Model-based design of regulatory DNA for cell type-specific gene expression

Christopher Yin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Georg Seelig, Chair

Eric Klavins

Sara Mostafavi

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Model-based design of regulatory DNA for cell type-specific gene expression

Christopher Yin

Chair of the Supervisory Committee:
Professor Georg Seelig
Department of Electrical and Computer Engineering

An important and largely unsolved problem in synthetic biology is how to target gene expression to specific cell types. Enhancers are a class of *cis*-regulatory element (CRE) that exist in the noncoding regions of the genome and are implicated as major drivers of cell type-specific gene expression, via a complex and incompletely understood sequence grammar. Next Generation Sequencing has enabled the interrogation of this grammar at high throughput via multiple paradigms. Massively Parallel Reporter Assays (MPRAs) directly measure enhancer activity for libraries of up to hundreds of thousands of sequences

at once, but are limited in terms of sequence length and experimentally compatible targets. Chromatin accessibility is commonly used as a surrogate metric indicating likely enhancer identity, and can be profiled genome-wide for a far greater range of biological targets compared to MPRAs using techniques such as ATAC-seq or DNase-seq; nonetheless, these cannot provide direct readout of enhancer activity.

In this dissertation I explore the capacity for deep learning models trained on either MPRA or chromatin accessibility data to design functional cell type-specific enhancers. In Chapter 2, I establish the viability of both approaches by designing and experimentally validating synthetic enhancers targeted to 2 human cancer cell lines; and in Chapter 3, I build upon this work by training models only on accessibility data, enabling me to take advantage of the greater coverage of biological diversity in accessibility datasets compared to MPRA compendia. I show successful enhancer design in 9/10 human cell lines, confirming the generalizability of this approach; and additionally show *in vivo* that enhancers targeted to a retinoblastoma line are active in mouse retinas. In both chapters I analyze the sequence determinants of enhancer specificity via enrichment-based and explainable AI techniques, exposing complex combinatorial relationships between discrete sequence elements corresponding to known Transcription Factor Binding Sites (TFBSs). Furthermore, I analyze why enhancers designed to achieve high predicted specific accessibility sometimes fail to exhibit correspondingly specific enhancer activity, which will inform future

expansions of this approach. This work shows that model-guided design of enhancers can help us decipher the *cis*-regulatory code governing cell type specificity and result in novel tools for selective targeting of human cell types.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thank you to my advisor Prof. Georg Seelig, who handed me the idea for this project and let me run with it. Thank you to Sebastian Castillo-Hair and Wouter Meuleman for your collaboration on all this work. Thank you to Gun Woo Byeon for running the scMPRA in Chapter 2; and to Leah VandenBosch and Timothy Cherry for the mouse retina work in Chapter 3. Thank you to Johannes Linder for helping get me started in the lab and for developing the techniques these projects rely on. Thank you to all of my labmates for everything else. Thank you also to the crows.

CHAPTER 1. INTRODUCTION

It has been estimated that the human body broadly comprises ~200 distinct cell types[1], with advances in single cell sequencing technologies providing an increasingly granular taxonomy[2]. Yet despite this diversity, every cell in the body contains an identical copy of the same genome. The mechanisms by which so many disparate functions and phenotypes arise from the same DNA are believed to reside significantly in the ~98% of the genome that does not code for proteins[3,4]. The term "*cis*-regulatory code" has been coined to encapsulate the complex, highly contextual, and incompletely understood rules by which differences in gene expression, epigenetic modifications, and ultimately cell type are enacted by processes contingent on noncoding DNA.

One of the most prominent features of this code are enhancers, a class of *cis*-regulatory elements (CREs) that significantly determine differential gene expression across cell types[5]. Transcription Factor Binding Sites (TFBSs), short sequence fragments recognized and bound by cognate transcription factors (TFs), are the core functional units within an enhancer sequence but occupy only a small fraction of the enhancer footprint, which typically spans hundreds or even thousands of base pairs[3,5]. TFs are proteins with DNA-binding Domains (DBDs) that can work individually or in coordination with other TFs and protein cofactors to regulate expression of a target gene or genes; this can be

accomplished via mechanisms such as chromatin remodeling or recruitment of RNA polymerase II[6]. Overall enhancer activity is determined by a complex grammar encompassing TFBS identity, abundance, relative positioning, flanking sequence, and combinatorial interactions; crucially, these elements can respond differently to different cellular contexts[3,7,8]. For instance, the same TF may have an activating or repressing effect on a given enhancer depending on other co-recruited TFs[9]. Because enhancers can therefore target gene expression to specific cell types they represent powerful tools for basic biology, and have significant potential for gene therapy applications where off-target expression must be minimized[10–12].

Despite concerted interest in and effort towards exploiting these desirable properties, there remain significant challenges to identifying strongly cell type-specific enhancers in the genome. A defining feature of enhancers is their ability to regulate distal genes, as well as to act upon multiple regulatory targets, making it difficult to causally associate enhancers with their effects upon gene expression in their native context[3,5]. Discrete enhancer elements have also been known to coordinate across the genome to form "super-enhancers" with potential positional dependencies between constituents, further contributing to the complexity[13,14]. Accordingly, deciphering this cis-regulatory enhancer code is considered one of the current frontiers of genome science[3,5,7].

In the absence of quantitative mappings from sequence to cell type-specific activity, proxy metrics are often relied upon to nominate candidate enhancers. For example, chromatin accessibility has shown to be a hallmark of regulatory elements[10,15,16], and can be mapped at genome-scale using DNase I digestion (DNase-seq[17]) or Tn5 insertion (ATAC-seq[18]) followed by high throughput sequencing. Maps of DNase I Hypersensitive Sites (DHSs) are providing increasingly detailed tissue and cell type-specific atlases of these elements[19,20]. Certain histone modifications – including H3K4 mono-methylation (H3K4me1) and H3K27 acetylation (H3K27ac) – have further been associated with enhancer identity[21,22]. However, while sequences carrying the requisite marks might be functional enhancers, their target genes and cell types are not always straightforward to identify; and our understanding of the mapping from chromatin state to enhancer activity remains incomplete[5]. More specifically, genomic regions with the expected chromatin marks for enhancer identity have a low success rate when tested for enhancer activity in isolation[23], necessitating extensive pre-screening and validation[24,25].

An alternative approach to identifying cell type-specific enhancers is via Massively Parallel Reporter Assays (MPRAs)[26,5,27] which provide a direct readout of the ability of a DNA fragment to drive gene expression of a minimal promoter in the context of a plasmid. Libraries of accessible genomic sequences and variants thereof are commonly screened in MPRAs. However, in spite of progress in lenti-viral delivery[28] and single cell readout[29,30] of

MPRAs, the assessment of large reporter libraries in complex tissues remains a bottleneck, and MPRA data are only available for a limited number of cell types compared to chromatin accessibility.

Recently, synthetic enhancer design has emerged as the state-of-the-art approach confronting the limitation of purely experimental screens. Rationally embedding TFBSs within inert sequences has produced cell type-specific enhancers, but this approach may be biased by our preconceived notions of enhancer grammar, has largely been limited to a confined set of well-studied cell lines for which strong candidate TFBSs have already been characterized, and can only explore a constrained combinatorial sequence space[31–34]. An alternative, data-driven strategy involves training deep learning models on large-scale genomic and/or MPRA datasets, which learn a sequence-to-function mapping that captures underlying biological principles[8,35–39] and can thereby guide the design of synthetic enhancers with targeted activity levels. Such an approach has seen success at targeting cell types in *Drosophila melanogaster*, generally starting from accessibility measurements and with validations based on individual fluorescent reporters. For example, Taskiran *et al.* trained a hybrid CNN-RNN classifier on pseudobulked single cell ATAC-seq data (~200k sequences) covering 15 total cell types, then used model-directed mutagenesis and Generative Adversarial Networks (GANs) to design enhancers targeting two distinct cell types, and validated them via fluorescent imaging of transgenic fly brains[40]. De Almeida et al. similarly

used CNNs individually trained on single cell ATAC-data (~460k peaks) to predict accessibility in five tissues, but finetuned on pre-existing enhancer activity annotations (~170k)[41]. 27 out of 40 designed enhancers validated via in-situ hybridization were specific to their targets.

More recent demonstrations of this model-guided design approach have used models trained on MPRA data[8,42]. Gosai *et al.* scaled synthetic enhancer design and validation in human cells using a CNN trained on functional MPRA measurements of >750k 200bp sequences, in conjunction with both gradient- and stochastic search-based generative methods, to design a library of 51k enhancers targeted to a set of three cell lines (HepG2, K562, SK-N-SH)[43]. DaSilva *et al.* and Lal *et al.* demonstrated synthetic cell type-specific enhancer design *in silico* using more recent generative models (diffusion, autoregressive transformer models), but did not experimentally confirm whether these produce functional enhancers[44,45].

This dissertation discusses work conducted alongside colleagues in the Seelig Lab and Altius Institute to advance the nascent field of deep learning-based enhancer design. First, I will describe a pilot project designing a library of synthetic enhancer sequences for maximal cell type-specificity in two human cancer cell lines, HEPG2 and K562 (Chapter 2). This project experimentally validates enhancers designed using models trained on either MPRA or chromatin accessibility data, establishing the viability of both approaches; though

as expected, enhancers designed from MPRA models exhibit greater success rate and higher specificity than those designed from accessibility models. In this chapter I further demonstrate an iterative design cycle wherein models are retrained with measurements from a first round of synthetic enhancers, enabling the design of a second round with improved performance; and I provide analysis seeking to characterize optimal model training and design algorithm practices. I conduct TFBS motif analysis to interrogate the sequence determinants of cell type-specificity in my designed enhancers, and identify a concise TFBS vocabulary and higher-density syntax driving superior performance compared to genome-derived controls. Additional experiments causally probe individual and pairwise contributions of TFBS sites to top-performing enhancers, and demonstrate the capacity of model-based design to explore nuanced objectives beyond simple maximization of sequence specificity. Finally, a single cell MPRA recapitulates enhancer specificity from bulk experiments and allows for correlation with differential TF expression.

Despite superior performance of enhancers designed from MPRA vs accessibility models, the dearth of publicly available MPRA data outside a fairly narrow set of well-studied and easily transfectable cell lines[46] limits the generalizability of this approach. Therefore, in Chapter 3 of this dissertation I focus on enhancer design purely from chromatin accessibility data, which is notably easier to profile genome-wide across diverse samples[47,48]. A chief aim of this chapter is to explore the broader potential of such an approach—namely,

the ability to leverage the vast diversity of accessibility datasets to design enhancers specific to a wide range of tissues. In addition, this work seeks to understand differences in the sequence determinants of genomic accessibility compared to those of enhancer function, e.g. due to the existence of pioneer factors and other chromatin remodelers that may not directly drive transcription[49,50]. Not only will this help bridge the performance gap between MPRA-based and accessibility-based enhancer design, but it can also further elucidate the mechanisms of the *cis*-regulatory code.

Specifically, Chapter 3 describes a deep learning model ("DHS64") trained on a subset of experiments collected in the ENCODE DNase I Hypersensitive Site (DHS) Index[20], covering a total of 64 distinct, biologically diverse cell and tissue types. This model is used to design a library of thousands of synthetic sequences, which are then experimentally validated in a representative panel of 10 diverse human cell lines via MPRAs, as well as *in vivo* in mouse retinas. This MPRA data is subsequently used to finetune the DHS64 model, enabling the application of explainable AI techniques to identify differences in the sequence grammar of genomic accessibility and enhancer activity. Additionally, having access to enhancer activity measurements in 5X more cell lines than the work discussed in Chapter 2 allows me to expose a correspondingly more complex TFBS grammar: while some TFBS motifs can be associated with specificity in a single cell line, many TFBSs exhibit multiple such associations. Interrogating this phenomenon, I find evidence of several TFBSs whose

contribution towards enhancer specificity can be converted between different cell lines in the presence of different partner motifs. Taken together, the work in this dissertation provides a foundation for ongoing and future studies to the practical application of data-driven design and understanding of synthetic regulatory elements.

CHAPTER 2. DESIGNING ENHANCERS IN 2 HUMAN CELL LINES FROM MPRA AND DHS MODELS

In this chapter I describe the validation and extensive characterization of a high-throughput, multimodal, and iterative enhancer design process. We train models on two different types of previously published data, MPRA and DHS, then apply several distinct strategies to design libraries of candidate enhancers that maximize cell type-specific enhancer activity. We test these synthetic enhancers in human cell lines (HepG2 and K562), then retrain our model with these data and repeat the design-build-test cycle to show that iterative retraining results in dramatic performance improvements. Strikingly, better performance can be achieved even with a relatively small amount of data. Such an iterative "small data" approach may be more readily adapted to selecting enhancers for specific cell types *in vivo*, rather than an approach that requires testing hundreds of thousands of sequences for model training. Model interpretation and motif analysis techniques reveal how the information content encoded in the enhancers evolves with each round of model training towards a more selective TFBS vocabulary and higher density syntax. Additionally, we perform motif ablations for a subset of enhancers to experimentally and computationally quantify the *cis*-regulatory code. We perform a single cell MPRA (scMPRA) to quantify cell-to-cell variation in enhancer activity and show correlated activity of enhancers with cognate TF expression

levels. Finally, we explore enhancer truncations and find that enhancers up to ~3x shorter than our baseline sequences can still exhibit high specificity. This work was published, in a slightly different form, as a research article titled "Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity" in *Cell Systems* on June 4th, 2025. It was conducted with the collaboration of Sebastian Castillo-Hair, Gun Woo Byeon, Peter Bromley, Wouter Meuleman, and Georg Seelig.

2.1 Deep learning design of cell type-specific enhancers from MPRA models

Enhancer MPRAs measure the ability of a DNA fragment to drive reporter gene expression and thus provide a functional readout of the regulatory activity of a putative enhancer. Here, we set out to capture sequence-encoded determinants of cell type-specific enhancer activity to guide the design of synthetic enhancers with tailor-made properties. To this end, we trained neural network predictors on the previously published Sharpr-MPRA dataset[51]. This dataset consists of two replicates of log2 fold-change (log2FC = $\log_2$(mRNA counts / DNA counts)) enhancer activity measurements from 467k 145 nt-long candidate enhancers extracted from accessible genomic regions in HepG2, K562, HUVEC and H1-hESC (**Figure 2.1A**), cloned into reporter plasmids upstream of a minimal promoter (minP), and assayed in HepG2 and K562 cells. We refer to enhancer activity in each cell line as $\text{log2FC}_{\text{HepG2}}$ and

$\log2FC_{K562}$, and define the differential activity or specificity as $\log2FC_{H2K} = \log2FC_{HepG2} - \log2FC_{K562}$. We filtered out sequences with low read coverage and low specificity, and retained 29,891 sequences for model training (**Supplementary Figure 2.1A, Appendix A**). We refer to this initial dataset as R0-MPRA (**Figure 2.1B**).

To avoid overreliance on any single modeling approach, we explored several variations of model training strategies. "Single" models were trained on the same data split but with different randomly initialized network weights, "boot" models with the same architecture were trained on randomly resampled bootstraps of the data, and "ensemble" models were formed by averaging the outputs of 10 boot models at a time. In total, we trained 120 multitask CNN models across these different approaches to predict enhancer activity in each cell type from one hot-encoded DNA sequence input (**Appendix A, Supplementary Figure 2.1B**). All model types achieved prediction-measurement correlation on par with the inter-replicate correlation of R0-MPRA (**Appendix A, Supplementary Figure 2.1C**).

We then applied three different design methods (Simulated Annealing, Fast SeqProp[52], Deep Exploration Networks[53], **Appendix A, Supplementary Figure 2.2A-C**) to generate 1,037 *de novo* candidate enhancers that maximize predicted differential enhancer activity. Each method optimizes sequences using a predictor as the oracle via a distinct approach: simulated annealing and Fast SeqProp optimize sequences individually

via stochastic search and gradient ascent, respectively, whereas in deep exploration networks, we train a new generative neural network to generate sequences that maximize predicted performance and minimize sequence similarity. Sequences were generated separately using each of the 120 models trained above. We verified that the designed sequences were highly dissimilar to one another and to the Sharpr-MPRA dataset (**Supplementary Figure 2.2D,E**). As controls, we re-synthesized 100 sequences from Sharpr-MPRA selected to span the entire enhancer activity range. The reverse complements of these 100 control sequences were synthesized as well to test the impact of enhancer orientation. Finally, we included 62 additional control sequences embedding known TFBS motifs at different multiplicities in a random backbone sequence (results discussed later). Collectively, these MPRA-guided designs and associated controls are referred to as R1-MPRA and the library composition is detailed in **Table 2.1**.

We experimentally quantified the activity of our synthetic enhancers and control sequences in the R1-MPRA library using the reporter construct and cell lines (HepG2 and K562) from Sharpr-MPRA[51] (**Appendix A**). Our measurements were highly consistent across replicates (**Supplementary Figure 2.2F**), and the activity of the selected control sequences and their reverse complements were well correlated with measurements in the original study (**Supplementary Figure 2.2G,H**). mRNA expression from designed enhancers overwhelmingly matched their target specificity: most sequences designed to

target HepG2 resulted in higher expression in HepG2 compared to K562 (median $log2FC_{H2K}$ = 3.28, 9.7x higher HepG2 expression), and vice versa (median $log2FC_{H2K}$ = -1.22, 2.3x higher K562 expression). Furthermore, many sequences achieved cell type-specificity not merely via weak activity in the off-target cell line, but via strong activity in the target cell line (**Figure 2.1C**). In general, HepG2-targeted designs exhibited higher differential activity than those targeting K562. 113/523 (22%) synthetic enhancers showed greater $log2FC_{H2K}$ than the most HepG2-specific control (4.80), but only 2/514 (0.004%) synthetic sequences surpassed the most K562-specific control sequence (-5.03).

2.2 Deep learning design of cell type-specific enhancers from chromatin accessibility models

Unlike enhancer MPRA data, DNA accessibility measurements are available for a plethora of cell and tissue types, providing a promising starting point for the selection and design of enhancer elements for biological contexts lacking coverage by MPRAs. We wanted to assess to what extent models trained on accessibility data alone can be used to generate functional enhancers that exhibit strong and specific activity in human cells.

To this end, we trained a Generative Adversarial Network (GAN)[54] model using a compendium of 918,057 endogenously accessible sequence elements obtained from large-scale chromatin accessibility assays encompassing >400 distinct cell and tissue types[20]

(**Appendix A**). Briefly, in these assays chromatin is digested using the DNase I endonuclease, preferentially cleaving sites that are accessible to DNA binding factors and thus allowing the genome-wide identification of DNase I Hypersensitive Sites (DHSs). The choice of a GAN was motivated by the goal of generating sequences that mimic natural DHSs. The GAN consists of a discriminator and generator model that are trained simultaneously. The discriminator is trained to distinguish between endogenously observed accessible sequence elements (i.e., "positive examples") and "fake" sequences generated by the generator, while the latter is trained to generate sequences that fool the discriminator, thereby increasingly resembling "real" ones. After training, the GAN thus produced *de novo* sequences resembling endogenous genomic accessible elements (**Supplementary Figure 2.3B,C**, **Appendix A**), albeit not explicitly geared towards any specific cellular contexts. To maximize cell type-specificity, we therefore subsequently tuned generated sequences using a separate classifier model trained on a subset of sequences specifically accessible in HepG2, K562, or otherwise (library referred to as R0-DHS, **Appendix A**). We find this results in a stable tuning process, with generated sequences largely retaining their identity across many tuning iterations (**Supplementary Figure 2.3D,E**, **Appendix A**); and with limited impact on sequence characteristics (**Supplementary Figure 2.3F,G**). Despite this, during tuning we observe a striking increase of the number of sequences containing TF motifs relevant[20] for our cellular contexts of interest

(**Supplementary Figure 2.3H,I**). After the tuning process, for each cell type we selected 300 designed sequences plus 37 positive-control genomic sequences for subsequent experimental validation; we refer to this library as R1-DHS (**Appendix A**). Enhancer activity was quantified as above and with equivalently high replicate correlation (**Supplementary Figure 2.3J,K**)

Despite being designed for chromatin accessibility only, we find that R1-DHS sequences strongly drive cell type-specific gene expression **Figure 2.1D**). As in the MPRA-based approach, HepG2 designs were more successful, with 46/300 (15%) designs exceeding the specificity of the best R0-DHS control (3.75), compared to 9/300 (3%) designs exceeding the specificity of the best K562 control (-3.45). The median $\log2FC_{H2K}$ of R1-DHS sequences was 1.71 (3.3x higher HepG2 expression) for HepG2 designs, and 0.04 for K562 designs.

R1-DHS sequences on average exhibited lower activity than R1-MPRA sequences (**Figure 2.1C,D,F, Supplementary Figure 2.4A**), consistent with R1-MPRA designs being generated from models trained directly on enhancer activity data. In HepG2-targeted designs the median $\log2FC_{H2K}$ was 1.71 in R1-DHS vs 3.28 in R1-MPRA (p=2.8e-16, Wilcoxon rank sum); and in K562-targeted designs the median $\log2FC_{H2K}$ was 0.04 vs -1.22 (p=2.5e-34, Wilcoxon rank sum). With this, we note the caveat that these sequences are derived from different data sources and design approaches. Indeed, the accessibility-based designs explicitly aim to minimize accessibility in cell types other than the two target cell

lines. No such constraint was used in the MPRA designs because activity measurements were only available for the two targets. Moreover, accessibility-based designs are regularized to be similar to native accessible elements, while the MPRA-based designs do not use such regularization and may in fact encourage the generation of sequences that look more extreme than those encountered in the training dataset.

R1-DHS sequences were approximately as diverse as R1-MPRA designs (**Supplementary Figure 2.4B**). As expected, R1-MPRA sequences are better predicted by M0 models than R1-DHS sequences (**Supplementary Figure 2.4C**), indicating complementary enhancer grammars are implemented by these libraries. This suggests a more complete model of enhancer activity could be achieved via integration of both sets of sequences.

2.3 Iterative retraining and design results in improved enhancer specificity and precision

We next asked whether design performance could be improved by retraining models with R1-MPRA and R1-DHS data. This was motivated by the observation that despite having high predicted specific activity or accessibility, not all designed sequences in the R1 libraries were found to be highly specific in our validation assay, suggesting the presence of regulatory information not captured by the initial models. Additionally, compared to the source

datasets, the R1 libraries contain a much higher fraction of positive (i.e. highly active) examples and thus carry the potential to improve model performance in the most relevant part of the design space. Practically, while it remains challenging to perform full-scale MPRAs in specific primary cell types at scale, testing up to a few thousand sequences in complex tissues using a single cell MPRA (scMPRA) assay is now within reach[29,30].

To understand if retraining with a relatively small dataset that is highly enriched for functional sequences can iteratively increase the success rate of enhancer designs, we trained a new set of models using two distinct strategies. First, we sought to improve performance of a CNN ensemble model by pretraining on R0-MPRA ("M0" models, **Appendix A**) and fine-tuning with additional training iterations on only R1-MPRA and R1-DHS ("M0+1" models, **Appendix A**). Second, we trained an identical CNN ensemble on both R1 datasets from scratch without R0-MPRA pretraining ("M1" models), to test whether a small dataset enriched in functional sequences can be sufficient for model-based design. We note a common alternative scenario would lack any initial MPRA data for model training and begin instead from DHS-based designs, therefore training models only on R1-DHS data. Nonetheless, our approach as implemented allows us to assess the value of pre-training. M1 models nearly matched M0+1 prediction performance on a held-out R1-MPRA test set, despite being trained on ~30× less data (**Supplementary Figure 2.5A**). While both M0+1 and M1 models perform worse on a held-out R0 test set compared with M0

models (**Supplementary Figure 2.5B**), we argue that any possible loss in generalizability is less relevant for our iterative design objective than a gain in prediction accuracy on sequences with higher specificity. We next separately applied Fast SeqProp to the M0+1 and M1 ensembles to collectively generate a new set of 690 sequences maximizing predicted target specificity (**Appendix A**). Library diversity was confirmed as above (**Supplementary Figure 2.5C,D**). As controls, we also included the top 5 enhancers in each cell line from R1-MPRA, as well as 200 randomly sampled R1-MPRA sequences. We refer to this set of designed and control sequences as R2 (**Table 2.2**).

We experimentally assayed R2 enhancers as before and with equivalently strong replicate correlation (**Supplementary Figure 2.5E**), observing dramatically higher median specificities of both HepG2-targeted and K562-targeted sequences in comparison to the R1 libraries (**Supplementary Figure 2.6A**). For HepG2-targeted sequences, median expression was 46.2-fold higher in HepG2 cells compared to K562 cells ($\log2FC_{H2K} = 5.53$); and for K562-targeted sequences median expression was 6.7-fold higher in the target cell line ($\log2FC_{H2K} = -2.74$) (**Figure 2.1E**). Evaluating the most specific sequences from each round, iterative improvement was observed in both cell types. For HepG2 designs the best $\log2FC_{H2K}$ increased from 4.87 (R0-MPRA) to 6.44/5.66 (R1-MPRA/R1-DHS) to 7.34 (R2), with significant improvement across all consecutive rounds. For K562 designs the best $\log2FC_{H2K}$ improved from -5.03 (R0-MPRA) to -5.47/-4.45 (R1-MPRA/R1-DHS) to -5.99

(R2), with significant improvement between R1 and R2 (**Figure 2.1F, bottom,** "lib" group).

In both cell types, improvements in specificity across design rounds were driven primarily but not exclusively by improvements in on-target cell-type activity. Comparing R0-MPRA to R1-MPRA to R2, in HepG2 designs, median log2FCHepG2 significantly increased (−1.85 to 1.86 to 3.23, p = 3e−191, 2e−37), while median log2FCK562 more weakly but still significantly decreased (−1.70 to −1.83 to −2.20, p = 5e−14, 5e−12). In K562 designs, median activity significantly increased in both cell types across rounds, but to a greater degree in log2FCK562 (−1.45 to −0.05 to 2.59, p = 7e−59, 4e−53) vs. log2FCHepG2 (−1.90 to −1.38 to −0.32, p = 3e−16, 1e−21).

One important constraint in a typical *in vivo* experimental setting is that of greatly reduced library throughput compared with cell line-based MPRAs[10]. To simulate such a scenario, we estimated the expected best enhancer performance for smaller library sizes (n=5, 20) using bootstrap sampling of R2 (**Appendix A**). Sampling in this way without *a priori* knowledge of enhancer activity allows us to estimate the relative success rate of each library at a range of measurement bottlenecks. Expected enhancer performance remained strong for R2 across all library sizes. Testing only 5 random HepG2 enhancers yielded on average a best $\log2FC_{H2K}$ of 6.31±0.57 (mean±standard deviation), exceeding all R0 enhancers (**Figure 2.1F, bottom**). We next sought to assess the likelihood for each library

to contain the most specific enhancer designed across all libraries when testing only a reduced number of enhancers. To this end, we conducted identical bootstrap simulations in R0 and R1 and found that for a library size of just 5 sequences, R2 enhancers outcompeted those in all other libraries 92.6% (HepG2) or 81.1% (K562) of the time. When randomly selecting 20 sequences, R2 enhancers outcompeted the other libraries 99.4% (HepG2) or 88.0% (K562) of the time; R1 enhancers were superior in the remaining cases, whereas R0-MPRA never outcompeted any of the others (**Figure 2.1F, top**). Thus, even for small sample sizes, synthetic enhancers greatly outperform those from the initial training data.

Additionally, we asked whether the improvement in model performance from iterative retraining would allow us to more precisely control enhancer specificity. Such an ability to tune target expression while maintaining specificity may be desirable in contexts where e.g. overexpression of a gene leads to deleterious and difficult to predict knock-on effects[55–57]. We used M0+1 models to design sequences targeting five different levels across a range of $\log2FC_{H2K}$ values and experimentally tested them in the R2 library, observing high accuracy ($r^2 = 0.93$) despite only designing twenty enhancers per intermediate target (**Figure 2.1G**). Although measured expression levels were generally slightly lower than predicted, we observe an almost perfect monotonic relationship between target and mean measured $\log2FC_{H2K}$.

2.4 Comparing sublibraries of synthetic enhancers suggests optimal design practices

Across all synthetic libraries we explored a range of model architectures, design algorithms, and design objectives. Here, we compare design performance under these approaches to identify best practices and develop recommendations for future work.

First, we found that design performance tracked model accuracy within each design round. Among R2 enhancers, M0+1 models achieved significantly better median specificity in both cell lines compared with M1 models (HepG2 median = 5.80 vs. 5.41, $p = 7.4e{-3}$; K562 median = −3.08 vs. −1.98, $p = 4.4e{-9}$; **Supplementary Figure 2.6B**). We observed a similar trend with R1-MPRA enhancers. As previously described, three model training variations were explored in this library, all implementing the same core architecture: single, boot, and ensemble. In HepG2-targeted designs, median specificity was higher for the most accurate ensemble models (4.27) compared with single models (3.93, $p = 1.2e{-2}$), which in turn was higher than the least accurate boot models (1.71, $p = 9.2e{-5}$), though no significant differences were observed between model types in K562-targeted designs (**Supplementary Figure 2.6C**). We further found that higher specificity M0+1-designed enhancers also exhibited greater Spearman correlation between predicted and measured $\log_2 FC_{H2K}$ compared with M0-designed enhancers (0.67 vs. 0.62 for HepG2 enhancers, 0.51 vs. 0.43 for K562 enhancers).

Second, we found that the performance of all sequence design methods used in R1-MPRA (DENs, Fast SeqProp, Simulated Annealing) was comparable (**Supplementary Figure 2.6D**). For HepG2 designs there was a significantly higher median $\log2FC_{H2K}$ in Fast SeqProp- vs DEN-generated enhancers (3.23 vs 1.71, p=4.5e-3), but otherwise no significant differences were observed between design methods. Taken together, these observations suggest that predictor performance was the "rate-limiting" factor in our design approach. This motivated the exclusive use of Fast SeqProp for the R2 designs: given similar performance, we chose the approach that was least computationally onerous.

Third, we initially hypothesized that enhancers designed to maximize specificity unbounded may result in unrealistic sequences with poor performance due to overfitting, but found this was not the case. In R2 we explicitly tested this by designing sequences that either maximized $|\log2FC_{H2K}|$ unbounded, or clipped to a threshold (1.1X the maximum value predicted by the M0+1 model on R1-MPRA sequences, **Table 2.2**). Sequences designed with the unbounded objective outperformed those designed with the clipped objective in both cell types (HepG2 median specificity = 5.80 vs 4.96, p=5.6e-9; K562 median specificity = -3.08 vs -2.59, p=8.0e-5) (**Supplementary Figure 2.6B**). These results suggest that our models are capable of extrapolating beyond the training data. That said, unbounded designs exhibited significantly higher overprediction of measured specificity

compared with clipped designs in both HepG2- (median residual $|\log_2 FC_{H2K}|$ = 2.75 vs. 1.79, p = 1.8e−10) and K562-targeted sequences (3.44 vs. 1.05, p = 7.2e−31).

Finally, we asked whether there is a trade-off between maximizing specificity and on-target activity. To investigate if enforcing low off-target activity limits on-target activity, we designed 20 enhancers each to maximize ("Max1") or minimize ("Min1") activity in only one cell type, regardless of the other (**Table 2.2**). Specificity was low for all Min1 designs, which had comparably low activity in both cell types (**Supplementary Figure 2.6E,F**). While target cell type activity was similarly high across all Max1 designs, the specificity of these sequences varied substantially (**Supplementary Figure 2.6E,G**), and this loss in specificity was not balanced out by a gain in absolute expression levels, as the most specific R2 enhancers obtained equivalent levels of target cell type activity as the most active Max1 design (**Supplementary Figure 2.6G-I**). Therefore, we conclude that designing for specificity did not limit the absolute expression levels of our enhancers.

2.5 Simulated experiments investigate the impact of training data on sequence design performance

To further explore optimal model training and sequence design practices, we performed several follow-up simulation analyses. First, we investigated the impact of the data source by fine-tuning M0 models on bootstraps of all R1-DHS data, all R1-MPRA data, or an

equal mixture of the two, and evaluating prediction performance on a held-out subset from R2 (**Appendix A**). We found that training with mixed or all-MPRA data yielded the best performance across the $\log2FC_{HepG2}$, $\log2FC_{K562}$, and $\log2FC_{H2K}$ prediction tasks, compared with training with all-DHS data (**Supplementary Figure 2.7A**). We then designed sequences *in silico* from each of these three model types and estimated their performance with an M2 ensemble obtained by fine-tuning the M0+1 models on crossfolds of R2 (**Appendix A**). Models trained with mixed data yielded enhancers with the highest median M2-predicted specificity in both cell types (**Supplementary Figure 2.7B**), indicating the value of sequence diversity in the training data—though we note this comparison is confounded by differences between the R1-MPRA and R1-DHS datasets, including sequence length and the distribution of enhancer activities.

As a follow-up, we investigated the impact of the proportion of strong enhancers in the training data. To remove possible confounders between datasets, we focused only on R1-MPRA data and trained models on bootstraps of data sampled from either weak enhancers, strong enhancers, or all enhancers (**Appendix A**). As expected, models trained with enhancers sampled from all R1-MPRA data performed best on all prediction tasks, though models trained on only strong data predicted $\log2FC_{H2K}$ with statistically equivalent correlation (**Supplementary Figure 2.7C**). We designed sequences *in silico* and evaluated them with M2 models as above, and we found that models trained on all or

exclusively strong enhancers yielded higher median M2-predicted specificity in both cell lines than models trained only on weak enhancers (**Supplementary Figure 2.7D**). While models trained on weak enhancers had slightly better prediction performance on $\log2FC_{HepG2}$ than models trained on strong enhancers, design performance was still inferior. This indicates that model performance in the sequence space most relevant to design can be prioritized over absolute predictive ability, which is further corroborated by the success of sequences designed from M1 models in the R2 library, despite these models being trained with a relatively small amount of data compared with M0+1 models.

Finally, we asked how many enhancer measurements are necessary to maximize the design performance of a retrained model. This is particularly relevant for applications where experimental throughput is limited, such as *in vivo* enhancer delivery. To this end, we fine-tuned M0 models on random samples of R1 data ranging in size from n = 100 to n = 1,750 sequences, training 5 ensembles per value of n. Evaluating these models on a test set of R2 measurements, we found that performance plateaued after n = 500 or n = 1,000 for all prediction tasks (**Supplementary Figure 2.7E**). For each ensemble, for each n value, we designed 100 sequences in silico as above and evaluated them with the M2 model. For HepG2 designs, models trained with $n \geq 200$ sequences produced enhancers with statistically equivalent predicted specificity, whereas for K562 designs, models trained with $n \geq 1{,}000$ yielded enhancers with equivalent predicted specificity

(**Supplementary Figure 2.7F**). This suggests that marginal improvements in prediction performance do not perfectly translate into improvements in design performance and that low amounts of synthetic enhancer measurements can be sufficient to retrain models with significantly increased design potential.

2.6 Designed sequences evolve a more compact TFBS motif grammar

To elucidate how synthetic enhancer grammar differed from its endogenous counterpart, we scanned all sequences for matches to the JASPAR2022 Core Vertebrate database of 137 TFBS motif clusters[58,59], discovering 67 (R0-MPRA), 50 (R0-DHS), 41 (R1-MPRA), 48 (R1-DHS), and 23 (R2) unique motif clusters in each respective library. We chose to represent each cluster by a metonymic TFBS motif or motif pair (the motifs in each cluster with highest representation in our sequence libraries). We estimated the total number of unique motifs in an equivalent number of sequences from each library via downsampling and found 44.06 +/- 2.38 (R0-MPRA), 35.60 +/- 1.71 (R1-MPRA), 27.21 +/- 1.08 (R2), 25.31 +/- 1.95 (R0-DHS), and 45.05 +/- 1.91 (R1-DHS) unique motif clusters in each respective library (**Appendix A**). Overall, we observed a narrowing of motif vocabulary across design iterations. Interestingly, the motif vocabulary size increases from R0-DHS to R1-DHS, possibly because the GAN was trained on a larger dataset beyond the HepG2- and K562-

specific DHSs in R0-DHS. Motif density increased significantly over successive design rounds, and was lower in DHS-based vs MPRA-based libraries, suggesting GAN designs are more similar to natural sequences than designs produced via the other methods (**Figure 2.2A**). Across all designed libraries motif density was positively associated with specificity. In R2, HepG2-targeted enhancers showed higher baseline activity at low motif densities, suggesting specificity was achieved with a smaller number of strongly activating motifs; whereas K562-targeted enhancers generally achieved specificity with a larger number of individually weaker motifs (**Figure 2.2B,C**).

Motif composition of the synthetic libraries significantly evolved over successive design iterations, and was markedly different from their genomic counterparts (**Figure 2.2D,E**). As an example, TP53 was modestly represented in the R0-MPRA dataset (41 TP53 motif hits across 29,891 sequences = 1.3e-3 average motifs/sequence) and entirely absent from the DHS-based libraries (R0-DHS and R1-DHS), but dramatically enriched in both R1-MPRA (712/1084 = 0.66 motifs/seq) and R2 (568/688 = 0.83 motifs/seq). Several motifs also progressively enriched in DHS- and/or MPRA-based designs include HNF4A/HNF4G, SPIB/ELK1, GATA2/GATA5, and NFE2/JUNB (**Figure 2.2E**). The HNF1A and GATA1::TAL1 motifs exhibited significant enrichment in R2 vs both R1-MPRA and R1-DHS, despite no or minimal enrichment in first round design libraries compared to their respective training libraries. These represent motifs with importance

emphasized in synthetic rather than genomic enhancers. Overall these results highlight that the models can identify and amplify strong motifs that are scarce in the training data.

Additionally, the synthetic enhancer libraries de-emphasized motifs deemed less relevant to cell type-specific activity. For instance, while the CTCF motif was significantly depleted in R1-MPRA vs R0-MPRA (3.0e-3 vs 0.21 motifs/seq) its cognate factor is not observed to have differential activity across these cell lines[60]. The SP/KLF motif was significantly depleted in R2 vs R1-MPRA and R1-DHS (3.3e-2 vs 0.76, 0.12 motifs/seq). This motif consists largely of C and/or G repeats, and belongs to a family of universal stripe factors (USFs), which are implicated in promiscuously cooperative activity across many cell types and with many TFBSs[61]. High SP/KLF prevalence in R0-MPRA sequences likely led to strong inclusion in R1-MPRA sequences, but R1 measurements discovered no strong association between these motifs and cell type-specific activity, resulting in R2 depletion.

2.7 Motif analysis elucidates sequence features of highly specific enhancers

To identify sequence features associated with strong specificity, we first compared motif occurrence in the most specific HepG2- and K562-targeted R2 designs ($|\log2FC_{H2K}| > 3$) (**Figure 2.2F**). HepG2 enhancers were dominated by TP53 motifs (660 motifs / 395 sequences = 1.67 motifs/seq average), followed by HNF4A/HNF4G (1.06), HNF1A (0.52),

HNF4A/NR2F1 (0.43), and TEF (0.29). K562 enhancers were dominated by NFE2/JUNB (247/168 = 1.47 motifs/seq), SPIB/ELK1 (1.38), GATA1::TAL1 (1.07), and GATA2/GATA5 (0.80) (**Figure 2.2F**). These findings largely agree with previous reports of cell line-specific TFs[37,51,60,62].

To dissect the effect of motif multiplicity, we focus on 6 motifs that occur with multiplicity > 1 in more than 50 R2 sequences. For 4 of these we observed a significant dose-response effect between multiplicity and specificity up to 2 motifs (TP53, HNF4A/HNF4G, SPIB/ELK1) or 3+ motifs (GATA1::TAL1), with the marginal effect of adding additional repeats progressively decreasing (**Figure 2.2G**). A significant *decrease* in specificity was observed between multiplicities 2 and 3+ for NFE2/JUNB and between 1 and 2 for GATA2/GATA5 (**Supplementary Figure 2.8A**). This effect is possibly explained by additional copies of these motifs decreasing the amount of sequence space available for other, stronger or non-redundant motifs which were not explicitly considered in this analysis, or by off-target activity increasing while on-target activity saturates.

As additional controls in R1-MPRA, we manually designed 62 homotypic enhancers consisting of 1-7 repeats of 9 known TFBS motifs from the CISBP2.0 database embedded in a background of fully random bases (**Appendix A**). Motifs were selected based on prior reporting of enhancing or repressing roles in HepG2 and K562[37,51], as well as enrichment analysis on the designed sequences. For TP53, GATA1::TAL1, and NFE2/JUNB, the

multiplicity effect was directly confirmed by these manually designed homotypic enhancers, which remove the confounding effect of other motif types (**Supplementary Figure 2.8B**). Saturation occurred at multiplicities 2, 3, and 3 for these motifs, respectively. For NFE2/JUNB, specificity decreased slightly at multiplicities > 3 because even though the on-target cell type activity continued to increase, the off-target cell type activity increased at a greater rate. This motif was deployed at high multiplicity in Max1 designs for both cell types, corroborating this pattern of differential but nonzero activity in both cell lines (**Supplementary Figure 2.8C,D**). One additional motif, HNF1A, exhibited a multiplicity effect (saturation at n=3) in the manual homotypic designs, which was not observed in the global R2 analysis due to low prevalence of high multiplicity sequences. Our findings on motif multiplicity are in good agreement with prior reporting[62,63]. For the remaining 5 motifs for which we tested manual homotypic enhancers, no multiplicity effect was observed (**Supplementary Figure 2.8B**).

Although repeats of the same motif were often associated with increased enhancer specificity, the best enhancers from each library and in each cell type contained more than one motif type. In fact, we observe a weak but significant linear association between motif diversity (number of unique motifs) and specificity in both HepG2 and K562-targeted sequences (**Supplementary Figure 2.8E**). Analysis of the top 5% of HepG2 enhancers in R2 reveals frequent co-occurrence of motifs identified by the differential enrichment

analysis, with the best enhancers primarily consisting of 1-2 TP53 motifs in combination with HNF4A/HNF4G and/or HNF1A (**Figure 2.2H**). The top 5% of K562 enhancers in R2 all contain 2-4 of the following motif types in varying combinations: NFE2/JUNB, SPIB/ELK1, GATA1::TAL1, and GATA2/GATA5 (**Figure 2.2H**). While illustrative, the ubiquity of these motif sets in designed enhancers precludes statistical association with enhancer activity. For instance, in HepG2 enhancers measured in R2, of the 334 enhancers that contain at least 1 instance of HNF4A, 326 also contain at least 1 instance of TP53 (98%), and only 20 out of 504 enhancers contain neither (0.04%). Accordingly, we calculated the significance of co-location using a metric described by Zhao et al.[61], but no significant p values were returned for any motif pairs. Subsequent ablation experiments probe this motif grammar more granularly.

As an alternative to enrichment analysis, we computed SHAP (Shapley Additive Explanations) values to interpret M0, M0+1, and M2 model predictions on our enhancer designs (**Appendix A**).[64] Comparing the relative importance of motifs across iterations of model training, we observe many similar motifs are highlighted by enrichment analysis in the designed sequences (**Supplementary Figure 2.8F**,**G**). This is to be expected, as design and interpretation both seek to identify input features that maximize model output. Among HepG2-associated motifs, we observe a moderate decrease in importance for many non-TP53 motifs from M0+1 to M2 models (**Supplementary Figure 2.8F**).

This suggests the potential limitation of further iteration in the HepG2 design space to improve either design or model performance, with subsequent designs likely converging toward reduced-diversity sequences increasingly relying on TP53 exploitation. For most K562-associated motifs, we observe largely similar importance according to M0+1 and M2 models (**Supplementary Figure 2.8G**). A small decrease in GATA1::TAL1 importance from M0+1 to M2 models may be attributed to the increase in GATA1::TAL1 instances in R2 (395) compared with the previous libraries (225 total instances in R0-MPRA, R1-MPRA, and R1-DHS), with the low sample size in R1 leading to overestimation of motif contribution in the R1-trained model. We also observe an increase in importance for STAT5A from M0+1 to M2, a motif that is not prominently highlighted by enrichment analysis, suggesting there may be capacity for even further iteration in the K562 design space beyond what is reported in this study.

2.8 Motif ablations in top R1 enhancers highlight different modes of motif interaction

In parallel with testing R2 sequences, we also performed perturbation experiments on the original R1 designs to better understand the underlying sequence grammar. Given the high motif density in synthetic enhancers and the possibility of redundant/non-causal motifs, we performed a feature ablation study on the top five R1 enhancers in each cell line. We ablated

all single motifs and all pairs by replacing them with randomized nucleotides (**Figure 2.3A, Appendix A**), then experimentally measured enhancer activity. For each single and double ablation, we calculated an ablation score (f_a) as the fraction of the original enhancer's activity reduced by the ablation (**Figure 2.3B, Appendix A**) The median single and double ablation scores were 0.13 and 0.30, respectively, validating that the identified motifs generally contributed to enhancer performance, with some ablations achieving much higher effects (max single $f_a = 0.92$, max double $f_a = 0.93$). Because the TP53 motif occurs so frequently in our designs, we were able to ask whether the motif position within the enhancer mattered. Indeed, variation in individual TP53 motif single ablations revealed that TP53 motifs closer to the promoter tended to have higher single ablation scores (**Figure 2.3C**).

Finally, we tested an additive model of motif interactions, comparing each double ablation score to the sum of the corresponding single ablation scores, and defining the difference as the deviation score, f_{dev} (**Figure 2.3D-E, Supplementary Figure 2.9, Appendix A**). The majority of motif interactions were additive ($f_{dev} \cong 0$), indicating motifs had independent activity/contribution to enhancer activity (e.g. R1 Seq 1099, **Figure 2.3F**). R1 Seq 433, which contained two TP53 and one HNF4A motifs, exhibited redundancy ($f_{dev} << 0$), wherein ablation of either of its two TP53 motifs resulted in only mild reduction of enhancer activity, but ablation of both TP53s knocked it out almost completely (**Figure 2.3G**). This sequence can be compared to R1 Seq 633, which consists

of three TP53 motifs, where no single or double ablation achieved more than moderate impact on enhancer activity (**Supplementary Figure 2.9F**). In contrast, from the R1 Seq 976 and 1041 ablations we uncovered an example of cooperativity ($f_{dev} >> 0$) between the GATA1::TAL1 and NFE2/JUNB motifs. Ablating either of these motifs reduces most of the enhancer activity, indicating both are necessary for the function of this enhancer (**Figure 2.3H, Supplementary Figure 2.9G**).

2.9 Optimizing non-motif sequence in top R1 enhancers can improve specificity

We next investigated whether we could improve enhancer specificity by redesigning sequences outside the identified motifs. We selected the top five R1 enhancers in each cell line and re-optimized all bases outside of an aligned motif with Fast SeqProp and the M0+1 model, generating 10 new designs per original enhancer. As controls, we generated 5 sequences where motif flanks were replaced with random bases. Re-optimizing non-motif sequence improved specificity for 5/5 HepG2 designs and 1/5 K562 designs (**Figure 2.3I, Supplementary Figure 2.10**). The majority of HepG2 designs were improved by embedding more non-TP53 motifs (e.g. **Figure 2.3J**), indicating initial designs reached saturation of TP53 multiplicity effects (**Figure 2.2G, Supplementary Figure 2.8E**). The most successful HepG2 redesign ($\log2FC_{H2K} = 7.35$) was one of the most specific

enhancers even compared to R2, representing a 1.33X improvement over the original sequence. Though as a comparison of design strategies, there was no statistical difference between the median specificities of *de novo* vs. re-optimized enhancers in either cell type. The successful K562 re-design originally consisted only of GATA1::TAL1 repeats, and was improved by adding either a SPIB/ELK1 or NFE2/JUNB motif, the latter completing the combinatorial pair implicated by ablation analysis (**Figure 2.3K**); the best redesign achieved ~1.73X improvement over the original enhancer. K562 and HepG2 redesigns were equally dissimilar from their original sequences on average (mean length-normalized edit distance = 0.348 vs. 0.345; **Supplementary Figure 2.11B**), and increase in specificity was not correlated with proportion of sequence changed (**Supplementary Figure 2.11C**). In general, K562 redesigns were likely less successful because of lower model accuracy on the K562 prediction task.

