Title

Deep-learning approaches to predicting molecular phenotypes using sequence-to-function models

ISBN

9798189270871

Author

Keivanfar, Ryan Luca

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Deep-learning approaches to predicting molecular phenotypes using sequence-to-function models

by

Ryan L. Keivanfar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Katherine S. Pollard, Co-chair
Professor Liana Lareau, Co-chair
Professor Max Staller

Spring 2026

Deep-learning approaches to predicting molecular phenotypes using sequence-to-function models

Abstract

Deep-learning approaches to predicting molecular phenotypes using sequence-to-function models

by

Ryan L. Keivanfar

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Katherine S. Pollard, Co-chair

Professor Liana Lareau, Co-chair

Noncoding sequences coordinate gene regulation through transcription factor binding and haplotype-specific expression, but translating regulatory sequence into quantitative predictions at the level of individual molecules and individual people remains challenging. This dissertation addresses two aspects of that problem using genomic deep learning, with a shared focus on how data representation shapes model behavior.

The first project asks whether representing DNA as both nucleotide sequence and local structure changes what a transcription-factor binding model can learn and explain. In Chapter 2, I develop DeepShape, a convolutional neural network that augments one-hot DNA sequence with five DNA structural attributes, including minor groove width, propeller twist, helical twist, roll, and electrostatic potential, to predict transcription factor (TF) binding across 919 regulatory targets in 148 cell types. Shape features contribute prediction accuracy independently of sequence motifs, and DeepLIFT-based attribution applied to shape inputs recovers structural binding signatures not apparent from sequence alone, providing a new axis of interpretability for genomic deep learning models.

The second project asks how representing expression data at haplotype resolution changes sequence-to-expression modeling. In Chapter 3, I fine-tune the Enformer model for allele-specific expression (ASE) prediction on phased allelic counts from the Genotype-Tissue Expression project (GTEx) whole blood, obtained using phASER (phasing and allele-specific expression from RNA sequencing), with raw-count targets and mask-based uninformative-donor filtering as design choices motivated by the structure of phased ASE data. The trained model captures structure in allelic-imbalance magnitude on held-out genes, but does not recover the direction of imbalance. These results characterize the reach and current limits

of Enformer-based ASE fine-tuning and motivate next-generation architectures that make haplotype-specific variation more explicit.

In Chapter 1, I establish the conceptual framework for both contributions. I review how regulatory sequences shape gene expression, summarize the role of DNA shape readout in transcription factor binding, and trace the development of sequence-to-expression models from early convolutional architectures to recent long-range models such as Enformer, Borzoi, and AlphaGenome. I discuss the cross-individual evaluation gap that has emerged as these models have been applied to personal genome prediction, recent attempts to close it through fine-tuning on personal expression data and through allele-specific learning, and the modality gap between chromatin and expression prediction that frames the limitations Chapter 3 confronts directly.

In summary, my work characterizes how the choice of input representation, including the addition of DNA shape channels and the use of phased diploid haplotype inputs, changes what genomic deep learning models can learn and explain about regulatory sequence. These contributions clarify both the reach and the current limits of sequence-to-function modeling for transcription factor binding and haplotype-resolved expression, and motivate next-generation architectures that make structural and haplotype-specific variation more explicit.

For Dexter

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors, mentors, collaborators, family, friends, babies, and especially, my partner — thank you for your love, for your patience, for loving me through a long and uncertain stretch of life, and for being the home I always come back to. None of this would have been mine to finish without you. This work bears the imprint of so many hands, voices, and acts of care, and it would not have been possible without the community that carried, challenged, and sustained me along the way.

Chapter 1

Introduction

1.1 Overview

Noncoding sequences coordinate gene regulation through transcription factor binding and haplotype-specific expression, but translating regulatory sequence into quantitative predictions at the level of individual molecules and individual people remains challenging. This introductory chapter establishes the conceptual framework for the dissertation. I first review how regulatory sequences shape gene expression, then discuss the role of DNA shape readout in transcription factor binding (the focus of Chapter 2), and finally trace the development of sequence-to-expression models, the cross-individual evaluation gap that has emerged as these models have been applied to personal genome prediction, and recent attempts to close it through fine-tuning on personal expression data and through allele-specific learning. Together, these threads motivate the haplotype-resolved evaluation framework developed in Chapter 3.

