4); AdjustedAccessibilityPrediction, scaled sum of DHSs accessibility to fit gene activity (detailed in Methods).

Table S3. ENCODE data used for cross-cell type activity analysis.

Collection of ENCODE mouse and human DNase-seq tracks employed to characterize cross-cell patterns. Name, cell or tissue label; DS and ENCODE, sample identifiers (see https://www.encodeproject.org); and Species.

Table S4. Investigated ENCODE tracks.

ENCODE tracks employed to compare baseline and fine-tune models. DS and ENCODE sample IDs are from https://www.encodeproject.org (Thurman et al. 2012; Vierstra et al. 2014; Meuleman et al. 2020). EnformerIndex lists track index among Enformer predictions. Replicate ID includes DS number of technical replicate; Other ID includes unique DS number for the other DNase-seq track to which the current was compared.

Name	Sample ID	ENCODE ID	Species	Enformer Index	Replicate ID	Other ID
H7_hESC	DS11909	ENCLB449ZZZ	human	20	DS13133	DS18873
GM12878	DS44618	ENCLB681AXO	human	12	DS49684	DS12436
K562	DS52908	ENCLB096YUZ	human	625	DS56591	DS56591
mESC_CJ7	DS13320	ENCLB163SYJ	mouse	5323	DS13094	DS21450
mESC_E14	DS21450	ENCLB890BTE	mouse	5324	DS18505	DS13320
mfLiver_F3	DS20937	ENCLB173ZFD	mouse	5322	DS20999	-
MEL	DS13036	ENCLB066PZV	mouse	5339	DS21963	-
mBcell	DS16168	ENCLB693PQW	mouse	5315	DS17866	-
lung	DS14479	ENCLB144OVE	mouse	5338	DS14480	-
kidney	DS13948	ENCLB388UBV	mouse	5333	DS17298	-