Randomizing non-motif sequence significantly, if modestly, reduced specificity for 9/10 tested enhancers compared with the original sequence (**Figure 2.3I**). We investigated the SHAP contribution of non-motif nucleotides in these 10 sequences using the M2 model and observed a greater mean predicted contribution of non-motif nucleotides in K562 enhancers vs. HepG2 enhancers (proportion of total predicted $|\log_2 FC_{H2K}|$ of sequence: 0.47 + −0.12 for K562 vs. 0.27 + −0.11 for HepG2, $p = 0.03$). Moreover, we found that nucleotides closer to a motif flank typically exhibited higher SHAP contribution than those

farther away, with this trend stronger in K562 vs. HepG2 enhancers (**Supplementary Figure 2.11D**).

Finally, we note that for R1 sequence 905, shuffling non-motif sequence resulted in minor but significantly improved specificity for 2/5 sequences, achieved via an increase in HepG2 activity. No close matches to known repressor motifs were discovered in the non-motif regions of this sequence, nor were our models able to accurately predict the ranking of these sequences, precluding the use of model interpretation to investigate this result. The strongest shuffled sequence was found to incorporate a match to an SP/KLF motif not present in the original sequence (**Supplementary Figure 2.10A**), though this motif is not implicated as driving enhancer activity in HepG2. We compared dinucleotide frequencies in the HepG2 enhancers that were subject to re-optimization and found that sequence 905 possessed a higher occurrence of the "GT" dinucleotide than the other sequences, which occurs 3 times in the aforementioned SP/KLF motif, but ultimately cannot establish a causal explanation.

2.10 Shorter enhancers display high specificity

To further refine our understanding of the core functional grammar of enhancers, we attempted to reduce enhancer size by designing 72-, 50-, and 25-bp-long sequences. In both cell lines, shorter enhancers down to 50 bp achieved statistically equivalent

specificity to those with a length of 145 bp by Wilcoxon rank sum (**Figure 2.4A**). In fact, the best 72 bp HepG2 enhancer had statistically equivalent specificity to that of the best 145 bp HepG2 enhancer (**Figure 2.4B**). A significant decrease was only observed at 25 bp (HepG2: 3.62 vs. 5.96, K562: 0.15 vs. −2.75).

Among HepG2 enhancers, activity was largely maintained at shorter lengths, with a significant increase in median $\log_2\text{FC}_{\text{HepG2}}$ only observed from 25 to 50 bp (0.54 vs. 3.06, $p =$ 5e−4; **Supplementary Figure 2.12A**). By contrast, among K562 enhancers, activity significantly increased with each length increment (median $\log_2\text{FC}_{\text{K562}} = -3.38$, 1.36, 2.93, and 3.63; **Supplementary Figure 2.12B**); however, median HepG2 activity of these K562 enhancers also increased with length up to 72 bp, hence the largely consistent specificity levels in these enhancers.

In HepG2 designs, 25-bp enhancers consisted of a single TP53 motif, which was sufficient to drive moderately specific activity. Conversely, 25 bp K562 designs consisted of a single GATA1::TAL1 motif, which could not drive enhancer activity alone, in agreement with the ablation analysis. As sequence length is increased, HNF4A/HNF4G and HNF1A motifs are the next motifs to be prioritized after TP53 in HepG2 enhancers, whereas NFE2/JUNB and SPIB/ELK1 are prioritized after GATA1::TAL1 for K562 (**Figure 2.4C**).

2.11 An scMPRA enables characterization of enhancer activity at the single-cell level

Finally, we sought to characterize cell-to-cell heterogeneity in enhancer activity and understand how it may depend on cell state and TF expression. To this end, we performed an scMPRA following an approach similar to Zhao et al.[29] with some modifications (**Figure 2.5A**, **Supplementary Figure 2.13A–S13F**; **Appendix A**). Briefly, we transfected our R1-MPRA library into a 1:1 mixture of K562 and HepG2 cells, followed by combinatorial split-pool indexing to tag mRNAs with cell barcodes[65] and dual sequencing of transcriptome and MPRA library cDNAs (**Supplementary Figure 2.13A**). We recovered 1,343 enhancers across 10,640 cells at a median of 4 unique enhancers per cell (**Supplementary Figure 2.13B**,**C**). We verified that our previous bulk MPRA measurements could be accurately reconstructed from pseudobulk analysis of scMPRA data (Pearson $R = 0.91$; **Figure 2.5B**). Subsampling cells shows that such high correlation is robust down to a pseudobulk size of ~1,000 cells (**Supplementary Figure 2.13D**). Thresholding by minimum expression levels in each subsample maintains the high correlation, indicating that the sampling noise is due to dropout events attributable, for example, to transfection efficiencies or mRNA capture sensitivities. To understand whether our synthetic enhancers show unexpected cell state specificity, we examined cell cycle phases and a known major "stem-like" differentiation/proliferation state within K562 cells marked by CD24

expression[66]. Examining differential pseudobulk enhancer activities, we found that our synthetic enhancers are largely robust across these substates (**Supplementary Figure 2.13G–O**).

If the TFBS motifs in our synthetic enhancers are indeed causal, we would expect to observe a relationship between individual TF expression and enhancer activities. To investigate this hypothesis, we binned cells by the expression levels of a given TF and obtained per-bin pseudobulk MPRA reporter expression levels for each enhancer. We then calculated a correlation value for every TF-enhancer pairing across all bins. We generally observe positive correlations between pairs of enhancers and TFs if the enhancer contains a motif corresponding to the TF in the pairing. By contrast, TF-enhancer pairs where the enhancer does not contain the corresponding TFBS show generally lower and negative correlation coefficients (**Figure 2.5C** and **Supplementary Figure 2.14**). For example, R1 Seq 976 is K562-specific and contains one GATA1::TAL1 binding site (**Supplementary Figure 2.15A**). The scMPRA recapitulates the expected GATA1 expression pattern (**Figure 2.5D**,**E**), along with the K562-specific activity of this enhancer (**Figure 2.5F**). We find a positive correlation between reporter levels from this enhancer and GATA1 expression across the bins (**Figure 2.5G**). We observed similar positive correlations for other GATA1::TAL1 motif-containing enhancers but not when considering other enhancers without GATA1::TAL1 motif against GATA1 expression

(**Figure 2.5H**). Similarly, R1 Seq 369 is HepG2-specific and contains a TP53 TFBS (**Supplementary Figure 2.15B**). Its activity is correlated with TP53 expression at the single-cell level, as are the activities of other TP53-containing enhancers (**Supplementary Figure 2.15C–G**). Lastly, enhancers whose cell-type-specific activities were shown to be strongly dependent on individual TFBS motifs according to their ablation scores in our previous perturbation experiments also display stronger correlations with the corresponding TF expression levels (**Supplementary Figure 2.15H**). Thus, cell-type-specific enhancer activities are predicted by levels of TFs that are differentially expressed across cells.

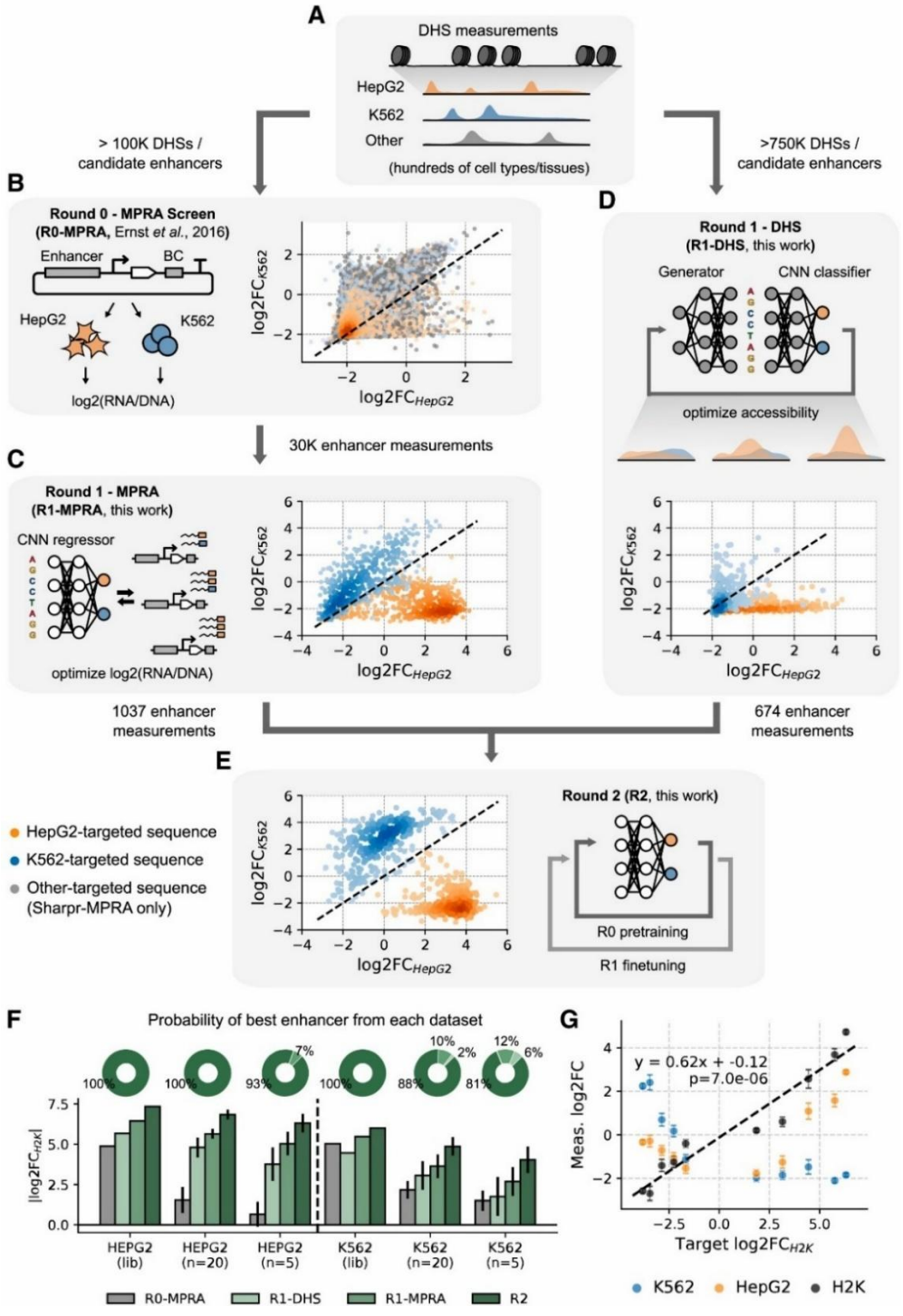


Figure 2.1. Multi-round model-based enhancer design results in cell-type-specific sequences with improved performance.

A-E, Overview of iterative design workflow. Chromatin accessibility measurements are first used to identify putative regulatory elements in the genome (**A**). The Sharpr-MPRA[51] assayed >100k of such elements in HepG2 and K562 via MPRA, which we filtered for read depth and enhancer activity (R0-MPRA, **B**). We designed and tested sequences using deep learning models trained on these MPRA data (R1-MPRA, **C**) or directly on accessibility (R1-DHS, **D**), and used the results to improve our models and further increase performance (R2, **E**). Individual points in scatterplots (C-E) correspond to measured enhancer activity in HepG2 (x-axis) vs K562 (y-axis). Dashed line: y = x. For Sharpr-MPRA sequences in **B**, target cell type is matched to the cell type of the accessibility dataset from which it was derived; cell types "Other" than HepG2 and K562 are colored gray (HUVEC and H1-hESC, see **Appendix A**). **F,** Activity and origin of the most specific enhancers across design rounds, as a function of library size. Top row, bootstrap-estimated probability that the most specific enhancer across all tested libraries will come from a given design round, when each entire library is considered (lib), or with simulated sizes of n=20 and n=5. Bottom row: $|\log2FC_{H2K}|$ of the most specific enhancer in each design round, library size, and target cell type. For "lib" columns, bar heights correspond to individual measurements of the best performing sequences; for "n=20" and "n=5" bar height and error bars are the mean and standard deviation across 10,000 bootstrap simulations. **G**, Performance of enhancers designed for intermediate target specificities. Markers and error bars are the mean and standard deviation across 20 (intermediate targets) or 110 (largest HepG2 or K562 target values) sequences designed for a given target value. The line of best fit is plotted (dashed black).

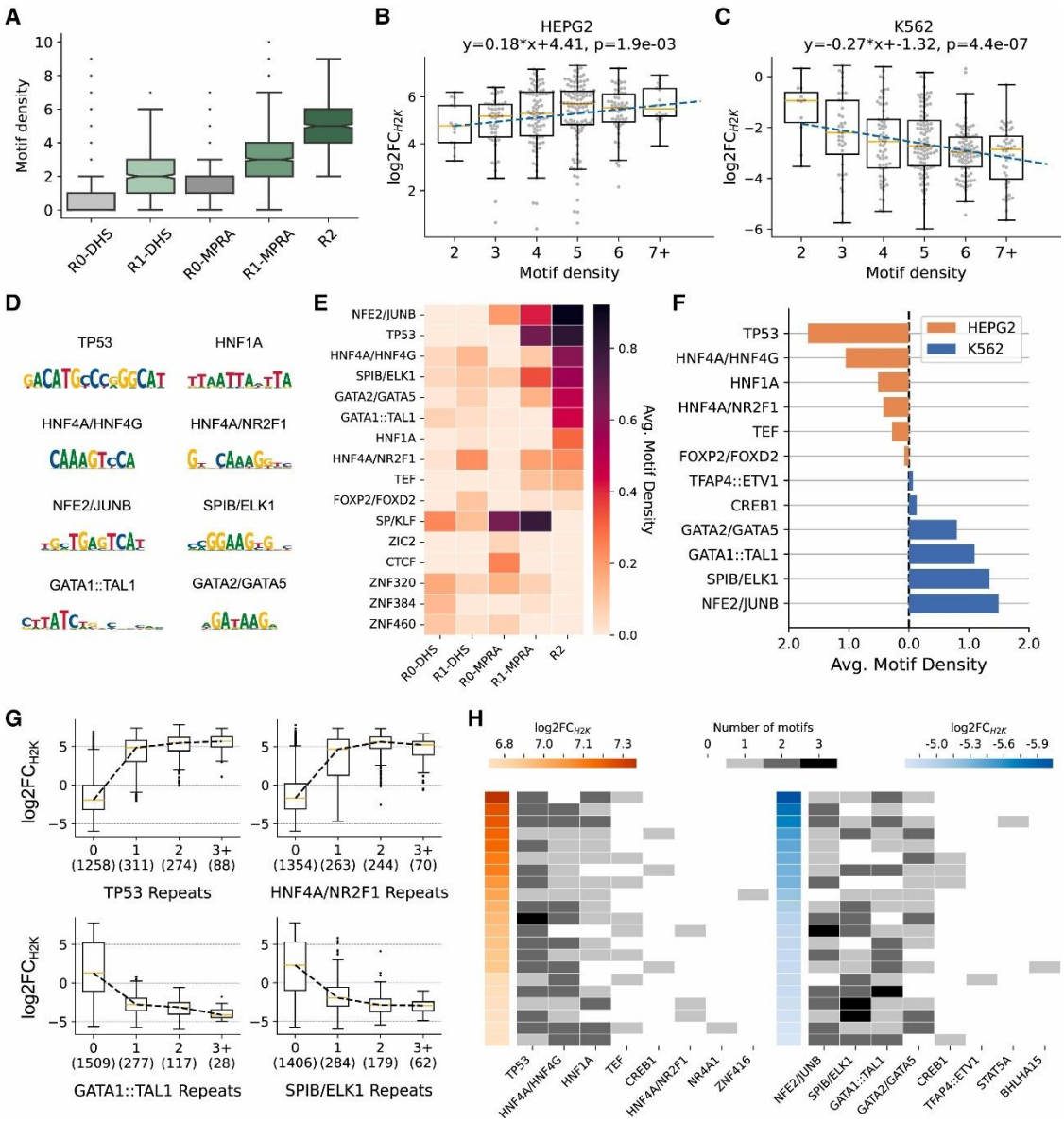


Figure 2.2. Synthetic enhancers exhibit more compressed TFBS motif grammar than natural enhancers.

A, Motif density (number of discrete motif matches per sequence, **Appendix A**) of synthetic enhancers exceeds the motif density of genome-derived enhancers and increases across consecutive rounds of design. **B-C,** In R2 designs, motif density is positively correlated with enhancer specificity. Each marker corresponds to an enhancer targeted to HepG2 (**B**) or K562 (**C**). Enhancers with motif density >=7 grouped together due to low sample size. No R2 enhancers had motif density < 2. Line of best fit plotted in blue. **D,** PWMs of motifs associated with cell type-specific enhancer activity. Motifs were extracted

from the most specific R1 ($|\log2FC_{H2K}|>2$) and R2 ($|\log2FC_{H2K}|>3$) sequences using STREME[67]. **E,** Average motif density across sequences within each design round for a subset of motifs with high enrichment in at least one library. **F,** Average motif density in the most cell type-specific R2 sequences ($\log2C_{H2K} > 3$, left; $\log2C_{H2K} < -3$, right) for the most differentially enriched motifs. **G,** Boxplots of $\log2FC_{H2K}$ measurements for R2 enhancers grouped by motif multiplicity (number of discrete instances of a given motif type in the same sequence), for a selection of motifs that show an association between multiplicity and increased specificity. Sequences with multiplicity >=3 are grouped together due to low sample sizes. For each x-value, the number of enhancers is reported in parentheses. **H**, Motif content of top 5% most-specific R2 enhancers for HepG2 designs (left), K562 designs (right). Each row represents a single enhancer sequence (sorted by descending specificity), columns indicate different motifs, cell color indicates motif multiplicity. $\log2FC_{H2K}$ plotted as a colored cell alongside corresponding sequence rows.

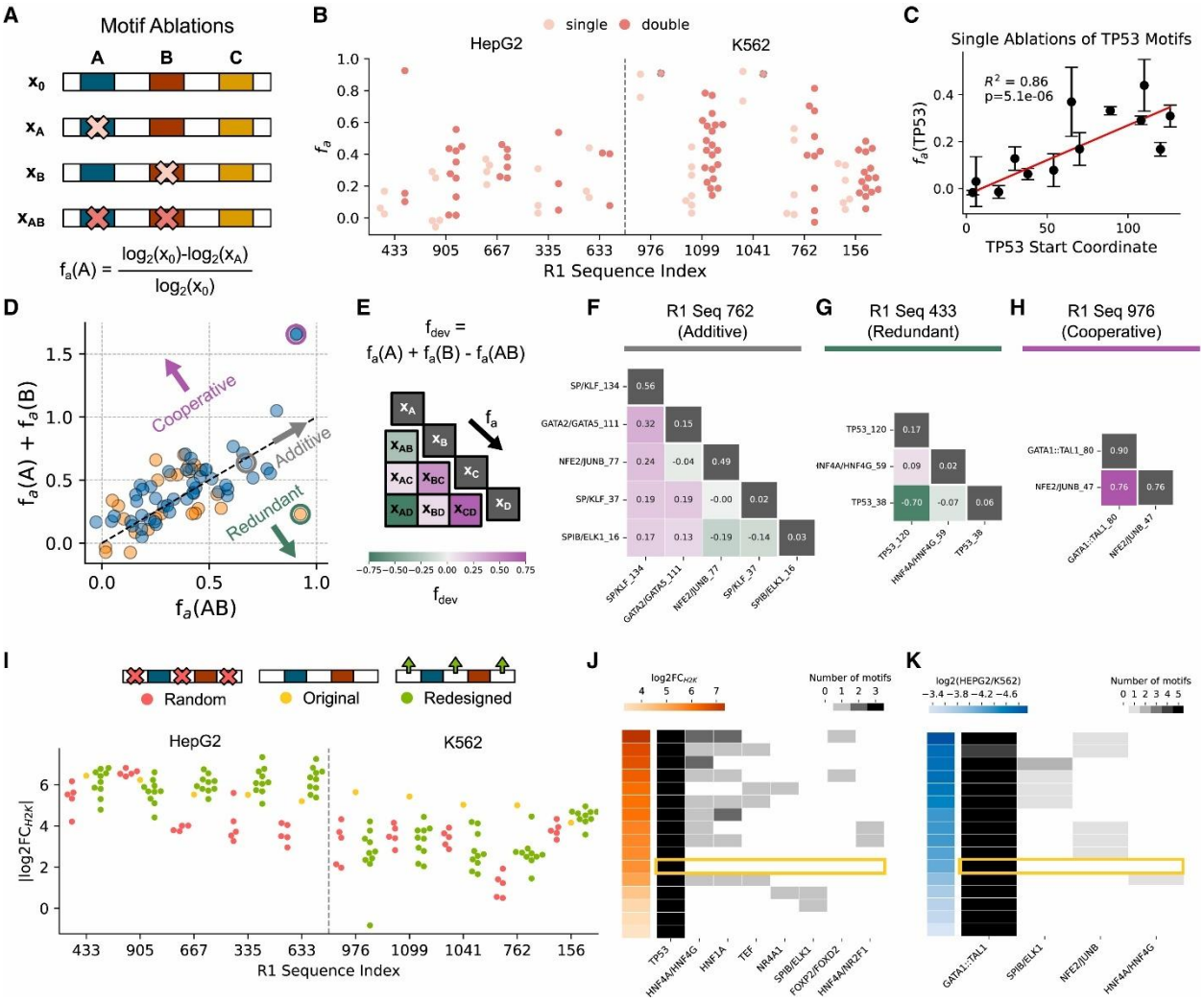


Figure 2.3. Perturbations of synthetic enhancers indicate causal features of cell type-specific activity.

A, Overview of single and double motif ablation procedure. Top 5 R1 enhancers in each cell line were perturbed by ablating all individual motifs and motif pairs via replacement with randomized nucleotides. Three randomizations were performed per ablation. Ablation scores were calculated for each single and double motif ablation as the reduction in specificity relative to the unmodified sequence. **B,** Ablation scores for single and double ablations for each R1 sequence. Each point is the mean f_a from the three randomizations for a given ablation. For Sequences 976 and 1041 double ablations were not directly measured as these would remove all motifs from the enhancer, and it was assumed this would remove the majority of enhancer activity. **C,** f_a regressed on motif position (distance between motif start and sequence start, 0-144) for the eleven single ablations of a TP53 motif. Markers and error bars show the mean and standard deviation from the

three randomizations of each ablation. Inverse variance-weighted least-squares regression line shown in red. **D**, An additive model of motif interaction ($f_a(A) + f_a(B)$, y-axis) is plotted against the measured double ablation score ($f_a(AB)$, x-axis). Each point represents an ablated motif pair, and is colored by the target cell type of the original enhancer. For R1 Sequences 976 and 1041 the $f_a(AB)$ was not measured directly, so we estimate with model predictions from M0+1 models fine-tuned on R2 data (**Appendix A**). Discrepancy between additive model and measurement indicate different modes of motif interaction, as annotated in plot. **E**, Schematic of deviation score calculation and heatmap. Single ablation scores were plotted along the heatmap diagonal, deviation scores for each motif pair were plotted on the off-diagonals. Each heatmap shows deviation scores for every motif pair in a single sequence. Motifs are named by motif type and starting position in the sequence. **F,** Deviation map for R1 Seq 762; most motifs interact additively, e.g. GATA1/GATA5_111 and NFE2/JUNB_77. **G,** Deviation map for R1 Seq 433; contains prominent example of redundancy (TP53_38 and TP53_120). **H,** Deviation map for R1 Seq 976; contains prominent example of cooperativity (GATA1::TAL1_80 and NFE2/JUNB_47). For the double ablation condition, the deviation score was obtained from model predictions. **I**, Top 5 R1 enhancers in each cell line were perturbed by preserving all FIMO-annotated motifs and either 1) randomly shuffling the non-motif sequence, or 2) re-optimizing the non-motif sequence with Fast SeqProp. Swarmplot shows measured $|\log2FC_{H2K}|$ for these sequence perturbations. **J,K** Motif composition of non-motif redesigns for R1 sequence 335 (**J**), 156 (**K**). Yellow box indicates original sequence, all other rows are redesigns (both Fast SeqProp and shuffled), sorted by descending specificity.

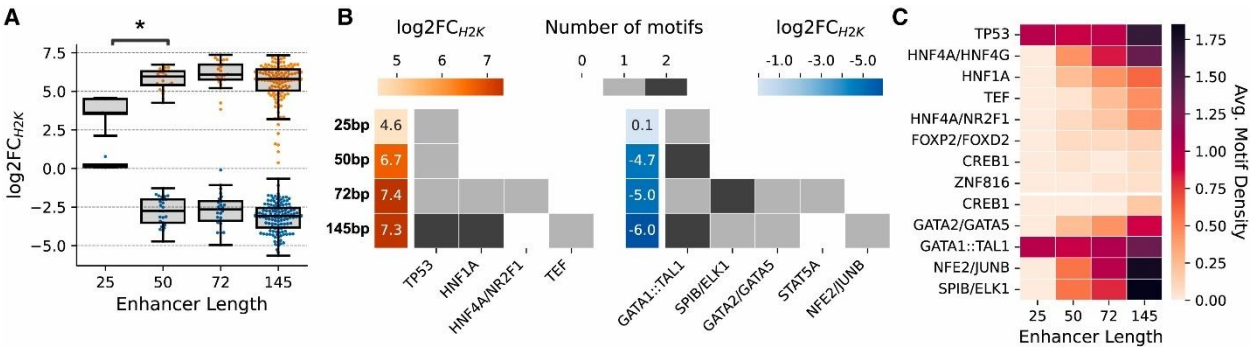


Figure 2.4. Shorter enhancers retain high specificity.

A, Boxswarm plots show log2FC$_{H2K}$ for R2 enhancer designs grouped by sequence length. Asterisks indicate significant difference in median specificity by Wilcoxon rank-sum ($p < 0.05$). **B**, Motif content and log2FC$_{H2K}$ of most specific enhancer of each sequence length for HepG2 designs (left) and K562 designs (right). Each row corresponds to a different enhancer length, denoted by text annotation to the left of plot. log2FC$_{H2K}$ colormaps computed separately within each cell type. **C,** Average motif density of R2 enhancers as a function of sequence length shown for motifs with average motif density >=0.05 at every sequence length. Horizontal white line separates HepG2- from K562-targeted enhancers, average motif density calculated separately within each cell type. Motifs sorted by enrichment in 145bp enhancers (descending in HepG2, ascending in K562).

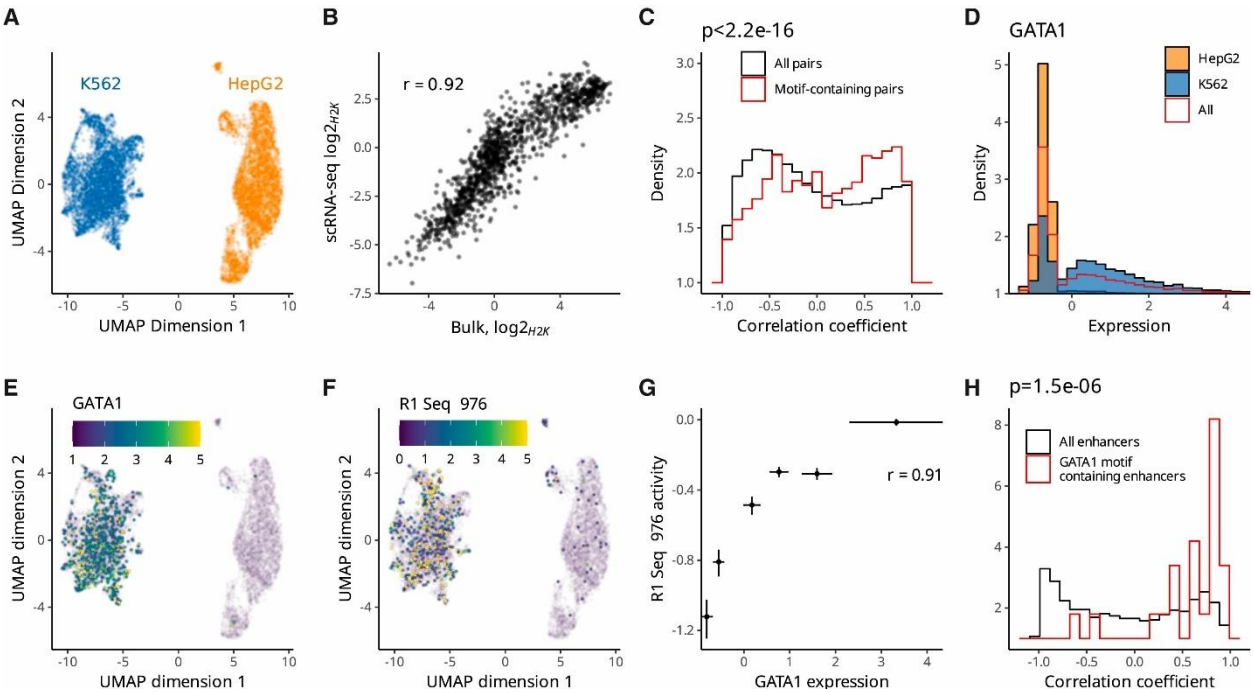


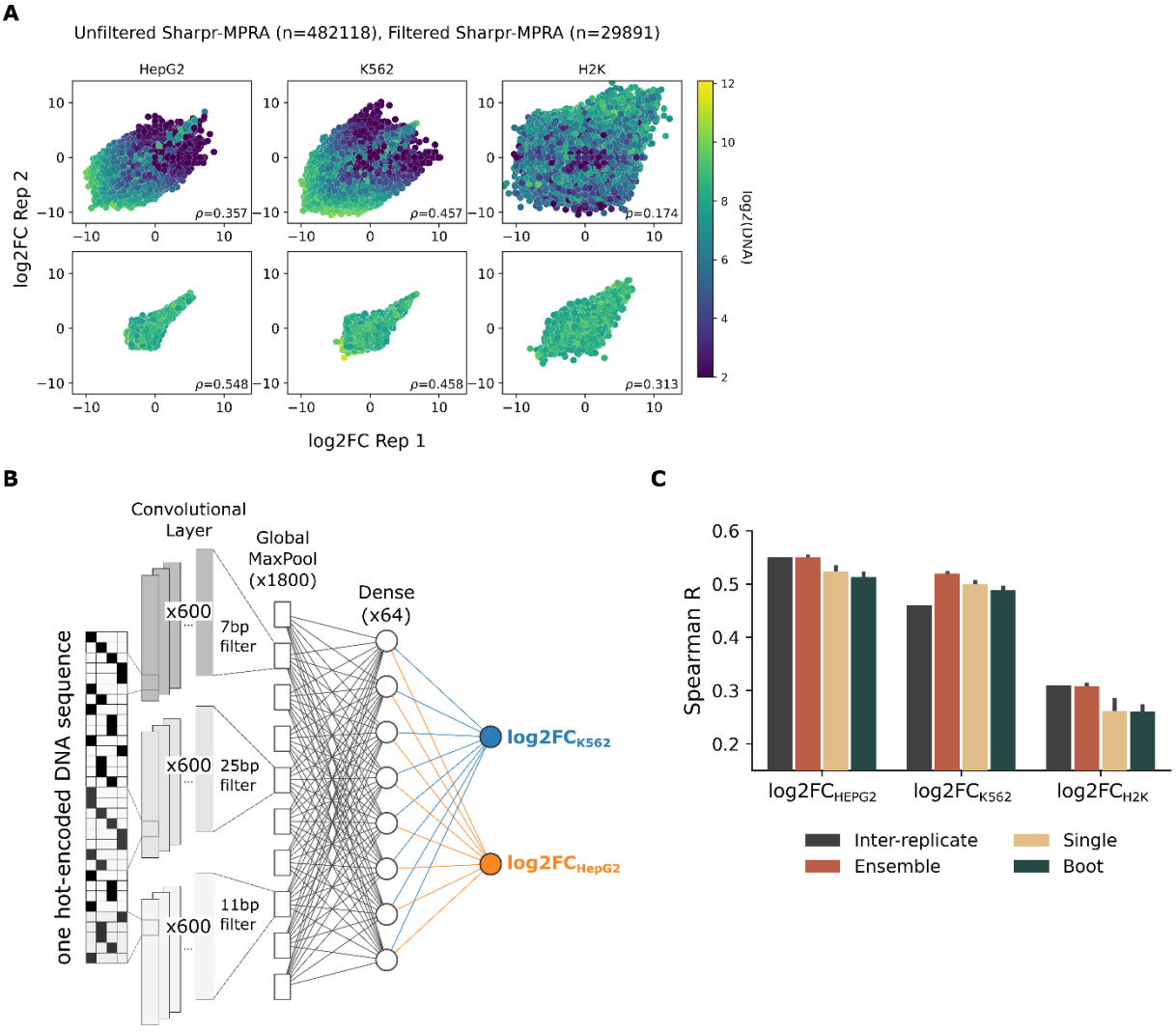
Figure 2.5. Synthetic enhancer activity confirmed at the single cell level.

A, Uniform Manifold Approximation and Projection (UMAP) projection of the single-cell transcriptomes (GEO: GSE269037). Clusters corresponding to K562 and HepG2 cell types are colored in blue and orange, respectively. **B,** Comparison of cell-type-specific activity of synthetic enhancers obtained by bulk (x axis, same as **Figure 2.1B**) vs. single-cell analysis (y axis), where the activity was calculated via pseudobulk aggregation of the clusters shown in (**A**). **C,** Distribution of all possible pairwise correlations between TFs and enhancer expression across pseudobulk of single cells binned by TF expression values. See (**H**) for an example of one such pair. In red are pairings where the enhancer contains the DNA sequence motif for the TF. p value from two-sided Kolmogorov-Smirnov test. **D,** Expression levels of GATA1 TF across all cells in red, HepG2 in orange, and K562 in blue. **E,** Expression levels of GATA1 TFs across single cells atop the UMAP projection of the transcriptomes. Color scale indicates Pearson residuals. **F**, Activities of R1 Seq 976, containing the GATA1 TF motif, across single cells atop the UMAP projection of the transcriptomes. Color scale indicates log2 UMI counts. **G,** Pseudobulk correlation of GATA1 TF expression levels and R1 Seq 976. The pseudobulks are binned by GATA1 expression levels. Vertical error bars represent bootstrap standard errors of enhancer activities. Horizontal error bars represent standard deviation of the TF expression levels across the cells in each pseudobulk bin. **H,** Distribution of correlations between GATA1 expression vs. all enhancers across pseudobulk of single cells binned by GATA1 expression values. In red are pairings where

the enhancer contains the DNA sequence motif for GATA1. p value derived from two-sided Kolmogorov-Smirnov test. Also see **Supplementary Figure 2.10**.

CHAPTER 2. SUPPLEMENTAL INFORMATION

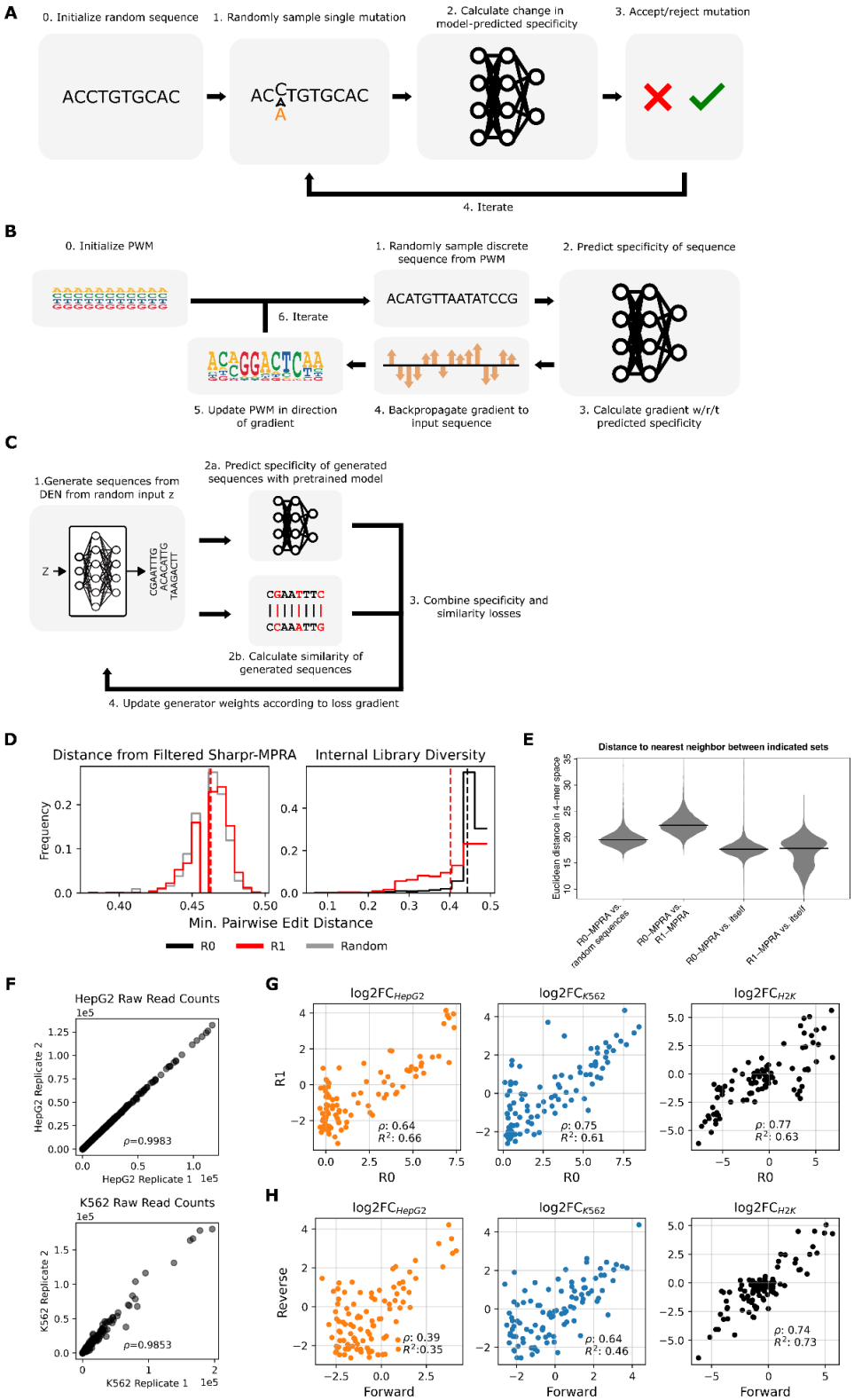
2.1 Chapter 2 – Supplemental Figures



Supplementary Figure 2.1. R0-MPRA processing and R1-MPRA generation

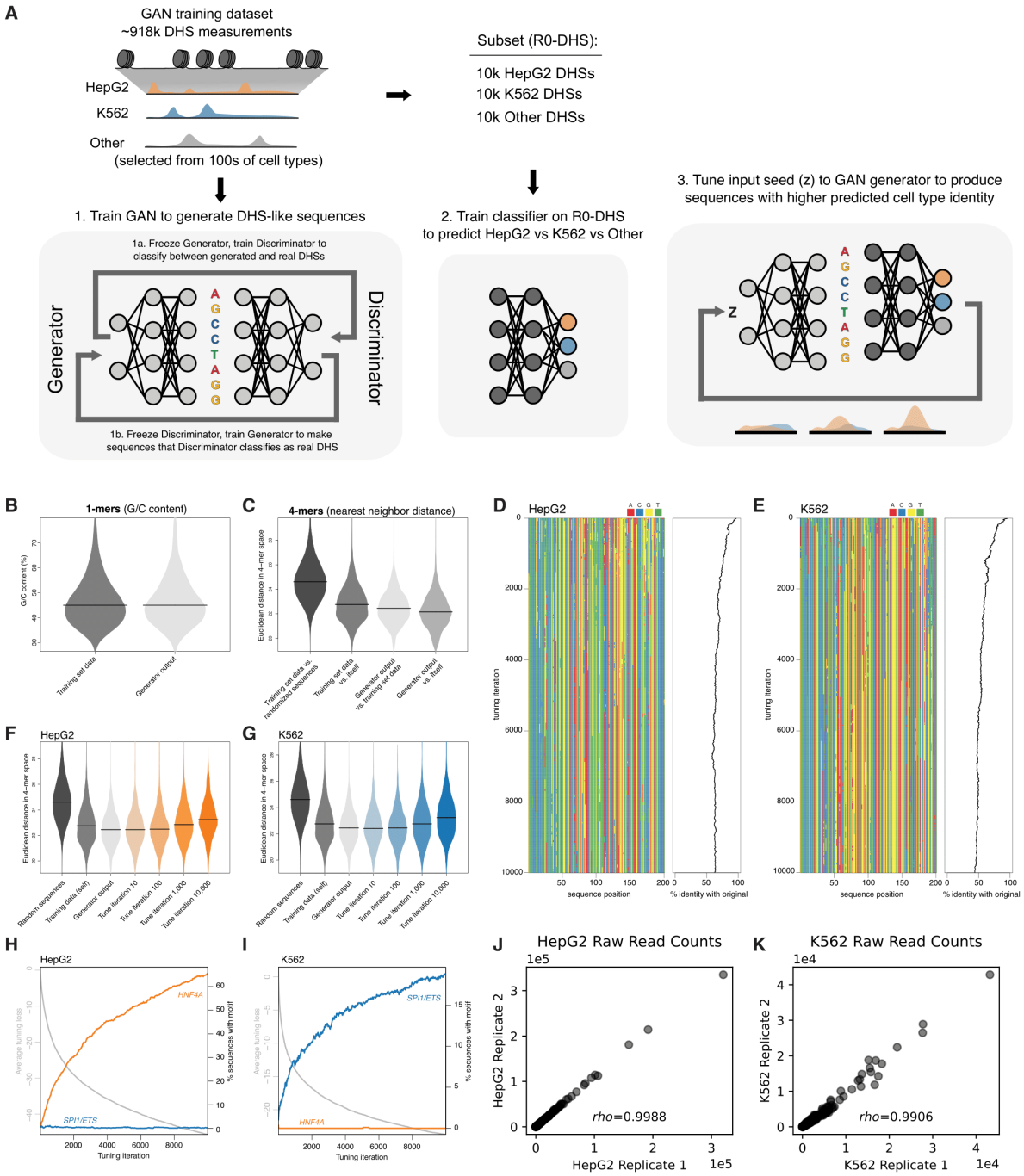
A, Inter-replicate correlation (Spearman, indicated with the letter "ρ") for $\text{log2FC}_{\text{HepG2}}$, $\text{log2FC}_{\text{K562}}$, and $\text{log2FC}_{\text{H2K}}$ in unfiltered Sharpr-MPRA (top row) vs. filtered Sharpr-MPRA (bottom row, R0), using total read count normalized measurements. Points are colored by log2(DNA count) . **B**, Architecture of multitask CNN model trained on R0-MPRA data and used to design R1-MPRA sequences. **C,** Prediction-measurement correlation and inter-replication correlation compared on $\text{log2FC}_{\text{HepG2}}$, $\text{log2FC}_{\text{K562}}$, and

$\log2FC_{H2K}$ held-out test set measurements for Single, Boot, and Ensemble model types trained on R0-MPRA sequences (10 models each). Mean and standard deviation shown.



Supplementary Figure 2.2. R1-MPRA sequence and measurement characteristics

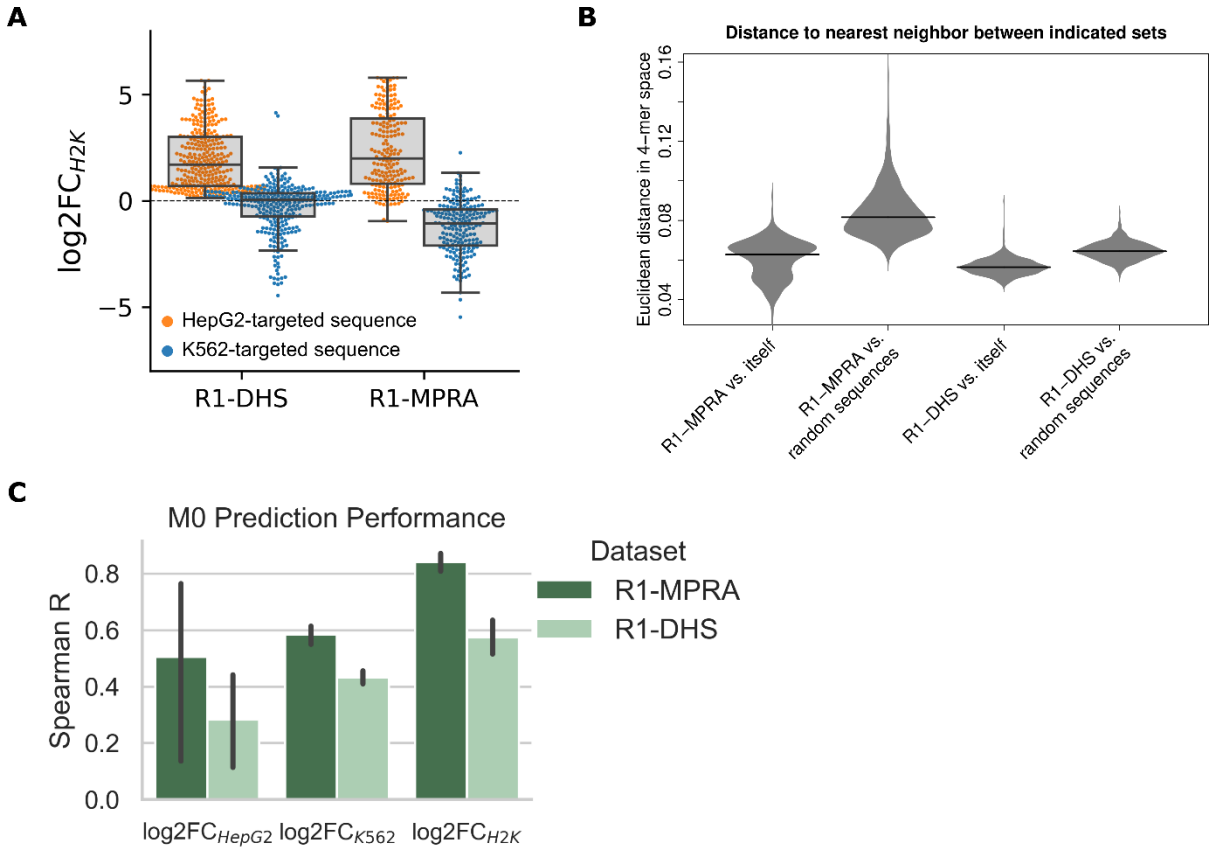
A-C, Schematic of design algorithms used in this study (**Appendix A**). **A**, Simulated annealing is a stochastic search algorithm that begins with a random sequence, randomly samples mutations, and probabilistically accepts or rejects the mutations based on improvement in model predictions. **B**, Fast SeqProp optimizes an entire sequence simultaneously in the direction of the gradient of the model predictions (Linder and Seelig, 2021). **C**, DENs are generative neural networks trained to produce sequences with high model predictions while penalizing the similarity of generated sequences to encourage exploration of the full design space (Linder et al., 2020). **D**, Left: Minimum pairwise edit distance between each R1-MPRA sequence and R0-MPRA sequences (with overlapping sequences tiled from the same accessibility window pruned to a single representative sequence), as well as between the pruned R0-MPRA sequences and a Random library generated by randomly shuffling the nucleotides of the pruned R0-MPRA sequences Right: Minimum pairwise edit distance between all R1-MPRA sequences compared to within the pruned R0-MPRA sequences. Means plotted as vertical dashed lines. **E**, Euclidean distances in 4-mer space between indicated pairs of sequence sets (**Appendix A**). Horizontal lines indicate medians. **F,** Inter-replicate correlation for mRNA raw read counts in each cell type for R1-MPRA library. **G,** Correlation between R0-MPRA (reprocessed from measurements in Ernst et al., 2016) and R1-MPRA (our) measurements of the 100 control sequences shared between libraries. **H**, Correlation between measurements of forward and reverse complement of 100 control sequences in R1-MPRA.