1.2 Regulatory sequences and gene expression

The vast majority of genetic variants identified in genome-wide association studies fall outside protein-coding regions, implicating noncoding regulatory sequences as the primary conduit through which inherited genetic variation influences phenotype [1, 2]. To understand these effects, we need models that connect DNA sequence to molecular function: models that predict which regulatory elements are active, which transcription factors bind them, and how much each haplotype of a gene is transcribed in a given individual. This dissertation presents two complementary contributions to that program through a common lens: how the representation of biological data changes model behavior. Chapter 2 contrasts nucleotide sequence with explicit DNA structural features; Chapter 3 contrasts haploid and diploid sequence inputs, and bulk-expression and allele-specific expression targets.

Gene expression is a natural intermediate phenotype: it is heritable, tissue-specific, and directly accessible from RNA sequencing at population scale. Large-scale catalogs such as

GTEx [3] and GEUVADIS (the Genetic European Variation in Disease) [4] have established that most inter-individual expression variation maps to local cis-regulatory regions, motivating sequence-based prediction as a tractable target. Transcriptome-wide association studies exploit this structure to infer genotype-to-disease relationships through predicted expression [5, 6], but the accuracy and cross-individual generalizability of sequence-based expression models remain active research frontiers.

1.3 DNA shape and transcription factor binding interpretability

Transcription factors are sequence-specific DNA-binding proteins that regulate when and where genes are transcribed. They bind short, recurring motifs in regulatory DNA — promoters, enhancers, silencers, and insulators — and from those sites recruit or block the molecular machinery that transcribes RNA from DNA. Because different cell types express different combinations of transcription factors, the same genome can produce distinct binding landscapes and distinct programs of gene expression across tissues and developmental states. Understanding where transcription factors bind, and why they bind some sequences and not others that look similar, is therefore central to understanding gene regulation. Sequence motifs — the canonical nucleotide patterns associated with a given factor — account for a large share of observed binding, but not all of it: many bound sites lack a strong motif match, many strong motif matches go unbound, and binding specificity often depends on the surrounding genomic context. One major source of the binding specificity not captured by sequence motifs alone is the three-dimensional structure of DNA itself, which the work in Chapter 2 examines directly.

Transcription factors recognize regulatory elements through both direct sequence readout of nucleotide identity and indirect readout of DNA structural properties, minor groove width, propeller twist, helical twist, roll, and electrostatic potential collectively termed DNA shape. In direct sequence readout, amino acid side chains physically interact with the accessible edges of base pairs in DNA. In indirect or shape readout, proteins interface with the double-helical shape of DNA rather than only with nucleobases through base-specific contacts [7, 8]. These shape features facilitate the recognition of DNA by DNA-binding proteins beyond conventional sequence motifs, with many transcription factors having been found to have shape motifs in both *in vivo* and *in vitro* data [9, 10].

The role of shape readout has been inferred from crystal structures of protein–DNA complexes, from binding assays, and from genome-scale analyses of DNA structural features. For instance, proteins with positively charged amino acids may favor a narrower minor groove width, which is associated with enhanced negative electrostatic potential [8]. Hox proteins directly use shape readout to recognize specific minor groove topographies [11, 12]. Genomic regions flanking E-box binding sites influence binding specificity of basic helix–loop–helix (bHLH) transcription factors through DNA shape [13], and proteins in the myocyte enhancer

factor-2 family use shape readout in the center of a degenerate motif [14]. Together, these studies show that the interplay of base and shape readout is a core feature of transcription factor–DNA interactions, and that a combination of these modes is used to achieve binding specificity.

Computational methods have made it possible to study DNA shape at genome scale. DNAshape and DNAshapeR predict structural features from nucleotide sequence using a sliding-window approach based on structural information from all-atom simulations, providing single-nucleotide resolution estimates of local DNA conformation [15, 16]. While protein–DNA recognition is a dynamic process, these predicted structural outputs represent preferred conformations intrinsic to a given DNA sequence. Expanded sets of DNA shape features, including electrostatic potential, provide a more detailed representation of the physical and electrostatic properties of DNA that influence interactions with proteins [17].