Supplementary Figure 2.3. R1-MPRA sequence and measurement characteristics

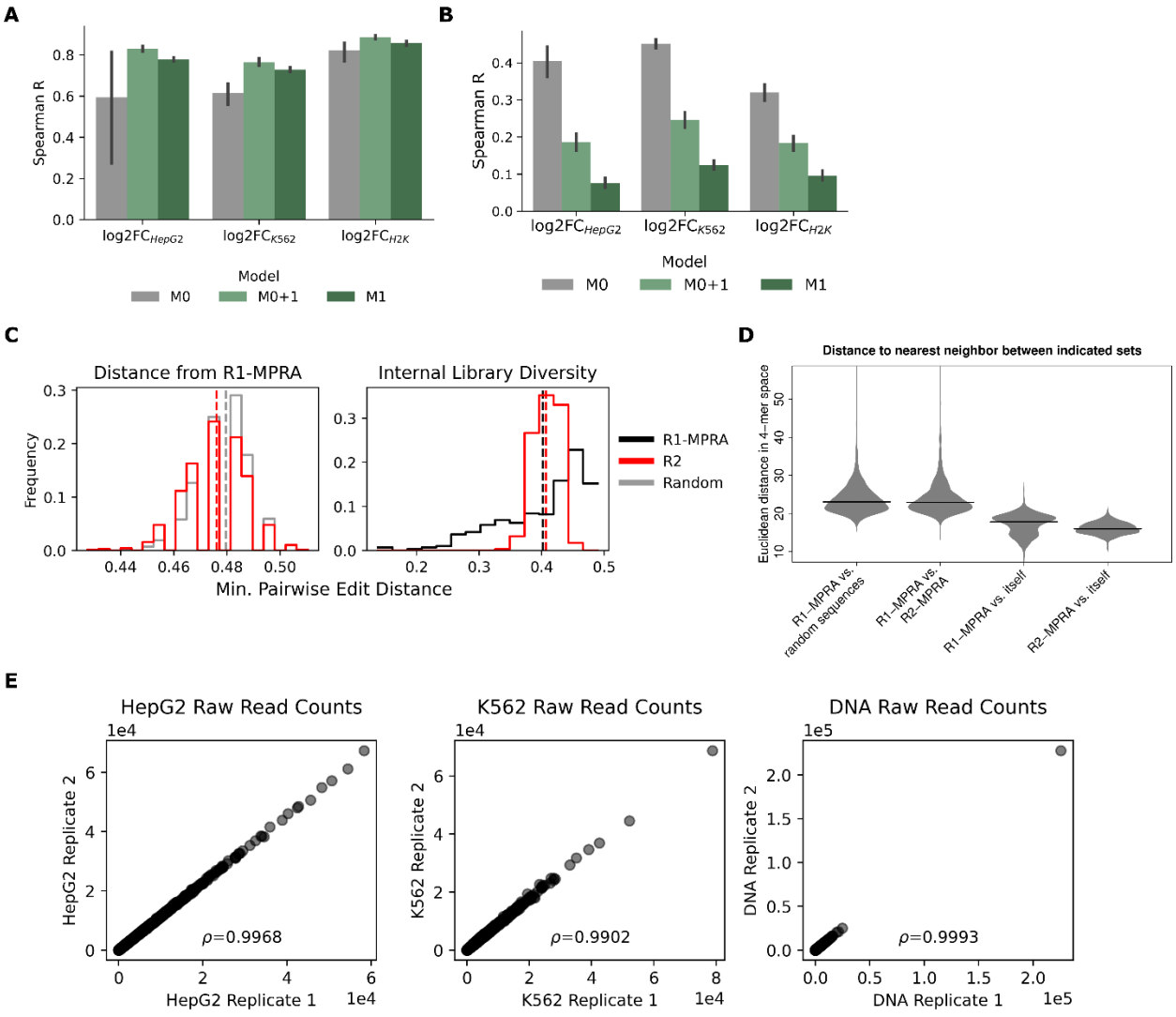
A, Overview of GAN training and R1-DHS design pipeline (**Appendix A**). **B-C,** Characteristics of endogenous accessible sequence elements used for training the GAN vs.

synthetic GAN-generated sequences w.r.t. G/C content percentage (**B**) and general 4-mer sequence content (**C**), the latter reporting Euclidean distances between indicated pairs of sequence sets. **D-E**, Example sequence tuning of a single GAN-generated sequence towards a HepG2 (**D**) and K562 (**E**) cellular context. Colormap shows the tuning process, starting from the same generated sequence, and slowly converging to cell type-specific sequences, while retaining a large fraction of the original sequence identity. **F-G,** Sequence characteristics during tuning iterations for HepG2 (**F**) and K562 (**G**), as shown by Euclidean distance in 4-mer space for indicated sequence sets, training data. **H-I,** Average tuning loss across tuning iterations versus percentage of sequences containing select transcription factor motifs. Shown are HNF4A (orange) and SPI1/ETS (blue), the former being enriched in HepG2 cells (**H**) and the latter in K562 cells (**I**). **J-K,** Inter-replicate correlation for mRNA raw read counts for the R1-DHS library in HepG2 (**J**) and K562 (**K**).



Supplementary Figure 2.4. Comparing MPRA vs DHS model and design characteristics

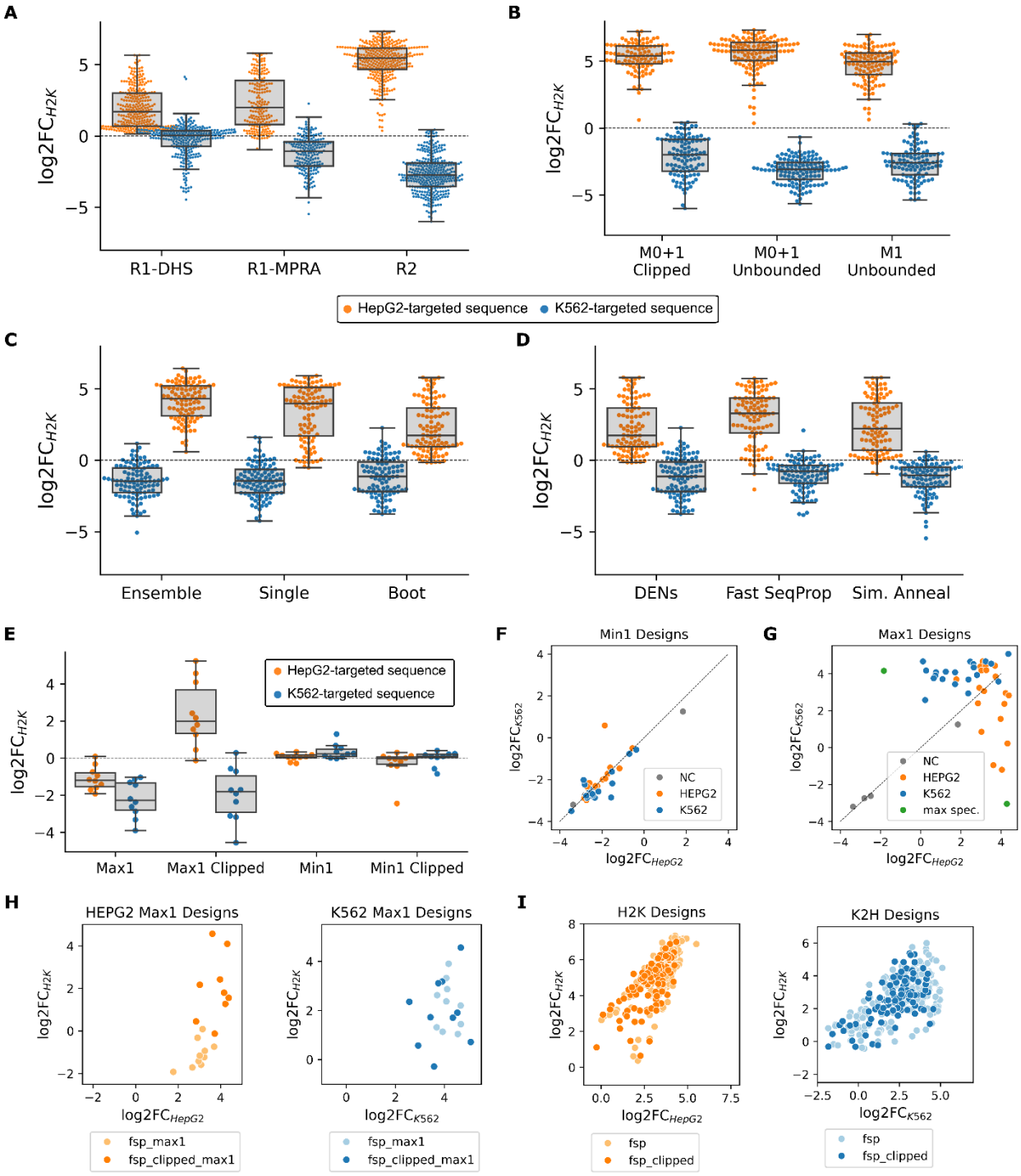
A, Comparing enhancers designed to maximize specificity between R1-MPRA and R1-DHS. **B**, Minimum pairwise Euclidean distances in 4-mer space within indicated sequence sets. Distances normalized by 4-mer vector norm (i.e. total number of 4-mers: 142 in R1-MPRA, 197 in R1-DHS). Horizontal lines indicate medians. **C**, Prediction-measurement correlation on indicated datasets, generating predictions from the 9 models of the M0 ensemble. Mean and 95% CI shown.



Supplementary Figure 2.5. R2 generation, sequence and measurement characteristics

A, Prediction performance on held-out test set of R1-MPRA data for models trained on R0-MPRA data only ("M0"), M0 models finetuned on R1-MPRA data ("M0+1"), and models trained only on R1-MPRA data ("M1"). Mean and 95% CI shown. **B**, Prediction performance on held-out test set of R0-MPRA data for the same models. **C,** Minimum pairwise edit distance between each R2 sequence and all R1-MPRA sequences, as well as between all R1-MPRA sequences and a Random library generated by randomly shuffling the nucleotides of the R1-MPRA sequences (left). Minimum pairwise edit distance between all R2 sequences compared to within all R1-MPRA sequences (right). Means plotted as dashed lines. **D**, Euclidean distances in 4-mer space between indicated pairs of

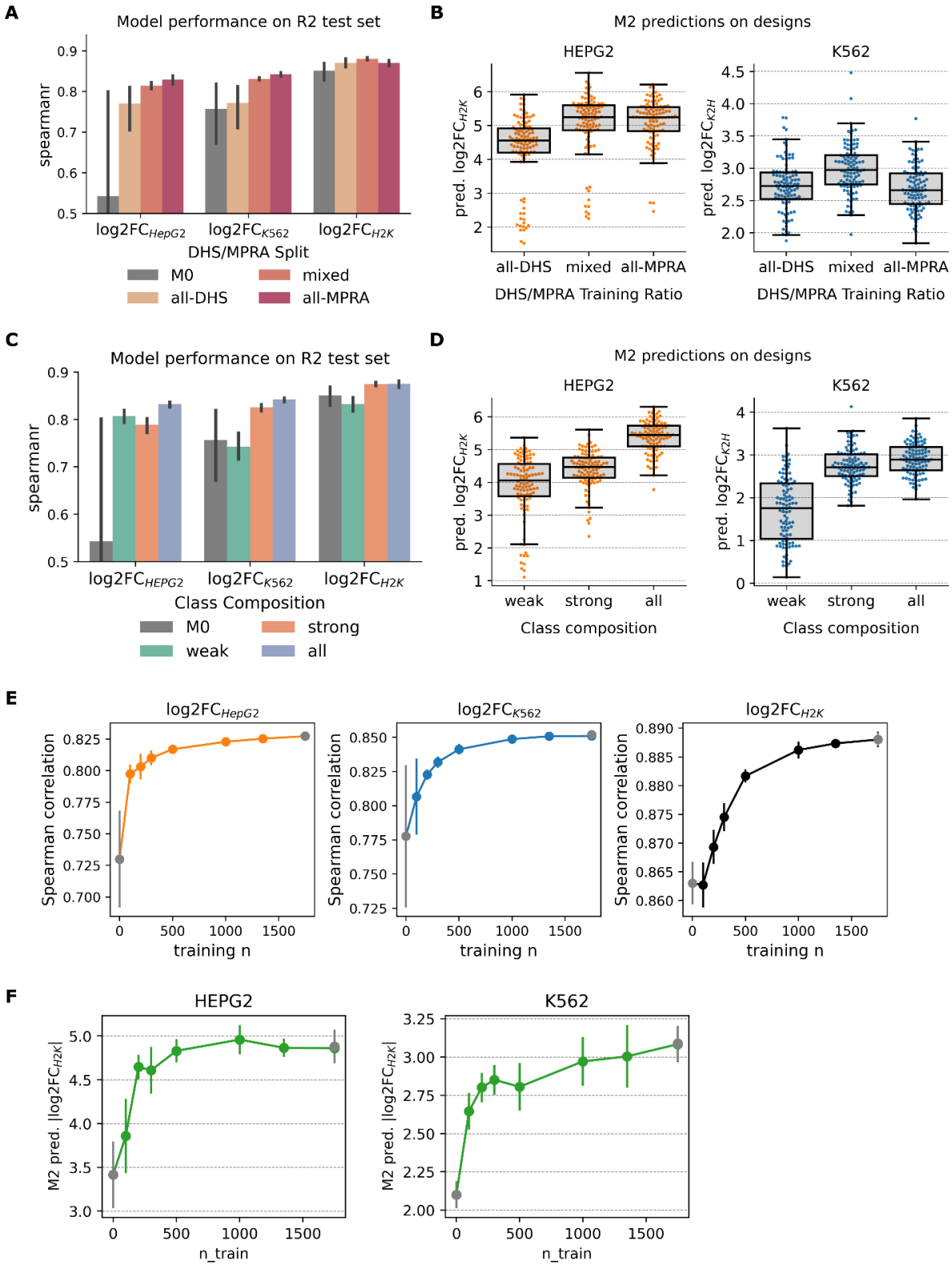
sequence sets. Horizontal lines indicate medians. **E,** Inter-replicate correlation (Spearman, indicated with the letter "ρ") for raw read counts in each cell type for R2 library.



Supplementary Figure 2.6. Comparing design practices

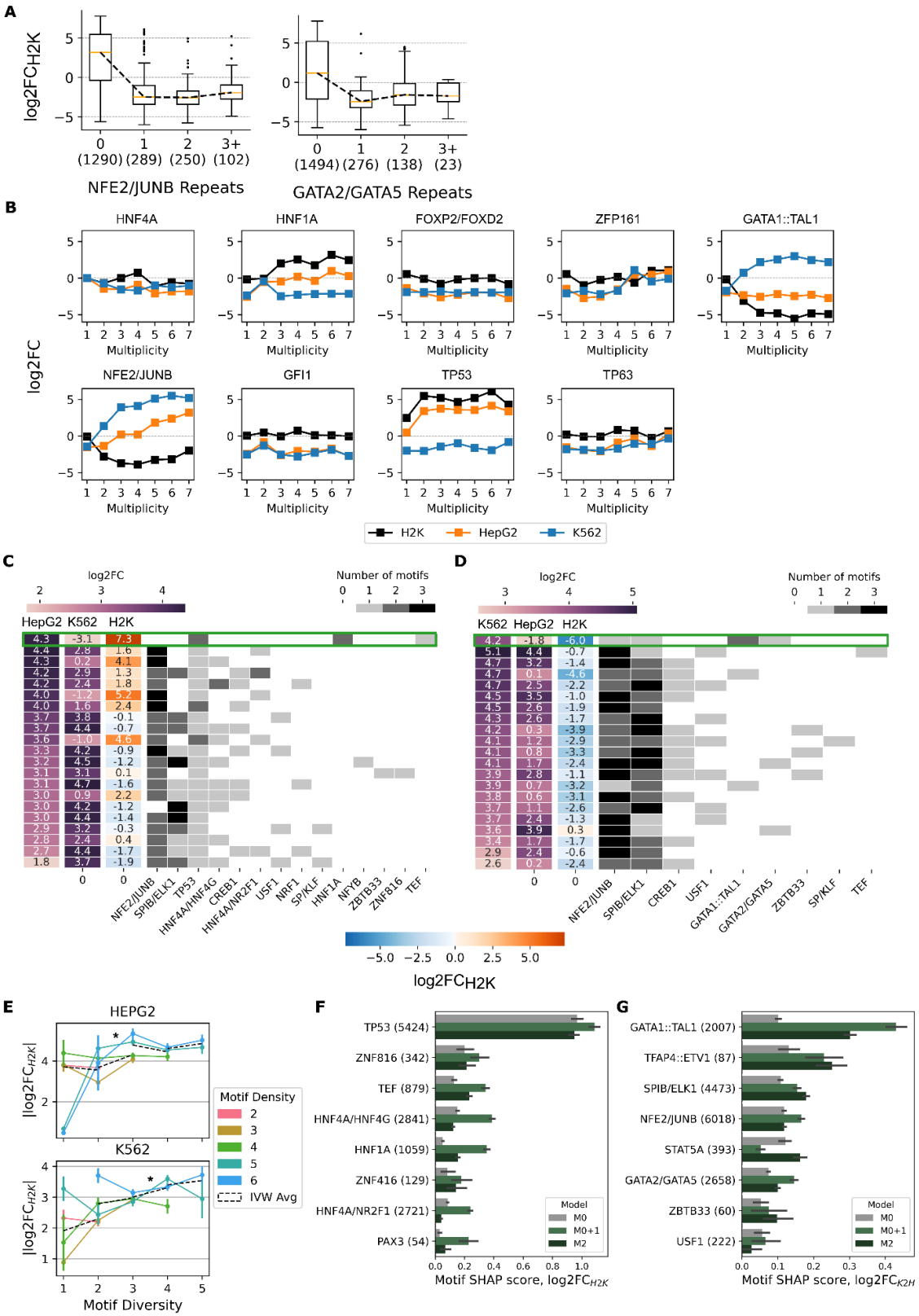
A-D, Comparing enhancer performance between sublibraries within R1 and R2. Each marker corresponds to an enhancer, colored by target cell type. **A,** Comparing enhancers designed to maximize specificity across R1-MPRA, R1-DHS, and R2. **B**, Comparing R2

designs. M0+1 models designed with clipped vs. unbounded objective (left), M0+1 models vs M1 models (right). **C**, Comparing model types in R1 for DEN-generated sequences (the only design method shared across all model types). **D**, Comparing design methods in R1 for Boot-generated sequences (the only model type with comparable numbers of sequences generated from each design method). **E**, Boxswarm plots of $\log2FC_{H2K}$ for Max1 and Min1 designs. Max1 designs in K562 had moderate specificity (median $\log2FC_{H2K}$ = -2.28, -1.81 for unbounded and clipped objectives); whereas Max1 designs in HepG2 had lower on-target specificity with the unbounded objective (-1.20) compared to moderate specificity with the clipped (1.98). **F**, Scatterplot of $\log2FC_{K562}$ vs $\log2FC_{HepG2}$ for Min1 designs (includes clipped and unbounded objectives), as well as 4 negative control (NC) sequences. Line of unity shown. **G**, Scatterplot of $\log2FC_{K562}$ vs $\log2FC_{HepG2}$ for Max1 designs (includes clipped and unbounded objectives), as well as 4 negative control (NC) sequences. Additionally, the most specific 145bp synthetic enhancer in each cell type is plotted in green (max spec.). Line of unity shown. **H**, Specificity vs target cell type strength for Max1 designs, compared to **I**, Specificity vs target cell type strength for R2 enhancers designed to maximize specificity, same as Figure 2.1E.



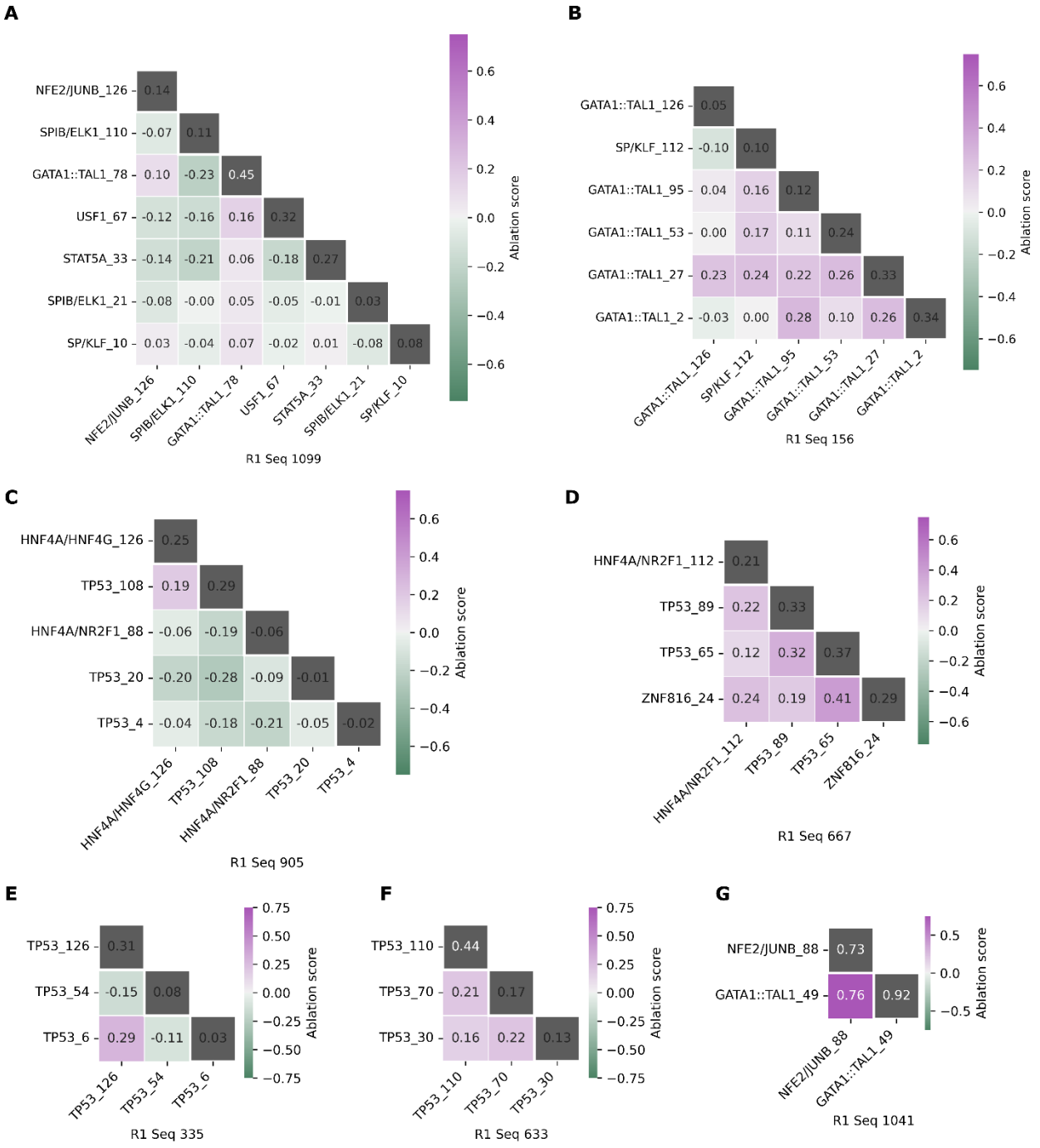
Supplementary Figure 2.7. In silico analysis of the impact of finetuning data on design performance

A, Mean prediction-measurement correlation on R2 test set across the 9 models of the indicated ensembles. Prediction performance for original (not finetuned) M0 model shown in gray. Bars indicate 95% CI. **B**, Predicted specificity of *in silico* enhancer designs generated from models finetuned on different data subsets, for HepG2-targeted designs (left) and K562-targeted designs (right). Each point represents an enhancer designed from a model trained with the data subset indicated by x-axis; y-axis indicates $|\log2FC_{H2K}|$ of enhancers as predicted by the M2 ensemble. **C**, Mean prediction-measurement correlation on R2 test set across the 9 models of the indicated ensembles. Prediction performance for not finetuned M0 model shown in gray. Bars indicate 95% CI. **D**, Predicted specificity of *in silico* enhancer designs generated from models finetuned on different data subsets, for HepG2-targeted designs (left) and K562-targeted designs (right). Each point represents an enhancer designed from a model trained with the data subset indicated by x-axis; y-axis indicates $|\log2FC_{H2K}|$ of enhancers as predicted by the M2 ensemble. **E**, Prediction performance as a function of number of training samples. Spearman R between prediction and measurement (y-axis) shown for M0 ensembles finetuned with varying number of R1 enhancer measurements (x-axis), for the quantity indicated by subplot title. Mean and standard deviation of correlation shown for 5 models per n-value. Gray point at n=0 indicates performance for model trained only on R0 data, without any finetuning on R1. **F**, Design performance as a function of number of training samples. Average M2-predicted $|\log2FC_{H2K}|$ (y-axis) shown for enhancers designed from M0 ensembles finetuned with varying number of R1 enhancer measurements (x-axis), for enhancers targeted to HepG2 (left) and K562 (right). 5 models trained per n-value, and 100 sequences generated per model; for each model the mean specificity is calculated, and the mean and standard deviations of these means is plotted at each n-value. Gray point at n=0 indicates performance for model trained only on R0 data, without any finetuning on R1.



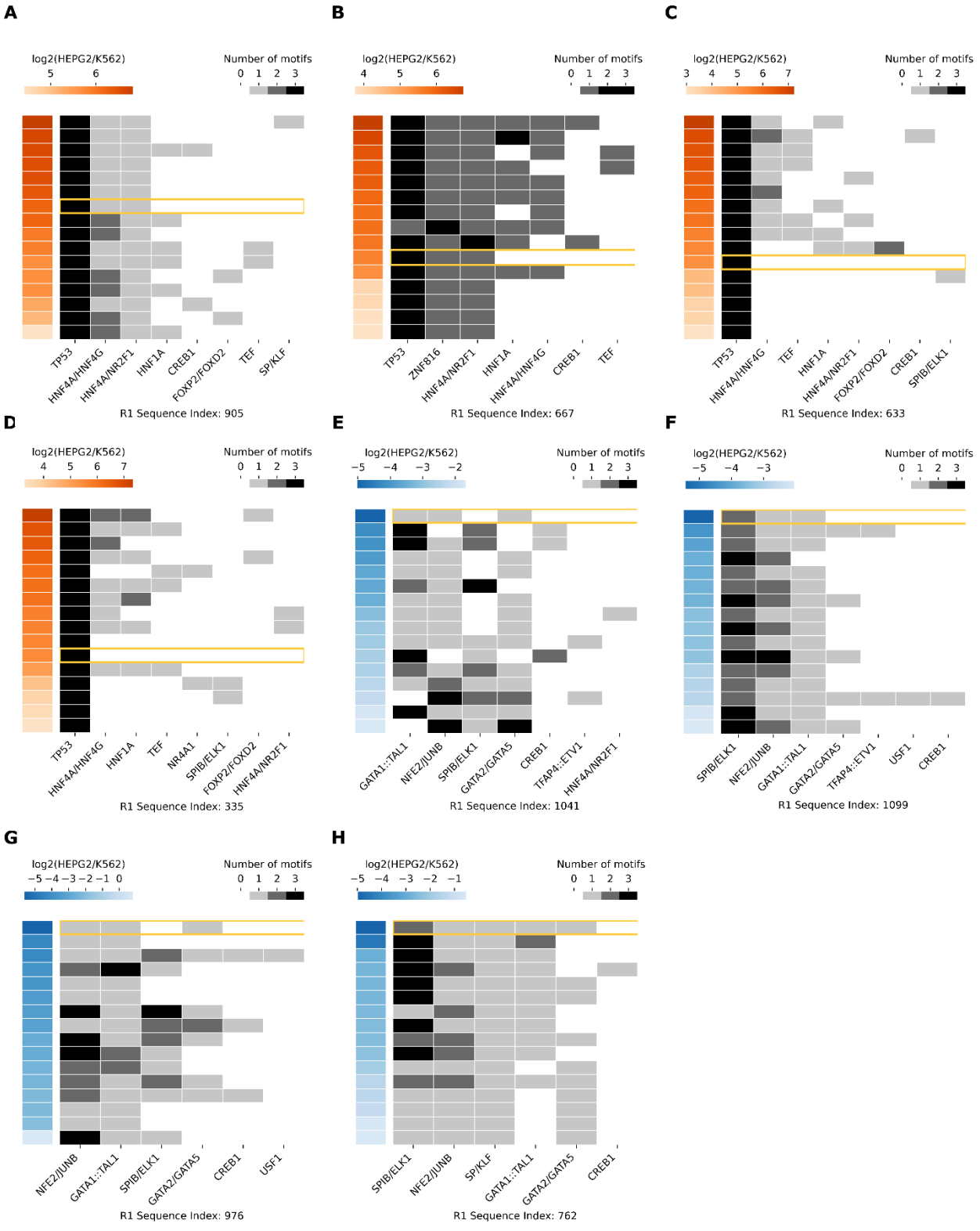
Supplementary Figure 2.8. Additional motif analysis

A, $log2FC_{H2K}$ vs multiplicity for motifs not featured in main text. **B**, $log2FC_{H2K}$, $log2FC_{HepG2}$, and $log2FC_{K562}$, vs multiplicity for manual homotypic enhancers measured in R1. Each point corresponds to a single sequence measurement. **C,D** Motif content of HepG2 (**D**) and K562 (**E**) Max1 designs plotted alongside $log2FC_{H2K}$, $log2FC_{HepG2}$, and $log2FC_{K562}$. The most specific 145bp synthetic enhancer from R2 is plotted at the top of each heatmap for reference (green rectangle), otherwise sequences sorted by descending strength in target cell type. NFE2/JUNB, SPIB/ELK1, CREB1, and USF1 motifs prominently deployed in Max1 designs targeted to both cell types, indicating some level of non-specific activity. **E**, Mean $log2FC_{H2K}$ vs motif diversity (number of unique motif types in a sequence) plotted for R2 enhancers stratified by motif density. Best fit lines shown for all stepwise increases in motif diversity, within each strata. Inverse variance-weighted average slopes ("IVW Avg") across strata plotted as dashed black lines. In HepG2 designs a significant positive IVW Avg slope is observed when increasing motif diversity from 2 to 3 (m = 0.645+-0.178, p = 3e-4, denoted by asterisk). In K562 designs a significant positive IVW Avg slope is observed when increasing motif diversity from 3 to 4 (m = 0.297+-0.117, p = 1e-2, denoted by asterisk), and when regressing across all stepwise increases in motif diversity (m = 0.244+-0.058, p = 2.7e-5, line not shown). **F**,**G** Motif-wise SHAP contribution (per sequence) for model predictions on enhancer libraries. 95% CI for mean SHAP score shown for each motif, for each model. Numbers in parentheses indicate total instances of motifs in sequences used for SHAP calculation. Motifs with the 8 highest mean SHAP score across all models shown for $log2FC_{H2K}$ (**F**), $log2FC_{K2H}$ (**G**).



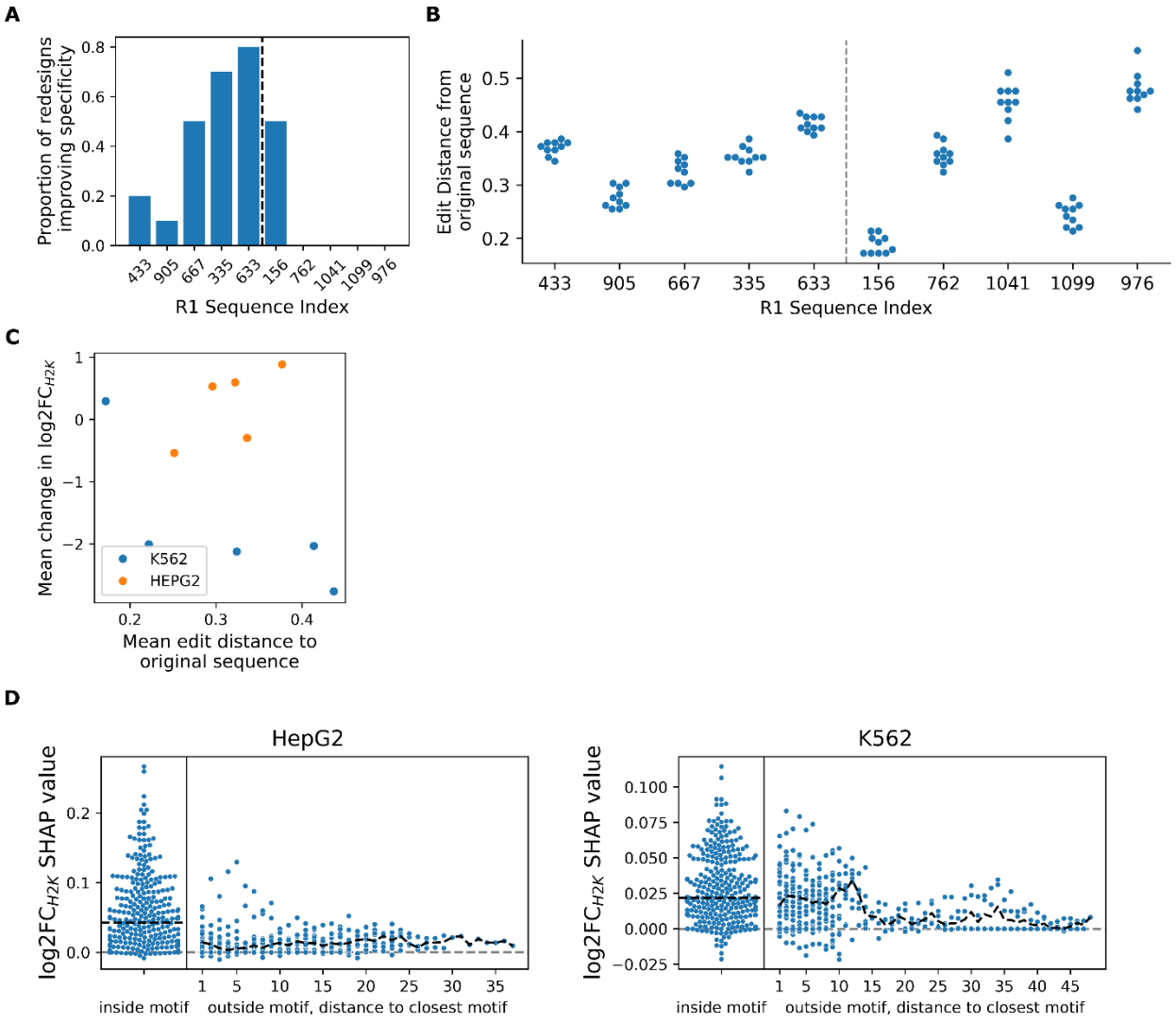
Supplementary Figure 2.9. Additional double ablation deviation score heatmaps for sequences in Figure 2.3.

A, R1 Seq 1099. **B**, R1 Seq 156. **C**, R1 Seq 905. **D**, R1 Seq 667. **E**, R1 Seq 335. **F**, R1 Seq 633. **G**, R1 Seq 1041. See also **Figure 2.3F-H**.


A
log2(HEPG2/K562)
5 6
Number of motifs
0 1 2 3
TP53
HNF4A/HNF4G
HNF4A/NR2F1
HNF1A
CREB1
FOXP2/FOXD2
TEF
SP/KLF
R1 Sequence Index: 905
B
log2(HEPG2/K562)
4 5 6
Number of motifs
0 1 2 3
TP53
ZNF816
HNF4A/NR2F1
HNF1A
HNF4A/HNF4G
CREB1
TEF
R1 Sequence Index: 667
C
log2(HEPG2/K562)
3 4 5 6 7
Number of motifs
0 1 2 3
TP53
HNF4A/HNF4G
TEF
HNF1A
HNF4A/NR2F1
FOXP2/FOXD2
CREB1
SPIB/ELK1
R1 Sequence Index: 633
D
log2(HEPG2/K562)
4 5 6 7
Number of motifs
0 1 2 3
TP53
HNF4A/HNF4G
HNF1A
TEF
NR4A1
SPIB/ELK1
FOXP2/FOXD2
HNF4A/NR2F1
R1 Sequence Index: 335
E
log2(HEPG2/K562)
−5 −4 −3 −2
Number of motifs
0 1 2 3
GATA1::TAL1
NFE2/JUNB
SPIB/ELK1
GATA2/GATA5
CREB1
TFAP4::ETV1
HNF4A/NR2F1
R1 Sequence Index: 1041
F
log2(HEPG2/K562)
−5 −4 −3
Number of motifs
0 1 2 3
SPIB/ELK1
NFE2/JUNB
GATA1::TAL1
GATA2/GATA5
TFAP4::ETV1
USF1
CREB1
R1 Sequence Index: 1099
G
log2(HEPG2/K562)
−5 −4 −3 −2 −1 0
Number of motifs
0 1 2 3
NFE2/JUNB
GATA1::TAL1
SPIB/ELK1
GATA2/GATA5
CREB1
USF1
R1 Sequence Index: 976
H
log2(HEPG2/K562)
−5 −4 −3 −2 −1
Number of motifs
0 1 2 3
SPIB/ELK1
NFE2/JUNB
SP/KLF
GATA1::TAL1
GATA2/GATA5
CREB1
R1 Sequence Index: 762

Supplementary Figure 2.10. Additional motif content of non-motif redesigned enhancers in Figure 2.3.

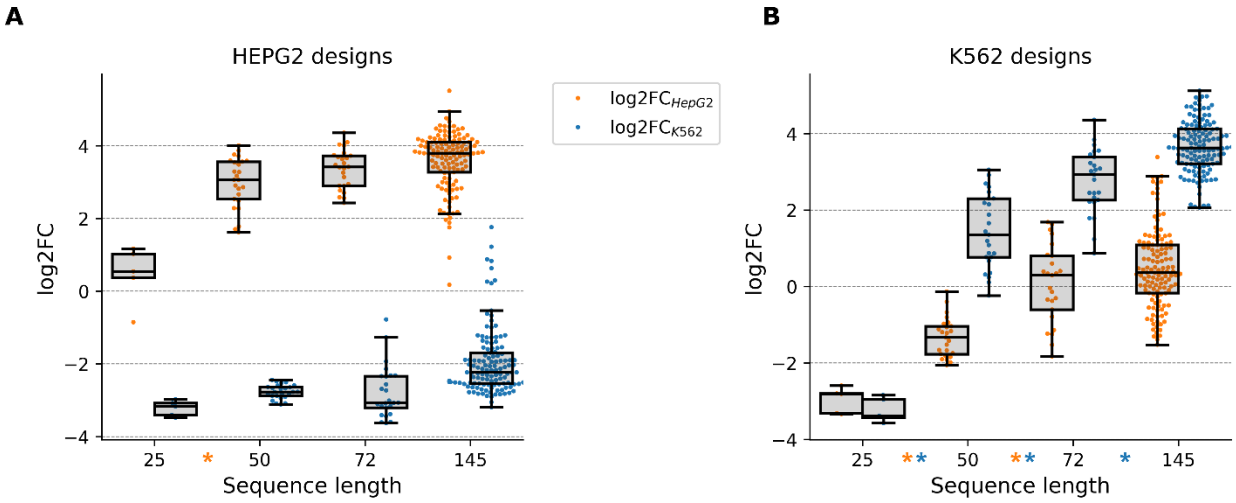
A, R1 Seq905. **B**, R1 Seq 667. **C**, R1 Seq 633. **D**, R1 Seq 335. **E**, R1 Seq 1041. **F**, R1 Seq 1099. **G**, R1 Seq 976. **H**, R1 Seq 762. See also **Figure 2.3I-K**.



Supplementary Figure 2.11. Additional analysis on enhancer perturbations

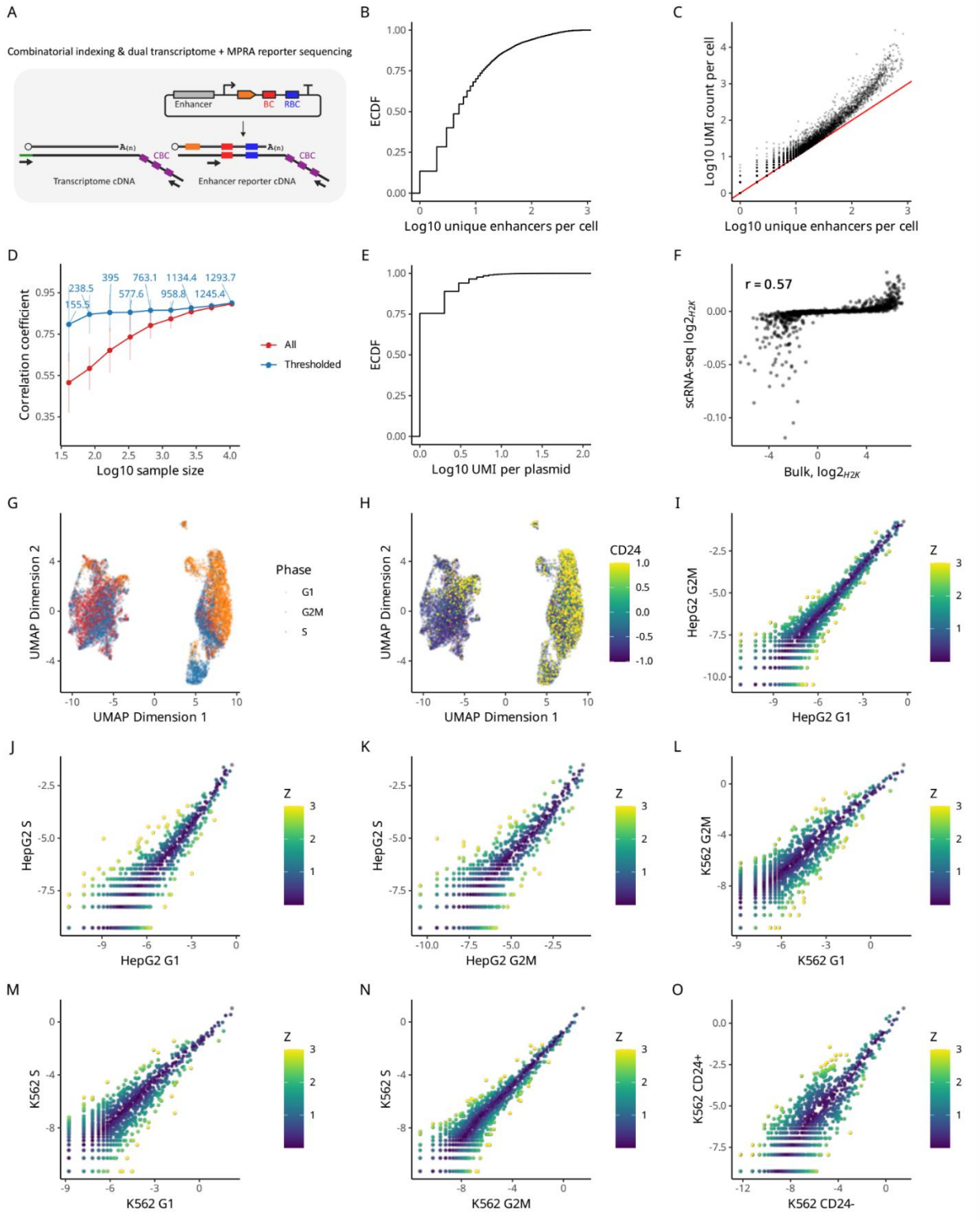
A, The proportion of enhancers designed by re-optimizing non-motif sequence with significantly improved specificity over the original R1 enhancer (out of 10 redesigns per original enhancer). Dashed black line separates HepG2-targeted enhancers (left) from K562-targeted enhancers (right). **B**, Length-normalized edit distance between original sequence and redesigned sequences (where non-motif portions of the enhancer have been redesigned with Fast SeqProp). Dashed black line separates HepG2-targeted enhancers (left) fromK562-targeted enhancers (right). **C**, Scatterplot of mean increase in $|\log2FC_{H2K}|$ (y-axis) vs average length-normalized edit distance of redesigned sequences to original enhancer (x-axis). No significant correlation observed. **D**, Relative $\log2FC_{H2K}$ SHAP contributions for nucleotides in the top 5 R1-MPRA enhancers per cell line, plotted vs. distance to the nearest motif flank. $\log2FC_{H2K}$ SHAP values calculated as the difference between SHAP values for

the $\log2FC_{HepG2}$ and $\log2FC_{K562}$ outputs for M0+1 models finetuned on R2 data; the SHAP value for each nucleotide is normalized by the sum of all SHAP values in its originating sequence, yielding its proportional contribution to the sequence's predicted $\log2FC_{H2K}$. Black dashed line indicates median SHAP value at given position; horizontal gray dashed line indicates SHAP value of 0.



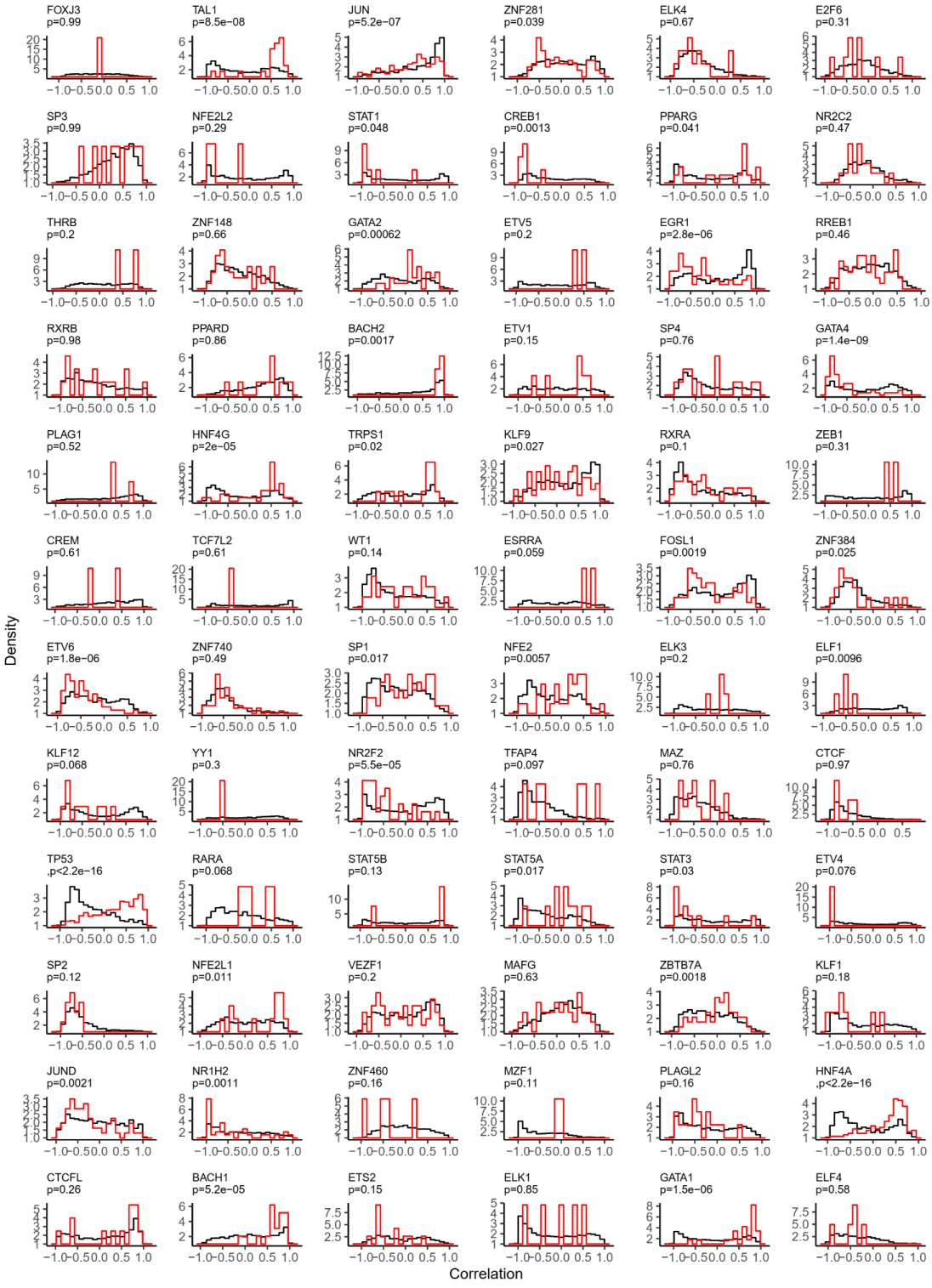
Supplementary Figure 2.12. Additional analysis on shorter enhancer design

A,**B** Absolute activities of shorter enhancer designs for HepG2-targeted enhancers (**A**) and K562-targeted enhancers (**B**). Colored asterisks between lengths indicate significant increase in $\text{log2FC}_{\text{HepG2}}$ (orange), $\text{log2FC}_{\text{K562}}$ (blue).