Previous studies have shown that incorporating DNA shape can improve the description and prediction of transcription factor binding sites. Shape profiles can reveal differential binding specificities among closely related transcription factors that are not apparent from nucleotide sequence alone [18]. DNA shape features improve transcription factor binding site prediction *in vivo* [19], and sequence-plus-shape kernels improve binding affinity prediction across PBM (protein binding microarray) and SELEX (Systematic Evolution of Ligands by Exponential Enrichment) datasets [20]. More recent neural network models have also incorporated sequence and shape features for transcription factor binding prediction [21, 22, 23, 24]. These studies motivate the use of DNA shape as an additional layer of genomic information, offering insights beyond what is obtained from sequence data alone.

While deep convolutional neural networks achieve strong predictive performance from one-hot encoded sequence [25, 26], their learned representations do not explicitly encode the continuous biophysical properties of the double helix. Because DNA shape is determined by local sequence context, sequence-only models may implicitly learn correlates of shape. However, standard attribution methods applied to one-hot sequence identify influential bases rather than the structural properties those bases induce. As a consequence, interpreting what a trained model has learned about shape-driven transcription factor binding is difficult, and shape-mediated binding at non-canonical sites may be captured implicitly without being extractable by standard attribution methods.

Chapter 2 addresses this limitation by developing DeepShape, a convolutional neural network that augments one-hot DNA sequence with precomputed DNA shape features as explicit input channels. Trained to predict TF binding across 919 regulatory targets in 148 cell types, DeepShape demonstrates that structural features contribute independently of sequence motifs. DeepLIFT-based attribution applied to shape input channels recovers structural binding signatures not apparent from sequence attributions alone, providing a mechanistic window into shape-mediated binding specificity. The approach connects the interpretability imperative of genomic deep learning to the biophysical reality of protein–DNA recognition.

1.4 Sequence-to-expression models and the cross-individual evaluation gap

Gene expression is the process by which the information in DNA is transcribed into RNA and translated into the proteins that carry out cellular function. The level at which each gene is expressed varies across cell types, across individuals, and across conditions, and that variation is shaped both by the local DNA sequence around the gene and by the broader regulatory context in which the gene sits. A long-standing question is whether gene expression — typically measured as RNA abundance — can be predicted directly from sequence: given a stretch of DNA, what level of expression will the corresponding gene reach, and how will that level shift when the sequence is altered? A model that does this well would let researchers ask what a never-before-seen variant does to expression without measuring it experimentally, prioritize variants in genome-wide association studies, and reason about regulatory mechanism in cases where the experimental signal is too sparse to interpret directly. The methods that follow — both the earlier genotype-based approaches and the more recent sequence-based deep-learning models — are different strategies for approximating this prediction problem, and they differ in what they take as input, what they are trained on, and how their performance is measured.

The state of the art for predicting individual gene expression has historically been genotype-to-expression prediction, usually from relatively common SNPs. PrediXcan [5], for example, trains penalized linear models for each gene using SNP dosages within a cis window and paired genotype–expression data from a reference cohort. Related transcriptome-wide association methods, including Functional Summary-based Imputation (FUSION) [6] and PrediXcan-S [27], extend this framework to summary-statistic settings. These models are effective because they are trained directly on genotype–expression associations in the target tissue or a closely matched reference tissue, but they are limited by the variants observed often enough in the training cohort to estimate stable effects.

Sequence-to-expression models have evolved through several architectural generations. Early convolutional models such as DeepSEA [25] and Basset [26] predicted chromatin accessibility and transcription factor binding from short input sequences of roughly one kilobase. These were epigenomic-track models rather than expression models, but they established the sequence-to-function framework later used for expression prediction. Basenji [28] expanded receptive fields to ∼131 kb and introduced quantitative prediction of cap analysis of gene expression (CAGE). Xpresso [29] predicted population-averaged mRNA abundance from a promoter-centered sequence window, and ExPecto [30] combined sequence-predicted chromatin features with spatial transformation functions and tissue-specific linear models. Enformer [31] incorporated transformer attention layers with a receptive field of ∼200 kb. More recently, Borzoi [32] predicts RNA-seq coverage profiles at 32 base pair resolution across a ∼524 kb input window, and AlphaGenome [33] accepts inputs of ∼1 Mb and predicts across multiple regulatory modalities. Across these architectural generations, cross-locus prediction from the reference genome has steadily improved, reflecting changes in model architecture,

training data scale, output resolution, and other aspects of model design and training.