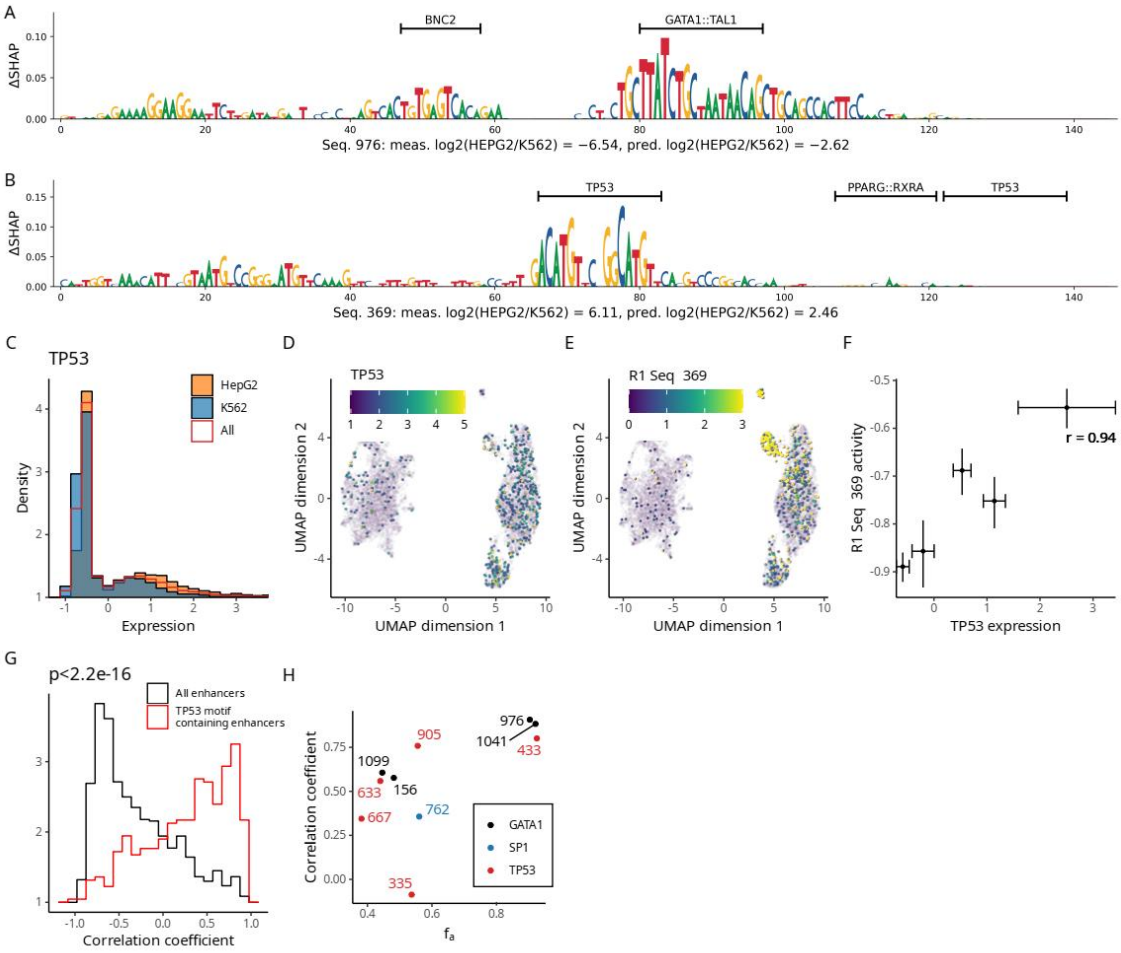
Supplementary Figure 2.13. scRNA quality analysis.

A, A schematic illustrating scMPRA via combinatorial indexing. Cells are transfected with a modified enhancer reporter library that has random barcodes to tag individual plasmid molecules. Single-cell RNA-seq is carried out via combinatorial indexing, with separate PCRs to amplify both the transcriptome and reporter library cDNA amplicons. BC: enhancer barcodes; RBC: random barcodes; CBC: cell barcodes. **B,** Empirical CDF of the distribution of the numbers of unique enhancer sequences detected per cell. **C,** A scatter plot illustrating total UMI counts per cell versus the numbers of unique enhancer sequences per cell. **D,** Subsampling analysis of bulk vs. pseudobulk correlations. Cells used for pseudobulk aggregation are subsampled at various cell numbers as shown on the X-axis. Mean and standard deviation of bulk vs. pseudobulk correlations from n=100 samples for each subsample sizes are plotted as dots and error bars. In blue are the correlations calculated with enhancers passing minimum reporter expression level threshold (>4 UMI / cell), with the text label indicating the average number of enhancers that were available to calculate the correlations for each subsample size at this threshold. **E,** Empirical CDF of the distribution of the numbers of UMIs expressed per unique plasmid molecule. **F,** A scatter plot comparing cell-type specific activity of synthetic enhancers obtained by bulk vs. single-cell analysis, using mean RBC-normalized single cell enhancer expression. **G,** UMAP projection of the single cell transcriptomes, colored by cell cycle phases. **H,** UMAP projection of the single cell transcriptomes, colored by CD24 expression, a differentiation marker in K562 cells. **I-O,** Pairwise scatter plots comparing enhancer expression between substates (cell cycle phases and K562 CD24+/- status) within each cell types. Colored by z-scaled log2 fold differences between the aggregated pseudobulk counts on X and Y axes.



Supplementary Figure 2.14. Correlation between TF expression and enhancer activity

Distribution of correlations between TF expression versus all enhancers across pseudobulk of single cells binned by TF expression values. In red are pairings where the enhancer contains the DNA sequence motif for each TF. p values from two-sided Kolmogorov-Smirnov test. Benjamini-Hochberg estimated FDR<0.05 at p<0.041.



Supplementary Figure 2.15. Single cell level enhancer activity.

A, A sequence logo for R1 Seq 976 with nucleotide height corresponding to the difference between log2FC$_{K562}$ SHAP value and log2FC$_{HepG2}$ SHAP value, according to the M0+1 ensemble (**Appendix A**). **B,** A sequence logo for R1 Seq 368 with nucleotide height corresponding to the difference between log2FC$_{HepG2}$ SHAP value and log2FC$_{K562}$ SHAP value, according to the M0+1 ensemble (**Appendix A**). **C,** Expression levels of TP53 transcription factor across: all cells in red; cells in HepG2 cluster in orange; cells in K562 cluster in blue. **D,** Expression levels of TP53 transcription factors across single cells atop the UMAP projection of the transcriptomes. **E,** Activities of an enhancer #369, containing a TP53 transcription factor motif, across single cells atop the UMAP projection of the transcriptomes. **F,** Pseudobulk correlation of TP53 transcription factor expression levels and enhancer #369. The pseudobulks are binned by TP53 expression levels. Vertical error bars represent bootstrap standard errors of enhancer activities. Horizontal error bars represent standard deviation of the transcription factor expression

levels across the cells in each pseudobulk bin. **G,** Distribution of correlations between TP53 expression versus all enhancers across pseudobulk of single cells binned by TP53 expression values. In red are pairings where the enhancer contains the DNA sequence motif for TP53. p value from Kolmogorov-Smirnov test. **H,** Comparison of the strength of dependency on individual TF motifs vs. correlations of enhancer activity with TF expression. Each point is an enhancer sequence, where the x-axis represents the maximum f_a from the ablation experiment for that enhancer sequence. The color represents the ablated TF motif that resulted in the highest ablation score. The y-axis represents the correlations between cell type specificity and the expression of the corresponding TF determined from the scMPRA experiment.

2.2 Chapter 2 – Supplemental Tables

Table 2.1. R1-MPRA Library Composition

	Fast SeqProp	Sim. Anneal	DEN	Other
Single	28	18	200	-
Boot	198	198	198	-
Ensemble	-	-	198	-
Control	-	-	-	200
Motif repeat	-	-	-	62

Table 2.2. R2 Library Composition

Model type	**Design type**	**Design objective**	**Count**
M1	Fast SeqProp	Unbounded H2K	210
M0+1	Fast SeqProp	Unbounded H2K	258
		Clipped H2K	220
		Target H2K	160
		Unbounded Max1	20
		Unbounded Min1	20
		Clipped Max1	20
		Clipped Min1	19
	Masked Fast SeqProp (nonmotif reoptimization)	Unbounded H2K	100
	Reduced length Fast Seqprop	Unbounded H2K	110
None	Dinucleotide shuffling	Nonmotif reoptimization	49
		Motif ablation	331
Control	Negative control	-	4
	Random	-	198
	Top enhancer	-	10

Table 2.3. Motif sequences used in hand-crafted motif repeat enhancers (R1-MPRA)

CISBP2.0 Motif Name	**Sequence (5'-3')**	**Putative Role**
Hnf4	AGGTTCAAAGGTCA	HepG2 Enhancer
Hnf1	GGTAATTATTAACC	HepG2 Enhancer
Foxa	TGTTTACTTAGG	HepG2 Enhancer
Zfp161	TGGCGCGCGCGCCTGA	K562 Enhancer
Gata	CTGGTGGGGACAGATAAG	K562 Enhancer
Nfe2l2	ATGACTCAGCA	K562 Enhancer
Gfi1	AAATCACAGC	K562 Repressor
Tp73	ACATGTC	HepG2 Enhancer
Tp63	ACATGCCCGGGCATG	HepG2 Enhancer

Table 2.4. Model architecture of GAN (R1-DHS)

Generator Layers:	**Dimension (BS x W x H x C)**
Input: noise / seed	N x 100
Linear	N x 1600
ReLU	N x 1600
BatchNorm	N x 1600
Transposed Conv (num_filters = 640, filter size = 10x1, stride = 10)	(N x 200 x 1 x 640)
ReLU	(N x 200 x 1 x 640)
BatchNorm	(N x 200 x 1 x 640)
Transposed Conv (num_filters = 1, filter size = 15x4, stride = 1)	(N x 200 x 4 x 1)
Softmax	(N x 200 x 4 x 1)

Discriminator Layers:	**Dimension (BS x W x H x C)**
Input: one hot sequence	(N x 200 x 4 x 1)
Conv (num_filters = 640, filter size = 15x4, stride = 1)	(N x 200 x 1 x 640)
Spectral Norm	(N x 200 x 1 x 640)
Max Pool (kernel size = 10x1)	(N x 20 x 1 x 640)
Leaky ReLU (slope=0.1)	(N x 20 x 1 x 640)
Conv (num_filters = 320, filter size = 11x1, stride = 1)	(N x 20 x 1 x 320)
Spectral Norm	(N x 20 x 1 x 320)
Max Pool (kernel size = 20x1)	(N x 1 x 1 x 320)
Leaky ReLU (slope=0.1)	(N x 1 x 1 x 320)
Linear (200)	N x 200
Spectral Norm	N x 200
Leaky ReLU (slope=0.1)	N x 200
Linear(1)	N x 1

(N = batch size, dimension format: batch size x width x height x channels)

Table 2.5. Model architecture of Classification model (R1-DHS)

Layers	**Dimension (BS x W x H x C)**
(input)	(N x 200 x 4 x 1)
Conv (num filters = 32, filter size = 15x4, stride = 1)	(N x 200 x 1 x 32)
ReLU	(N x 200 x 1 x 32)
Dropout (dropout rate = 0.5)	(N x 200 x 1 x 32)
Max Pool (kernel size = 200)	(N x 1 x 1 x 32)
Linear (50)	N x 50
ReLU	N x 50
Dropout (dropout rate = 0.5)	N x 50
Batch Normalization	N x 50
Linear (3)	N x 3
Softmax	N x 3

(N = batch size, dimension format: batch size x width x height x channels)

Table 2.6. PCR primer sequences corresponding to Chapter 2

Primer Name	Primer Description	Primer Sequence (5'-3')
CY01	Add Gibson flank to 5' of R1-MPRA and R2 145bp enhancer, for overlap with pMPRA1 backbone	gaacatttctctGGCCTAACTGGCCGCTTCACTG
CY02	Add Gibson flank to 3' end of R1-MPRA enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgccgtGGCCCCGCTCCTGTATAGCTG
CY03	Add Gibson flank to 5' of R1-DHS enhancer, , for overlap with pMPRA1 backbone	gaacatttctctGGCCTAACTGGCCCTTCGCTG
CY04	Add Gibson flank to 3' end of R1-DHS enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgCCGTGGCCTCAGTTCACCGCGTC
CY05	Reverse transcriptase primer (adds UMI to mRNA transcript)	AAGCAGTGGTATCAACGCAGAGTACATGGGNNNNNNNNN Nccaaactcatcaatgtatcttatcatgt
CY06	qPCR forward R1-MPRA enhancer	AATGATACGGCGACCACCGAGATCTACACGGCTCTGAtctc attaaggccaagaagggc
CY07	qPCR forward R1-DHS	AATGATACGGCGACCACCGAGATCTACACAGGCGAAGtctc attaaggccaagaagggc
CY08	Custom read1	ggcggcaagatcgccgtgtaataattCTAGA
CY09	Custom Index 2	tgccgcccttcttggccttaatgaga
CY10	Add Gibson flank to 5' of R2 25bp enhancer, for overlap with pMPRA1 backbone	gaacatttctctGGCCTACCTATGCCCACGTCCC
CY11	Add Gibson flank to 3' of R2 25bp enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgccgGTGATACGTGTGTCTGGC
CY12	Add Gibson flank to 5' of R2 50bp enhancer, for overlap with pMPRA1 backbone	gaacatttctctGGCCTATCGGAATCGGTAACGGC
CY13	Add Gibson flank to 3' of R2 50bp enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgccgGTACACCACTGTCCACTG
CY14	Add Gibson flank to 5' of R2 72bp enhancer, for overlap with pMPRA1 backbone	gaacatttctctGGCCTATTCCATCCGCCTGACC
CY15	Add Gibson flank to 3' of R2 72bp enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgccgGTCTGTGAGCATCGACC
SC_01264	Add Gibson flank to 3' end of R2 145bp enhancer, for overlap with pMPRA1 backbone	cccgactagcttggccgccgATCTACCTGGTCCGGCA
Custom_read_2		AAGCAGTGGTATCAACGCAGAGTACATGGG
Custom_index_1		CCCATGTACTCTGCGTTGATACCACTGCTT
GB_39		GGCAAGATCGCCGTGTAATAATTCTAGA
GB_42		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCAAG ATCGCCGTGTAATAATTCTAGA
GB_43		GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

GB_29		cggccaagctagtcggggcggccggTAATACGACTCACTATAGGGATGC gcgaagctcgtctgtactatggccNNNNNNNNNNNNNNNNNNNNATGCcc ggccgcttcgagcagacatgata
GB_34		tatcatgtctgctcgaagcggccggGCAT

Table 2.7. Batch correction regression coefficients

	Cell type	**slope**	**intercept**
R1 to R2	HepG2	0.626018	-1.524307
	K562	0.614353	-1.288952
R0-MPRA to R1	HepG2	0.897912	-0.691838
	K562	0.883498	-1.052554

CHAPTER 3. GENERALIZING ENHANCER DESIGN THROUGH EXPANDED DHS MODELS

This chapter directly builds upon the previously described work. Our success with accessibility-based synthetic enhancers suggested a promising path towards a more general design workflow, not limited by the sparsity of MPRA-assayed (or assayable) targets. Nonetheless, the relatively inferior performance of the DHS- vs. MPRA-based enhancers indicated room for improvement. In the previous chapter R1-DHS sequences were designed with a more conservative approach than R1-MPRA/R2 sequences, utilizing a GAN alongside additional heuristics evaluated on sequence properties to encourage similarity to genomic DHSs. We found evidence that this may have limited enhancer performance. Therefore, in this chapter we sought to (1) more fully exploit the wealth of publicly available chromatin accessibility datasets to design enhancers across a greatly expanded set of cell- and tissue-types, and (2) optimize our modeling and design techniques for a purely accessibility-based workflow. We do so by training Neural Network (NN) models on experiments collected in the ENCODE DHS index to predict genomic accessibility in 64 different cell- or tissue-types, and designing sequences to maximize predicted specific accessibility to each model output. By measuring enhancer function for all sequences in a panel of 10 cell lines subsetted from the model outputs, we are able to more rigorously

confirm the feasibility, generalizability, and potential of synthetic enhancer design. The manuscript corresponding to this work is under submission. It has been conducted in collaboration with Sebastian Castillo-Hair, Wouter Meuleman, Leah VandenBosch, Timothy Cherry, and Georg Seelig.

3.1 NN MODELS PREDICT GENOMIC ACCESSIBILITY ACROSS MULTIPLE CELL TYPES

We first trained NN models to predict genomic accessibility across a diversity of cell types and tissues.[68] We initially selected 64 biosamples with high-quality experimental metrics representing all 15 previously described major tissue origins ("components") in the ENCODE database (**Appendix B**). Our model, christened "DHS64", implements a deep residual neural network architecture to predict continuous $\log_{10}$ accessibility signals (i.e. normalized peak DNase-seq read density) and peak call probabilities for all 64 biosamples, given an input sequence of up to 500bp (**Figure 3.1B**, **Supplementary Figure 3.2A**, **Appendix B**). While large transformer models have been effective at predicting multiple data modalities from a >100kb genomic input window[69,70], we opted to train more compact models compatible with sequence design methods that require repeated model and gradient evaluations[71,72], and that used a shorter input window so that designed enhancers functioned independently of sequence context.

In our initial tests, we found that naïve evaluation on an unfiltered set of DHSs held-out from training resulted in overly optimistic performance metrics with high accuracy on strong, non-specific DHSs at the expense of cell type-specific sequences with weaker signal (**Supplementary Figure 3.2B-G**)—an issue likely also present in large transformer models[73]. To better assess the ability of our model to capture cell type-specific activity, we chose to evaluate performance on a filtered set of highly specific DHSs with enhancer-like chromatin and genomic features; and found that filtering the training data for DHSs active in 10 or fewer biosamples resulted in best performance on this test set (**Appendix B**). On a per-biosample basis, DHS64 predicted accessibility signals with a Pearson R of 0.56±0.11 (median ± interquartile range) and peak call probabilities with an area under the precision-recall curve (AUPRC) of 0.49±0.13 (**Figure 3.1C**, **Supplementary Figure 3.2H-I**). For a given sequence, DHS64 could predict variation in accessibility signals across biosamples with a per-DHS Pearson R of 0.60±0.26, where lower performance corresponded to DHSs with lower signal variation (**Figure 3.1D**, **Supplementary Figure 3.2J**), as observed previously in a model of immune cell accessibility[74]. Finally, we confirmed that DHS64 accurately predicted the accessibility levels of the most cell type-specific DHSs in the test set (**Supplementary Figure 3.2K-L**).

3.2 Sequences optimized for accessibility show cell type-specific enhancer activity

We next used DHS64 to generate synthetic sequences with cell type-specific accessibility. Specifically, we sought to maximize the difference between predicted $\log_{10}$ accessibility in a selected target biosample and the average $\log_{10}$ accessibility across all others (**Figure 3.2A**). We used Fast SeqProp[71], which optimizes sequences via gradient descent, to design 250 sequences with a length of 145 nt targeted to each of the 64 modeled biosamples. Similarly, we used Deep Exploration Networks (DENs)[72] – generative NNs that produce sequences with high predicted performance – and generated 250 additional sequences per target. We successfully generated sequences predicted to be specific to almost every individual biosample (**Figure 3.2B-C**). For example, sequences targeted to the retinoblastoma cell line WERI-Rb-1 showed strong on-target signal and weak predictions in all other cell types, including related biosamples from the "neural" component (**Figure 3.2C**, top). However, cross-over within the same component was sometimes difficult to avoid, for example within three fetal lung tissue biosamples that differed only in the originating subject (**Figure 3.2B**, bottom), as well as between the B-lymphocyte-derived GM12878 and GM12865 cell lines (**Figure 3.2C**). Sequence features of synthetic designs, such as GC content and sequence diversity, were similar to those from DHSs (**Supplementary Figure 3.3A-E**). No major performance differences were

predicted between Fast SeqProp and DEN-generated sequences (**Supplementary Figure 3.3F**). Predicted performance was substantially higher than the most specific DHSs (**Supplementary Figure 3.3G**), even though our sequences were shorter than the median DHS length (196 nt)[68].

To experimentally validate these sequences as functional enhancers, we selected 10 cell lines from the 64 modeled biosamples in which to perform MPRAs. These target cell lines were chosen for their compatibility with established MPRA protocols while also spanning a representative range of tissue origins covered by the model. (**Figure 3.2D**). For each of these 10 biosamples we tested 300 synthetic sequences designed (as described above) to be active in only the target cell line and inactive in the remaining 63 biosamples. As controls, we also tested 110 DHSs with the highest biosample-specific accessibility, truncated to their central 145 nt to accommodate oligo pool synthesis, which had a small effect on predicted accessibility and specificity (**Supplementary Figure 3.4**). We cloned candidate enhancers into a plasmid reporter library, transfected into all cell lines, and extracted and sequenced mRNA, yielding good data quality and replicate correlation (**Supplementary Figure 3.5**, **Appendix B**). Enhancer activity was quantified in each cell line as $\log_2$(mRNA counts from cell line/DNA counts in plasmid library) (hereafter $\log_2$FC), and overall enhancer performance was assessed via the

stringent mingap score[75], defined as the difference between $\log_2$FC in the target cell line and the maximum $\log_2$FC across all non-targets (**Appendix B**).

Designed sequences exhibited the desired expression patterns across all but one cell line; and had higher on-target expression and mingap scores than corresponding DHSs, which were largely weak or inactive (**Figure 3.2E-G, Supplementary Figure 3.6, Supplementary Figure 3.7A**). These differences were particularly pronounced among the top performing sequences (**Figure 3.2G**): e.g. median mingap scores of the top 20% of synthetic sequences and DHSs were 5.57 and 0.21 for WERI-Rb-1, 4.35 and 2.74 for K562, 2.39 and -0.07 for the embryonal NT2-D1, and 2.24 and 0.10 for the neural-derived SK-N-SH. In fact, DHS-derived sequences achieved significantly positive median scores in only 1 cell line (GM12878, **Supplementary Figure 3.7A**) when considering all sequences, or in 2 additional cell lines (K562 and HepG2, **Figure 3.2F**) when considering only the top 20%. GM12878 was the only target cell line where DHSs had equivalent performance to synthetic sequences. In some cases, such as HepG2 and WERI-Rb-1, most synthetic sequences showed high specificity and strong on-target expression (**Supplementary Figure 3.6, Supplementary Figure 3.7A**). In others, performance varied depending on the design method: for the muscle-derived SJCRH30 and the cervix-derived HeLa, DENs produced sequences with significantly higher on-target activity and mingap scores, whereas for GM12878 and NT2-D1 Fast SeqProp performed slightly better

(**Supplementary Figure 3.6, Supplementary Figure 3.7A**). For the kidney-derived 786-O, all design methods failed to produce enhancers with positive specificity. These conclusions remained consistent when using alternate specificity metrics, including one based on average non-target activities as well as the tissue specificity index τ[76] (**Supplementary Figure 3.7B-C**). These results demonstrate that synthetic enhancers optimized with DHS64 for accessibility can function as effective cell type-specific enhancers across diverse cell types, outperforming endogenous DHS-derived sequences.

We next compared our design strategy based on genomic accessibility predictors to NN-based methods we previously used to design HepG2- and K562-specific enhancers[77]. To this end, we measured the activity of the highest performing HepG2- and K562-specific enhancers from our previous work in all 10 cell lines (**Appendix B**). Our results showed that K562 enhancers designed with DHS64 had stronger specificity than both Round 1 sets, and equivalent performance to the further optimized Round 2 sequences (**Supplementary Figure 3.8A**). When targeting HepG2, DHS64 enhancers were comparable to Round 1-MPRA and superior to Round 1-DHS, but were outperformed by Round 2 enhancers. We next examined the specificity of our previous enhancers beyond HepG2 and K562, by using the measurements we collected of 20 top-performing Round 1-MPRA enhancers across all 10 cell lines. Notably, HepG2-targeted enhancers showed substantial activity in cell lines beyond HepG2 and K562, which were not considered in

our previous work (**Supplementary Figure 3.8B**). In contrast, DHS64-designed sequences did not show comparable off-target effects. These results indicate that enhancers designed with DHS64 can have comparable performance to those designed using models trained on enhancer MPRA data, and can retain specificity across a larger number of cell types.

Finally, to provide evidence of specificity at a larger scale we assayed, across the same 10 cell lines, five DHS-sourced and five Fast SeqProp-designed sequences targeted to each of the 54 remaining biosamples. As expected, average enhancer activity across cell lines was low (avg. $\log_2 FC < 1$) for all DHSs and Fast SeqProp sequences designed for most targets (46/54) (**Supplementary Figure 3.9C**). Enhancers occasionally exhibited off-target activity in a cell line related to the target biosample, which was well-captured by DHS64 predictions (**Supplementary Figure 3.9G,H)**. Only for a few target biosamples (e.g. the prostatic adenocarcinoma cell line PC3, the glioblastoma cell line A-172, and the variant mammary epithelial cell line vHMEC) did designed enhancers exhibit broad off-target activity that was not predicted by the model. Notably, a similar failure mode consisting of broad expression was observed for HeLa-targeted sequences when designed with Fast SeqProp but not DENs (**Supplementary Figure 3.6**), suggesting that DEN-designed sequences for these targets may also show reduced off target activity. Overall, sequences designed for accessibility specific to non-assayed cell types typically

exhibited low enhancer activity, occasionally showed activity in a cell line related to the target cell type, and, only in rare cases, displayed unexpected broad expression.

3.3 Accessibility NN predictors enable programming complex functions into synthetic enhancers

We next explored enhancer design objectives beyond maximizing specificity towards a single cell type. First, we sought to design cell type-specific enhancers with tunable (intermediate) target activities (**Figure 3.2A**, **Appendix B**). Using Fast SeqProp, we designed 120 sequences per tested cell line, with target accessibility setpoints ranging from inactive to maximally active, and validated them experimentally via MPRAs (**Appendix B**). For 9/10 cell lines, target enhancer activity was tunable and significantly positively correlated with predicted target accessibility across the entire range (**Figure 3.2B, Supplementary Figure 3.10A**). Mingap scores were also significantly positively correlated with accessibility predictions for 8/10 cell lines (**Supplementary Figure 3.10B**). The exceptions aligned with previous Fast SeqProp failures in the context of designing maximally specific enhancers (**Supplementary Figure 3.7A**): HeLaS3, where Fast SeqProp failed but DENs succeeded, and 786-O, where no design method was successful. These results show that optimizing for submaximal predicted accessibility is a feasible strategy to design cell type-specific enhancers with tunable activity.

Next, to assess the orthogonality of our approach, we sought to design and validate enhancers with specificity towards two (**Figure 3.2C-E, Supplementary Figure 3.11**) and three (**Figure 3.2F-H, Supplementary Figure 3.12**) unrelated cell types. As targets, we randomly selected eight pairs and eight triplets from the 10 assayed cell lines, ensuring that each target group contained cell types from different biological components. Using Fast SeqProp, we generated 117 sequences optimized for high accessibility in all target cell types within each group (**Supplementary Figure 3.11A, Supplementary Figure 3.12A, Appendix B**). We additionally sought to include, for each target group, 50 DHSs with enhancer-like chromatin marks and positive peak calls in all target cell types. However, such DHSs were scarce, and insufficient numbers were found for 1/8 pairs and 5/8 triplets; for one triplet no DHSs met our criteria (**Supplementary Figure 3.11B, Supplementary Figure 3.12B, Appendix B**). We scored these designs via an extended mingap score obtained by subtracting the minimum $\log_2$FC across targets from the maximum nontarget $\log_2$FC (**Figure 3.2C**, top). When considering the top 20% performing sequences from each source and target group, synthetic enhancers frequently exhibited the expected expression patterns whereas DHSs were almost always inactive (**Figure 3.2C** and **F**). Furthermore, synthetic sequences had significantly positive mingap scores in 5/8 target pairs and 4/8 target triplets (Wilcoxon test, one-sided, Bonferroni-corrected p-value < 0.05), and significantly higher scores than corresponding

DHSs in 5/8 pairs and 3/8 triplets (Mann-Whitney U test, one sided, Bonferroni-corrected p-value < 0.05, **Figure 3.2D** and **G**). The top performing designs showed high expression in all targets, whereas the best-performing DHSs showed noticeable expression in at most one (**Figure 3.2E** and **H**). Nevertheless, mingap scores declined as the number of target cell types increased. Across all designs, both predicted accessibility and measured $\log_2$FC in a target cell type were lower in sequences designed for multiple targets compared to those optimized for one (**Supplementary Figure 3.11C-F, Supplementary Figure 3.12C-F**). One explanation is that as the number of targets increases, sequence length becomes insufficient to accommodate all regulatory elements required for strong activation in all cell types. Supporting this hypothesis, longer sequences designed for multiple targets achieved higher predicted target accessibilities (**Supplementary Figure 3.11G-H, Supplementary Figure 3.12G-H**). In summary, DHS64 enabled the predictable design of enhancers with complex functions, including tunable target activities and specificity for multiple cell types. However, as the number of target cell types increases, designing longer enhancers may be necessary to maintain robust activity.

3.4 Enhancers designed for retinal targets are active *in vivo* in mouse retinas

To evaluate whether our enhancers function *in vivo*, we performed a retinal tissue MPRA by transfecting our enhancer library in P0 mouse retinas, extracting mRNA, and sequencing as above (**Figure 3.4**, **Appendix B**). Although we did not design enhancers specifically for mouse retinas, two of our modeled biosamples – the retinoblastoma cell line WERI-Rb-1 and fetal eye tissue – originate from retinal or ocular sources, leading us to hypothesize that sequences targeting these biosamples would be active. Consistent with this expectation, synthetic enhancers targeting WERI-Rb-1 (median $\log_2$FC = 4.88) and fetal eye tissue (median $\log_2$FC = 3.93) showed the highest activity among all targets (**Figure 3.4B**). Across the hundreds of sequences targeting each of the 10 previously assayed cell lines (**Figure 3.2**), only synthetic enhancers designed for WERI-Rb-1 exhibited consistent strong activity, whereas WERI-Rb-1-targeted DHSs and sequences targeting all other cell lines were inactive (**Figure 3.4C**). Furthermore, enhancers designed for tunable submaximal WERI-Rb-1 activity displayed graded $\log_2$FC levels in the mouse retina, which was highly correlated with their WERI-Rb-1 accessibility setpoints (**Figure 3.4D**). Additionally, two of the five synthetic sequences targeting the fetal eye biosample showed high activity in mouse retina and low activity in all previously assayed cell lines including WERI-Rb-1 (**Figure 3.4E**), suggesting that they leverage

regulatory grammar specific to *in vivo* retinal tissue that is not captured by this cell line. Conversely, DHSs targeting the same biosample were weak (**Figure 3.4E**). In conclusion, our accessibility predictor-based approach enables the design of functional, cell type-specific enhancers that outperform their genomically derived counterparts, even *in vivo*.

3.5 Analyzing sequence determinants of successful enhancer function

Having established that our designed sequences function as cell type-specific enhancers, we next investigated their underlying regulatory grammar, how it compares to that of endogenous enhancers, and which sequence features distinguish high- from low-performing designs. We first scanned thousands of known TF binding sites (TFBSs) across all selected DHSs and designed sequences (**Appendix B**), identifying 730 TFs with binding sites present in at least one biosample. As observed in our previous work[78], on average synthetic sequences contained more TFBSs per base pair than their DHS counterparts (**Supplementary Figure 3.13A**), and utilized TFBSs in different proportions (**Supplementary Figure 3.13B**). We found that TFBSs present in DHSs from fewer cell types (**Supplementary Figure 3.13C-D**) or corresponding to TFs expressed with greater cell type or tissue specificity (**Supplementary Figure 3.14**) were generally enriched in the synthetic sequences; whereas TFBSs that were ubiquitous in DHSs or

corresponded to broadly expressed TFs were generally de-emphasized (**Supplementary Figure 3.13, Supplementary Figure 3.14, Supplementary Figure 3.15**). Among the TFBSs enriched in synthetic sequences, usage patterns across cell types tended to correlate between DHSs and synthetic sequences (**Figure 3.5A**, **Supplementary Figure 3.15**), indicating that synthetic sequences recapitulated and/or exaggerated the TFBS grammar present in the most specific DHSs. Notably, many TFBSs were not uniquely associated with a single cell type or even a single biological component (**Figure 3.5A**, **Supplementary Figure 3.15**). This points to a complex TFBS-driven enhancer grammar where specificity is not simply conferred by individual TFBSs.

Next, we investigated which of these sequence features could explain MPRA-measured enhancer performance. Corroborating our previous work[78], a simple yet strong predictor of enhancer activity was the total number of TFBSs per sequence: synthetic sequences had higher average TFBS counts than DHS-sourced controls (**Supplementary Figure 3.13A**), which generally showed weaker performance (**Figure 3.2**, **Figure 3.3**). Furthermore, when including enhancers designed for tunable expression, TFBS counts clearly correlated with both on-target activity and mingap score (**Supplementary Figure 3.16**). To identify specific TFBSs that enhanced or hindered performance, we compared TFBS usage between the top and bottom 10% of all tested single-target synthetic enhancers ranked by mingap score (**Figure 3.5B**, **Appendix B**). For many

target cell types – including HepG2 (liver), WERI-Rb1 (retina), SJCRH30 (muscle), SK-N-SH (neuron), and MCF7 (breast) – top performing sequences used the same TFBSs as low-performing ones, but at higher densities (**Figure 3.5B**). However, higher TFBS density was not always beneficial. For example, while GATA1/2 and TAL1 were required for specificity in K562 (myeloid), higher density was associated with broad off-target expression.

However, for most cell types a single TFBS was not implicated as driving cell type-specific activity alone; rather, combinations of TFBSs were generally required. For example, HNF4A/G sites did not drive HepG2 expression unless accompanied by SOX8/9/10 or FOXA1/2 (**Supplementary Figure 3.17A**). Similarly, MYOD1/MYOG induced SJCRH30 activity only in combination with SOX8/9/10 (**Supplementary Figure 3.17B**). Combinations of NANOG + TEAD (NT2-D1) and GRHL1/2 + FOXA1/2 (MCF7) were associated with specificity, even when NANOG and GRHL1/2 alone were insufficient (**Supplementary Figure 3.17C-D**). Conversely, AP-1 TFBSs showed both beneficial and detrimental effects. In GM12878 (lymphoid), while top enhancers included AP-1 motifs, bottom-ranked sequences contained them at an even higher frequency (**Figure 3.5B, Supplementary Figure 3.17E**). In HeLa (cervix), however, AP-1 TFBSs were almost exclusively associated with poor performance and broad off target expression (**Supplementary Figure 3.17F-G**). For 786-O (kidney),

where no successful enhancers were generated, candidate sequences included either AP-1, HNF1A/B, or PAX TFBSs, each linked to undesirable profiles: AP-1 led to broad off-target activity, HNF1A to off-target activity in HepG2, and PAX motifs to little to no activity (**Figure 3.5B**). To test causality, we assayed sequences with select TFBSs embedded into inert background sequences (**Supplementary Figure 3.18A**). These experiments recapitulated several of the observed effects: broad activity from AP-1 except in GM12878 (**Supplementary Figure 3.18B-C**), HNF1A-driven expression in HepG2 but not 786-O despite its enrichment in both digestive and kidney DHSs (**Supplementary Figure 3.18D**), specificity conferred by GATA1::TAL1 in K562 and NEUROD1 in WERI-Rb1 (**Supplementary Figure 3.18E-F**), insufficiency of GRHL2, MYOD1, and HNF4A to drive MCF7, SJCRH30, and HepG2 activity (**Supplementary Figure 3.18G-I**), and the ability of FOXD3 and FOS to rescue HNF4A activity in HepG2 (**Supplementary Figure 3.18J-M**). Overall, these results highlight the varied and complex modes of TFBS effects on gene expression across cell types.

Finally, we used explainable AI methods to investigate what aspects of the regulatory grammar learned from genomic accessibility successfully translated into enhancer function. We developed DHS64-MPRA, a DL predictor of MPRA-measured enhancer activity, by finetuning DHS64 on our MPRA data (**Supplementary Figure 3.19**, **Appendix B**). Using both DHS64 and DHS64-MPRA, we then computed the

contributions of each nucleotide in every tested sequence towards accessibility and enhancer activity predictions in all assayed cell lines (**Appendix B**). Sequence regions with high contributions generally aligned with known TFBSs, suggesting that this approach could independently identify TFBSs and their cell type-specific effects (**Figure 3.5C-D**). Moreover, in high-performing sequences, we observed strong concordance between contributions from the accessibility and MPRA models (**Figure 3.5C**), indicating that the regulatory grammar learned from DHSs translated effectively into enhancer function. In contrast, poorly performing sequences showed noticeable divergence between the two models, highlighting regulatory features that failed to drive expression or drove expression too broadly despite predictions of cell type-specific accessibility (**Figure 3.5D**). To characterize these features, we calculated the average contribution of each TFBS towards accessibility and enhancer activity in each cell type (**Figure 3.5E**, **Supplementary Figure 3.20A-B**, **Appendix B**). While many TFBSs featured strong agreement between accessibility and enhancer contributions (e.g. MYOG/MYOD1 for SJCRH30 (muscle), HNF4A/G for HepG2 (liver), and POU5F1/NANOG for NT2-D1 (embryonal)), other TFBSs – such as those for AP-1 and the TP53 family – exhibited broad contributions to enhancer activity even when their accessibility contributions were more cell type-specific. Alternatively, TFBSs such as PAX contributed strongly to accessibility (e.g., in 786-O) but had negligible impact on enhancer activity, helping

explain the lack of functional enhancers in that cell line. Another class of TFBSs, typified by GRHL1/2 (enriched in the best MCF7 (breast) enhancers), showed strong qualitative agreement, but their relatively weaker contributions to enhancer activity explained the unexpectedly low expression in these otherwise well-performing enhancers. Despite these discrepancies, accessibility contributions of TFBSs were generally predictive of their cell type-specific effects on enhancer activity. Overall, our analysis indicates that our design approach effectively captures and amplifies much of the regulatory grammar underlying cell type-specific accessibility into functional enhancer activity, though certain motif elements remain challenging to translate effectively.

3.6 Cell type-specificity of the same TFBS can vary with the presence of different partner motifs

As previously noted, many TFBSs deployed in our sequence designs were not uniquely associated with accessibility or enhancer activity in a single cell type (**Figure 3.5A,E**). Of particular interest are the TFBSs for the TEAD family, which show strong SHAP contribution towards the activity of most measured cell lines, as well as enrichment in enhancers specific to both HeLaS3 and NT2-D1 (**Figure 3.5B,E**). Therefore, we sought to understand if and how sequence context influences the contribution of these TFBSs to

enhancer specificity in different cell lines, as well as to what extent these complex dependencies are captured by our models.

As a first step, we identified a strongly HeLaS3-specific sequence (HeLaS3 DEN #133, HeLaS3 mingap score = 3.99) containing two TEAD sites with high predicted HeLaS3 contribution. Keeping these TEAD sites constant, we iteratively mutated the remaining sequence to decrease the predicted HeLaS3 contribution of the TEAD TFBSs, until convergence was reached (**Appendix B**). We found that mutations disrupting three distinct KLF TFBSs largely destroyed the HeLaS3 contribution of both TEAD instances, and hypothesized that the KLF and TEAD sites are involved in a cooperative relationship contributing to HeLaS3 specificity (**Supplementary Figure 3.21A-C**). In support of this, we find that among all sequences measured in our library, those containing TEAD and KLF exhibit significantly higher HeLaS3 specificity compared to those containing either without the other (**Supplementary Figure 3.21D**). Additionally, the average SHAP contribution of TEAD sites towards HeLaS3 specificity is significantly greater in sequences containing KLF sites, whereas contribution towards all other cell types decreases or remains comparable; and the same is true of KLF contribution in sequences with or without TEADs (**Supplementary Figure 3.21E-F**).

To explain TEAD enrichment in NT2-D1-specific sequences, we hypothesized the existence of a different partner motif capable of altering the activation patterns of TEAD

sites. We find strong enrichment of both NANOG and SOX alongside TEAD in top NT2-D1 enhancers. However, among sequences excluding SOX motifs, those containing both TEAD and NANOG have significantly higher median NT2-D1 specificity than sequences containing only one motif of the pair (**Supplementary Figure 3.21G**). The reverse is not true when examining TEAD and SOX associations in sequences excluding NANOG. Moreover, only NANOG has positive average SHAP contribution to NT2-D1 specificity; and TEAD contribution towards NT2-D1 mingap is significantly greater in sequences containing NANOG sites than those without (**Supplementary Figure 3.21H-I**). Thus, our model and data provide evidence for two distinct partner TFBSs—KLF and NANOG—that can convert the cell type-specificity of the TEAD TFBS in opposing directions.

CHAPTER 3. FIGURES

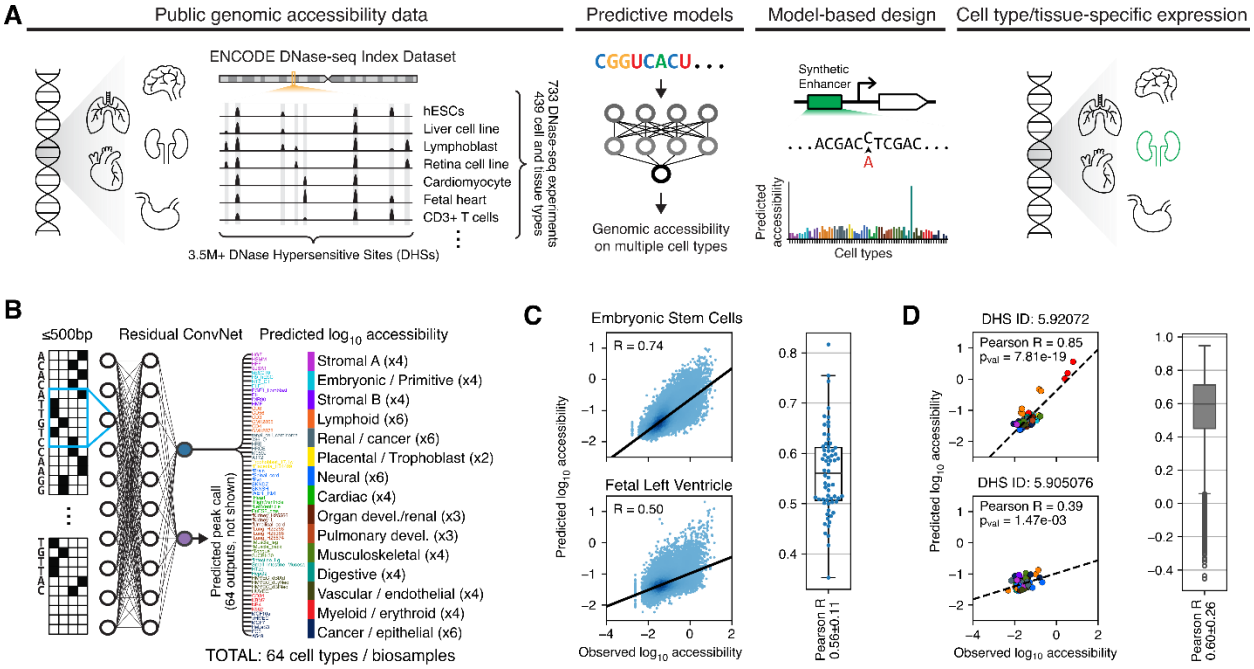


Figure 3.1. Neural network models of genomic accessibility enable programming cell type-specific gene expression.

A, Outline of our sequence design method. NN models trained on public genomic accessibility data are used to design enhancers with cell and tissue type-specific activity. **B,** High-level schematic of the DHS64 model. Predicted output cell types are colored and categorized by their associated "DHS component"[68], roughly corresponding to distinct originating tissue types within the DHS Index (**Appendix B**). See **Supplementary Figure 3.2** for a more detailed depiction of the model architecture. **C,** Per-biosample model performance. Left: examples of regression (top) and classification (bottom) performance for two biosamples. Right: Distribution of regression (Pearson R) and classification (AUPRC) across all 64 biosamples. **D,** Per-DHS model performance. Left: performance on two example DHSs, where each dot corresponds to a biosample. Right: Pearson R distribution across all DHSs in the test set. In **(C)** and **(D)**, we evaluate the model against a test set of held-out DHSs active in ≤10 biosamples and with enhancer-like genomic and chromatin annotations (**Appendix B**).

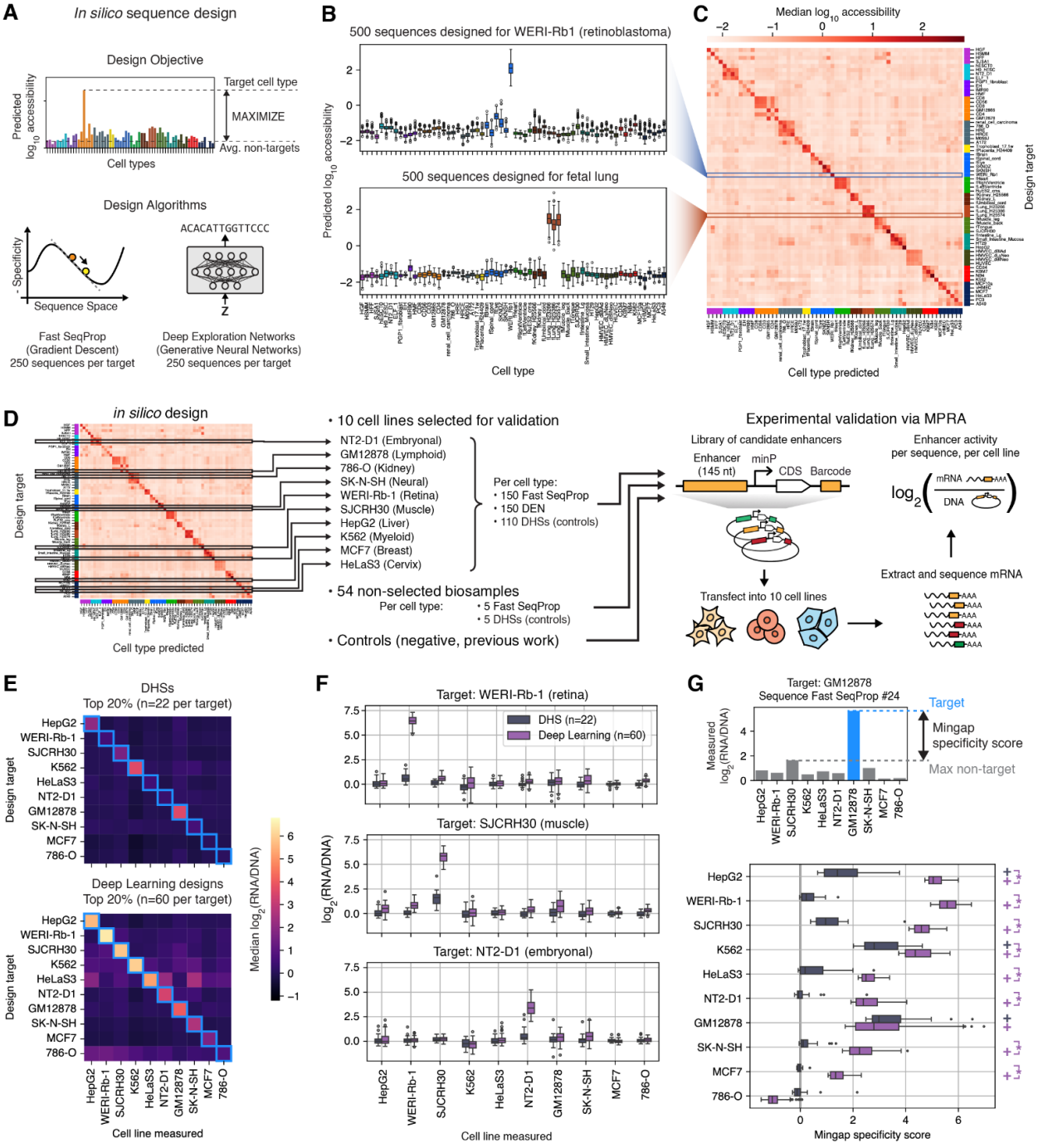


Figure 3.2. Synthetic sequences optimized for accessibility function as cell type-specific enhancers.