Other model families broaden this sequence-to-expression landscape. GraphReg [34] and EPInformer [35] incorporate enhancer–promoter contacts, epigenomic tracks, or chromatin interaction maps to improve population-level expression prediction. Models trained on nascent transcription assays, including ProCapNet [36] and Convolutionally Learned, Initiation-Predicting NETwork (CLIPNET) [37], predict promoter-proximal initiation from precision run-on and capped RNA sequencing (PRO-cap) or related data. These assays are closer to local transcription initiation than steady-state RNA abundance, making them a related but distinct prediction problem.

These models are typically trained on a single reference genome and evaluated on their ability to predict expression differences across genomic loci, often called the cross-gene or cross-locus setting. In this framework, the model learns features of gene structure and promoter architecture that distinguish highly expressed genes from lowly expressed ones, including CpG island content, core promoter elements, and the density of nearby regulatory elements [29, 38]. Cross-gene correlations are substantial. Enformer achieves Pearson r of $\sim$0.85 across genes for CAGE-based expression in matched cell types, and $\sim$0.58 for high-variance genes, reflecting the greater difficulty of predicting cell-type-specific expression [39, 40].

A different evaluation framework asks whether these models can predict expression differences across individuals for a given gene. This distinction is important because cross-gene and cross-individual variation emphasize different sequence features. Cross-gene variation is strongly associated with the fixed architecture of each gene's promoter and nearby regulatory context, which is represented once per locus in reference-genome training. Cross-individual variation instead reflects segregating genetic variants that modulate cis-regulatory activity, such as a single nucleotide variant that disrupts a transcription factor binding site in an enhancer or a variant that alters splicing efficiency. A model can learn promoter architecture well while remaining insensitive to variant-driven effects, since a reference-genome trained model has only seen one sequence per locus and receives little direct training signal about how alternative alleles change output.

In practice, personal genome prediction is performed by applying an individual's variants to the reference genome, or by constructing whole-genome-sequencing-derived haplotype windows, running the model, and comparing predicted expression across individuals to observed total RNA-seq expression pooled across both haplotypes. This setup requires extrapolation from sequences not encountered during training. For diploid genomes, predictions from the two haplotypes are typically averaged. This strategy has been applied to phased whole genome sequencing cohorts with matched RNA-seq data, including Geuvadis Consortium lymphoblastoid cell lines with approximately 421 individuals [4], ROSMAP (The Religious Order Study and Memory and Aging Project) dorsolateral prefrontal cortex samples with approximately 859 individuals [41], and GTEx [3].

While the application of genomic deep learning models to variant effect prediction has been reviewed in detail elsewhere [42], here we focus specifically on the use of such models for personal transcriptome prediction, which is distinct from their broader role in predicting

variant effects on molecular phenotypes. Variant effect prediction usually scores one variant at a time by comparing reference and alternate alleles in an otherwise fixed sequence. Personal transcriptome prediction instead requires integrating all variants carried by an individual in their haplotype contexts. A detailed review of progress and challenges in this specific application area is needed, since recent studies have revealed important limitations that are not apparent from standard variant effect benchmarks.

Cell-type and cell-state prediction are also related but distinct. Some supervised and self-supervised transcriptomic models attempt to improve prediction of cell-type differences or learn cell-state representations. TranscriptFormer [43], for example, is a cross-species single-cell expression foundation model rather than a DNA sequence model. Such approaches are relevant to cell-state modeling, but they do not directly test whether personal genome sequence predicts cross-individual expression.

1.4.1 Benchmarking cross-individual prediction from reference-trained models

Two independent benchmarking studies published in 2023 evaluated how well reference-trained models perform on this extrapolation task.

One benchmark assessed Enformer on 6,825 cortex-expressed genes in 839 individuals from ROSMAP with paired whole genome sequencing and RNA sequencing [39]. Per-gene Pearson correlations between predicted and observed expression across individuals ranged from -0.76 to 0.84, with a mean of 0.01. Among 598 genes reaching statistical significance (false discovery rate [FDR] <0.05), 195 genes, or 33 percent, were anti-correlated, meaning the predicted direction of variant effects was reversed. In contrast, PrediXcan [5], a linear elastic net model trained directly on genotype–expression data, predicted 921 of 1,570 testable genes with no sign-flipped predictions and a mean correlation of 0.26.

A second benchmark reached similar conclusions using a broader set of models including Enformer, Basenji2, ExPecto, and Xpresso [40]. They evaluated 3,259 genes with significant cis-expression quantitative trait loci (eQTLs) in 421 individuals from the GEUVADIS Consortium. Mean cross-individual Spearman correlations were near zero for all deep learning models, with Enformer at 0.045 ± 0.183, Basenji2 at 0.037, ExPecto at 0.034, and Xpresso at 0.007. PrediXcan achieved 0.249 ± 0.161. Cross-gene correlations were largely preserved when personal sequences were used (Enformer: 0.55 vs. 0.57 on reference), confirming that the models capture gene-level expression patterns but not inter-individual variant effects.

The direction-of-effect issue is especially informative. A model that localizes the relevant regulatory region but predicts the wrong sign has identified where regulation occurs, yet has not learned how sequence changes alter output. This is consistent with what we would expect from reference-genome training: the model can learn that a particular motif is associated with regulatory activity because it sees the motif in context across many loci, but it has limited basis for learning the quantitative relationship between motif perturbation and expression change because it has only observed each motif instance once, in its reference

allele form. Both benchmarking studies found that variant effects were concentrated near the transcription start site (TSS), consistent with earlier observations that these models under-rely on distal regulatory elements [38]. The first benchmark further showed that a shallow single-layer convolutional neural network (CNN) exhibited the same pattern, supporting the interpretation that the problem reflects reference-genome training generally, not any particular architecture.

A related uncertainty analysis examined this problem from the perspective of prediction uncertainty [44]. A natural question is whether the direction-of-effect errors are systematic, meaning the model has learned the wrong mapping, or stochastic, meaning the model has not learned a stable mapping at all. To distinguish these possibilities, the study trained a deep ensemble of five Basenji2 replicates with different random seeds, with identical architecture and data but differing in initialization and mini-batch order. If the model had learned a stable but incorrect mapping, all replicates would agree. Instead, 55% of fine-mapped eQTLs had inconsistent sign predictions across replicates (pooled across 13 GTEx tissues), and for cross-individual expression in Geuvadis, 67% of genes showed sign disagreement. For high-uncertainty genes, even the specific driver variants identified differed across replicates. These findings indicate that many sign errors are stochastic. The models have not converged on a stable variant-to-expression mapping, which supports the view that reference genome training provides insufficient signal to learn variant effects rather than providing an incorrect one.

1.4.2 Fine-tuning on personal genome and transcriptome data

The benchmarking results above suggest that the problem is not architectural alone but stems from the training data; models trained on a single genome have little opportunity to learn how sequence variation maps to expression variation. A straightforward fix is to provide this signal directly, by fine-tuning pre-trained models on datasets where personal genomes are paired with individual expression measurements. Four studies have pursued this approach at increasing sample sizes.

One study fine-tuned Enformer on GEUVADIS data (421 individuals, lymphoblastoid cell expression) [45]. Among several tested approaches, pairwise regression performed best. This objective predicts expression differences between pairs of individuals and directly optimizes inter-individual variation. On a random split of 200 genes, pairwise regression achieved a mean per-gene Pearson r of 0.284, compared to 0.055 for baseline Enformer and 0.283 for elastic net. By focusing the loss on between-individual differences rather than absolute levels, the model is encouraged to attend to variants that differ across genomes instead of shared promoter features.

Variformer was developed by replacing Enformer's output head with a single linear layer and fine-tuning on GTEx whole blood (n=670) with a combined mean squared error (MSE) and pairwise difference loss [46]. On 301 training genes evaluated across held-out individuals, Variformer matched elastic net performance (mean R^2 difference of 2.2%). Variformer also achieved the highest mean AUROC (0.737) for predicting CAVIAR (CAusal Variants Iden-

tification in Associated Regions) fine-mapped eQTLs, compared to ridge regression (0.707) and pre-trained Enformer (0.611). CAVIAR fine-mapping identifies candidate causal variants within linkage disequilibrium blocks, so this benchmark asks whether model scores prioritize variants statistically resolved as likely drivers of an eQTL signal. However, multi-gene performance failed to generalize to 100 held-out genes not seen during fine-tuning. Applying the same approach to Borzoi yielded comparable results.

SAGE-net (small and good enough CNN) took a different architectural approach, using a framework built around compact CNNs ($\sim$2.7M parameters) trained from scratch on $\sim$40 kb sequences [47]. Their personal-genome model, p-SAGE-net, uses a contrastive architecture that decomposes gene expression into two components: a population-mean prediction from the reference sequence, and an individual-deviation prediction from the difference between personal and reference representations. On 734 training genes evaluated across held-out ROSMAP and GTEx cortex individuals, p-SAGE-net achieved cross-individual correlations comparable to PrediXcan at roughly 70$\times$ faster inference.