A, Sequence design procedure, illustrating the specificity objective function and the two NN-based sequence generation methods. **B,** Predicted accessibility signals for 500 sequences (250 generated with Fast SeqProp + 250 with DEN) designed to be specific to WERI-Rb-1 (a retinoblastoma cell line, top) and a fetal lung tissue (bottom). Predictions were obtained using a separately trained model from the ones used during sequence design (**Appendix B**). **C,** Predicted accessibility of sequences targeted to each modeled biosample, with rows showing the median of 500 sequences per target. **D,** Experimental validation via MPRAs in a panel of 10 cell lines. **E,** Measured enhancer activity of DHS-sourced (top) and NN-designed (bottom) sequences targeted to each assayed cell line. Rows represent the median of sequences targeted to each cell line. Light blue squares indicate the intended targets. **F,** Distribution of measured enhancer activities for sequences targeting three representative cell lines. **G,** Mingap specificity scores[75] of DHS-sourced and NN-designed sequences. Top: illustration of mingap score calculation from enhancer activity measurements of an example sequence. Bottom: score distributions by target cell line and sequence source. Plus signs indicate significantly positive median scores (Wilcoxon test, one-sided, Bonferroni-corrected p-value < 0.05). Brackets with asterisks denote whether NN-generated sequences have significantly higher medians than DHS-derived sequences (Mann-Whitney U test, one sided, Bonferroni-corrected p-value < 0.05). In **(E)**, **(F)**, and **(G)**, only the top 20% of sequences by mingap score are used. See **Supplementary Figure 3.6** and **Supplementary Figure 3.7A** for distributions of all sequences and controls.

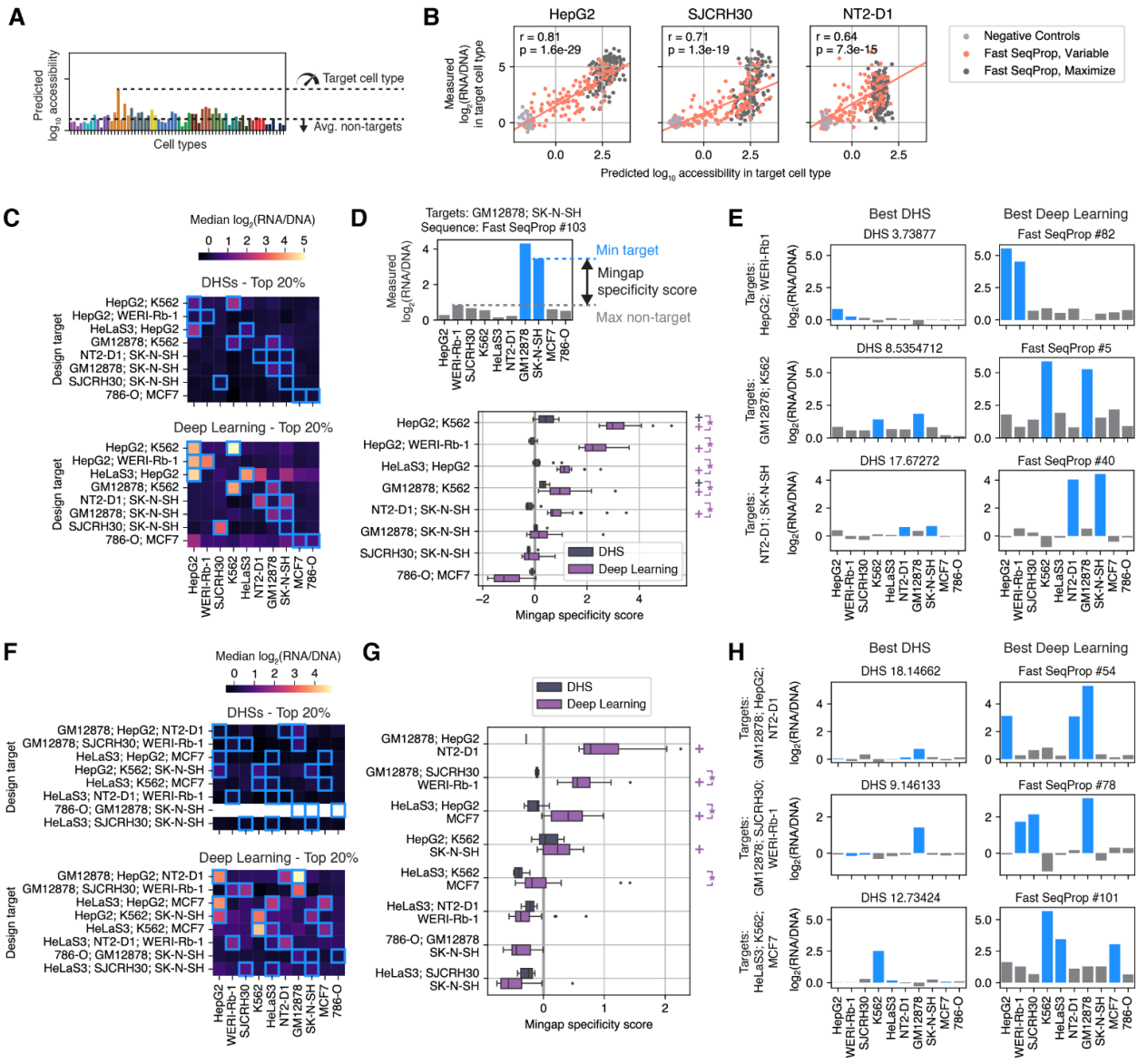


Figure 3.3. Synthetic enhancers achieve complex design objectives such as tunable activity and multiple cell type targets.

A, Schematic of the objective function used to design enhancers with tunable target activity. **B,** Relationship between predicted target accessibility and measured target enhancer activity for sequences designed for maximal (i.e. **Figure 3.2**) and tunable target activity, alongside negative controls, for three target cell lines. Linear regression fits, Pearson r coefficients, and p-values were calculated using tunable enhancers only (pink markers). See **Supplementary Figure 3.10** for expression and mingap scores of sequences targeting every cell line. **C,** Median expression of DHSs-sourced and NN-designed sequences targeting the indicated cell line pairs. **D,** Distribution of mingap

scores. Top: mingap score calculation with multiple target cell types. Bottom: score distributions by target pair and sequence source. Plus signs indicate significantly positive median scores (Wilcoxon test, one-sided, Bonferroni-corrected p-value < 0.05). Brackets with asterisks denote significantly higher medians for NN-generated sequences compared to DHS-derived sequences (Mann-Whitney U test, one sided, Bonferroni-corrected p-value < 0.05). **E,** Enhancers with the highest mingap scores for the indicated cell line pairs. Light blue bars correspond to target cell lines. **F,** Median expression of sequences targeted to the indicated cell line triplets. For the 786-O; GM12878; SK-N-SH triplet, no suitable DHSs were found. **G,** Distribution of mingap scores for the indicated cell line triplets. Markers on the right follow the conventions in **(D)**. **H,** Enhancers with the highest mingap scores for the indicated cell line triplets. In **(C)**, **(D)**, **(F)**, and **(G),** only the top 20% by mingap score on each set are shown.

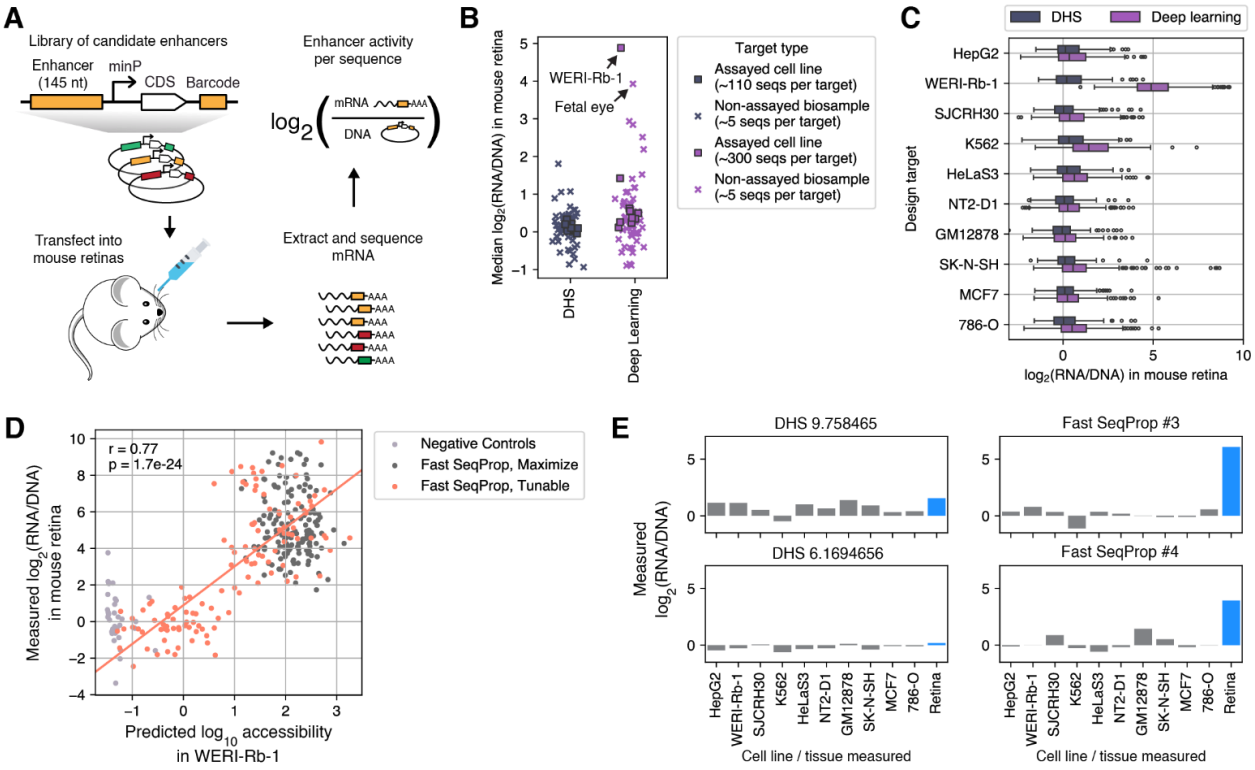


Figure 3.4. Synthetic enhancers designed for eye-related biosamples function in mouse retinas.

A, Schematic of the mouse retina MPRA. **B,** $\log_2$(RNA/DNA) of DHS-sourced and NN-designed sequences targeted to all DHS64 modeled biosamples. Markers indicate the median across sequences with the same target. For the 10 targets corresponding to assayed cell lines, we measured hundreds of sequences per target and sequence source (**Appendix B**). For the remaining 54 biosamples, ~5 sequences per target and sequence source were included (**Supplementary Figure 3.9**). The exact number of sequences per target varies slightly due to sequencing dropout (**Appendix B**). **C,** Distribution of mouse retina $\log_2$(RNA/DNA) measurements for sequences targeting the 10 cell lines. Each marker represents an individual sequence. **D,** Tunable WERI-Rb-1 enhancers: comparison between predicted WERI-Rb-1 accessibility and measured mouse retina $\log_2$(RNA/DNA). Each marker represents an individual sequence. **E,** $\log_2$(RNA/DNA) of four sequences targeting the fetal eye biosample, measured across all assayed cell lines and the mouse retina. A total of five DHSs and five synthetic sequences were tested. Plots show the two DHSs (left) and two synthetic sequences (right) with the highest difference between retinal activity and the maximum observed cell line activity.

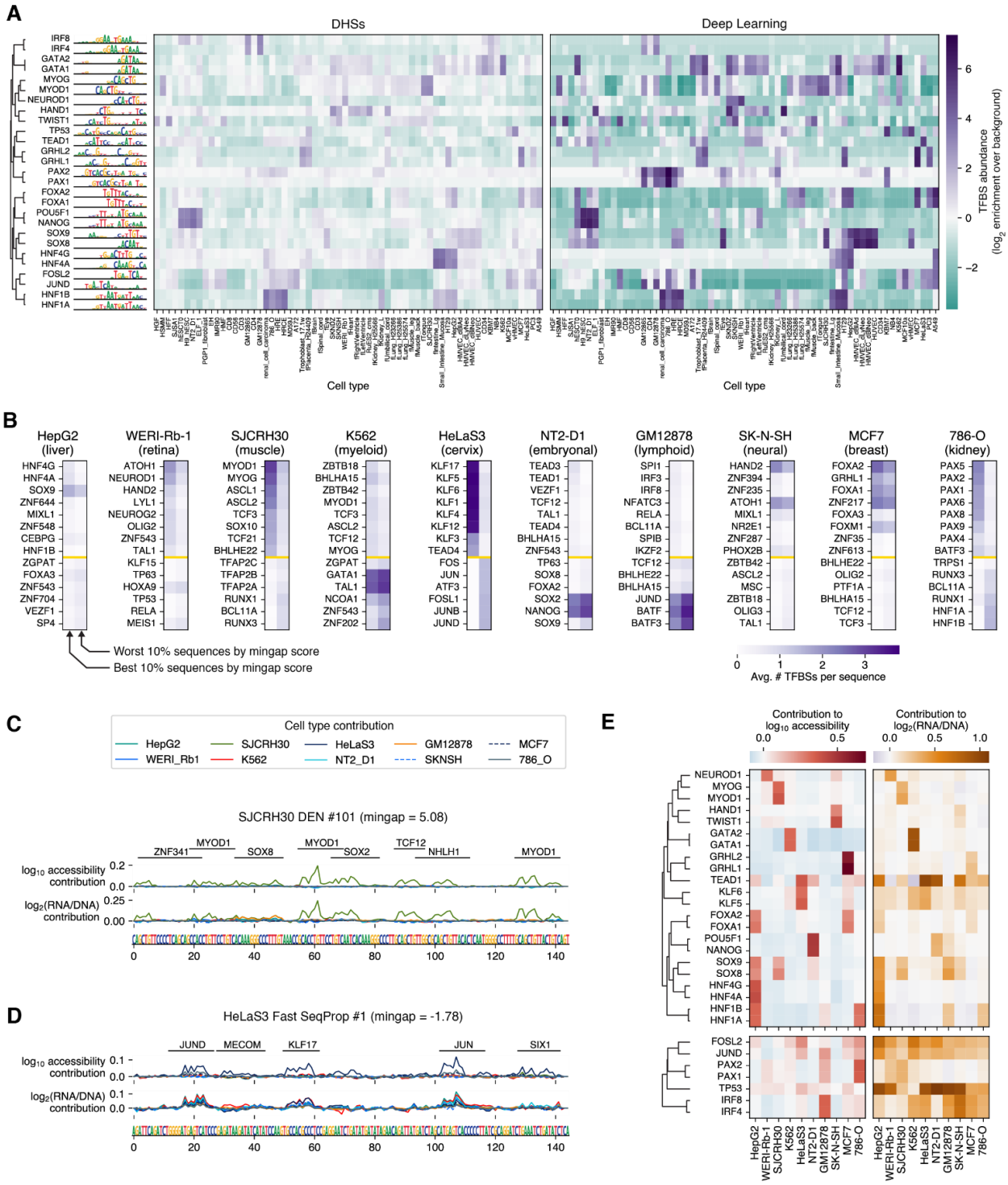
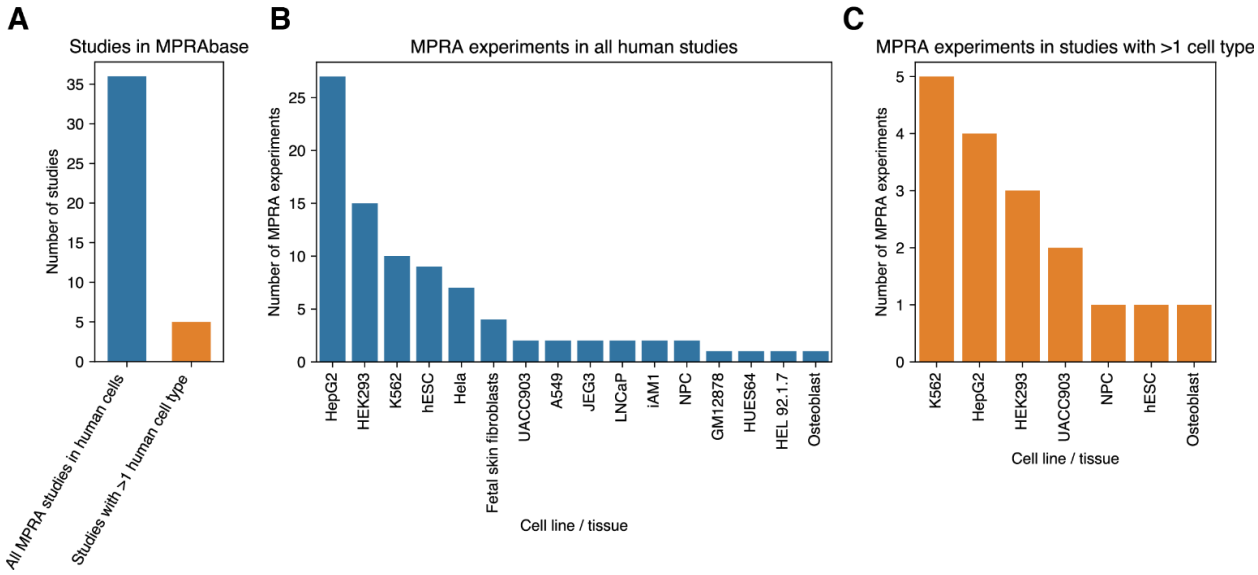


Figure 3.5. Transcription Factor Binding Site (TFBS) grammar from DHSs is captured and amplified in synthetic enhancers.

A, TFBS utilization in DHSs (top 1,000 per cell type by specificity, left) and in Deep Learning-designed sequences (500 per cell type, right) targeted to each DHS64-modeled biosample. Utilization was calculated as the $\log_2$ enrichment in the number of TFBSs per sequence relative to a background set of 1,000 randomly sampled DHSs (**Appendix B**). TFBSs (rows) were manually selected and clustered by PWM similarity (**Appendix B**). See **Supplementary Figure 3.15** for an expanded set. **B,** TFBSs associated with the best and worst-performing enhancers in cell line MPRAs (**Figure 3.2D**). For each target cell line, 300 designed sequences were stratified by mingap score, and the top and bottom 10% were compared. Shown are TFBSs with the largest differences in their frequency per sequence. TFBSs above the yellow line are enriched in top-performing sequences; those below are enriched in poorly performing ones. **C,D** Sequence features of two example sequences. From top to bottom: TFBSs identified via FIMO (**Appendix B**), nucleotide-level contributions to predicted accessibility in all 10 cell lines, contributions towards predicted enhancer activity (i.e. $\log_2$(RNA/DNA)), and sequence. Contributions were calculated using the DHS64 model for accessibility and a DHS64-MPRA model finetuned on MPRA measurements for enhancer activity (**Appendix B**). **E,** Average contribution of TFBSs towards accessibility (left) and enhancer activity (right). Top panels: TFBSs for which accessibility and activity contributions are positively correlated across cell lines. Bottom panels: TFBSs where this relationship is weak or inconsistent. See **Supplementary Figure 3.20** for an expanded analysis.

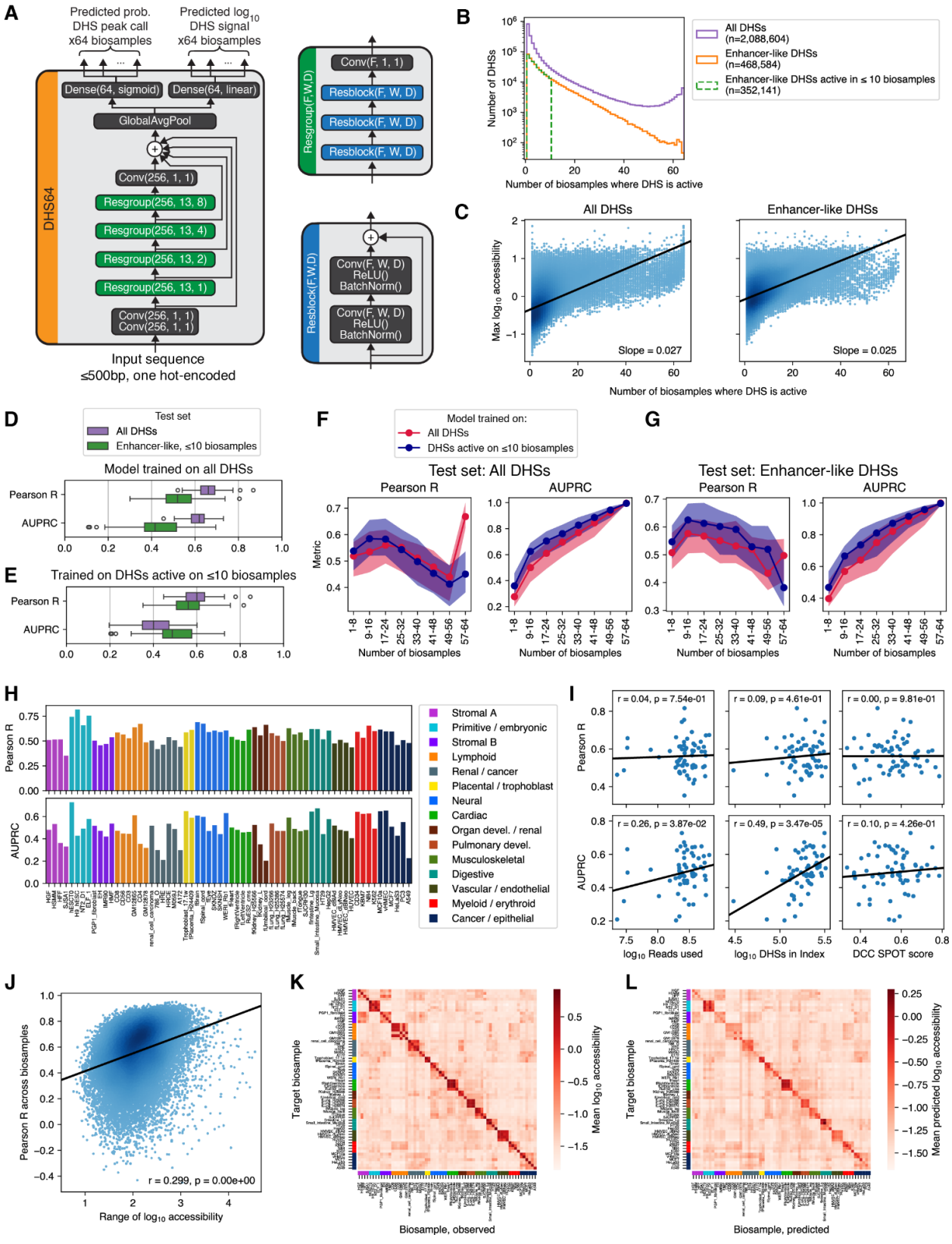
CHAPTER 3. SUPPLEMENTAL INFORMATION

3.1 Supplemental Figures



Supplementary Figure 3.1. Statistics on publically available MPRAs collected in MPRAbase.

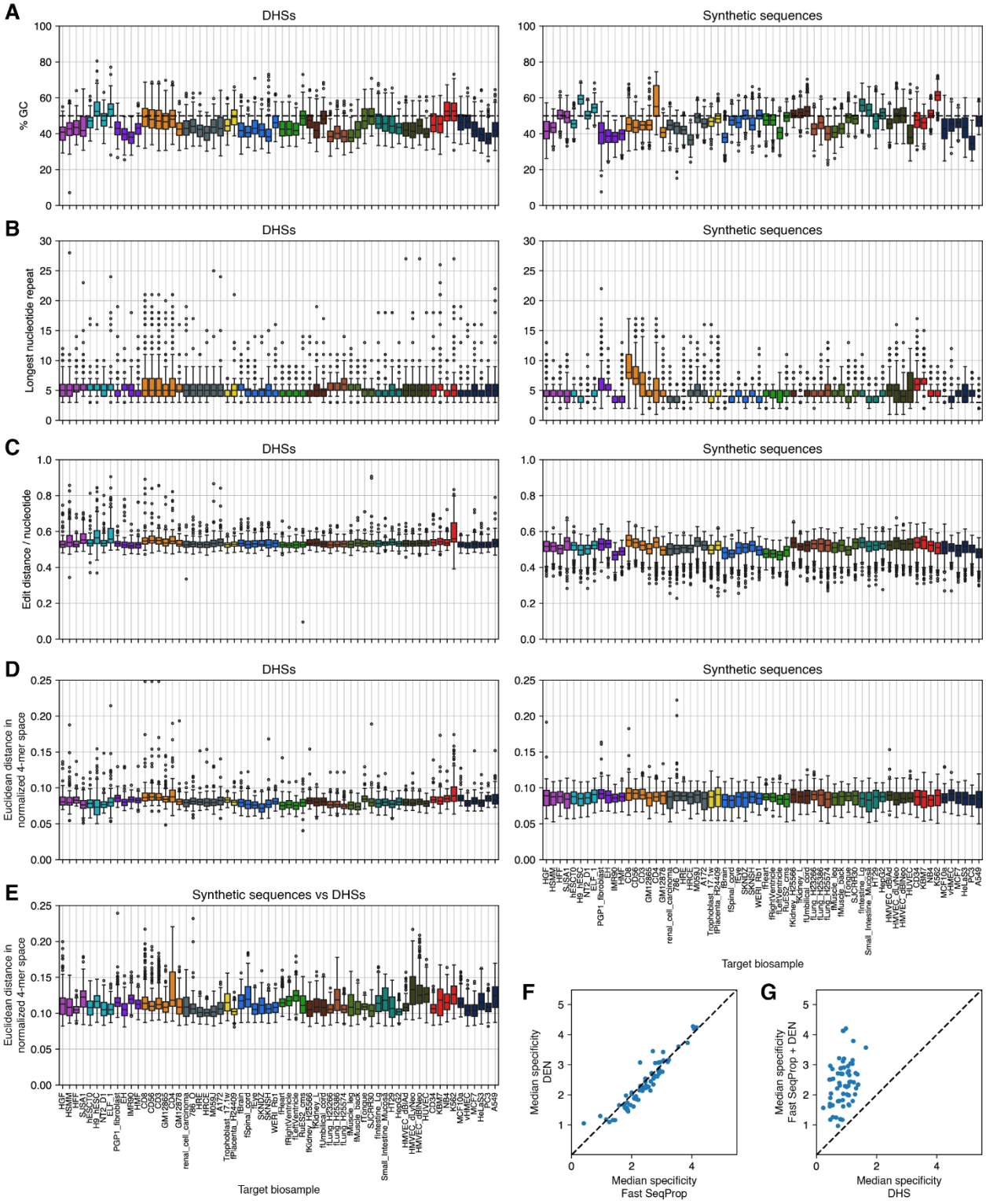
(A) Comparison of number of MPRA experiments collected in MPRAbase assayed in only 1 cell lines vs. in more than one cell line. **(B)** Number of MPRA experiments collected in MPRAbase assayed in different cell types. **(C)** Same as **(B)**, but restricted to only the experiments conducted simultaneously in more than 1 cell line.s



Supplementary Figure 3.2. DHS64 model performance.

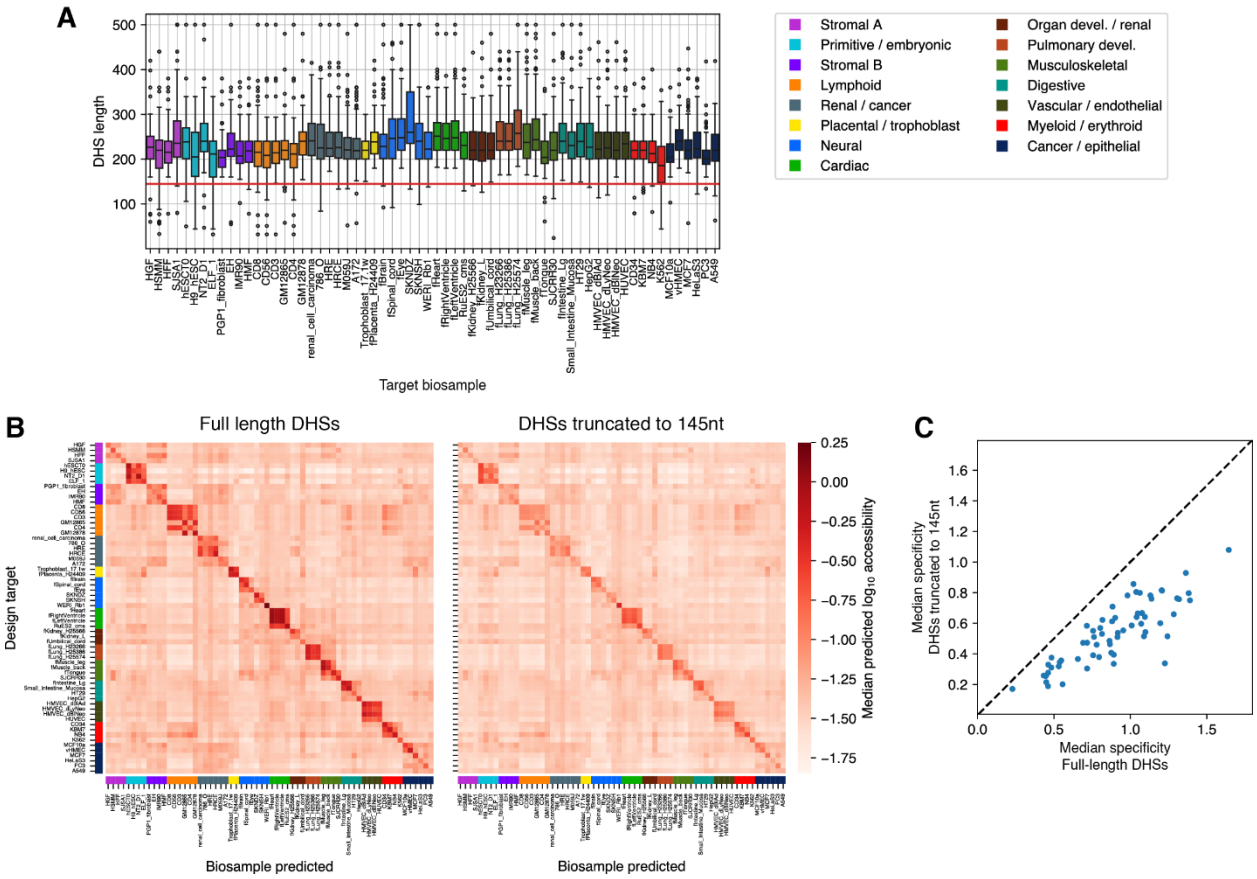
Plots were generated on models trained on chromosome split #3 (**Appendix B**). Results are similar for the other two DHS64 models used in this manuscript (splits 0 and 1). **(A)** Model architecture. Input is the sequence of a candidate DNase-hypersensitive site (DHS). Outputs are peak call probabilities (classification) and $\log_{10}$ accessibility signals (regression) for all 64 modeled cell types / biosamples. Conv(F, W, D): Convolutional layer with F filters of width W and dilation rate D. Dense(U, A): Dense layer with U units and activation function A. **(B)** Distribution of the number of biosamples in which DHSs are accessible (i.e. positive peak call) in any of the 64 modeled biosamples. Distributions are shown for all DHSs, "enhancer-like" DHSs (high "mean_signal" annotation, far from annotated transcription start sites, "enhancer" ChromHMM annotations, **Appendix B**), and enhancer-like DHSs active in ≤10 biosamples. Enhancer-like DHSs have a higher proportion of cell type-specific DHSs (low # biosamples) compared to broadly-accessible DHSs (high # biosamples). **(C)** Maximum accessibility signal across biosamples (y axis) for each DHSs, against the number of biosamples in which the DHS is active. Accessibility is generally higher in broadly accessible DHSs, even in the "enhancer-like" subset. Linear regression slopes shown are in units of $\log_{10}$(accessibility signal) per additional biosample. **(D)** Per-biosample performance of a model trained on all available DHSs, tested against all held-out DHSs or enhancer-like DHSs active in ≤10 biosamples. This model does worse on the latter set, possibly in part because broadly accessible DHSs have a higher signal, contributing disproportionally to error metrics. **(E)** Per-biosample performance of a model trained on DHSs filtered to be active in ≤10 biosamples. There is a decrease in performance when evaluating against all held-out DHSs, but an improvement when testing on enhancer-like DHSs active in ≤10 biosamples. This is our DHS64 model, and the results reported on enhancer-like DHSs active in ≤10 biosamples correspond to **Figure 3.1C**. **(F)** Per-biosample performance of both models on DHSs binned by their number of active DHSs. As expected, the model trained on DHSs active in ≤10 biosamples has better performance on DHSs active in fewer biosamples (i.e. more cell type-specific). **(G)** Per-biosample performance on enhancer-like DHSs binned by their number of active DHSs. The difference in performance is more dramatic. **(H)** Regression (top) and classification (bottom) per-biosample performance for each modeled biosample. **(I)** DNase-seq quality metrics of each biosample, such as the number of sequencing reads, number of called peaks, and Signal Portion of Tags (SPOT) scores, generally do not correlate with performance metrics. The exception is the number of called DHSs with respect to AUPRC, as the floor of this metric corresponds to the fraction of positive samples. **(J)** Per-DHS performance, as measured by the Pearson R across biosamples, is higher for DHSs in which their accessibility values span a wider range (max - min across all biosamples). **(K)**

Accessibility of the most specific DHSs. For each biosample, we selected 50 DHSs from the test set with the highest specificity score (signal in the target biosample minus average signal on all other biosamples, **Appendix B**), and plotted their average observed accessibility as a row. **(L)** Predicted accessibility of DHSs in **(K)**.



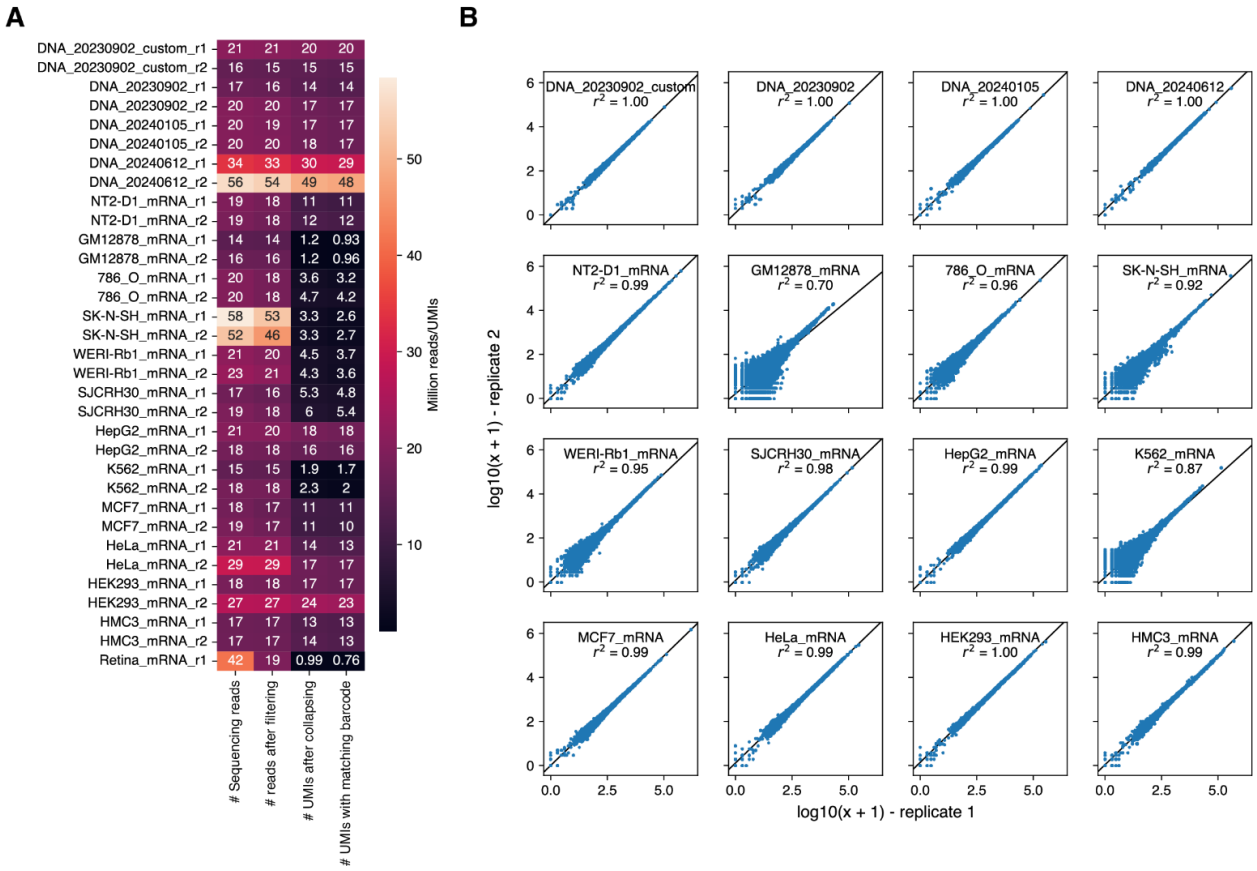
Supplementary Figure 3.3. Sequence features of DHSs and DHS64 designs.

Panels **(A)** to **(E)** compare, for each target biosample, the top 100 DHSs with the highest specificity (measured difference in $\log_{10}$ accessibility in the target biosample versus the average across non-targets) selected from DHSs with enhancer-like chromatin annotations and peak calls in ≤10 modeled biosamples, against 500 sequences (250 per method) designed via Fast SeqProp or DENs for optimized DHS64-predicted specificity (**Appendix B**). Boxes show the distribution of each metric across sequences selected or designed for a given target. **(A)** GC content. **(B)** Longest single nucleotide stretch. **(C)** Length-normalized edit distance between sequences within their own method- and target-specific set (**Appendix B**). **(D)** Euclidean distance of sequence 4-mer counts between sequences within their own method- and target-specific set (**Appendix B**) **(E)** Euclidean distance of sequence 4-mer counts between DHSs and Fast SeqProp-designed sequences targeting the same biosample (**Appendix B**) **(F)** Median predicted specificity of 250 Fast SeqProp- (x axis) and DEN- (y axis) designed sequences, for each of the 64 target biosamples. Predictions were obtained with the DHS64 model trained on data split #0, which was not used during sequence design. **(G)** Median predicted specificity of the top 100 enhancer-like DHSs per biosample (x axis) versus 500 NN-designed sequences (y axis).



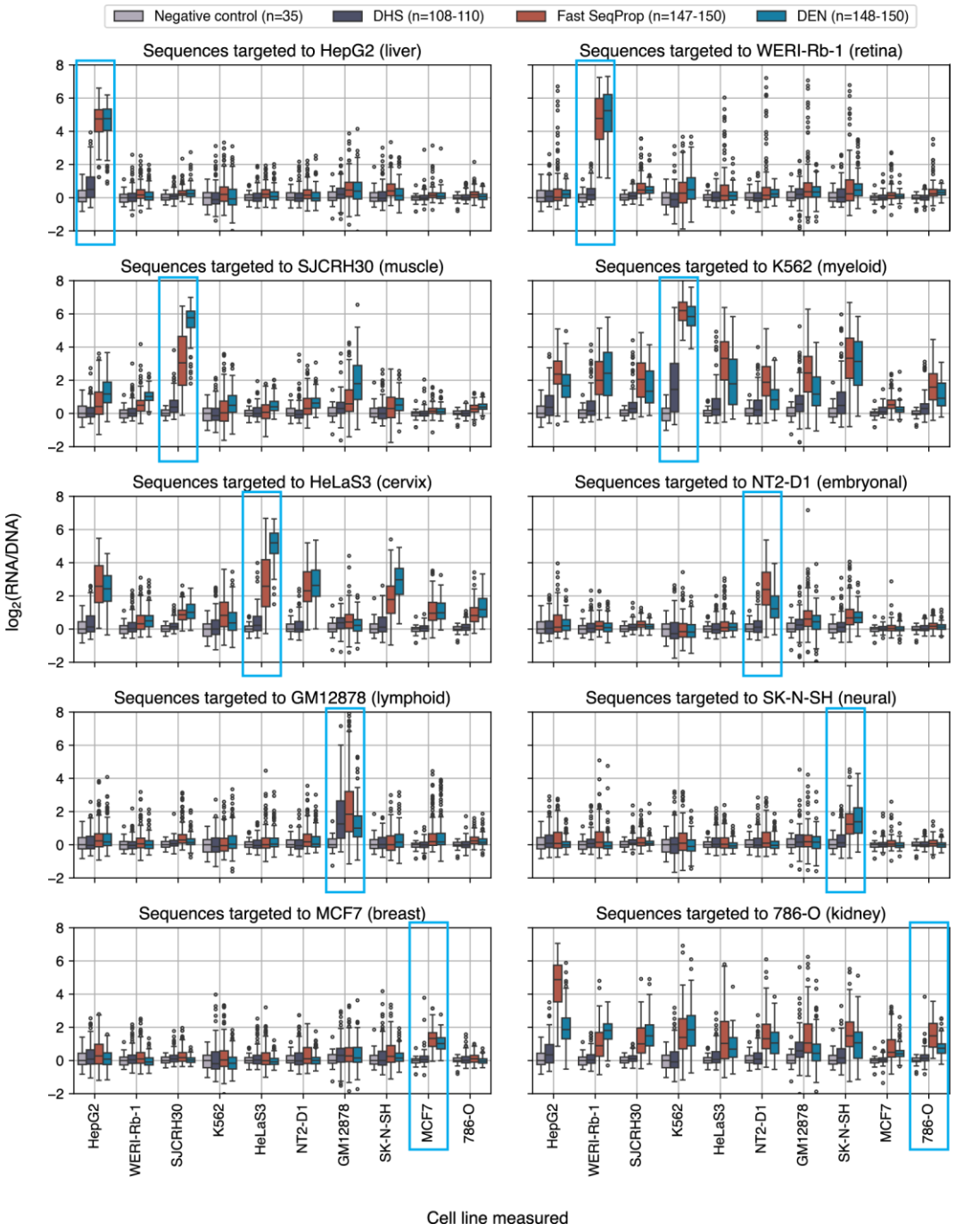
Supplementary Figure 3.4. Predicted effects of truncating DHSs to 145 nt.

For each target biosample, we selected the top 100 DHSs by specificity (measured difference in $\log_{10}$ accessibility in the target biosample versus the average across non-targets) from DHSs in the DNase I Index with enhancer-like chromatin annotations and peak calls in ≤10 modeled biosamples (**Appendix B**). **(A)** Annotated DHS lengths. The horizontal bar marks y=145 nt. **(B)** Predicted median $\log_{10}$ accessibility for full-length DHSs (left) and the same DHSs truncated to 145 nt (right). Each row represents 100 DHSs for a given target. **(C)** Predicted specificity of full length (x axis) versus truncated (y axis) DHSs. Each dot represents the median predictions for DHSs targeting a biosample. The dashed line indicates y=x.



Supplementary Figure 3.5. Sequencing quality metrics of all DNA and mRNA libraries sequenced for enhancer MPRAs.

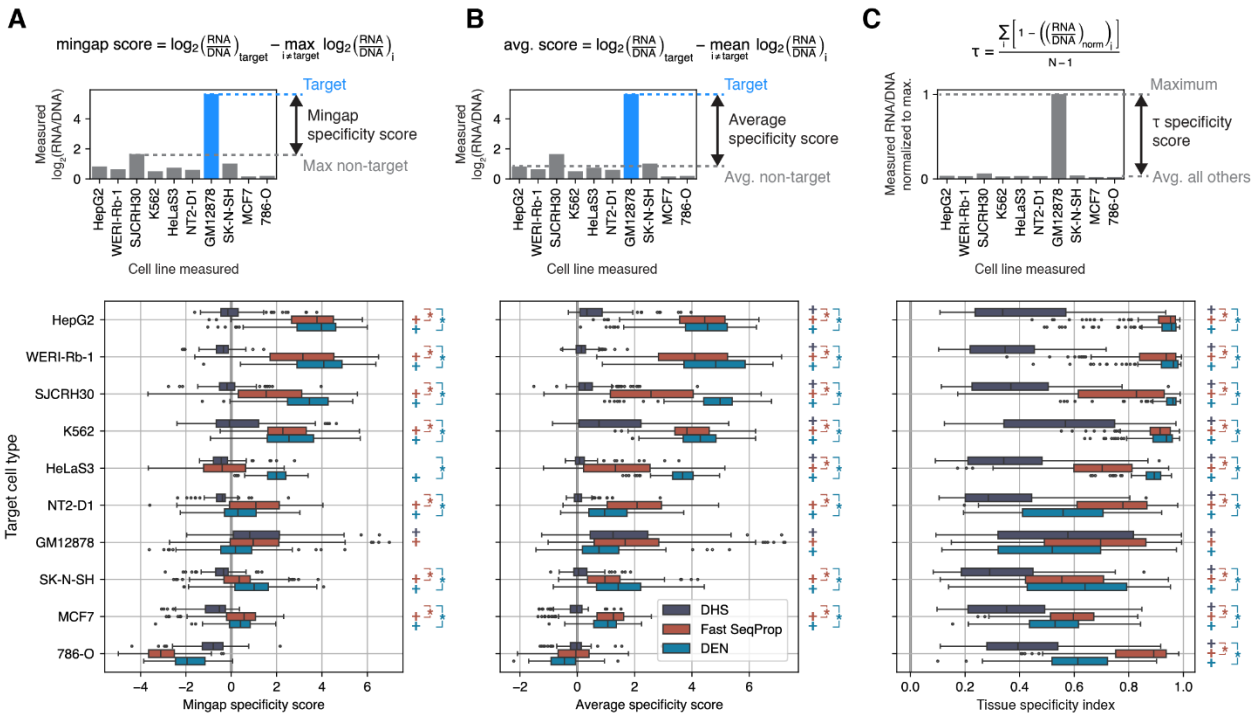
Four DNA library sequencing runs were performed: three for distinct DNA library preparations (DNA_20230902, DNA_20240105, and DNA_20240612) which were transfected into different cell lines, and one where custom Illumina adaptors and sequencing primers were used (DNA_20230902_custom), which were also used for some mRNA libraries. Two replicates of each DNA library (_r1 and _r2) were sequenced. For each tested cell line, mRNA extracts from two biological replicates were sequenced. For mouse retina, only one replicate was performed and sequenced (**Appendix B**). **(A)** Sequencing read counts, read counts after amplicon quality filtering, UMIs after collapse and clustering, and "useful" UMIs (barcodes matching expected sequences), in millions, for each replicate library. **(B)** Inter-replicate correlation of DNA and mRNA log-transformed counts. Least squares regression line and coefficient of determination are shown.



Supplementary Figure 3.6. Enhancer activity measurements of sequences targeting cell lines where enhancer MPRAs were performed.