These three studies converge on a consistent result that fine-tuned deep models match but do not surpass linear genotype–expression models on seen genes, and do not generalize to unseen genes. The Enformer fine-tuning study reports a mean Pearson r of $\sim$0.05 on 670 unseen genes, indistinguishable from baseline Enformer. The SAGE-net study observes per-gene correlations centered at zero for 116 test-chromosome genes, and finds that increasing training epochs actually degrades validation-gene performance. Variformer reports a higher but not significant correlation on held-out genes (Pearson r=0.13, p=0.18), despite testing longer input sequences and larger training gene sets.

A biobank-scale Borzoi study tested whether these conclusions hold at larger sample sizes [48]. They fine-tuned Borzoi on UK Biobank proteomic data (up to 54,219 individuals with paired whole-genome sequencing and plasma protein levels) for 150 genes, using a loss function combining MSE with a pairwise difference term similar to the approaches above. The results differ from the smaller-cohort studies in two respects. First, fine-tuned Borzoi outperformed elastic net on 130 of 150 genes (mean Pearson correlation coefficient [PCC] = 0.78 vs. 0.77), whereas at n$\approx$200, comparable to the earlier studies, elastic net was equal or better. This suggests that sequence-based models may continue to improve with more data in settings where genotype-based models have less room to improve. Second, in silico mutagenesis of the fine-tuned model revealed that 67% of its high-scoring variants were rare (minor allele frequency [MAF] <0.01), whereas 65% of high-scoring elastic net variants were common (MAF >0.05). Masking rare variants substantially reduced prediction accuracy, consistent with the advantage deriving from rare variant effects that linear models cannot capture with genotype-level features alone.

This rare variant effect highlights a setting where deep sequence models have a principled advantage. An elastic net can only assign nonzero weight to a variant it has observed in enough training individuals to estimate an effect. A sequence model can, in principle, predict the effect of a never-before-seen variant from the surrounding sequence context, for instance recognizing that a rare single nucleotide variant (SNV) disrupts a transcription factor binding motif. This advantage is not apparent at small sample sizes, where rare

variants are too infrequent to evaluate, and for common-variant-dominated genes, where linear models already capture most of the signal. It emerges only with large cohorts that include enough carriers of rare variants for the model to learn and be evaluated on their effects.

These biobank-scale results come with caveats. The phenotype is plasma protein abundance rather than RNA expression, and protein levels reflect post-transcriptional processes absent from RNA sequencing analyses. The covariates included in the model (age, sex, BMI, genotype PCs) may contribute substantially to prediction accuracy. On the generalization question, multi-gene fine-tuned models did outperform pre-trained Borzoi on unseen genes, a contrast with the earlier studies, but absolute accuracy remained low (Pearson $r<0.2$).

The limited generalization of fine-tuned models appears to reflect locus-specific learning. Each gene's expression variation across a cohort is typically associated with a small number of variants, often just one or two lead QTL SNPs with relatively high minor allele frequencies, even if rare variants further alter the expression levels of a subset of individuals. During fine-tuning, the model learns to weight these specific variants, effectively performing variable selection at each training locus, but the identity of the causal variants at one gene provides no information about which variants matter at another. This is the same reason linear eQTL models are trained per gene: the predictive features are locus-specific.

The Enformer fine-tuning study provided direct evidence for this interpretation. A linearized version of the fine-tuned Enformer, using *in silico* mutagenesis scores as additive variant weights, achieved nearly identical performance to the full nonlinear model ($r=0.274$ vs. 0.278). The deep model's capacity for learning complex sequence interactions is not being used; its predictions reduce to a weighted sum of individual variant effects. At typical human common variant densities, roughly one heterozygous SNP per kilobase, variants are sparse enough that interactions between them within a regulatory element may be rare, and the training data at current cohort sizes may be too small to detect such interactions even when they exist. Whether the larger cohort sizes available in biobanks change this, by providing enough examples of rare variant effects to learn general motif-disruption rules, is suggested by the biobank-scale multi-gene results but not yet resolved.

Despite these limitations, fine-tuned sequence models offer advantages beyond raw prediction accuracy. Both Variformer and the fine-tuned Enformer identify high-scoring variants enriched for active chromatin states, transcription factor motif disruptions, and fine-mapped eQTLs, at levels exceeding elastic net [45, 46]. The biobank-scale Borzoi study reports that fine-tuned Borzoi high-scoring variants overlap ENCODE regulatory elements at roughly twice the rate of elastic net variants and show stronger transcription factor motif disruption signals. The SAGE-net study shows that personal genome training shifts model attribution toward more distal sequence elements, and demonstrates in at least one case (GSTM3) that this recovers a regulatory motif missed by reference-trained models. The Variformer study reports related evidence, showing that its high-scoring variants include distal candidates at loci such as BTNL3 and are enriched for motif disruption overall, including rare and common ZFP57 variants predicted to create TF motifs. Linear model weights reflect statistical association mediated by linkage disequilibrium, whereas sequence model scores reflect learned

regulatory grammar. Even when prediction accuracy is similar, the latter provides a clearer path to mechanistic interpretation.

1.4.3 Allele-specific learning as an alternative training paradigm

The fine-tuning studies described above rely on expression measured across individuals, each carrying a distinct combination of cis and trans variants in varied cellular and environmental contexts. Individual-level expression reflects not only cis-regulatory effects but also trans-acting variation, environmental influences, and technical noise. As a result, the signal-to-noise ratio for learning cis effects is limited, particularly for genes with low cis-heritability.

An alternative strategy is to learn variant effects from allele-specific measurements within a single diploid organism. In a heterozygous individual, the two alleles of a gene share the same trans-regulatory environment and cellular context. Any consistent difference in their expression must therefore be cis-encoded. This design controls for many confounders that complicate cross-individual analyses and offers a cleaner signal for learning cis-regulatory grammar. The approach also has costs: it requires phased genomes, separates read counts across alleles, and loses reads that cannot be assigned unambiguously to one haplotype, reducing power for lowly expressed genes.

DeepAllele pursued this direction using F1 hybrid mice from crosses between inbred strains [49]. These crosses offer ideal properties for allele-specific analysis: high heterozygosity (∼20–40M variants per cross) and perfect haplotype phasing from the parental reference genomes. DeepAllele processes both haplotype sequences through shared convolutional layers and uses a dedicated nonlinear head to predict the log allelic ratio.

While a standard single-input CNN trained on one or both reference genomes performs comparably to DeepAllele on total count prediction, the paired-allele architecture substantially outperforms these baselines for allelic ratio prediction. The gain is larger for chromatin accessibility and gene expression than for TF binding, and single-input models reach near-zero correlation for gene-expression allelic ratio prediction. At well-predicted loci with substantial allelic imbalance ($|\log FC| \geq 1$), single "main" variants explain the majority of predicted allelic ratios ($R=0.80$ for ChIP-seq [chromatin immunoprecipitation sequencing], $R=0.86$ for ATAC-seq [assay for transposase-accessible chromatin using sequencing]); summing additive ISM effects across all variants accounts for most of the remaining variance (total $R=0.85$ and 0.96, for ChIP-seq and ATAC-seq, respectively). DeepAllele also identifies biologically coherent motifs at nearly twice the rate of single-input models.

DeepAllele has not been validated for human cross-individual prediction, and mouse F1 crosses differ from human personal genomes in ways that simplify the problem: much higher heterozygosity, perfect phasing, and controlled genetic backgrounds. Human genomes have far lower heterozygosity, imperfect phasing, and the connection between allele-specific and cross-individual expression prediction is indirect; the former measures within-individual cis effects, whereas the latter requires predicting absolute expression levels across people. Whether allele-specific training signals from human data, such as GTEx allele-specific expression or phased single-cell RNA-seq, can support similar approaches is an open question.

The conceptual value is in demonstrating that a training signal exists that is more directly informative about cis-regulatory variant effects than population-level expression.

1.4.4 A modality gap between chromatin and expression prediction

The most recent evaluation tested AlphaGenome and Borzoi on personal prediction of both gene expression and chromatin accessibility [50].