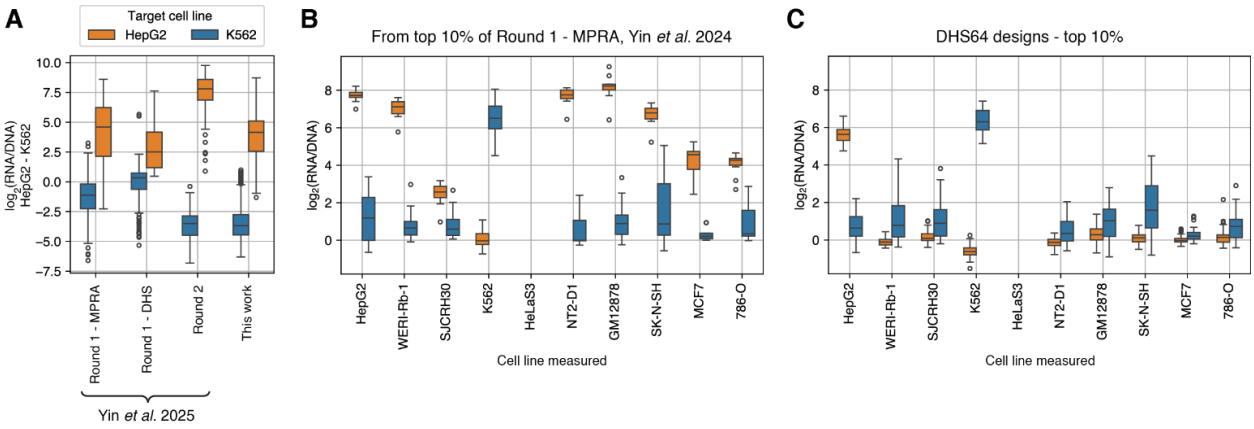
Each panel shows measurements across 10 cell lines (x axis) of sequences targeted towards one cell line only. Box colors indicate sequence source (DHS, Fast SeqProp, DEN, negative control). Due to sequencing depth filters (**Appendix B**), some panels contain

slightly fewer sequences than shown in **Figure 3.2D**. Thus, number of sequences per sequence source are indicated as ranges in the legend. Negative controls are the same across all panels. Light blue rectangles highlight the target cell type in each panel.



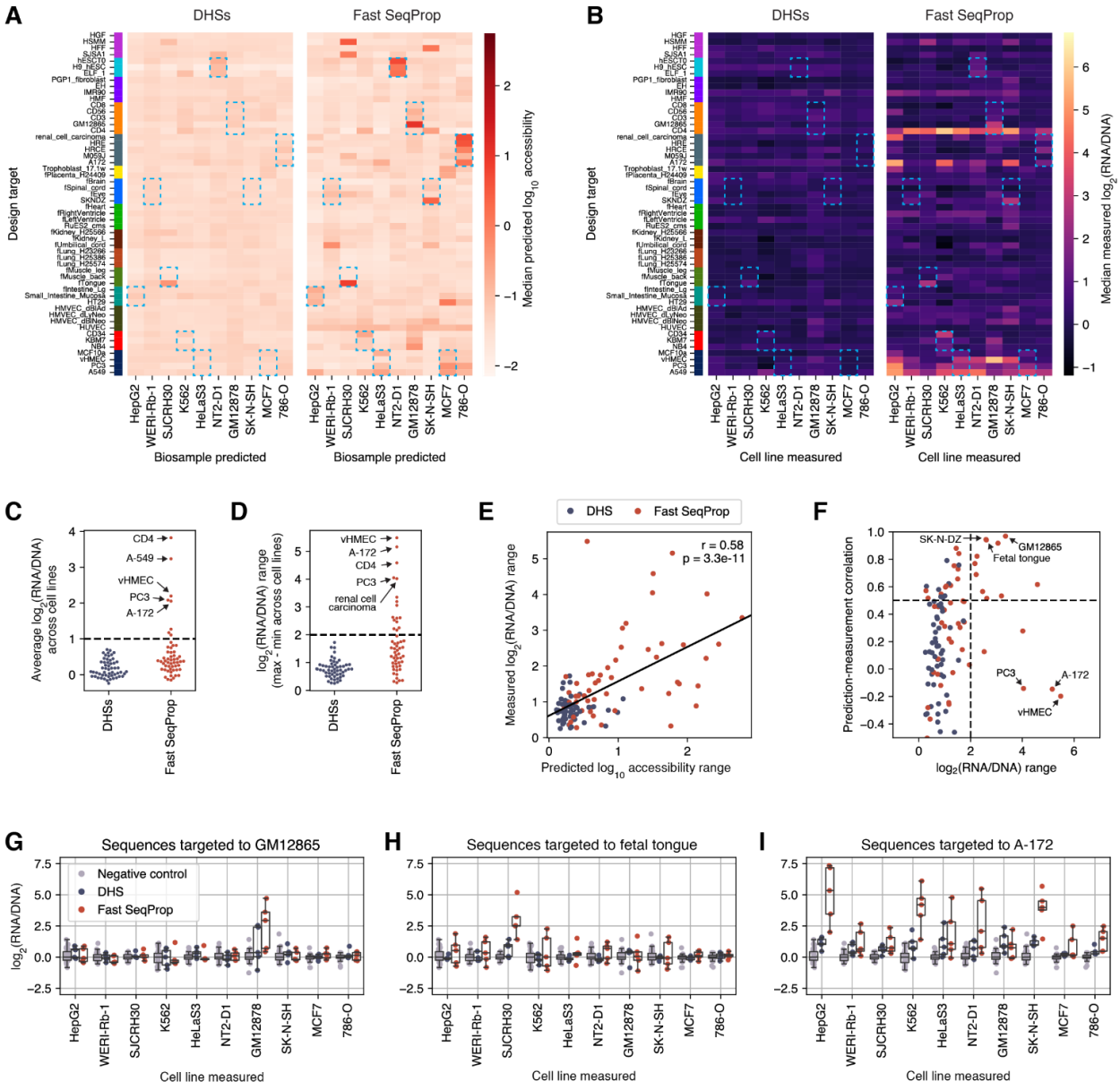
Supplementary Figure 3.7. Specificity scores of deep learning-designed and DHS-sourced enhancers.

Top: mathematical definition and graphical representation of specificity score calculations given cell line enhancer measurements of a sequence. Bottom: score distribution of all sequences grouped by source and target cell type. Plus signs on the right indicate whether the median of each box is significantly positive (Wilcoxon test, one-sided, Bonferroni-corrected p-value < 0.05). Brackets with asterisks denote whether Fast SeqProp- (red) or DEN- (blue) generated sequences have significantly higher medians than DHS-derived sequences (Mann-Whitney U test, one sided, Bonferroni-corrected p-value < 0.05) **(A)** Mingap scores[75]. Compared to **Figure 3.2G**, here we show all synthetic and DHS-sourced sequences, with synthetic sequences further separated by design method. **(B)** Average scores, computed using the mean non-target $\log_2$FC instead of the maximum, aligning more closely with the design objective we used to optimize sequences for biosample-specific accessibility (**Figure 3.2A-C**) **(C)** Tissue specificity index τ[76], calculated using RNA/DNA count ratios rather than $\log_2$-transformed ratios. Note that τ is based on the maximum expression across cell types, regardless of whether it corresponds to the intended target, and can therefore be misleading when specificity is obtained towards the incorrect cell type. For example, sequences targeting 786-O show high τ values despite higher off-target expression in HepG2 and other cell lines (**Supplementary Figure 3.6**).



Supplementary Figure 3.8. Comparison of enhancer designs from this study with those from Yin et al.[78].

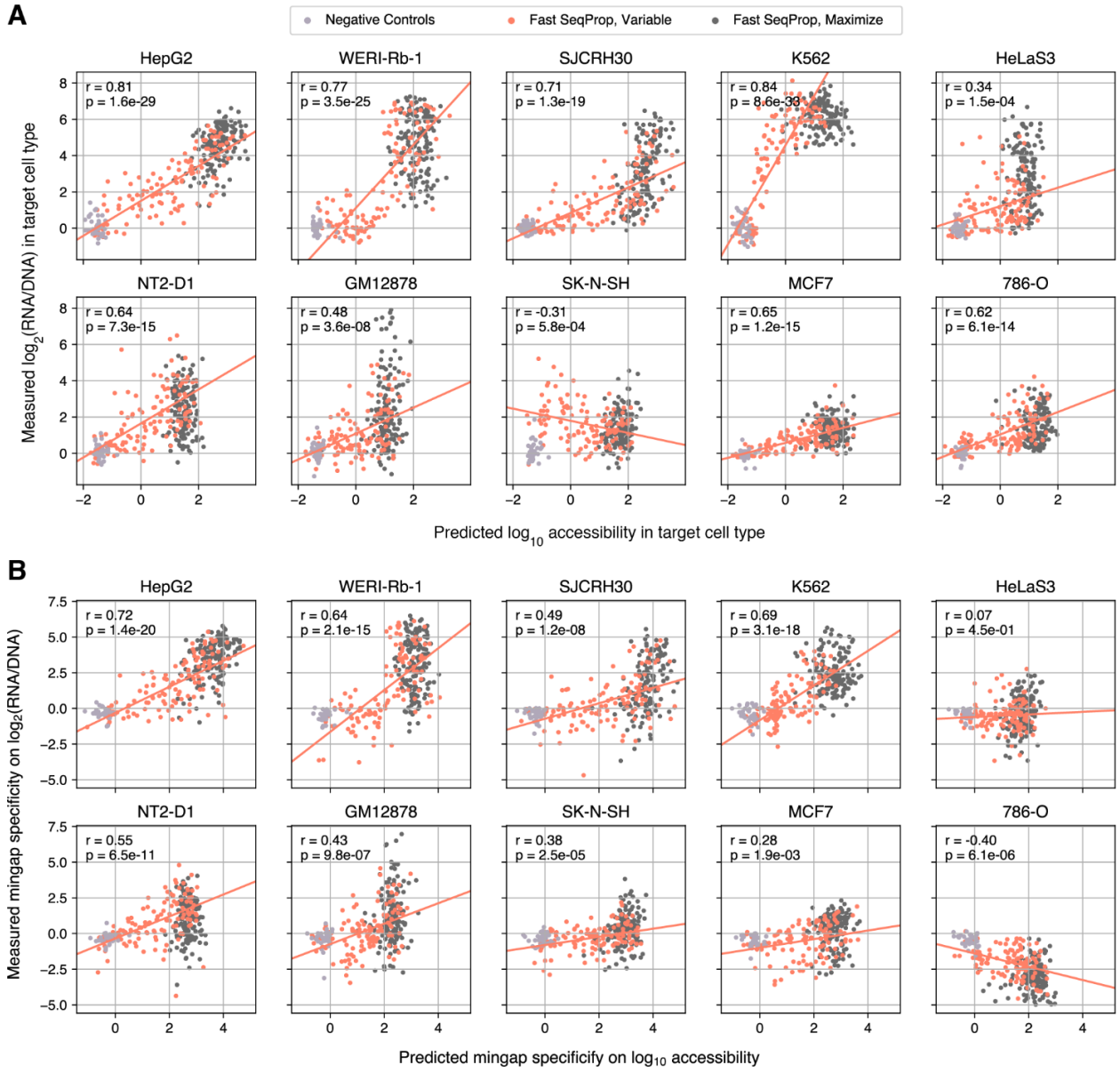
40 sequences from our previous study were re-synthesized and tested across 10 cell lines as part of the DHS64 MPRAs. These included 20 highly specific sequences from the SHARPR-MPRA dataset[79] used as CNN training data in Yin *et al.* 2025, and 20 sequences from the top 10% by specificity designed with those models (Round 1-MPRA), evenly split between HepG2- and K562-targeted sequences. New measurements of these sequences were used to batch correct the data from Yin *et al.* 2024 for comparability (**Appendix B**). **(A)** Distribution of HepG2/K562 specificity (difference in $\log_2$(RNA/DNA)) for sequences designed using various methods. From Yin *et al.* 2024: "Round 1-MPRA": sequences designed with CNNs trained on the SHARPR-MPRA dataset. "Round 1-DHS": sequences designed with a generative adversarial network (GAN) trained to produce DHS-like sequences, and a classifier of HepG2/K562 peak calls. "Round 2": sequences designed with CNNs trained on SHARPR-MPRA data and finetuned on both Round 1 measurements. "This work": sequences designed with DHS64, Fast SeqProp, and DENs to target HepG2 and K562. **(B)** Enhancer activity of 20 re-synthesized sequences from Round 1-MPRA in Yin *et al.* 2025 measured across 10 cell lines. HepG2-targeted enhancers show high off-target activity across multiple cell lines. **(C)** Enhancer activity of a comparable set of DHS64-designed sequences (top 10% by HepG2/K562 specificity only). These exhibit reduced off-target activity, though HepG2-targeted designs show lower on-target activity than Round 1-MPRA.



Supplementary Figure 3.9. Sequences targeting DHS64 cell types other than those tested in MPRAs.

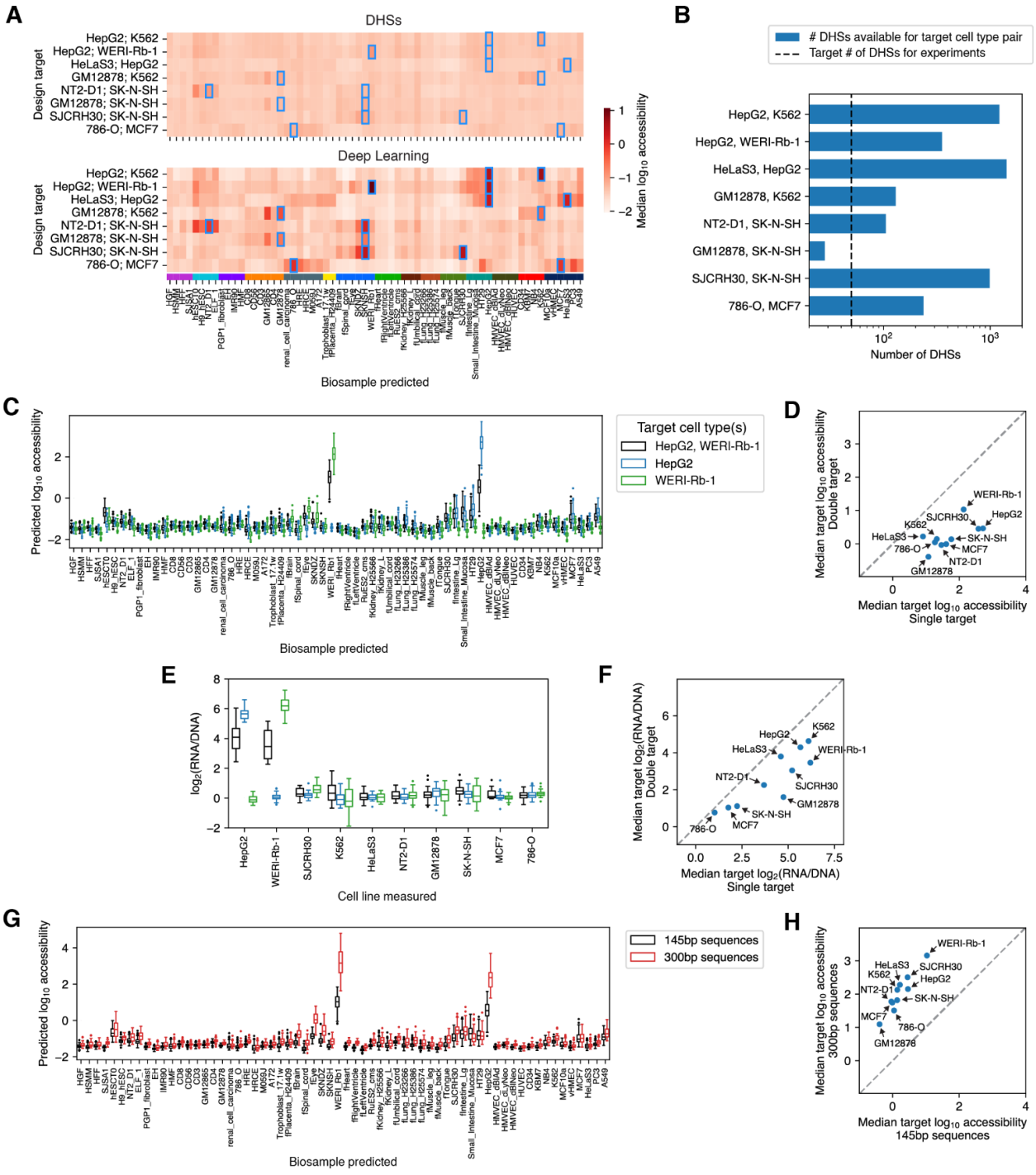
We included in our library 5 DHS-derived and 5 Fast SeqProp-designed sequences for each of the 54 targets. MPRA measurements were obtained for all but one DHS-sourced sequence targeting HT29 cells. **(A)** DHS64-predicted $\log_{10}$ accessibility. **(B)** Measured $\log_2$(RNA/DNA). In **(A)** and **(B)**, each row corresponds to the median across sequences for a given target. Blue dashed rectangles show where cell lines (columns) correspond to the same biological component as the target biosample (rows). **(C)** Average $\log_2$(RNA/DNA) across cell lines. **(D)** $\log_2$(RNA/DNA) range (maximum - minimum) across cell lines, indicating expression variability. In **(C)** and **(D)**, dots represent

sequences targeted to each biosample, calculated from the average **(C)** or range **(D)** of each row in **(B)**. Dots marking the highest Fast SeqProp values are labeled with their target biosample. **(E)** Correlation between predicted $\log_{10}$ accessibility range and observed $\log_2$(RNA/DNA) range. x and y coordinates of each dot were calculated from the range within each row in **(A)** and **(B)**. Least squares linear regression fit, Pearson correlation, and p-value are indicated. **(F)** Prediction-measurement correlation as a function of the $\log_2$(RNA/DNA) range across cell lines. Dots correspond to each target biosample. y values were obtained by performing linear regression on median predicted $\log_{10}$ accessibility and median measured $\log_2$(RNA/DNA) for each target, i.e. rows in **(A)** and **(B)**. Relevant thresholds and target biosamples from the main text are highlighted. **(G-I)** Measured $\log_2$(RNA/DNA) of sequences targeting biosamples with high variability across cell lines. Dots correspond to individual sequences. **(G-H)** Cases where variation was well predicted, and high expression was present in cell lines corresponding to the same biological component as the target. **(I)** A case where widespread high expression was not predicted.



Supplementary Figure 3.10. Synthetic enhancers designed for tunable target expression.

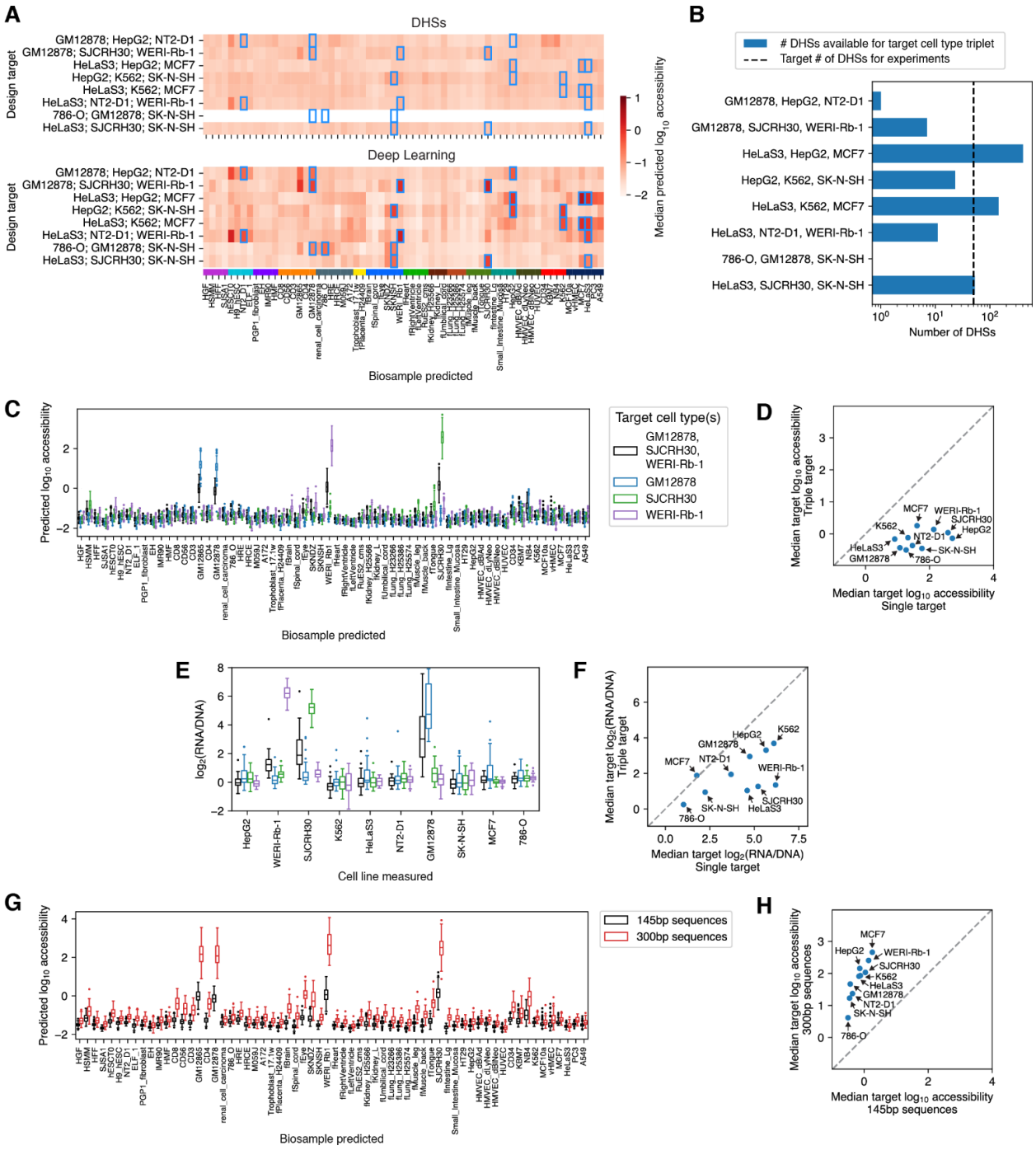
Relationship between predicted target accessibility and measured target $\log_2$(RNA/DNA) **(A)**, and between DHS64-predicted and observed mingap scores **(B)** for sequences designed for maximal (i.e. **Figure 3.2**) and tunable target activity, alongside negative controls, for all 10 target cell lines. Mingap scores predictions were calculated from accessibility predictions across the 10 measured cell lines. Observed scores were calculated from $\log_2$(RNA/DNA) values as in **Figure 3.2G**. Linear regression fits, Pearson r coefficients, and p-values were calculated using tunable enhancers only (salmon markers).



Supplementary Figure 3.11. Additional analysis on dual-target enhancers.

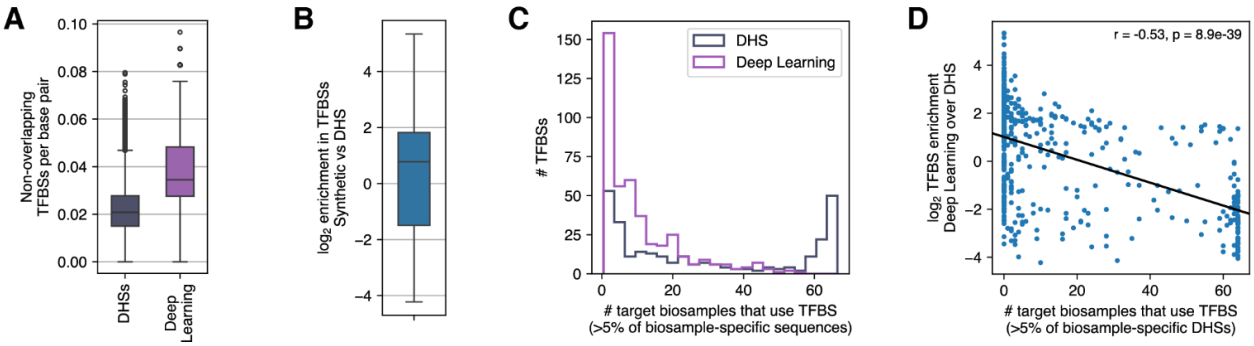
(A) DHS64-predicted $\log_{10}$ accessibility for DHS-derived (top) and Fast SeqProp-designed sequences (bottom), with rows representing median values for each target cell line pair. **(B)** Number of DHSs in the DNase I Index meeting our filtering criteria (enhancer-like chromatin annotations, peak calls in the target cell lines, and in $\leq$10 of the modeled

biosamples) for each target. **(C)** Predicted accessibility for synthetic sequences targeting the HepG2;WERI-Rb-1 pair and each cell line individually. **(D)** Median predicted accessibility on each cell line, for synthetic sequences targeting cell lines individually (x axis) or as part of a dual-target design (y axis). Markers below the diagonal indicate weaker predicted accessibility in a given cell line for sequences designed for multiple targets. **(E)** Measured enhancer activities of synthetic sequences targeted to the HepG2;WERI-Rb-1 pair and to each cell line individually. **(F)** Median enhancer activity of sequences designed to target each cell line, either individually (x axis) or as part of a dual-target design (y axis). In **(E)** and **(F)**, only sequences within the top 20% by mingap score per target are considered, as in **Figure 3.2** and **Figure 3.3**. **(G)** Predicted accessibility for 145 nt and 300 nt-long synthetic sequences targeting the HepG2;WERI-Rb-1 pair. **(H)** Median predicted accessibility on each cell line for dual-target synthetic sequences with 145 nt (x axis) and 300 nt (y axis). Markers above the diagonal indicate higher predicted accessibility for longer sequences.

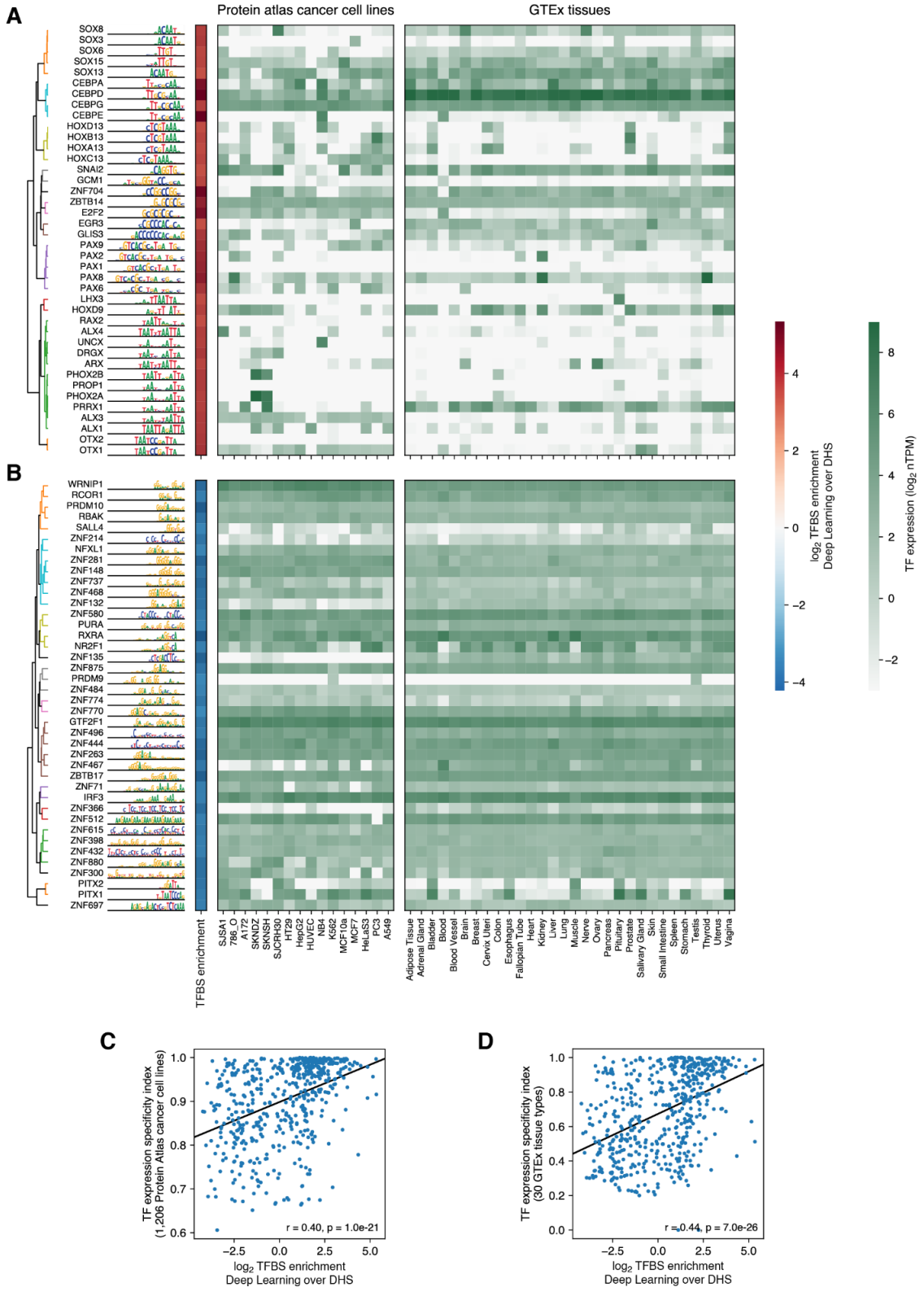


Supplementary Figure 3.12. Additional analysis on triple-target enhancers. **(A)** DHS64-predicted $\log_{10}$ accessibility for DHS-derived (top) and Fast SeqProp-designed sequences (bottom), with rows representing median values for each target cell line triplet. **(B)** Number of DHSs in the DNase I Index meeting our filtering criteria (enhancer-like chromatin annotations, peak calls in the target cell lines, and in ≤10 of the modeled biosamples) for each target. **(C)** Predicted accessibility for synthetic sequences targeting

the GM12878;SJCRH30;WERI-Rb-1 triplet and each cell line individually. **(D)** Median predicted accessibility on each cell line, for synthetic sequences targeting cell lines individually (x axis) or as part of a triple-target design (y axis). Markers below the diagonal indicate weaker predicted accessibility in a given cell line for sequences designed for multiple targets. **(E)** Measured enhancer activities of synthetic sequences targeted to the GM12878;SJCRH30;WERI-Rb-1 triplet and to each cell line individually. **(F)** Median enhancer activity of sequences designed to target each cell line, either individually (x axis) or as part of a triple-target design (y axis). In **(E)** and **(F)**, only sequences within the top 20% by mingap score per target are considered, as in **Figure 3.2** and **Figure 3.3**. **(G)** Predicted accessibility for 145 nt and 300 nt-long synthetic sequences targeting the GM12878;SJCRH30;WERI-Rb-1 triplet. **(H)** Median predicted accessibility on each cell line for triple-target synthetic sequences with 145 nt (x axis) and 300 nt (y axis). Markers above the diagonal indicate higher predicted accessibility for longer sequences.

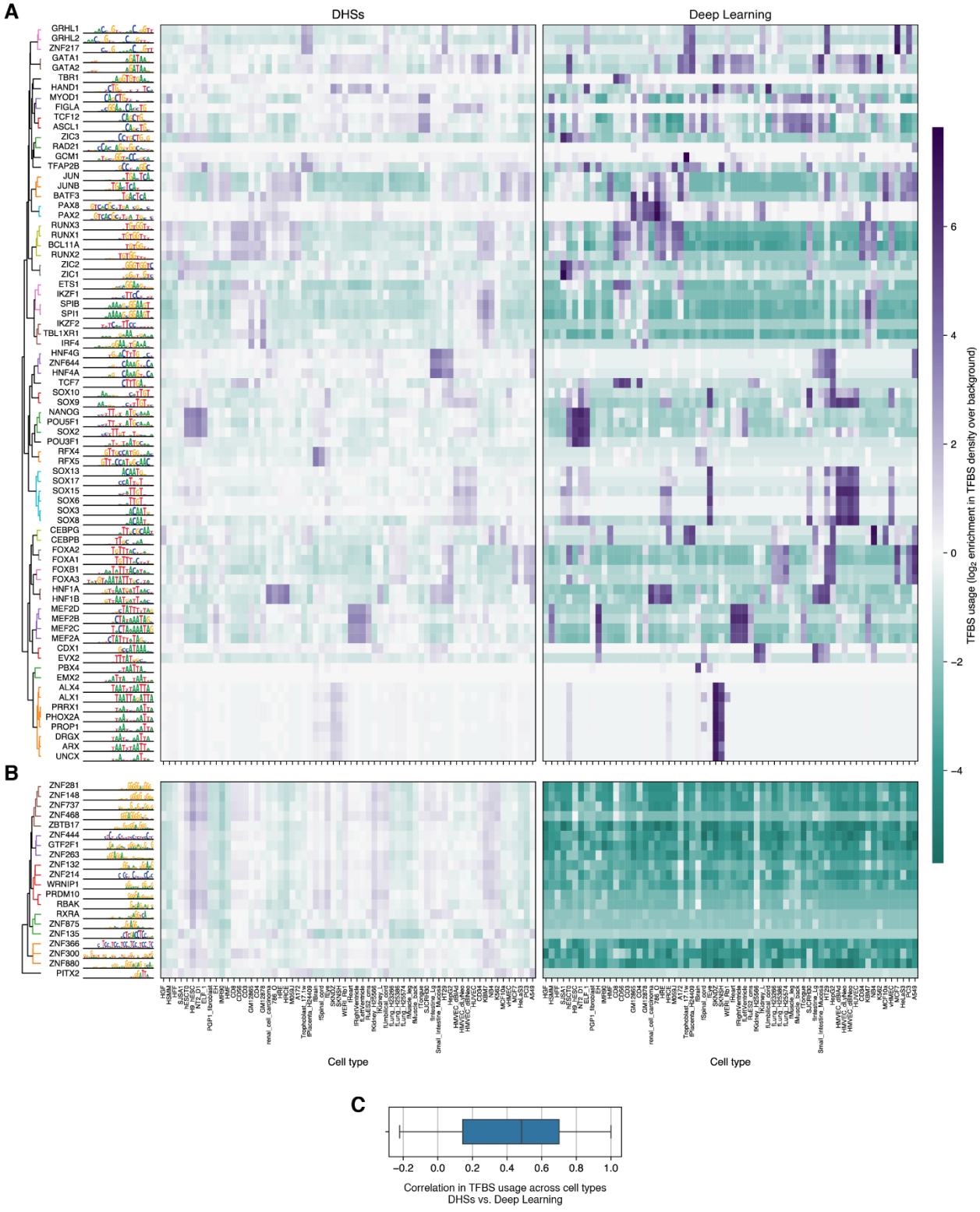


Supplementary Figure 3.13. DHS64-designed sequences are enriched for specific transcription factor binding sites (TFBSs) compared to DHSs. We compare DHSs with enhancer-like annotations and high accessibility towards each DHS64-modeled biosample (1,000 per target, 64,000 total, **Appendix B**) against DHS64-designed sequences (500 per target, 32,000 total). **(A)** Distribution of TFBS density (number of non-overlapping TFBSs per base pair) across both sets. **(B)** Enrichment of each TFBS in designed sequences relative to DHSs, measured as the change in the average number of occurrences per sequence. **(C)** Distribution of the number of target biosamples in which each TFBS is found (present in >5% of biosample-specific sequences), either in the most specific DHSs or in DHS64-designed sequences. **(D)** Correlation between the number of biosamples where a TFBS is present in DHSs, and its enrichment in DHS64-designed sequences compared to DHSs.



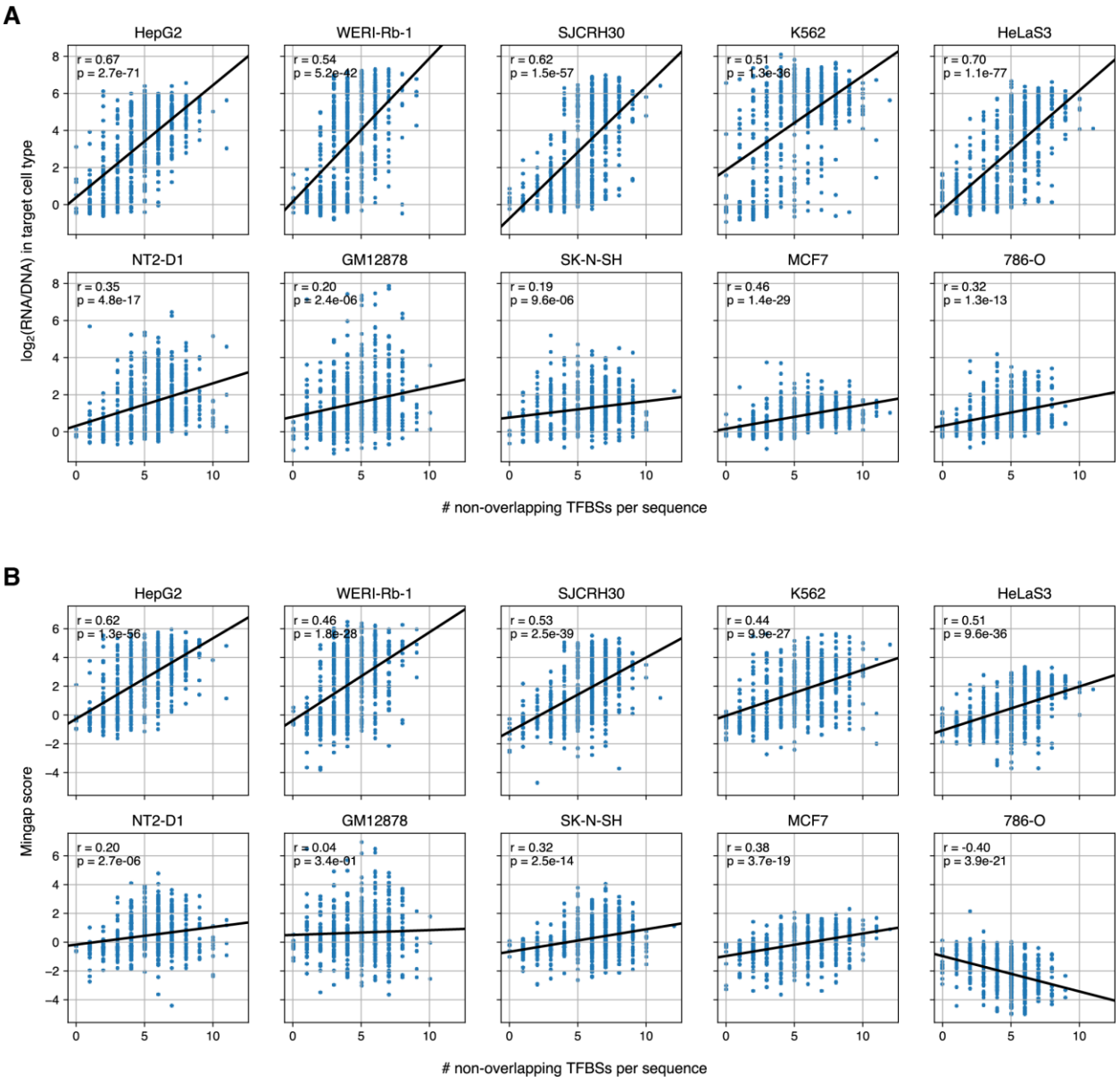
Supplementary Figure 3.14. TFBSs enriched in DHS64-designed sequences correspond to transcription factors (TFs) with cell type-specific expression.

(A and B) The 40 most enriched **(A)** and 40 most depleted **(B)** TFBSs in designed sequences compared to DHSs. For each TFBS, we show its position weight matrix (PWM), enrichment value, TF expression in selected cancer cell lines (Human Protein Atlas[80], **Appendix B**), and TF expression across tissues (GTEx[81], **Appendix B**). For cancer cell line TF expression, only cell lines corresponding to those modeled by DHS64 are shown. **(C and D)** Correlation between TFBS enrichment in designed sequences and the tissue specificity index[76] of the corresponding TF's expression across 1,206 cell lines in the Human Protein Atlas dataset **(C)** or across 30 GTEx tissue types **(D)**.



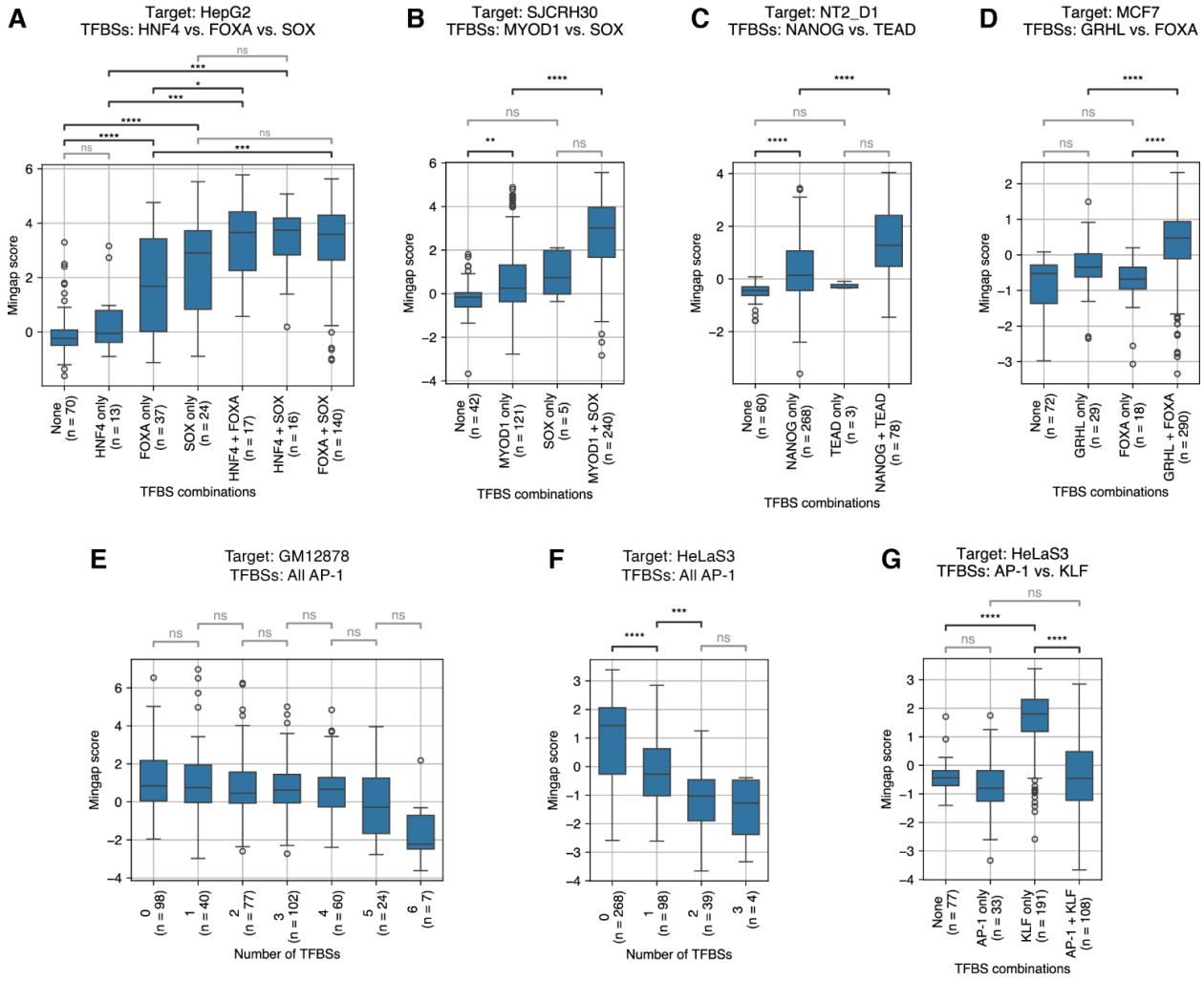
Supplementary Figure 3.15. Changes in TFBS utilization across cell type-specific sequences.

(A and B) TFBS utilization in DHSs (top 1,000 per cell type by specificity, left) and in deep learning-designed sequences (500 per cell type, right) targeting each DHS64-modeled biosample. Utilization is reported as the $\log_2$ enrichment in TFBSs per sequence compared to a background set of 1,000 randomly sampled DHSs (**Appendix B**). **(A)** The 75 TFBSs with the highest Pearson correlation between cell type-specific utilization in DHSs versus designed sequences. **(B)** The 20 TFBSs most strongly depleted in designed sequences compared to DHSs. **(C)** Distribution of Pearson correlations comparing TFBS utilization in DHSs versus designed sequences, for all 730 TFBSs detected in any sequence.



Supplementary Figure 3.16. The number of TFBSs in a sequence correlates with its enhancer activity and specificity.

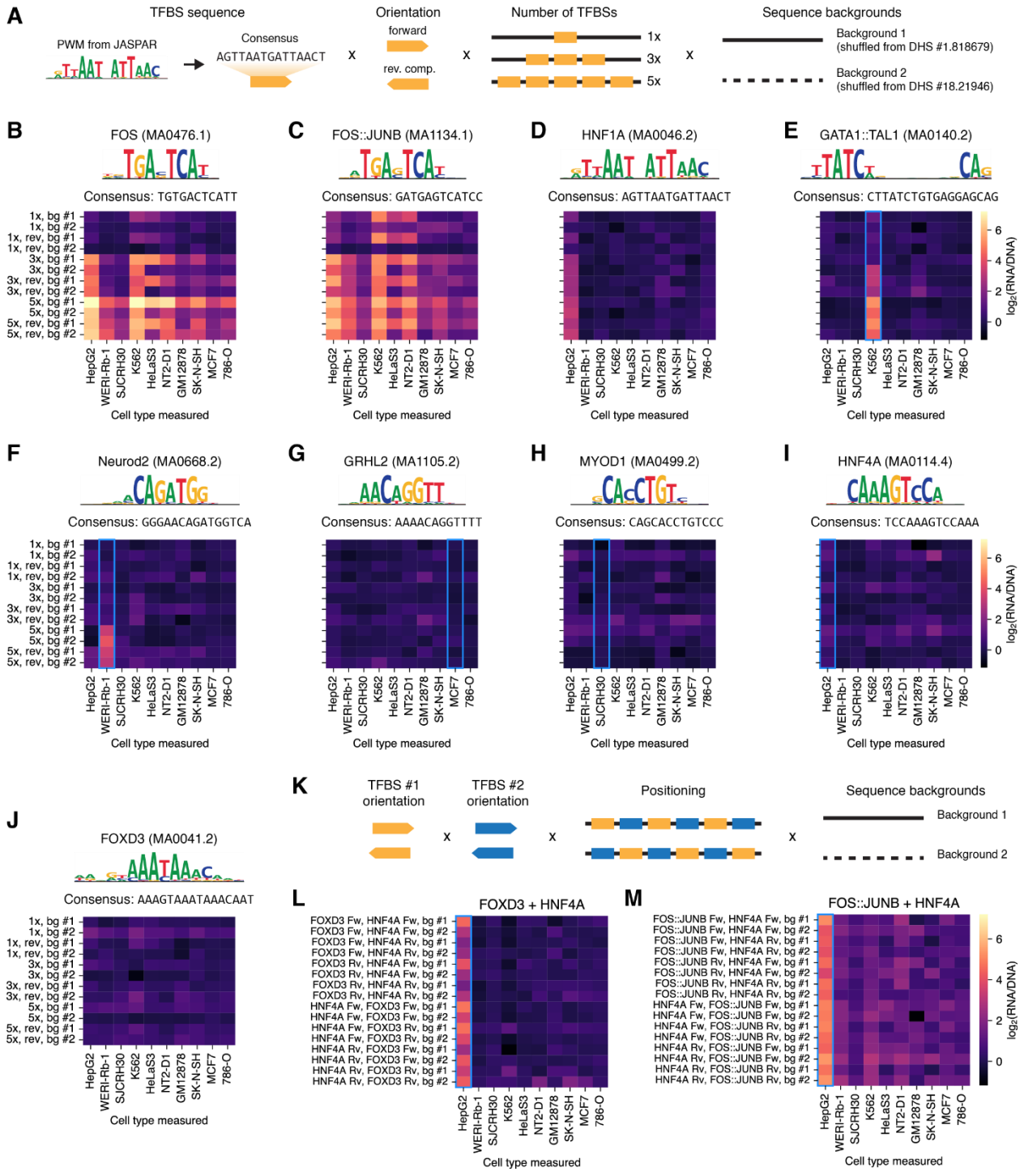
Each panel includes sequences targeting the indicated cell line, including ~110 DHSs, ~300 deep learning-designed sequences optimized for maximal specificity, and ~120 sequences designed for tunable activity (**Figure 3.3A-B**, **Supplementary Figure 3.10**), totaling ~530 per panel (exact counts range from 526 to 529 due to sequencing dropout). **(A)** Target enhancer activity versus number of non-overlapping TFBSs. All panels show statistically significant correlations. **(B)** Mingap score versus number of TFBSs. Correlation is not significant for GM12878 and is negative for 786-O, for which no cell type-specific enhancers were successfully designed.



Supplementary Figure 3.17. Combinations of TFBSs determine cell type-specific activity.

Each panel includes sequences targeting the indicated cell line, including ~110 DHSs and ~300 deep learning-designed sequences optimized for maximal specificity, totaling ~410 per panel. All tests were Mann-Whitney two-sided with Bonferroni corrections. ns: p value > 0.05; *: $0.01 < p < 0.05$; **: 1e-3 $< p <$ 1e-2; ***: 1e-4 $< p <$ 1e-3; ****: $p <$ 1e-4. In panels where combinations of TFs are evaluated, statistical tests are performed between sequences with each TFBS in isolation against neither TFBSs, and between sequences containing TFBS pairs against sequences containing each TFBS individually. **(A)** Effect of HNF4 (HNF4A or HNF4G), FOXA (FOXA1 or FOXA2), and SOX (SOX8, SOX9, or SOX10) combinations over HepG2 specificity. **(B)** Effect of MYOD1 and SOX combinations over SJCRH30 specificity. **(C)** Effect of NANOG (NANOG, POU5F1, or SOX2) and TEAD (TEAD1, TEAD3, or TEAD4) combinations over NT2-D1 specificity. **(D)** Effect of GRHL (GRHL1 or GRHL2) and FOXA (FOXA1 or FOXA2) combinations

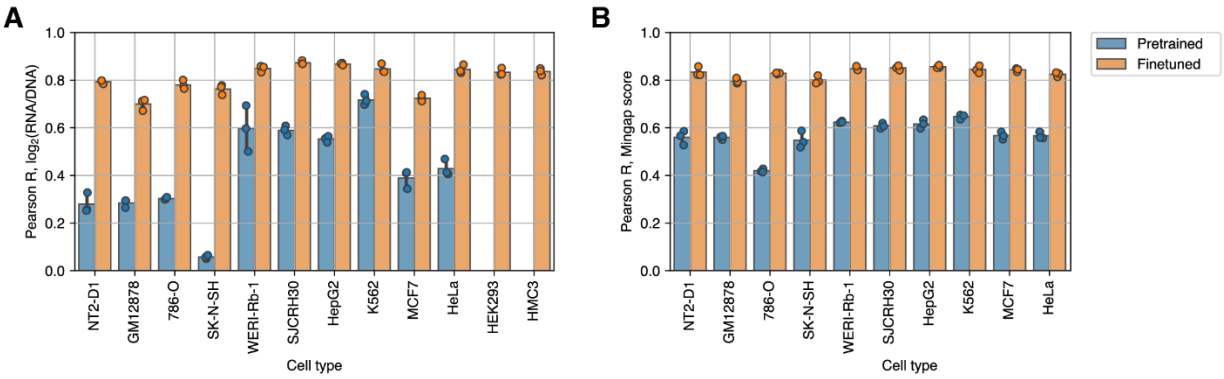
over MCF7 specificity. **(E)** Effect of the number of AP-1 occurrences (FOS, JUND, etc. See **Appendix B** for full list) over GM12878 specificity. **(F)** Effect of the number of AP-1 occurrences over HeLa specificity. **(G)** Effect of KLF (KLF1, KLF3, KLF4, KLF5, KLF6, KLF12, or KLF17) and AP-1 combinations over HeLa specificity.



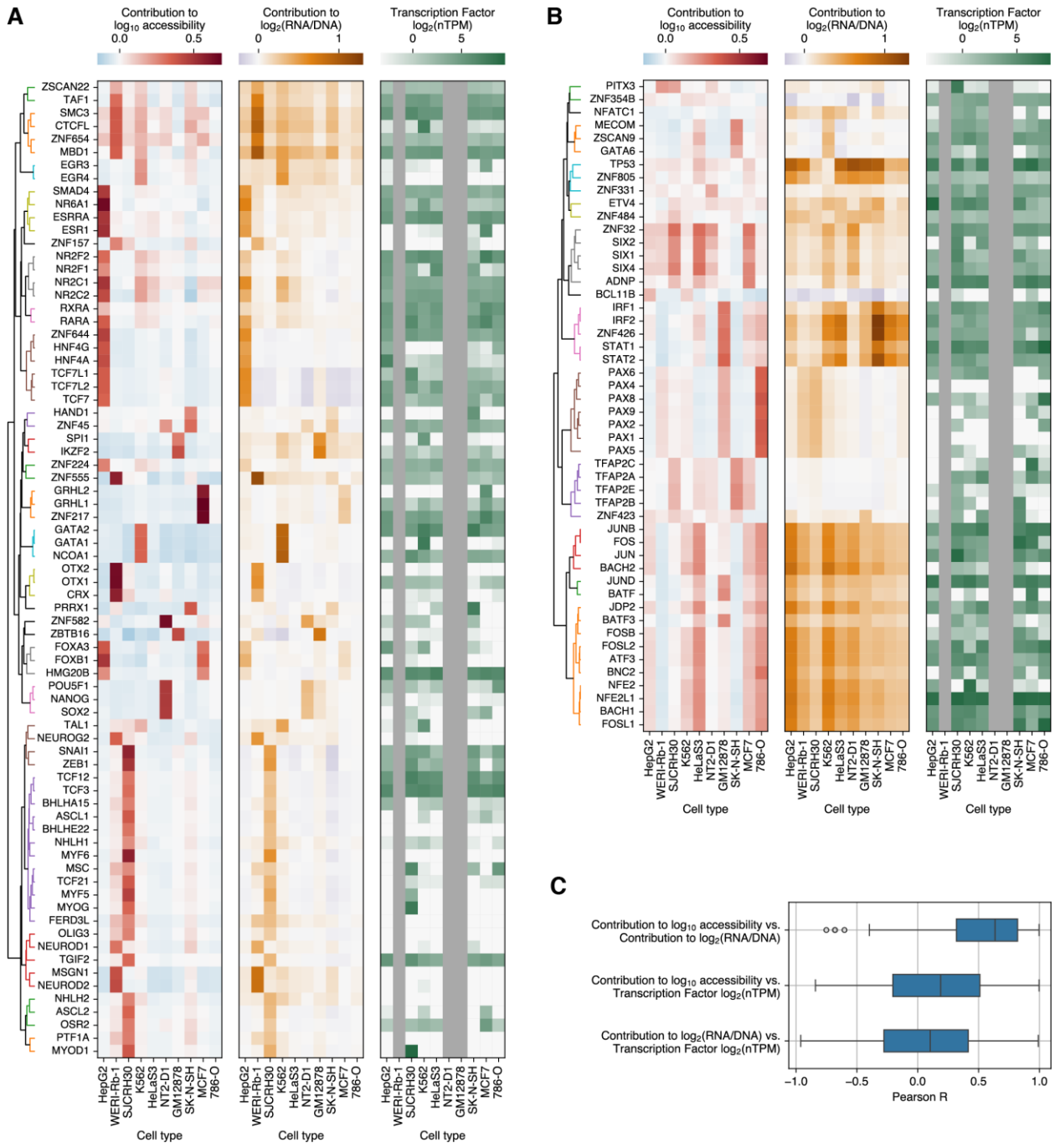
Supplementary Figure 3.18. Handcrafted sequences with embedded TFBSs reveal determinants of cell type-specificity.

(A) Design of sequences with one or multiple copies of a single TFBS, starting from PWMs from the JASPAR 2022 vertebrate database (**Appendix B**). **(B and C)** Enhancers with FOS and FOS::JUNB TFBSs drive gene expression in multiple cell types,

starting from 3 TFBS copies. **(D)** Enhancers with up to 5 copies of HNF1A drive gene expression in HepG2 only. **(E)** Enhancers with the composite GATA1::TAL1 drive K562-specific expression. **(F)** Enhancers with 5 copies of Neurod2 drive WERI-Rb1-specific expression. **(G-I)** Enhancers with GRHL2, MYOD1, and HNF4A do not drive gene expression in any cell line. Blue rectangles indicate cell lines where expression was expected. **(J)** The transcription factor FOXD3 does not drive gene expression in isolation. **(K)** Design of sequences with multiple copies of two different TFBSs (**Appendix B**). **(L)** Combinations of FOXD3 and HNF4A drive HepG2-specific expression. **(M)** Combinations of FOS::JUNB and HNF4A drive HepG2-specific expression, although with broad background expression.



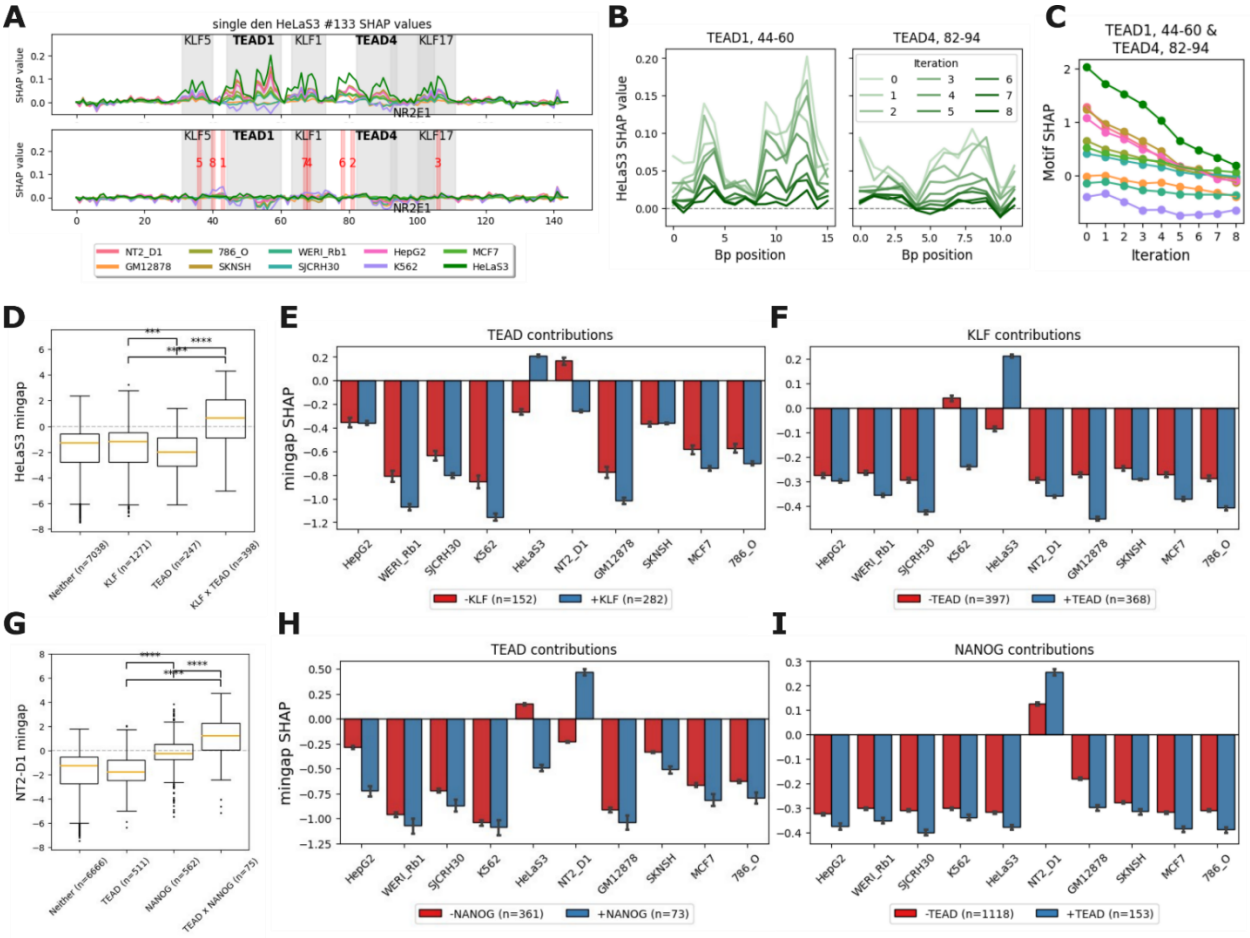
Supplementary Figure 3.19. Prediction performance of DHS64-MPRA. DHS64-MPRA was obtained by finetuning DHS64 on MPRA measurements from our entire enhancer library across 12 cell lines. These included the 10 cell lines used as enhancer design targets (**Figure 3.2D-G**) as well as two additional lines, HEK293 and HMC3, which are not part of the DNase I Index dataset. Each marker shows the performance of a separate model trained on a distinct MPRA data split and evaluated on held-out data (**Appendix B**). Bars and error bars represent the average and standard deviation across three models. Performance is compared to the original (pretrained) DHS64 model when evaluated against MPRA data. **(A)** Correlation between model predictions and measured $\log_2$(RNA/DNA) for each cell line. Note that the pretrained model lacks outputs for HEK293 and HMC3. **(B)** Correlation between predicted and observed mingap scores, calculated on every sequence by treating every cell type as the target independently.



Supplementary Figure 3.20. Average TFBS contributions towards accessibility and enhancer activity, and their relationship to TF expression across cell lines.

(A and B) For the TFs shown, each panel displays: (1) the average contribution of each TFBS to predicted $\log_{10}$ accessibility across assayed cell lines (via DHS64), (2) the average contribution to predicted enhancer activity ($\log_2$(RNA/DNA), via DHS64-MPRA), and

(3) TF expression from the Human Protein Atlas cancer cell line dataset[80]. Gray indicates cell lines lacking TF expression data. TFs (rows) were hierarchically clustered based on PWM similarity (**Appendix B**) **(A)** 75 TFs whose TFBS contributions toward accessibility and enhancer activity are most positively correlated across cell lines. **(B)** 50 TFs whose TFBS contributions toward accessibility and enhancer activity are least correlated. **(C)** Distribution of Pearson correlation coefficients between accessibility and enhancer activity contributions, and between each of these and TF expression, for all 730 TFs with binding sites identified in any evaluated sequence.



Supplementary Figure 3.21. TEAD TFBS contributes specifically to at least 2 distinct cell types in the presence of different partner TFBS sites.
(A) Per-nucleotide SHAP contribution scores from DHS64-MPRA plotted for Single DEN HeLaS3 #133 for all 10 cell lines. TFBS sites are annotated with gray boxes. The two TEAD sites are annotated in bold text; these sites are frozen, and the remaining sequence is iteratively mutated to reduce the total HeLaS3 contribution of these two sites. The original sequence is shown in the top row, and the mutated sequence after convergence is shown in the bottom row, with the locations and ordering of the mutations denoted by red bars and text. **(B)** HeLaS3 SHAP contribution of the nucleotides in the 2 TEAD TFBSs after each mutation. **(C)** Total SHAP contribution of the TEAD TFBSs to each cell line after each mutation. **(D)** Comparison of HeLaS3 mingap in sequences that contain KLF and TEAD sites, vs sequences that contain one without the other, vs sequences that contain neither. **(E)** Average SHAP contribution of TEAD sites towards mingap in each cell line, in sequences also containing KLF sites vs sequences without KLF sites. **(F)** Average SHAP contribution of KLF sites towards mingap in each cell line, in sequences also containing TEAD sites vs sequences without TEAD sites. **(G)** Comparison

of NT2-D1 mingap in sequences that contain NANOG and TEAD sites, vs sequences that contain one without the other, vs sequences that contain neither. **(H)** Average SHAP contribution of TEAD sites towards mingap in each cell line, in sequences also containing NANOG sites vs sequences without NANOG sites. **(I)** Average SHAP contribution of NANOG sites towards mingap in each cell line, in sequences also containing TEAD sites vs sequences without TEAD sites.

CHAPTER 4. DISCUSSION AND FUTURE DIRECTIONS

In Chapter 2 we demonstrate the iterative deep learning-based design and experimental validation of functional, cell type-specific enhancers in human cell lines. Across two generations of sequence design we achieve enhancers that drive cell type-specific gene expression more strongly and with a higher relative success rate than putative enhancers sourced from accessible regions of the genome, either selected directly from accessibility measurements (R0-DHS) or screened via MPRA (R0-MPRA). We also find that sequences generated in a second round of design were more specific than those from the first round despite the use of only a relatively small number of sequences for model retraining. Our models are able to both interpolate (design for intermediate levels of differential expression) and extrapolate (design beyond the range of measured expression activities), via *ab initio* discovery and implementation of a condensed TFBS motif grammar.

Analysis of synthetic enhancers and corresponding follow-up experiments implicated several sequence features driving stronger specificity compared to genome-sourced enhancers: most prominently, higher motif density coupled with enrichment of a concise vocabulary of TFBS motifs. (This behavior is recapitulated in Chapter 3.) While top-performing enhancer designs often featured motif repeats, they were also heterotypic,

suggesting that deep learning models capture an impact of motif diversity for optimizing sequence specificity.

A scMPRA supports a relationship between TF mRNA abundance and enhancer activity. However, while these data are consistent with a causal relationship between TF expression and the activity of enhancers with cognate TFBSs, there are several caveats. First, TF mRNA levels may not be predictive of TF activity. For example, while enhancers containing the TP53 motif have no activity in K562 cell lines, TP53 mRNA levels are only two-fold lower in K562 than HepG2. However, previous work found that both TP53 alleles in K562 encode a non-functional protein, but the relevant sequence variation is not captured in a typical scRNA-seq workflow[82]. More generally, the activity and localization of TFs is often controlled through post-transcriptional modification (e.g. phosphorylation) not captured by RNA-seq. Moreover, we find that for many TFs, expression differences between cell types are larger than expression variation within each cell type. As a result, observed correlations may be driven primarily by the expression differences between the two cell types, resulting in spurious correlations due to the fixed activity ratio between different TFs at the bulk level. Nevertheless, the capability to measure full pairwise variations in enhancer activities and TF expression levels is an important feature of scMPRAs. Future screens on heterogeneous cell populations or using multiplexed knockdown/overexpression

perturbations will be effective experimental approaches to infer causal relationships between enhancers and TFs in a highly parallelized, pooled manner.

An important caveat to comparisons with genomic sequences in this and other work[43] is the confounder of sequence length. Natural enhancers have been argued to reach up to kilobase lengths[5], and excerpting only small windows of genomic sequence may therefore capture incomplete enhancers. This could theoretically disadvantage them against synthetic enhancers explicitly designed to be <145 bp. Accordingly, we find a relatively low motif density in both R0 libraries, especially when compared to our designed enhancers (**Figure 2.2A**). Thus, a "fair" comparison to genomic enhancers would require large DNA fragment synthesis. However, ~1kb genome-sourced enhancers have been characterized in some contexts and found to have a relatively low success rate[10,30,83]; compositions of multiple such elements may be required to form a fully functional super-enhancer. Furthermore, designing shorter functional enhancers is useful for gene therapy applications, where AAV vectors are limited by the size of their payload[84]. Here, we show that enhancers can be shortened to 50 bp without significant reduction in activity, suggesting core functional grammar can be greatly compressed, at least in directly promoter-adjacent enhancers.

In Chapter 3, we more fully realize the potential of synthetic enhancer design using deep learning models by (1) training models solely on chromatin accessibility data, which is far more broadly available than MPRA data, and (2) designing and experimentally

validating synthetic enhancers in a greatly expanded set of cell and tissue types, including *in vivo* mouse retina. We achieve a nearly perfect success rate in targeting enhancers to be specifically active in 9/10 cell lines; and while there remains more work to do in understanding the differences in sequence grammars driving chromatin accessibility vs. enhancer activity, these results establish the viability of expansion into arbitrary design targets for which chromatin accessibility data is available. This includes the remaining 54 model outputs not explicitly assayed in our experiments, as well as the 438 total unique biological targets collected in the ENCODE DHS index.

With the work presented in this dissertation as the foundation, there are several important extensions that future projects might explore. To begin with, we have validated synthetic enhancer designs almost exclusively in cancer cell lines, which may not be perfectly representative of normal human biology. Additionally, we have only validated enhancers in an episomal (plasmid-based) MPRA format. Therefore, future design efforts should focus on more biologically and/or clinically relevant targets—e.g. primary cells, differentiation trajectories—and integrate synthetic enhancers into the genome via lentiMPRA or related techniques, which may be necessary anyway for design targets incompatible with transient transfection assays.

More broadly, our efforts thus far have only explored one aspect of the *cis*-regulatory code by focusing on a single data modality: the DNA sequence. While we have demonstrated

that these models can successfully extrapolate from sequence to *cis*-regulatory activity, the mapping is imperfect. This is to be expected, as enhancer activity is a function not just of sequence, but of the factors that interact with the sequence, and the accessibility of the sequence to these factors. Currently our models only capture the first of these elements, and multimodal models integrating DNA sequence, epigenetic profile, and transcriptome information are required to completely encompass, comprehend, and deploy the mechanisms of cell type-specific gene expression.

BIBLIOGRAPHY

1. Khan, Y.S., and Farhana, A. (2025). Histology, Cell. In StatPearls (StatPearls Publishing).

2. Osumi-Sutherland, D., Xu, C., Keays, M., Levine, A.P., Kharchenko, P.V., Regev, A., Lein, E., and Teichmann, S.A. (2021). Cell type ontologies of the Human Cell Atlas. Nat. Cell Biol. *23*, 1129–1135. https://doi.org/10.1038/s41556-021-00787-7.

3. Kim, S., and Wysocka, J. (2023). Deciphering the multi-scale, quantitative *cis*-regulatory code. Mol. Cell *83*, 373–392. https://doi.org/10.1016/j.molcel.2022.12.032.

4. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351. https://doi.org/10.1126/science.1058040.

5. Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. *21*, 292–310. https://doi.org/10.1038/s41576-019-0209-0.

6. Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81. https://doi.org/10.1016/j.gde.2016.12.007.

7. Jindal, G.A., and Farley, E.K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev. Cell *56*, 575–587. https://doi.org/10.1016/j.devcel.2021.02.016.

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

9. Friedman, R.Z., Ramu, A., Lichtarge, S., Wu, Y., Tripp, L., Lyon, D., Myers, C.A., Granas, D.M., Gause, M., Corbo, J.C., et al. (2025). Active learning of enhancers and silencers in the developing neural retina. Cell Syst. *16*, 101163. https://doi.org/10.1016/j.cels.2024.12.004.

10. Graybuck, L.T., Daigle, T.L., Sedeño-Cortés, A.E., Walker, M., Kalmbach, B., Lenz, G.H., Morin, E., Nguyen, T.N., Garren, E., Bendrick, J.L., et al. (2021). Enhancer

viruses for combinatorial cell-subclass-specific labeling. Neuron *109*, 1449-1464.e13. https://doi.org/10.1016/j.neuron.2021.03.011.

11. Kohn, D.B., Chen, Y.Y., and Spencer, M.J. (2023). Successes and challenges in clinical gene therapy. Gene Ther. *30*, 738–746. https://doi.org/10.1038/s41434-023-00390-5.

12. Kussick, E., Johansen, N., Taskin, N., Wynalda, B., Martinez, R., Groce, E.L., Reding, M., Liang, E., Shulga, L., Huang, C., et al. (2024). Enhancer AAVs for targeting spinal motor neurons and descending motor pathways in rodents and macaque. bioRxiv, 2024.07.30.605864. https://doi.org/10.1101/2024.07.30.605864.

13. Blayney, J.W., Francis, H., Rampasekova, A., Camellato, B., Mitchell, L., Stolper, R., Cornell, L., Babbs, C., Boeke, J.D., Higgs, D.R., et al. (2023). Super-enhancers include classical enhancers and facilitators to fully activate gene expression. Cell *186*, 5826-5839.e18. https://doi.org/10.1016/j.cell.2023.11.030.

14. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319. https://doi.org/10.1016/j.cell.2013.03.035.

15. Gross, D.S., and Garrard, W.T. (1988). NUCLEASE HYPERSENSITIVE SITES IN CHROMATIN. Annu. Rev. Biochem. *57*, 159–197. https://doi.org/10.1146/annurev.bi.57.070188.001111.

16. McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., and Felsenfeld, G. (1981). A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. Cell *27*, 45–55. https://doi.org/10.1016/0092-8674(81)90359-7.

17. Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M.O., et al. (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc. Natl. Acad. Sci. U. S. A. *101*, 16837–16842. https://doi.org/10.1073/pnas.0407387101.

18. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218. https://doi.org/10.1038/nmeth.2688.

19. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82. https://doi.org/10.1038/nature11232.

20. 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. https://doi.org/10.1038/s41586-020-2559-3.

21. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112. https://doi.org/10.1038/nature07829.

22. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318. https://doi.org/10.1038/ng1966.

23. Lambert, J.T., Su-Feher, L., Cichewicz, K., Warren, T.L., Zdilar, I., Wang, Y., Lim, K.J., Haigh, J.L., Morse, S.J., Canales, C.P., et al. (2021). Parallel functional testing identifies enhancers active in early postnatal mouse brain. eLife *10*, e69479. https://doi.org/10.7554/eLife.69479.

24. Mich, J.K., Graybuck, L.T., Hess, E.E., Mahoney, J.T., Kojima, Y., Ding, Y., Somasundaram, S., Miller, J.A., Kalmbach, B.E., Radaelli, C., et al. (2021). Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. Cell Rep. *34*. https://doi.org/10.1016/j.celrep.2021.108754.

25. Ben-Simon, Y., Hooper, M., Narayan, S., Daigle, T., Dwivedi, D., Way, S.W., Oster, A., Stafford, D.A., Mich, J.K., Taormina, M.J., et al. (2024). A suite of enhancer AAVs and transgenic mouse lines for genetic access to cortical cell types. Preprint at bioRxiv, https://doi.org/10.1101/2024.06.10.597244 https://doi.org/10.1101/2024.06.10.597244.

26. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175. https://doi.org/10.1038/nbt.1589.

27. Fleur, A.L., Shi, Y., and Seelig, G. (2024). Decoding biology with massively parallel reporter assays and machine learning. Genes Dev. *38*, 843–865. https://doi.org/10.1101/gad.351800.124.

28. Gordon, M.G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., Feng, S., Zhao, J., Ashuach, T., Ziffra, R., et al. (2020). lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. Nat. Protoc. *15*, 2387–2412. https://doi.org/10.1038/s41596-020-0333-5.

29. Zhao, S., Hong, C.K.Y., Myers, C.A., Granas, D.M., White, M.A., Corbo, J.C., and Cohen, B.A. (2023). A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. Nat. Genet. *55*, 346–354. https://doi.org/10.1038/s41588-022-01278-7.

30. Lalanne, J.-B., Regalado, S.G., Domcke, S., Calderon, D., Martin, B.K., Li, X., Li, T., Suiter, C.C., Lee, C., Trapnell, C., et al. (2024). Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. Nat. Methods *21*, 983–993. https://doi.org/10.1038/s41592-024-02260-3.

31. Chen, W., Choi, J., Li, X., Nathans, J.F., Martin, B., Yang, W., Hamazaki, N., Qiu, C., Lalanne, J.-B., Regalado, S., et al. (2024). Symbolic recording of signalling and cis-regulatory element activity to DNA. Nature *632*, 1073–1081. https://doi.org/10.1038/s41586-024-07706-4.

32. Frömel, R., Rühle, J., Martinez, A.B., Szu-Tu, C., Pastor, F.P., Corral, R.M., and Velten, L. (2024). Synthetic enhancers reveal design principles of cell state specific regulatory elements in hematopoiesis. Preprint at bioRxiv, https://doi.org/10.1101/2024.08.26.609645 https://doi.org/10.1101/2024.08.26.609645.

33. Fu, Z.-H., He, S.-Z., Wu, Y., and Zhao, G.-R. (2023). Design and deep learning of synthetic B-cell-specific promoters. Nucleic Acids Res. *51*, 11967–11979. https://doi.org/10.1093/nar/gkad930.

34. Wu, M.-R., Nissim, L., Stupp, D., Pery, E., Binder-Nissim, A., Weisinger, K., Enghuus, C., Palacios, S.R., Humphrey, M., Zhang, Z., et al. (2019). A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). Nat. Commun. *10*, 2880. https://doi.org/10.1038/s41467-019-10912-8.

35. Maslova, A., Ramirez, R.N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., and Mostafavi, S. (2020). Deep learning of immune cell differentiation. Proc. Natl. Acad. Sci. U. S. A. *117*, 25655–25666. https://doi.org/10.1073/pnas.2011795117.

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

37. Movva, R., Greenside, P., Marinov, G.K., Nair, S., Shrikumar, A., and Kundaje, A. (2019). Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLOS ONE *14*, e0218073. https://doi.org/10.1371/journal.pone.0218073.

38. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

39. Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D.R. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nat. Genet., 1–13. https://doi.org/10.1038/s41588-024-02053-6.

40. Taskiran, I.I., Spanier, K.I., Dickmänken, H., Kempynck, N., Pančíková, A., Ekşi, E.C., Hulselmans, G., Ismail, J.N., Theunis, K., Vandepoel, R., et al. (2024). Cell-type-directed design of synthetic enhancers. Nature *626*, 212–220. https://doi.org/10.1038/s41586-023-06936-2.

41. de Almeida, B.P., Schaub, C., Pagani, M., Secchia, S., Furlong, E.E.M., and Stark, A. (2024). Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. Nature *626*, 207–211. https://doi.org/10.1038/s41586-023-06905-9.

42. Li, Z., Zhang, Y., Peng, B., Qin, S., Zhang, Q., Chen, Y., Chen, C., Bao, Y., Zhu, Y., Hong, Y., et al. (2024). A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. Nucleic Acids Res. *52*, 13447–13468. https://doi.org/10.1093/nar/gkae912.

43. Gosai, S.J., Castro, R.I., Fuentes, N., Butts, J.C., Mouri, K., Alasoadura, M., Kales, S., Nguyen, T.T.L., Noche, R.R., Rao, A.S., et al. (2024). Machine-guided design of

cell-type-targeting cis-regulatory elements. Nature *634*, 1211–1220. https://doi.org/10.1038/s41586-024-08070-z.

44. DaSilva, L.F., Senan, S., Patel, Z.M., Reddy, A.J., Gabbita, S., Nussbaum, Z., Córdova, C.M.V., Wenteler, A., Weber, N., Tunjic, T.M., et al. (2024). DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements. Preprint at bioRxiv, https://doi.org/10.1101/2024.02.01.578352 https://doi.org/10.1101/2024.02.01.578352.

45. Lal, A., Garfield, D., Biancalani, T., and Eraslan, G. (2024). regLM: Designing realistic regulatory DNA with autoregressive language models. Preprint at bioRxiv, https://doi.org/10.1101/2024.02.14.580373 https://doi.org/10.1101/2024.02.14.580373.

46. Zhao, J., Baltoumas, F.A., Konnaris, M.A., Mouratidis, I., Liu, Z., Sims, J., Agarwal, V., Pavlopoulos, G.A., Georgakopoulos-Soares, I., and Ahituv, N. (2025). MPRAbase a Massively Parallel Reporter Assay database. Genome Res., gr.280387.124. https://doi.org/10.1101/gr.280387.124.

47. Song, L., and Crawford, G.E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. Cold Spring Harb. Protoc. *2010*, pdb.prot5384. https://doi.org/10.1101/pdb.prot5384.

48. Grandi, F.C., Modi, H., Kampman, L., and Corces, M.R. (2022). Chromatin accessibility profiling by ATAC-seq. Nat. Protoc. *17*, 1518–1552. https://doi.org/10.1038/s41596-022-00692-9.

49. Balsalobre, A., and Drouin, J. (2022). Pioneer factors as master regulators of the epigenome and cell fate. Nat. Rev. Mol. Cell Biol. *23*, 449–464. https://doi.org/10.1038/s41580-022-00464-z.

50. Barral, A., and Zaret, K.S. (2024). Pioneer factors: roles and their regulation in development. Trends Genet. *40*, 134–148. https://doi.org/10.1016/j.tig.2023.10.007.

51. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. *34*, 1180–1190. https://doi.org/10.1038/nbt.3678.

52. Linder, J., and Seelig, G. (2021). Fast activation maximization for molecular sequence design. BMC Bioinformatics *22*, 510. https://doi.org/10.1186/s12859-021-04437-5.

53. Linder, J., Bogard, N., Rosenberg, A.B., and Seelig, G. (2020). A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. Cell Syst. *11*, 49-62.e16. https://doi.org/10.1016/j.cels.2020.05.007.

54. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

55. Flynn, M.J., Mayfield, A.M.H., Du, R., Gradinaru, V., and Elowitz, M.B. (2024). Synthetic dosage-compensating miRNA circuits allow precision gene therapy for Rett syndrome. Preprint at bioRxiv, https://doi.org/10.1101/2024.03.13.584179 https://doi.org/10.1101/2024.03.13.584179.

56. Gibson, T.J., Seiler, M., and Veitia, R.A. (2013). The transience of transient overexpression. Nat. Methods *10*, 715–721. https://doi.org/10.1038/nmeth.2534.

57. Prelich, G. (2012). Gene Overexpression: Uses, Mechanisms, and Interpretation. Genetics *190*, 841–854. https://doi.org/10.1534/genetics.111.136911.

58. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. https://doi.org/10.1093/nar/gkab1113.

59. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. *45*, e119. https://doi.org/10.1093/nar/gkx314.

60. Agarwal, V., Inoue, F., Schubach, M., Penzar, D., Martin, B.K., Dash, P.M., Keukeleire, P., Zhang, Z., Sohota, A., Zhao, J., et al. (2025). Massively parallel characterization of transcriptional regulatory elements. Nature *639*, 411–420. https://doi.org/10.1038/s41586-024-08430-9.

61. Zhao, Y., Vartak, S.V., Conte, A., Wang, X., Garcia, D.A., Stevens, E., Jung, S.K., Kieffer-Kwon, K.-R., Vian, L., Stodola, T., et al. (2022). “Stripe” transcription factors

provide accessibility to co-binding partners in mammalian genomes. Mol. Cell *82*, 3398-3411.e11. https://doi.org/10.1016/j.molcel.2022.06.029.

62. Georgakopoulos-Soares, I., Deng, C., Agarwal, V., Chan, C.S.Y., Zhao, J., Inoue, F., and Ahituv, N. (2023). Transcription factor binding site orientation and order are major drivers of gene regulatory activity. Nat. Commun. *14*, 2333. https://doi.org/10.1038/s41467-023-37960-5.

63. Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C.O., et al. (2022). Sequence determinants of human gene regulatory elements. Nat. Genet. *54*, 283–294. https://doi.org/10.1038/s41588-021-01009-4.

64. Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

65. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182. https://doi.org/10.1126/science.aam8999.

66. Litzenburger, U.M., Buenrostro, J.D., Wu, B., Shen, Y., Sheffield, N.C., Kathiria, A., Greenleaf, W.J., and Chang, H.Y. (2017). Single-cell epigenomic variability reveals functional cancer heterogeneity. Genome Biol. *18*, 15. https://doi.org/10.1186/s13059-016-1133-7.

67. Bailey, T.L. (2021). STREME: accurate and versatile sequence motif discovery. Bioinformatics *37*, 2834–2840. https://doi.org/10.1093/bioinformatics/btab203.

68. 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. https://doi.org/10.1038/s41586-020-2559-3.

69. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

70. Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D.R. (2023). 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 https://doi.org/10.1101/2023.08.30.555582.

71. Linder, J., and Seelig, G. (2021). Fast activation maximization for molecular sequence design. BMC Bioinformatics *22*, 510. https://doi.org/10.1186/s12859-021-04437-5.

72. Linder, J., Bogard, N., Rosenberg, A.B., and Seelig, G. (2020). A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. Cell Syst. *11*, 49-62.e16. https://doi.org/10.1016/j.cels.2020.05.007.

73. Kathail, P., Shuai, R.W., Chung, R., Ye, C.J., Loeb, G.B., and Ioannidis, N.M. (2024). Current genomic deep learning models display decreased performance in cell type-specific accessible regions. Genome Biol. *25*, 1–22. https://doi.org/10.1186/s13059-024-03335-2.

74. Maslova, A., Ramirez, R.N., Ma, K., Schmutz, H., Wang, C., Fox, C., Ng, B., Benoist, C., Mostafavi, S., and Immunological Genome Project (2020). Deep learning of immune cell differentiation. Proc. Natl. Acad. Sci. *117*, 25655–25666. https://doi.org/10.1073/pnas.2011795117.

75. Gosai, S.J., Castro, R.I., Fuentes, N., Butts, J.C., Mouri, K., Alasoadura, M., Kales, S., Nguyen, T.T.L., Noche, R.R., Rao, A.S., et al. (2024). Machine-guided design of cell-type-targeting cis-regulatory elements. Nature *634*, 1211–1220. https://doi.org/10.1038/s41586-024-08070-z.

76. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659. https://doi.org/10.1093/bioinformatics/bti042.

77. Yin, C., Castillo-Hair, S., Byeon, G.W., Bromley, P., Meuleman, W., and Seelig, G. (2024). Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity. Preprint at bioRxiv, https://doi.org/10.1101/2024.06.14.599076 https://doi.org/10.1101/2024.06.14.599076.

78. Yin, C., Castillo-Hair, S., Byeon, G.W., Bromley, P., Meuleman, W., and Seelig, G. (2025). Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. Cell Syst. *0*. https://doi.org/10.1016/j.cels.2025.101302.

79. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. *34*, 1180–1190. https://doi.org/10.1038/nbt.3678.

84. Wang, J.-H., Gessler, D.J., Zhan, W., Gallagher, T.L., and Gao, G. (2024). Adeno-associated virus as a delivery vector for gene therapy of human diseases. Signal Transduct. Target. Ther. *9*, 1–33. https://doi.org/10.1038/s41392-024-01780-w.

85. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811. https://doi.org/10.1101/gr.144899.112.

86. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

87. Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999. https://doi.org/10.1101/gr.200535.115.

88. Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A.K., et al. (2021). De novo protein design by deep network hallucination. Nature *600*, 547–552. https://doi.org/10.1038/s41586-021-04184-w.

89. Hao, G.-F., Xu, W.-F., Yang, S.-G., and Yang, G.-F. (2015). Multiple Simulated Annealing-Molecular Dynamics (MSA-MD) for Conformational Space Search of Peptide and Miniprotein. Sci. Rep. *5*, 15568. https://doi.org/10.1038/srep15568.

90. Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. Preprint at arXiv,

https://doi.org/10.48550/arXiv.1802.05957
https://doi.org/10.48550/arXiv.1802.05957.

91. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

92. Zorita, E., Cuscó, P., and Filion, G.J. (2015). Starcode: sequence clustering based on all-pairs search. Bioinformatics *31*, 1913–1919. https://doi.org/10.1093/bioinformatics/btv053.

93. Kaminow, B., Yunusov, D., and Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at bioRxiv, https://doi.org/10.1101/2021.05.05.442755 https://doi.org/10.1101/2021.05.05.442755.

94. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://doi.org/10.1186/s13059-019-1874-1.

95. Abugessaisa, I., Noguchi, S., Hasegawa, A., Kondo, A., Kawaji, H., Carninci, P., and Kasukawa, T. (2019). refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. J. Mol. Biol. *431*, 2407–2422. https://doi.org/10.1016/j.jmb.2019.04.045.

96. Vu, H., and Ernst, J. (2022). Universal annotation of the human genome through integration of over a thousand epigenomic datasets. Genome Biol. *23*, 1–37. https://doi.org/10.1186/s13059-021-02572-z.

APPENDIX A – METHODS IN CHAPTER 2

5.1 Library synthesis and assembly

To test our synthetic library, we largely followed the same protocol used to generate the SHARPR-MPRA dataset[85]. Briefly, we filtered our designed enhancers to remove and replace any containing KpnI, XbaI, or SfiI recognition sites. Enhancers were placed next to their corresponding 3'UTR barcodes separated by KpnI and XbaI sites. We used 5 distinct barcodes per enhancer, which were generated using the dna-barcodes package (https://github.com/feldman4/dna-barcodes) (length 10, minimum edit distance = 2). Oligos containing enhancers and barcodes were ordered as a pool from Twist Biosciences. We amplified the oligo library with primers CY01 and CY02 (**Table 2.6**) to add Gibson overhangs; digested and gel purified the plasmid backbone (pMPRA1, Addgene #49349) with SfiI (NEB R0123) to remove its promoter and reporter sequence; then inserted the library cassette into the digested pMPRA1 backbone using the NEBuilder HiFi DNA Assembly kit (NEB E2621). This intermediate plasmid was purified with Kapa Pure Beads (Roche KK8002), transformed into NEB 10-beta electrocompetent cells (NEB C3020K), and extracted using the Qiagen Plasmid Maxi Kit (Qiagen 12162). We then insert a separate promoter and reporter cassette (minP, luciferase) excerpted from pMPRAdonor2 (Addgene #49353) in between the enhancer and barcode elements, by digesting both plasmids with KpnI (NEB R3142) and XbaI (NEB R0145), gel-purifying the appropriate fragments, dephosphorylating the intermediate plasmid fragment with Antarctic Phosphatase (NEB M0289) and ligating with T4 ligase (NEB M0202) for 16h. This ensures that barcodes will appear in RNA transcripts which can be mapped back to their corresponding enhancer. This final library plasmid was again purified, transformed, and Maxiprepped as above. The pooled library and 12 individual clones were sequence-verified via Sanger sequencing.

R1 and R2 libraries were synthesized via Twist and cloned as above. Given low variation observed among barcodes in R1, in R2 we reduced the number of barcodes per sequence from 5 to 2. To avoid PCR bias towards shorter sequences in R2, we used different primer sequence for 25 bp, 50 bp, 72 bp, and 145 bp enhancers. Each sublibrary was cloned separately into pMPRA1, then recombined according to the following proportions: 85% 145 bp, 5% 72 bp, 5% 50 bp, and 5% 25 bp. The second cloning step was performed on the recombined library.

5.2 Library transfection

HepG2 cells were obtained from ATCC (HB-8065) and cultured in EMEM (ATCC 30-2003) + 10% FBS + 1% Pen/Strep at 37°C and 5% CO2. For each transfection, 600k (R1 libraries) or 1 million (R2) cells were seeded in 6-well plates or 6 cm plates respectively. 24 hours later, cells were transfected with Lipofectamine 3000 (Invitrogen L3000001) and 5 (R1) or 10 (R2) ug of library DNA following the manufacturer's instructions. For R2, media was replaced 6 hours later. Cells were then grown for 48 hours after transfection. Transfection efficiency, estimated by transfecting a GFP plasmid in a parallel culture and performing flow cytometry, was ~20%.
K562 cells were obtained from ATCC (CCL-243) and cultured in RPMI (Gibco 11875093) + 10% FBS + 1% Pen/Strep at 37°C and 5% CO2. For each transfection, 1.5 million K562 cells per replicate were electroporated with 10 ug DNA using the Neon Transfection System (Invitrogen MPK5000), transferred to 6-well plates (R1) or 6 cm plates (R2) with RPMI + FBS (no Pen/Strep), and incubated for 48 hours. Transfection efficiency, estimated by transfecting a GFP plasmid in a parallel culture and performing flow cytometry, was ~90% (R1) and 96% (R2).

RNA was extracted 48h after transfection using the Monarch total RNA miniprep kit (NEB T2010S) and stored at -80C until further processing. Two independently-transfected replicates were performed per cell line, per DNA library.

5.3 Library preparation and sequencing

mRNA was isolated from total RNA extracts using the magnetic mRNA Isolation Kit (NEB S1550S) following the manufacturer's instructions. mRNA was then reverse transcribed with the Maxima H Minus Reverse Transcriptase (Thermo EP0753) using UMI-containing RT primer CY05 (**Table 2.6**), followed by digestion with RNase I (Thermo AM2294) and RNase H (NEB M0297) and purification with the DNA Clean & Concentrator - 5 (Zymo D4014), 7x binding buffer. To estimate the optimal number of PCR cycles for cDNA amplification, we first ran a small-scale qPCR for 30 cycles with 1-2 uL cDNA. qPCRs were ran with KAPA HiFi HotStart ReadyMix PCR Kit (Roche 07958935001), EvaGreen (Biotium 31000), and primers CY06/CY07 (fw) and Bri0xx (rev) containing P5 and P7 adapters and index sequences respectively (**Table 2.4**). The optimal cycle before the end of exponential amplification was determined from the amplification curves, and a new PCR reaction was run using this number with all remaining cDNA and scaling up all

volumes accordingly. Amplification products were gel extracted using the Monarch® DNA Gel Extraction Kit (NEB T1020).

For DNA libraries, 500 ng plasmid library was amplified with HiFi HotStart ReadyMix PCR Kit and primers CY06/CY07 (fw) and CY05 (rev) for two cycles, then purified with the Zymo DNA Clean & Concentrator, 5x binding buffer. 50 ng of the resulting product were used in a small-scale qPCR with P5 and Bri0xx primers to determine the optimal amplification cycle as above. Finally, the remaining template was amplified via PCR using the optimal number of cycles determined via qPCR. Reaction products were gel purified as above.

For R1 libraries, we prepared 1 DNA replicate per enhancer library (R1-MPRA and R1-DHS). For R2 we prepared 2 DNA replicates. In all cases we prepared two RNA replicates per cell line, per enhancer library. Libraries were quantified using the KAPA Library Quantification Kit (Roche 7960140001) before mixing. Sequencing was performed in an Illumina NextSeq 550 with the following settings: Number of cycles for R1: >10 (enhancer barcode), R2: 10 (UMI), index 1: 8, index 2: 8, custom primers: Read 1: CY08, Read 2: Custom_read_2, Index 1: Custom_index_1, Index 2: CY009 (Custom_Index_2).

5.4 Sequencing raw data processing

fastq file preprocessing, UMI deduplication, and barcode matching was performed using custom python scripts. UMI counts from barcodes corresponding to the same enhancers were pooled together. R1, R1-DHS, and R2 sequence measurements were processed with the DESeq2 package[86], following the example of Zhao et al[29]. R0 sequence measurements were also re-processed with DESeq2 instead of total read count normalization prior to model retraining for R2 design. Enhancers were treated as genes, with cell type and replicate sublibraries treated as samples.

5.5 Batch correction across design rounds

To improve our ability to compare measurements from different experiments, which covered a slightly different range in log2FC due to experimental variation, we performed a simple batch correction procedure as follows. Weighted linear regression was performed separately for each cell line on R2 measurements of R2 control sequences vs R1 measurements of the same sequences, using $\sqrt{{SE_{R1}}^2 + {SE_{R2}}^2}$ as the weights. For each cell type the following equation was obtained, where x represents $\log2FC_{HepG2}$ or $\log2FC_{K562}$ and

the subscript denotes library measurement: $x_{R2} = m_{21}x_{R1} + b_{21}$. Regression slopes and intercepts (m_{10}, b_{10}) were identically calculated for R1 vs R0.

Then, within each cell type batch-corrected R1 measurements were calculated as $x_{R1*} = m_{21}x_{R1} + b_{21}$; and batch-corrected R0 measurements were calculated as $x_{R0*} = m_{21}m_{10}x_{R1} + (m_{21}b_{10} + b_{21})$. Standard errors were propagated appropriately, and $\log2FC_{H2K}$ scores recalculated on the batch-corrected measurements. All batch correction coefficients are reported in **Table 2.7.**

Finally, batch-corrected control sequence measurements were averaged together.

5.6 Combinatorial indexing scMPRA library preparation

For HepG2, 2.5 million cells were electroporated with 11.5ug R1 library plasmid DNA in 250uL Resuspension Buffer R using the Neon Transfection System at 1230V, pulse width 20ms and 3 pulses total. 2 million cells were seeded in 2mL total media volume in a 6-well plate and fixed 24 hours later. For K562, 2.5 million cells were electroporated with 11.5ug R1 library plasmid DNA in 250uL Resuspension Buffer R using the Neon Transfection System at 1450V, pulse width 10ms and 3 pulses total. 2 million cells were seeded in 10mL total media volume in a T25 flask and fixed 24 hours later. We targeted a total of 10,000 transfected HepG2/K562 cells mixed at 1:1 ratio and performed 3-round combinatorial indexing using Parse Bio Evercode Mini v2 kit with modifications to the kit protocol for cDNA amplification steps to prepare an additional sequencing library enriching for MPRA amplicons. First, we spiked-in a MPRA amplicon targeting primer GB39 to initial cDNA amplification reaction at the final concentration of 400nM. We performed a second PCR, starting from 50ng of the cDNA from the initial amplification round, using GB42/GB43, with the following conditions: 0.02U/uL NEB Q5 HotStart Polymerase, 1x NEB Q5 Reaction Buffer, 300nM each primers, 20uM dNTPs, 1x Biotium EvaGreen dye; 98°C 20s - 67°C 10s - 72°C 30s. The cycling was stopped after the exponential phase, and the reaction products were double size-selected using Roche KAPA Pure beads at 0.8x-1.5x. Index PCR was performed according to the kit protocol alongside the whole transcriptome library. The final indexed libraries were quantified using KAPA Library Quantification Kit (Roche 7960140001) for pooling and sequenced at 2x150 cycles on NovaSeq X Plus by Novogene.

5.7 Preparation of filtered Sharpr-MPRA (R0) for R1-MPRA model training

We downloaded the raw read count-level Sharpr-MPRA dataset from Gene Expression Omnibus, accession number GSE71279, via the Dropbox hosting link provided by Movva et al.: https://www.dropbox.com/sh/wh7b30dauxuajcw/AABQsvfmG65knGbFv0UsIcv1a?dl=0. The Sharpr-MPRA library contained a few duplicated sequences, thus measurements for 4912 enhancers are included in the dataset 2-3 times. These were not explicitly removed for the initial model training. We filtered out all sequences with a minP DNA read depth < 200, log2(HEPG2) < -2.4, or log2(K562) < -2.4 in either replicate. log2(RNA/DNA) was calculated using raw RNA and DNA counts with a pseudocount of 1. This resulted in 6,694 sequences derived from HepG2 DNase sites, 8,329 sequences derived from K562 sites, and 14,868 sequences derived from other cell type sites (Huvec and H1hesc).

5.8 R1-MPRA model training

The model architecture used in generating R1-MPRA is constructed as follows. The one hot-encoded Input Layer (145x4) is passed in parallel to three branches of convolutional layers. Each branch contains 600 filters of size 7, 11, or 25; and feeds the output of the convolution through BatchNormalization, ReLU activation, Global MaxPooling, and Dropout (0.075). The outputs of each branch are concatenated and passed to a Dense layer with 64 nodes and ReLU activation, before finally feeding into the 2-node output layer with linear activation predicting log2(HEPG2) and log2(K562). We trained our models following the multilayer CNN architecture implemented by [37]/Basset [87]. We used a grid search approach to determine optimal filter sizes, number of filters, dense layer size, dropout rate, and learning rate for our specific datasets. Models were trained in TensorFlow using the ADAM optimizer (learning rate = 2e-4, beta_1 = 0.9, beta_2 = 0.999) and a batch size of 64, for 60 epochs with early stopping (patience = 8, min_delta = 1e-6). For these models the R0-MPRA data was randomly split into training, validation, and test sets containing 19,973, 4,961, and 4,957 sequences, respectively. All data splits were augmented with reverse complements.

To avoid overreliance on any single model for sequence design, we trained a total of 120 models divided in three categories. First, we trained 10 “Single” models on the same random training/validation/test data split but with different randomly-initialized network

weights. Second, we trained 100 CNN models with the same architecture on randomly resampled bootstraps of the merged training and validation data splits ("Boot" models). Third, we formed 10 ensemble predictors by splitting the bootstrap models into groups of 10 and averaging the outputs of every model within the group ("Ensemble" models). Evaluated on the same held-out test dataset, prediction-measurement correlations from all models were on par with inter-replicate correlations (**Supplementary Figure 2.1C**), with Ensemble models having the best performance, followed by Single and Boot models.

5.9 R1 DESIGN METHOD IMPLEMENTATION

For the R1-MPRA library, we used three design methods based on different general principles: simulated annealing (Monte Carlo sampling), Fast SeqProp (gradient descent) and Deep Exploration Networks (DENs) (deep generative models). Sequences were designed to maximize or minimize $\log2FC_{H2K}$, clipped to a threshold.

5.9.1 *Deep Exploration Networks*

In Deep Exploration Networks [53] a generative neural network is trained to produce sequences that simultaneously maximize a fitness score derived from a predictor model while penalizing the similarity of generated sequences. At a high level, the cost function for training a DEN has 3 components: a fitness cost, diversity cost, and entropy cost, each with an associated weight. Here, the fitness cost was $\max(\log2FC_{K562} - \log2FC_{HepG2} + \text{fitness\_target},0)$ for HEPG2-targeted designs, and $\max(\log2FC_{HepG2} - \log2FC_{K562} + \text{fitness\_target},0)$ for K562-targeted designs. This loss function was implemented because empirically it was found that maximizing the predicted model output unbounded yielded sequences with unnatural characteristics that we initially hypothesized to be undesirable. Without this loss clipping, generated sequences were predicted to have cell type-specificity several orders of magnitude beyond the training data, and exhibited low diversity due to strong single or di-nucleotide repeats. The following parameters were used in the cost function for training all DENs: fitness target = 2.5, fitness weight = 0.0075, entropy_min_bits = 1.8, entropy weight = 0.5, similarity margin = 0.5, similarity weight = 5.0.

For each generator trained, we generated 1000 sequences and took the top 10 highest HEPG2- or K562-specific predicted sequences. DENs were used to generate 10 sequences per model per cell type from Single models 0-9, Boot models 0-9, and Ensemble models 0-9 (600 sequences total).

5.9.2 *Fast SeqProp*

Fast SeqProp[52] is a technique previously developed by our lab in which a PWM is optimized via gradient descent in conjunction with the straight-through gradient estimator, such that sequences sampled from the PWM maximize the output of a predictor model. For each sequence generated with Fast SeqProp, we ran the algorithm for 200 gradient updates to generate 10 sequences, then included only the sequence with the highest (HepG2 target) or lowest (K562 target) $\log2FC_{H2K}$ score in our designed library. We use n_sequences=10, n_samples=1, n_epochs=1, steps_per_epoch=200, pwm_target_bits=1.8, pwm_entropy_weight=0, fitness_target=3, and ADAM optimizer with learning_rate=1e-3, beta_1=0.9, and beta_2=0.999. The loss function used is $\max(\log2FC_{K562} - \log2FC_{HepG2} + \text{fitness\_target},0)$ for HEPG2-targeted designs, and $\max(\log2FC_{HepG2} - \log2FC_{K562} + \text{fitness\_target},0)$ for K562-targeted designs. Fast SeqProp was used to generate 2 HEPG2-targeted sequences and 1 K562-targeted sequence from Single models 0-9 (2 K562 sequences per model intended, the 2nd omitted by error), and 1 sequence per model per cell type for Boot models 1-99 (model 0 omitted by error). 228 sequences total.

5.9.3 *Simulated Annealing*

In simulated annealing[88,89], sequences are randomly initialized, then at each step a random mutation is introduced at a random position and probabilistically accepted based on the current temperature value and the change in predicted fitness score. The temperature value is decayed with each iteration so that the algorithm is less likely to accept deleterious mutations as the step number increases. Here, we run simulated annealing for 1000 iterations per sequence, decaying the temperature from an initial value of 0.1 to a minimum value of 0.05 after every 100 iterations, using an exponential scale factor of 0.143. The loss function minimized was $\log2FC_{K562} - \log2FC_{HepG2}$ for HepG2 targets, and $\log2FC_{HepG2} - \log2FC_{K562}$ for K562 targets. Simulated Annealing was used to generate 1 sequence per model per cell type for Single models 0-9 and Boot models 1-99 (model 0 omitted by error). 218 sequences total.

5.9.4 *Hand-crafted motif repeats*

Handcrafted homotypic motif repeat enhancers were designed for 9 known TFBS motifs from the CISBP2.0 database, which were embedded in randomly generated sequences with multiplicity values of 1-7 for all motifs except HNF4A (multiplicity=1 omitted by

error). 62 sequences total. Motifs were selected based on prior reporting of HepG2 and K562 specificity, as well as enrichment analysis on the designed sequences (**Table 2.3**).

Random sequences were generated using the following background distribution from R0: A=0.230, C=0.254, G=0.284. T=0.231. Each sequence used a different random background.

5.9.5 *R0 control sequence selection*

In total 100 sequences from the filtered Sharpr-MPRA dataset were re-measured in R1: 25 selected from both the 99th and 1st percentiles of $\log2FC_{H2K}$ scores, and 50 selected at uniform intervals along the $\log2FC_{H2K}$ range. The reverse complements of these 100 control sequences were synthesized as well.

5.10 R1-DHS DESIGN AND MODEL TRAINING

5.10.1 *Generative Adversarial Network (GAN) training*

We use a training set of 758k DHSs located on chromosomes 3-X to train a Generative Adversarial Network (GAN) [54] for generating synthetic sequences with characteristics similar to endogenous human accessible genomic regions. Specifically, we obtained delineations and annotations for 3.5M+ DNase I Hypersensitive Sites (DHSs) [20], and filtered out DHSs with length less than 200bp and an annotated "mean signal" confidence score less than 0.5. We truncated each DHS to 200bp, centered on their annotated "centroid" position. In case a DHS summit position is less than 100bp from its annotated start or end position, we compensate by including additional length at the other end of the summit. This results in a set of 918,057 DHS sequences of length 200bp. We split these into general training (758,692 sequences, 82% of total, chromosomes 3-X), validation (78,108, 9%, chromosome 2) and test (81,257, 9%, chromosome 1) sets.

The GAN model consists of basic generator and discriminator functions using both convolutional and fully connected layers (**Table 2.4**). To improve training stability and maintain diversity of generated sequences, we use hinge loss and apply spectral normalization [90] to the convolutional and fully connected layers of the discriminator. After training, the generator provides a mapping from random input seeds (length 100) to synthetic DNA sequences (length 200bp). Generated sequences generally reflect

characteristics of endogenously accessible sequences, as shown by matching nucleotide composition of G/C content and 4-mer sequence patterns (**Supplementary Figure 2.3B,C**).

5.10.2 *Classifier training*

We tune generated sequences *in silico* to increase specificity to cellular contexts of interest by way of a separately trained multi-class classification model (**Figure 2.1D,** Table 2.5) discriminating between three classes of accessible elements: 1) 'K562', 2) 'HepG2' & 3) 'Other'.

These classes are defined using a combination of DHS component annotations[20] and observed accessibility in K562 or HepG2 cells. To maximize the amount of component-relevant signal, minimize inclusion of experimental noise and promote a high-contrast classification task, we select DHSs for each class to be relatively component-selective, as follows. We define the "purity" of each DHS as the proportion of its dominant component annotation divided by the sum of all component annotations. We then define the 'Other' class of DHSs by selecting the 625 highest purity DHSs from each of 16 NMF components, excluding any DHSs that occur in HepG2 or K562 biosamples, for a total of 10k DHSs. For the other two classes, we focus on the 'Digestive' component for HepG2 and the 'Myeloid/Erythroid' component for K562. Specifically, we co-rank Digestive-component purity and the number of HepG2 biosamples in which a DHS is accessible to select the top ranking 10k 'HepG2' DHSs. We follow the same procedure for the 'K562' class, using the 'Myeloid/erythroid' component. We apply this procedure to chromosomes 3-X to create a dataset of 30k DHSs with high-confidence component labelings, to be used as a training set. We follow the same procedure to select 1k sequences per class from chromosome 2 as a validation set for tuning model hyperparameters, and similarly from chromosome 1 as a test set.

5.10.3 *Sequence tuning*

We use the output nodes of our trained classification model to guide the sequence tuning process. Specifically, we search the latent space of the trained generator network for sequences that maximally activate the classification node corresponding to the class of interest. For each tuning run, one class is chosen as a target, and the final classification node (pre-softmax) corresponding to that class is multiplied by -1 and used as loss. We then backpropagate this loss through the networks to the generator's latent space, and update the input seed to generate a slightly more context-specific sequence. For each generated

sequence, we perform 10,000 tuning iterations, converging to an input seed that results in high activation of our target class.

To generate a tuned set of sequences for subsequent experimental validation (R1-DHS), we generated and tuned 300 sequences to the K562 and HepG2 target classes. For each sequence and each target class, we selected a single iteration after convergence as the final representative tuned sequence for subsequent experimental validation. The median selected iterations were 4750 (HepG2) and 3010 (K562), well within range of any notable sequence divergence relative to the training set (**Supplementary Figure 2.3F,G**), yielding sequences that are primed for accessibility in the selected cellular contexts, while preserving identity with the originally generated untuned sequence. We refer to this library as R1-DHS.

5.11 Second generation model training

To train models for R2 enhancer design, R0 data was reprocessed with DESeq2, sorted by descending $\log2FC_{H2K}$, then split into 10 crossfolds snakestyle to ensure equivalent distributions of enhancer activities in all the folds. R1-MPRA and R1-DHS were split into crossfolds according to the same scheme, with only 5 crossfolds used for R1-DHS due to the smaller library size.

We pre-trained 9 independent models ("M0") on the DESeq2-processed R0 splits, holding out Crossfold 0 as a constant test set and rotating through the remaining crossfolds as the validation set. Before retraining, we added L2 regularization to the dense layer weights after model interpretation indicated overemphasis on a single TP53 motif per sequence even in sequences with high TP53 multiplicity; while this may be an accurate representation of the biology, we deemed this more likely a model pathology that might also obscure the importance of non-TP53 motifs. We then re-optimized model hyperparameters using the keras_tuner implementation of the Hyperband strategy, resulting in minor tweaks to layer and filter sizes. The re-optimized architecture consisted of 3 parallel branches of 608 filters each of size 11, 15, and 21, trained with a dropout rate of 0.181; followed by a dense layer of size 224 with an L2 weight of 1e-3; trained with the Adam optimizer (learning rate = 3e-4, beta_1 = 0.9, beta_2 = 0.999) and early stopping monitoring the validation loss (min_delta = 1e-6, patience = 5). To accommodate eventual training on R1-DHS data, the input layer was expanded to size 200; any sequences shorter than this were 0-padded symmetrically on both sides.

To train M0+1 models, we held out Crossfold 0 from the R1 and R1-DHS datasets, then finetuned each M0 model on the remaining data, rotating through the R1 crossfolds as the validation set; all R1-DHS splits were used for training. A learning rate of 1e-4 was

used, in addition to a ReduceLROnPlateau callback (factor = 0.25, min_delta = 1e-2, patience = 4). Models were finetuned for a maximum of 250 epochs using early stopping (min_delta = 1e-3, patience = 5), batch_size = 64.

Finally, we trained 9 M1 models using the same architecture and optimizer hyperparameters as M0 models, but using only the R1 and R1+DHS splits without a pre-training step with R0.

5.12 R2 Design

All R2 designs were Fast SeqProp-based, and implemented an additional loss term not used for R1 designs (pwm_loss) penalizing single nucleotide repeats. We used target_weight = 1, pwm_weight = 2.5, entropy_weight = 1e-3, learning_rate = 1e-3, n_iter_max = 1000, and default for all other parameters.

5.12.1 *Unbounded objective*

An unbounded Fast SeqProp objective was used with both M0+1 and M1 models. 200 sequences were designed to maximize both HepG2- and K562-specificity unbounded, and the top 130 (M0+1) or 105 (M1) predicted sequences for each cell type were selected. The loss function minimized was $\text{log2FC}_{\text{K562}}$ - $\text{log2FC}_{\text{HepG2}}$ for HEPG2-targeted designs, and $\text{log2FC}_{\text{HepG2}}$ - $\text{log2FC}_{\text{K562}}$ for K562-targeted designs. 470 sequences total.

5.12.2 *Clipped objective*

200 sequences were designed to maximize both HepG2- and K562-specificity with a clipped Fast SeqProp objective, and the top 110 predicted sequences for each cell type were selected, using the M0+1 model. The loss function minimized was $\text{log2FC}_{\text{K562}}$ - $\text{log2FC}_{\text{HepG2}}$ - target for HEPG2-targeted designs, and $\text{log2FC}_{\text{HepG2}}$ - $\text{log2FC}_{\text{K562}}$ - target for K562-targeted designs, where target = 1.1X the highest predicted specificity on R1 sequences according to the M0+1 model. 220 sequences total.

5.12.3 *Max1 and Min1 designs*

100 sequences each were designed to maximize $\text{log2FC}_{\text{HepG2}}$ or $\text{log2FC}_{\text{K562}}$, regardless of the other cell type ("Max1"), using the M0+1 model. The top 10 predicted sequences for

each cell type were selected for inclusion in R2. The same procedure was applied using a clipped objective as above. 40 sequences total.

10 clipped and 10 unbounded designs per cell type ("Min1") were also generated as above, this time minimizing $\log2FC_{HepG2}$ or $\log2FC_{K562}$, regardless of the other cell type. 40 sequences total.

5.12.4 *Ablations*

Ablations were performed on the top 5 most specific enhancers measured in the R1 library. All possible single and pairwise double ablations of motifs were performed, unless a sequence only had 2 motifs, in which case the double ablation was not performed. Motif identification and coordinates were obtained via FIMO on the R1 library. To ablate a motif, the motif sequence was replaced with an equal-length subset of a dinucleotide-shuffling of the entire enhancer sequence. Dinucleotide shuffling was not performed on the motif sequence itself due to the high risk of motif-like elements emerging from the shuffling of a short motif sequence. 3 dinucleotide shuffles were performed per ablation. A 1 bp buffer upstream and downstream of the motif coordinates from FIMO were included in the ablation.

5.12.5 *Re-optimized R2 sequences*

A masked version of Fast SeqProp was used to optimize all non-motif sequence while preserving motif sequence unchanged. The same enhancers and motif annotations used for the ablations were used here. 10 re-optimizations of each enhancer were generated. As a control, non-motif sequence was dinucleotide shuffled 5 times per enhancer.

5.12.6 *Target designs*

4 target values were uniformly chosen in each cell type with a minimum $|\log2FC_{H2K}|$ of 2 and maximum of the highest predicted specificity on R1 sequences using the M0+1 model. 100 sequences were generated per target, and the top 20 sequences with the closest predicted specificity to the target value were selected for inclusion in the R2 library. The Fast SeqProp objective was the same as for the Clipped designs. 160 sequences total.

5.12.7 *Shorter enhancer design*

Enhancers with length <145bp were generated similarly to the Unbounded designs above, using 0-padding for the 200 bp model input size, and setting the gradient to 0 in the FSP algorithm for all unused bp positions. For each cell type, 25 enhancers of length 72bp, 25 enhancers of length 50bp, and 5 enhancers of length 25bp were designed.

5.12.8 *Control sequence selection*

The top 5 most specific sequences from R1 for each cell line were including in the R2 library. 100 sequences each were randomly selected from HepG2-targeted and K562-targeted R1 designs. 210 sequences total.

Additionally, the 2 sequences from R1 with lowest measured $\log2FC_{HepG2}$ + $\log2FC_{K562}$ were selected as negative controls. A dinucleotide shuffled variant of each of these sequences was also included in R2.

5.13 scMPRA DATA PROCESSING

For both MPRA and transcriptome libraries, we initially performed anchored alignments to detect and filter the reads for correct amplicon structures using cutadapt [91]. For MPRA libraries, we extracted and concatenated only the barcode sequences from the reads and clustered them using starcode [92] in order to remove PCR chimerism artifacts. We used STARsolo [93] to align the reads to the human genome (hg38) and enhancer barcode references and to quantify the read counts. We filtered the cells based on a knee plot of per-cell transcriptome UMI counts. For the transcriptome data used for clustering and pseudobulk binning, we use normalized expression matrix following SCTransform scaling as implemented in Seurat [94].

We observed that random barcodes were not an effective proxy for DNA copy numbers as indicated by: 1) the sparse, singleton-dominated per-plasmid expression levels (**Supplementary Figure 2.13E**); and 2) the poorer correlation of cell-type specific enhancer activities with the bulk data calculated using averages of random barcode-normalized single cell expressions compared to the correlation calculated using pseudobulk (**Supplementary Figure 2.13F**). Thus, we used pseudobulk quantifications in our downstream analysis.

5.14 Sequence diversity analysis

Each sequence is represented as a 256 element vector with each element corresponding to the count of a corresponding 4-mer in the sequence. For two sequence libraries, the Euclidean distance between a member of the first library and all sequences in the second is computed, and the minimum (i.e. "nearest neighbor") distance is recorded. This process is repeated for all sequences in the first library.

5.15 Motif calling

For all libraries (R0, R0-DHS, R1, R1-DHS, and R2), we separately ran FIMO with the default p-value threshold of 1e-4, scanning sequences for matches to the JASPAR 2022 Non-redundant Vertebrate database. We then applied an additional q-value filter, discarding all motif hits with q-values exceeding a threshold of 5e-2. Finally, we applied a custom position-wise clustering algorithm to collapse overlapping motif hits. First, within each sequence we collapsed all overlapping motifs with the same motif ID to the instance with the lowest p-value. Then, we scanned across all the positions in a sequence from 5' to 3' and for any position included in multiple motif hits, we kept the motif hit with the lowest p-value and discarded the rest. We allow an overlap of 3 bp without collapsing motifs together.

For subsequent analysis, we additionally clustered motifs using the JASPAR Core Vertebrates RSAT clusters (841 motifs, 137 clusters). (JASPAR - JASPAR CORE Vertebrates clustering (genereg.net)).

For the purpose of **Supplementary Figure 2.3H,I,** we obtained motif files HNF4A_nuclearreceptor_3 and SPI1_ETS_1 from https://www.vierstra.org/resources/motif_clustering and scanned sequences using FIMO with a threshold of 1e-5.

5.16 Motif vocabulary size downsampling estimation

To estimate the total number of unique motifs in each library, we downsampled each library to the smallest number of sequences across all libraries (688). We randomly sampled 1000 bootstraps of 688 sequences from each library and calculated the mean and standard deviation of the total number of unique motifs across all bootstraps.

5.17 Model interpretation

M0, M0+1, and M2 ensemble model predictions were interpreted on all R1-DHS, R1-MPRA, and R2 sequences, as well as 3000 randomly sampled R0 sequences, using the DeepExplainer library of the SHAP[64] package, 100 dinucleotide shuffles per sequence. SHAP values were obtained for both $\log2FC_{HepG2}$ and $\log2FC_{K562}$ predictions, and the difference between these values for a given nucleotide was taken as the $\log2FC_{H2K}$ or $\log2FC_{K2H}$ SHAP value. To compare SHAP values across models we rescaled using the batch correction regression coefficients. To calculate motif-wise SHAP scores, we sum the SHAP scores for all nucleotides of the motif.

5.18 Reduced library size estimations

To estimate the best enhancer for a given library (**Figure 2.1F**), cell type, and library size n, n sequences were sampled with replacement 10,000 times, and the best $\log2FC_{H2K}$ for each bootstrap was then averaged together. For this analysis we used a subset of R0, filtered for sequences annotated with chromatin state = 5 ("Enhancer", from a 25-state ChromHMM model[51]) to ensure the strongest comparison with synthetic enhancers; and treat the cell type of origin as the target cell type. For each of the 10,000 bootstraps, we also calculate which library (R0, R1, R1-DHS, R2) yields the best enhancer in each cell type, and report the percentage of bootstraps in which the best enhancer comes from each library.

5.19 Investigation of R1-DHS vs R1-MPRA data for model finetuning and design

For each model retraining condition, 1000 sequences were bootstrap resampled either all from R1-DHS, all from R1-MPRA, or equally from both datasets. Within each dataset, sequences were sampled equally from HepG2- and K562-targeted designs. For each bootstrap, 15% of sequences were allocated to a validation set, and the remaining used for training. For this analysis, each model of the M0 ensemble was finetuned on a separate bootstrap, for 9 bootstraps total. Performance was evaluated for each of the 9 models per training condition on a test set consisting of all *de novo* R2 enhancers.

To estimate design performance, the 9 models per training condition were combined into an ensemble, and Fast SeqProp was used to design 100 sequences per cell type to maximize predicted specificity. M2 model predictions were computed for all designs.

5.20 Investigation of proportion of strong enhancers for model finetuning and design

The R1-MPRA dataset was split into quartiles along the $\log2FC_{H2K}$ measurement. For the "weak" training set, the 2nd and 3rd quartiles were combined. For the "strong" training set, the 1st and 4th quartiles were combined. The "all" training set consisted of the entire R1-MPRA data. For each training set, 1000 sequences were bootstrap resampled. Models were trained as and sequences designed as in the previous section.

5.21 Investigation of training data size for model finetuning and design

For these experiments, we first retrained 5 M0 ensembles using the same 10-fold split of R0 data originally used for M0 training, but rotating the crossfold used as the test set from index 0-4. Then, for each of these 5 M0 ensembles, we sampled n total sequences from R1-MPRA and R1-DHS (without replacement), maintaining the same ratio between datasets as used for M0+1 training, split these sequences into 9 crossfolds, and finetuned each of the constituent models of the M0 ensemble on 8 crossfolds at a time, using the 9th as a validation set. This yielded 5 ensembles per n value, with each ensemble composed of 9 models. We additionally trained 5 new M0+1 models by finetuning the test set-rotated M0 models trained above, using an analogous test set rotation of the R1 data.

To estimate design performance, Fast SeqProp was used with the 5 ensembles per n value to design 100 sequences per cell type to maximize predicted specificity. M2 model predictions were computed for all designs.

5.22 Ablation analysis

For each single and double ablation, 3 different dinucleotide shufflings of the target motif(s) were performed. The measurements for each ablation were averaged together, and the ablation score calculated on these averages. The deviation score was also calculated on the averaged measurements. For R2 Seq 976 and 1041, we estimated the double ablation score *in silico*. We finetuned the 9 individual M0+1 models on all R2 measurements (randomly split into 10 crossfolds) using the same hyperparameters used for R1 finetuning, then obtained ensemble predictions on 3 ablations per enhancer.

APPENDIX B – METHODS IN CHAPTER 3

6.1 Data preprocessing for DHS64

ENCODE DHS Index data was downloaded from Zenodo (https://doi.org/10.5281/zenodo.3838751) as indicated in the original publication[68]. These data comprises four main elements: 1) a matrix of normalized DNase-seq read density values (i.e. continuous accessibility signals) with 3,591,898 rows corresponding to DHSs and 733 columns corresponding to DNase-seq experiments (biosamples); 2) a similar matrix containing binary values corresponding to peak calls, 3) annotations for all DHSs including their genomic coordinates (e.g. chromosome, summit, start, and end positions); and 4) annotations for all biosamples including germ layer, system, and organ of origin, ENCODE experiment and protocol IDs, and sequencing quality metrics such as number of reads and DNase-seq Signal Portion of Tags (SPOT) scores.

Previously, the DHS Index data was found to decompose into 16 biological "components": 15 strongly associated with samples of different tissue of origin and with DHSs specifically represented in those samples, and one component corresponding to broadly accessible (i.e. non-specific) DHSs. For the DHS64 model, we chose 64 biosamples – including primary tissues *ex vivo*, primary cultured cells, and immortalized cell lines – with strong association to each of the 15 tissue-specific components and with high DNase-seq quality metrics. To preprocess the continuous accessibility matrix, we filtered its columns to retain the 64 selected biosamples, performed quantile normalization across biosamples, and log-transformed values via $\log_{10}(x + pc)$, where the pseudocount pc was chosen as the smallest non-zero value in each biosample. The binary accessibility matrix was similarly filtered to retain only the 64 relevant columns.

DHS sequences were obtained by mapping the annotated coordinates to the GRCh38 genome assembly. If the annotated DHS was longer than 500 bases, and if the start (end) position was closer than 250 bases from the summit, the end (start) position was adjusted to reach a total of 500 bases, otherwise 500 bases centered at the summit would be used. DHS annotations were further augmented with the following: 1) Distance to the closest transcription start site (TSS) extracted from the Gencode v42 basic annotations, where the TSS positions were obtained by filtering annotations with "feature" = "transcript" and "gene_type" = "protein_coding" or "lncRNA", and extracting the start or end positions depending on the strand orientation. 2) Distance to the closest TSS annotated in refTSS v3.3[95]. 3) Chromatin state annotations from the "full

stack" version of the ChromHMM model[96], trained on >1000 chromatin annotation tracks across >100 cell and tissue types, and thus expected to generalize to a variety of cell types. Finally, the following DHSs were removed: 1) 31 DHSs with sequences containing non-canonical bases, 2) 126,840 DHSs not in the autosomes, and 3) 1,453,085 DHSs not active in any of the 64 selected biosamples (i.e. all-zero row in the binary matrix). Thus, only 2,011,942 DHSs were used to train and evaluate DHS64 models.

6.2 DHS64 MODEL TRAINING AND EVALUATION

DHS64 is a deep residual neural network that accepts a zero-padded, one hot-encoded input sequence of up 500nt and predicts both the $\log_{10}$-transformed continuous accessibility signal as well as the peak call probability of all 64 selected biosamples. Predicting both data modalities was found to slightly improve performance in preliminary testing. A detailed schematic of model architecture can be found in **Supplementary Figure 3.2A**. Chromosome-aware data splits (i.e. training, validation, and test sets) were designed using the prtpy python package (https://github.com/coin-or/prtpy) such that each split comprised a rough 80/10/10 (train/validation/test) ratio of DHSs, and DHSs from the same chromosome were contained in the same set. Three different splits – arbitrarily named #0, #1, and #3 – were used in this manuscript to train independent models. The following data augmentation techniques were used: 1) both the DHS sequence and its reverse complement were fed into the model during each epoch, and 2) left- or right-zero padding were randomly selected for each sequence. The overall training loss was comprised of a mean squared error term between the observed and predicted continuous $\log_{10}$ accessibility signals and a binary cross-entropy term between the observed peak calls and their predicted probabilities, weighted equally. We used the Adam optimizer, a learning rate of 2e-4, and a batch size of 256. Early stopping was used by monitoring the validation loss with a patience parameter of 2. All training and model analysis were performed in tensorflow 2 (specific versions between 2.4 and 2.10) in python 3 (specific versions between 3.7 and 3.10). Model training was performed on Amazon AWS EC2 g5.2xlarge instances, and other model-related analyses were performed on g5.xlarge instances.

Since our primary modeling goal was to capture the determinants of cell type-specific activity, we evaluated model performance mainly on two data sets: 1) all held-out test DHSs, and 2) a subset with DHSs likely to be cell type-specific enhancers selected via the following criteria: i) annotated "mean_signal" greater than 0.5. This corresponds to the mean signal across all DHS Index biosamples with a positive peak call, serving as a proxy of the "strength" of this DHS when active. ii) a distance greater than 2kb from the

closest gencode- and refTSS-annotated TSS, to remove promoters which tend to have strong but non-specific signals. iii) Full-stack ChromHMM annotations corresponding to the state groups "enhancers", "weak enhancers", and "transcribed and enhancer". iv) Active in 10 or fewer biosamples out of the 64 selected (i.e. at most 10 ones in the corresponding row of the binary matrix). While the latter set is dramatically smaller (32,151 DHSs compared to 186,404 in the full test set for chromosome split #3), it is a more interesting evaluation set for our purposes. A model naively trained on all available training DHSs (n = 1,636,099 for chromosome split #3) performed well on the larger test set of all DHSs, but significantly worse on the reduced cell type-specific enhancer-like set (**Supplementary Figure 3.2D**). We found that restricting the training dataset to those DHSs active in 10 or fewer biosamples (n = 1,401,497) improved performance on the cell type-specific enhancer-like test set while reducing it in the larger set of all DHSs (**Supplementary Figure 3.2E**). More detailed analysis revealed that the latter model performed better on more cell type-specific DHSs, independently of whether they had the mean_signal, TSS, and ChromHMM enhancer annotations (**Supplementary Figure 3.2F-G**). Further restricting the training set to those with mean_signal, TSS, and ChromHMM annotations did not improve performance, likely because the number of samples decreased too much (n = 278,266). With the exception of **Supplementary Figure 3.2**, all DHS64 models in this manuscript refer to models trained on cell type-specific DHSs only, and the performance metrics are given in relation to the cell type-specific enhancer-like test set.

6.3 DHS64-GUIDED DESIGN OF BIOSAMPLE-SPECIFIC ENHANCERS

Let $n = 64$ be the total number of biosamples modeled by DHS64 and $x_{i\in\{1\ldots n\}}$ be the predicted $\log_{10}$ accessibility signal for the i-th modeled biosample. To optimize sequences for one target biosample with index $i = t$, we maximize the objective function $x_t - \frac{1}{n}\sum_i x_i$. This is equivalent to $\left(\frac{n-1}{n}\right)\left[x_t - \frac{1}{n-1}\sum_{i\neq t} x_i\right]$, i.e. the difference between the target signal and the average signal across the non-targets – the specificity score defined in the main text up to a constant. To optimize sequences specific to two or three target biosamples, where $T = \{t, u\}$ or $T = \{t, u, v\}$ is the set of target biosample indices, we maximized the objective function $\min_{i\in T} x_i + \frac{a}{m}\sum_{i\in T} x_i - \frac{1}{n-m}\sum_{i\notin T} x_i$, where $m = 2$ or 3 is the number of target biosamples and $a = 0.2$ is a constant. Maximizing a combination of the minimum and average target signal gave the best results, as maximizing only the average sometimes resulted in high signal in only one target whereas maximizing only the minimum sometimes resulted in low overall target signal values.

To avoid generating sequences that overfit a particular model, we used two strategies. First, we optimized sequences against predictions from a "pessimistic" ensemble of two independently trained DHS64 models (data splits #1 and #3), where the predicted target signal was defined as the minimum across both models, and the non-target predictions were taken from their maximum. Second, after sequence generation, accessibility predictions and performance metrics were recalculated using a third "validation" DHS64 model (split #0). All plots related to predictions of designed sequences (e.g. fig 1F-G and Supp fig 3) were generated with this model. The predicted peak call probabilities were not used. To design sequences specific to one biosample with intermediate target accessibility $x_t = y$, we maximize the objective function $-|x_t - y| - \frac{1}{n}\sum_i x_i$.

We reimplemented Fast SeqProp[71] (https://github.com/castillohair/corefsp) to simplify the API and enable compatibility with TensorFlow 2. The following design parameters were used: target weight: 1, PWM weight: 3, entropy weight: 1e-3, learning rate: 1e-3, number of iterations: 2500. Similarly, we updated the DEN[72] codebase (https://github.com/castillohair/genesis) to enable TensorFlow 2 compatibility. We trained individual DENs for each target biosample, each of which required different parameters to at least match the predicted performance of Fast SeqProp-designed sequences.

6.4 DHS64-GUIDED DESIGN OF ENHANCERS WITH COMPLEX DESIGN OBJECTIVES

To optimize enhancers specific to two or three targets, where $T = \{t, u\}$ or $T = \{t, u, v\}$ is the set of target biosample indices, we maximized the objective function $\min_{i \in T} x_i + \frac{a}{m}\sum_{i \in T} x_i - \frac{1}{n-m}\sum_{i \notin T} x_i$, where $m = 2$ or 3 is the number of target biosamples and $a = 0.2$ is a constant. Maximizing a combination of the minimum and average target signal gave the best results, as maximizing only the average sometimes resulted in high signal in only one target whereas maximizing only the minimum sometimes resulted in low overall target signal values. Target pairs and triplets were chosen by sampling uniformly from the list of target cell lines used in MPRAs. 117 sequences per target pair or triplet were obtained via Fast SeqProp using the objective function just described and a similar model ensemble strategy as with single-target enhancers, followed by removing sequences with KpnI (`GGTACC` and `GTACC` at the 5' end) and XbaI (`TCTAGA`) restriction sites.

To design sequences with tunable activity, we maximized the objective function $-|x_t - x_t^*| - \frac{1}{n}\sum_{i \neq t} x_i$, where x_t^* is the accessibility setpoint value on target biosample t. We selected 120 x_t^* values uniformly spaced between biosample-specific lower and upper bounds spanning the range achieved by Fast SeqProp in the "maximize specificity" design

task described above. Specifically, the lower bound was set to the minimum average prediction value that a biosample reached when optimizing sequences for all 64 targets, and the upper bound was taken as 1.5 times the maximum average value. In addition, we explicitly penalized short sequences corresponding to the KpnI (`GGTACC` and `GTACC` at the 5' end) and XbaI (`TCTAGA`) restriction sites. For each target cell line used in MPRAs, we designed 120 sequences with Fast SeqProp, the objective function and setpoint values just described, and a model ensemble where non-target predictions were taken from the maximum across two models as described above, but target predictions were taken from their average.

6.5 Selection of DHS-sourced enhancer controls

To perform *in silico* analysis of synthetic enhancers across all 64 modeled biosamples, a set of matched "enhancer-like" DHSs were selected as follows: 1) DHSs in the Index dataset were filtered for i) annotated "mean_signal" greater than 0.5, ii) a distance greater than 2kb from the closest gencode- or refTSS-annotated TSS to remove promoters, iii) full-stack ChromHMM annotations corresponding to the state groups "enhancers", "weak enhancers", and "transcribed and enhancer", iv) peak calls in a number of modeled biosamples between 1 and 10. Note that these are similar criteria as those used to construct the DHS64 test set. 2) For a given target biosample, DHSs without positive peak calls in that biosample were discarded. 3). The remaining DHSs were sorted by the difference between target biosample accessibility and average accessibility across non-targets, and the top sequences were selected. For controls used for MPRAs, DHSs were truncated to their central 145bp, and an additional filtering criterium to exclude sequences with KpnI (`GGTACC` and `GTACC` at the 5' end) and XbaI (`TCTAGA`) restriction sites was used.

DHS controls targeting two and three cell types were chosen similarly with the following differences: In step 2, positive peak calls in all target cell types were required; In step 3, sorting was performed using the difference between the average target and average non-target accessibilities. Note that in some cases these stringent criteria resulted in insufficient or even zero sequences.

Negative control DHSs were chosen by randomly sampling from the set of enhancer DHSs ("mean_signal" > 0.5, $>$2kb from the closest TSS, enhancer-related full-stack ChromHMM annotations) with no peak calls in any of the modeled biosamples and no KpnI and XbaI restriction sites. An additional set of negative controls was generated via dinucleotide shuffling of the DHS-sourced negative controls.

6.6 Enhancer library design

The enhancer library used for MPRA experiments contained a large number of sequences selected or designed for specificity towards the 10 cell lines where MPRAs were conducted, with a minority of enhancers targeting the remaining 54 modeled biosamples, and additional controls. Specifically, our library included: 1) 1370 DHS-sourced enhancers selected to target one biosample (110 per MPRA cell line + 5 per remaining biosample). 2) 1770 synthetic enhancers designed with Fast SeqProp to target one biosample (150*10 + 5*54). 3) 1500 synthetic enhancers designed with DENs to target one MPRA cell line (150*10). 4) 1200 synthetic enhancers designed with Fast SeqProp for tunable specific activity towards each MPRA cell line (120*10). 5) 378 DHS-sourced enhancers selected to target 8 pairs of MPRA cell lines (50 for each of 7 pairs, 28 for SK-N-SH + GM12878). 6) 936 synthetic enhancers designed with Fast SeqProp to target the same cell line pairs (117 per pair). 7) 192 DHS-sourced enhancers selected to target 8 MPRA cell line triplets (up to 50 per triplet but often fewer). 8) 936 synthetic enhancers designed with Fast SeqProp to target the same cell line triplets (117 per triplet). 9) 20 DHS-sourced negative controls. 10) 20 dinucleotide-shuffled negative controls. 11) 40 highly specific HepG2- and K562-targeted enhancers from our previous publication[6], including 20 from the SHARPR-MPRA dataset[75] we used as training data as well as 20 *de novo* designed enhancers. 12) 740 enhancers with manually embedded TF motifs. This resulted in a total of 9102 enhancers. However, the DHS control selection process sometimes resulted in identical sequences for different targets, resulting in a final number of 8989 unique enhancers.

Two barcodes, to be cloned into the 3'UTR of the reporter gene to identify enhancers from their transcripts, were used per enhancer. Barcodes were generated with the `barcode_design` package (https://github.com/feldman4/dna-barcodes/blob/master/barcode_design.py) with options `--length 10 --distance 2 --limit 50000 --exclude "GGTACC|TCTAGA|ATC$|AATAAA|ATTAAA"`. The final oligo pool had the following structure: `ACTGGCCGCTTCACTG[145nt enhancer]GGTACCTCTAGA[10nt barcode]TGCCGGACCAGGTAGAT`.

6.7 Enhancer library cloning

The library described above was ordered as a ssDNA oligo pool (200nt, <=18k oligos) from Twist Biosciences. Cloning into a plasmid library was performed similarly to our previous work[6]. This process includes Gibson assembly of the oligo library into the backbone of the pMPRA1 plasmid (Addgene# 49349), restriction of the resulting plasmid to separate the enhancer from the barcode, and insertion of a pMPRAdonor2-derived

(Addgene# 49353) fragment containing the minP promoter and the luciferase gene via T4 ligation.

To assemble the oligo pool into the pMPRA1 backbone, we first resuspended the pool in ddH2O to 10ng/uL. 10ng were then used as PCR template with primers containing Gibson overhangs, using 25uL Phusion polymerase master mix (NEB M0531), 2.5uL 10uM primer CY01, 2.5uL 10uM primer SC_01264, and water to 50uL, with the following thermocycler program: denature at 98°C for 30s; denature/anneal/extend at 98°C for 10s, 62°C for 30s, and 72°C for 15s for 15 cycles; and finally extend for 10 minutes. The amplified library was purified using KAPA pure beads (Roche KK8002) per manufacturer's instructions using a beads-to-product ratio of 1.5x, and resuspended in 15uL ddH2O. In parallel, pMPRA1 was digested using SfiI (NEB R0123) and the 2.5kb-long backbone was gel-purified. 158ng amplified library and 325ng pMPRA1 backbone (1-to-5 molar ratio) were assembled with the NEBuilder HiFi DNA Assembly kit (NEB E2621) in a 40uL reaction at 50°C for 60 minutes. The assembly reaction was purified with KAPA pure beads (2x bead-to-product ratio) and resuspended in 12uL. This reaction was transformed into NEB 10-beta electrocompetent cells (NEB C3020K) via two electroporations (2.5uL assembly product + 40uL cells), recovered in 1mL SOC at 37°C for 1h, and incubated in 200mL LB + Ampicillin overnight. 40uL of SOC was extracted after recovery to plate serial dilutions and estimate CFU numbers. The next day, plasmid was purified from the LB culture using the Qiagen Plasmid Maxi Kit (Qiagen 12162). The estimated CFU number was 1.53e9, or ~3.9k per library member.

To insert the minP and luciferase reporter cassette, we first digested the purified pMPRA1/library plasmid with KpnI (NEB R3142) and XbaI (NEB R0145) by preparing two identical 50uL reactions with the rCutSmart buffer and 2ug plasmid each, adding 1uL KpnI per reaction and incubating at 37°C for 2h, then adding 1uL XbaI and incubating for 37°C for 6h, followed by heat-inactivating at 65°C for 20 min. We then dephosphorylated with Antarctic phosphatase (NEB M0289) by adding 6uL of phosphatase buffer, 1uL water, and 3uL phosphatase directly to the 50uL reaction (final volume = 60uL), incubating at 37°C for 1h and heat inactivating at 80°C for 2 min. Finally, the resulting fragment was gel purified. In parallel, the pMPRAdonor2 plasmid was digested with KpnI and XbaI using an identical protocol without the Antarctic Phosphatase step, and the 1.78kb-long fragment was gel purified. Ligation was performed using T4 ligase (NEB M0202) via two identical 50uL reactions each with 2.5uL T4 ligase, 5uL 10x ligase buffer, 150.96ng pMPRA1/library-derived fragment and 187.5ng pMPRAdonor2-derived fragment (molar ratio 1-to-2, maximum ligated product: 250ng per reaction). Reactions were incubated at 16°C for 16h, 65°C for 10 min, and 4°C forever, purified with KAPA pure beads and a 1.5X beads-to-product ratio, and resuspended in

12uL water as above. This was then transformed into NEB 10-beta cells via two electroporations (2.9uL assembly product + 40uL cells each) as above. The next day, before plasmid purification, 6 glycerol stocks were prepared (800uL overnight culture + 800uL 50% glycerol) and stored at -80°C. The estimated CFU number was 1.79e9, or ~7.5k per library member. We used the Qiagen Maxi Kit with the remaining overnight culture to obtain the final purified plasmid library, hereafter referred to as library preparation DNA_20230902. At later dates, glycerol stocks were revived by thawing and aliquoting their entire contents in 200mL LB + 200uL 1000X Carbenicillin, incubating overnight, and purifiying with the Qiagen Maxi Kit as above. This resulted in library preparations DNA_20240105 and DNA_20240612.

Successful cloning and diversity of each library preparation was verified mainly by amplifying and sequencing a fragment from the plasmid library containing the enhancer region, the promoter, reporter gene, and barcode. First, a 20uL qPCR with Phusion polymerase was prepared with 10ng plasmid, 1uL 10uM primer MPRA_seq_F, 1uL 10uM primer MPRA_seq_R, 10uL Phusion master mix, and 0.2x EvaGreen (Biotium 31000). The reaction was run for 30 cycles and a cycle number before the end of exponential amplification was selected (usually ~10). The reaction was then scaled up to 100-150uL total volume without EvaGreen and ran as a regular PCR with the previously determined number of cycles. The resulting product was verified via gel electrophoresis, purified with the DNA Clean & Concentrator - 5 (Zymo D4014), and submitted to Plasmidsaurus (previously Primordium sequencing) for their long-read Premium PCR sequencing service. Each run resulted in 5000 long reads, which we verified for the expected constant regions such as promoter and reporter gene, for enhancers with the appropriate length and expected sequence, and for enhancer/barcode matching, within sequencing quality limits. In addition, we submitted for Sanger sequencing individual colonies from the plates used for CFU quantification, for both the intermediate plasmid (pMPRA1/library, 10 colonies) and final plasmid preparation (DNA_20230902, 10 colonies). We used primers MPRA_Seq_F and MPRA_Seq_R, which should cover the enhancer and barcode regions, respectively, and verified that enhancer sequences and enhancer/barcode matching was as expected.

6.8 Cell culture and transfection

Cells were cultured at 37°C and 5% CO2 using standard cell culturing techniques. All indicated media was supplemented with 10% FBS (Cytiva SH30396.03) and 100 U/ml Penicillin/Streptomycin (ThermoFisher 15140122). Cell lines listed as “adherent” were detached using Trypsin-EDTA 0.25% (ThermoFisher 25200056) except for NT2-D1, which

was detached with TrypLE (Fisher 12605028). During MCF7 passaging, both adherent and floating cells were retained as suggested by ATCC.

Library transfection was performed via lipofection or electroporation. Lipofection was performed using Lipofectamine 3000 (ThermoFisher L3000001) as follows: On day 1, cells were seeded at the indicated number. On day 2, lipofectamine reagent and DNA solutions were prepared as indicated, mixed, incubated for 15 minutes, and added slowly to the culture, after which the culture was gently swirled to mix and returned to the incubator. 4-6 hours later the media was changed. On day 4, cells were detached and RNA extraction was performed. Electroporation on suspension cultures was performed using the Neon Transfection System (Invitrogen MPK5000) as follows: On day 1 cells were counted, the indicated number of cells were aliquoted, then centrifuged and resuspended twice in PBS with no Ca2+ and Mg2+. Cells were then resuspended in 100uL Resuspension Buffer R with plasmid DNA and electroporated with a 100uL Neon Tip, and a Neon tube with E2 Electrolytic Buffer. Finally, cells were placed in an appropriate container with media and FBS but no Penicillin/Streptomycin, and returned to the incubator. On day 3, RNA extraction was performed. Electroporation on adherent cells was performed as follows: On day 1 cells were seeded at the indicated number. On day 2, cells were detached, then centrifuged and resuspended twice in PBS with no Ca2+ and Mg2+, electroporated and replated as indicated above. On day 4, cells were detached and RNA extraction was performed. Two replicates per cell line were performed. With each transfection, two parallel cultures in a 24-well format were maintained, one of which was transfected with a GFP plasmid under identical but downscaled conditions, in order to estimate transfection efficiency via flow cytometry on the last day. Total RNA was purified from cells using the Monarch Total RNA Miniprep Kit (NEB, T2010S) and either processed immediately or stored at -80°C for later processing.

VITA

Christopher Yin received his B.S. in Electrical Engineering from the University of California, San Diego in 2018. He attended the PhD program in Electrical and Computer Engineering at the University of Washington, Seattle, from 2020-2025, where he was advised by Professor Georg Seelig. His research focuses on applications of deep learning to understand and exploit the sequence grammar underlying cell type-specific gene expression, in conjunction with experimental approaches to validate model-based hypotheses. Chris is also a graduate of the Clarion West Six Week Writers' Workshop ('24), and will begin attending the Iowa Writers' Workshop in Fall 2025, where he will earn his MFA.