For chromatin accessibility (ATAC-seq in lymphoblastoid cell lines, n=100), AlphaGenome achieved a mean per-peak Pearson r of 0.627, approaching 83% of the heritability ceiling, defined here as the performance of a cis-genotype elastic net baseline trained and evaluated on the same phenotype (elastic net baseline: 0.758). Borzoi reached 0.448. For gene expression (RNA-seq in ROSMAP dorsolateral prefrontal cortex, n=859, 500 heritable genes), AlphaGenome achieved 0.151 and Borzoi 0.058, far below the ceiling of 0.670. The same models that approach the heritability ceiling for one regulatory readout capture only ∼23% of heritable variation for the other.

This gap is informative because it localizes the bottleneck. Chromatin accessibility is a comparatively local molecular phenotype: whether a genomic region is open is associated with nucleosome remodeling proteins and transcription factors, including pioneer factors, acting near that region. Expression is an integrated phenotype: a gene's transcription rate depends on the combined activity of its promoter, proximal regulatory elements, and potentially dozens of distal enhancers connected through three-dimensional chromatin architecture. The cis-regulatory variants that drive inter-individual expression variation reflect this complexity. Chromatin accessibility QTLs act locally (median lead variant distance of 2.3 kb from peak center), while expression QTLs act more distally (median d90 of 34 kb from TSS). Accordingly, AlphaGenome's expression predictions are worse for distally regulated genes (P=0.004), while accessibility shows no such distance dependence. Context truncation experiments confirm: accessibility saturates at ∼4 kb windows, while expression improves with longer context but remains far below the ceiling at all sizes up to 1 Mb.

The implication is that current models can learn local sequence grammar well enough to predict personal chromatin accessibility, but cannot effectively integrate these local signals across enhancer–promoter distances to predict expression. The models have large receptive fields (AlphaGenome accepts 1 Mb), but receptive field is necessary, not sufficient. The model must also learn which distal elements regulate which genes, and how their effects combine. This enhancer–promoter wiring problem is distinct from learning local regulatory grammar, and current training on a single reference genome may not provide sufficient signal to solve it.

The same evaluation provided additional evidence for this interpretation through a supervised readout experiment. The authors trained locus-specific readout heads on model-derived chromatin features to predict expression. This approach recovered substantial predictability, and even features from an untrained, random-weight model supported non-trivial prediction.

Genotype-dependent information relevant to expression is present in the intermediate representations, but the pre-trained RNA output heads do not decode it effectively. The models appear to encode information about how personal variants affect local chromatin at regulatory elements; they do not know how to combine this information across elements to predict a gene's expression.

1.4.5 Evaluation frameworks and their limitations

The studies above draw attention to limitations in the frameworks currently used to evaluate personal transcriptome prediction.

Cross-gene versus cross-individual correlation. A model can achieve high cross-gene correlation by learning gene-level features (promoter strength, CpG content, gene length) that are shared across all individuals and thus irrelevant to inter-individual variation. Cross-individual evaluation, computing per-gene correlations across people, is the appropriate metric for the personal prediction task. Yet most model development papers still report only cross-gene performance, which is not informative about whether a model can predict the effects of personal variants [43].

Direction-of-effect accuracy. Models may identify regulatory loci but predict the wrong sign of variant effects. This is invisible to standard cross-gene benchmarks and is only partially captured by pooled variant-effect metrics, such as overall Pearson correlation between predicted and observed effect sizes across variants. As noted previously [50], a model can achieve a positive pooled correlation while predicting the wrong direction for many individual genes, because genes with large correct effects dominate the pooled statistic. Per-gene direction-of-effect accuracy is a more informative metric.

Heritability ceiling estimation. Not all genes have substantial cis-heritable expression variation. Evaluating models without reference to the heritability ceiling conflates model limitations with biological limits on predictability. Recent evaluations conditioned analyses on genes with significant cis-heritability [47, 50]. This is useful but introduces selection bias: genes with high linear heritability are those where linear models perform best, potentially understating deep learning models' relative contribution for genes with more complex regulation.

Seen versus unseen genes. Fine-tuned models generally do not generalize to unseen genes at cohort sizes of hundreds of individuals, though biobank-scale fine-tuning reports modest generalization on unseen genes [48]. This distinction is critical. Results on seen genes show that a model can learn locus-specific genotype–expression mappings. Results on unseen genes test whether the model has learned transferable grammar. Only the latter represents progress toward a general personal transcriptome predictor. Evaluations that report only seen-gene performance, analogous to reporting training accuracy, are not informative about this question.

Comparison to linear baselines. PrediXcan-style elastic net models serve as the standard comparison, but are typically trained with cross-validation on the evaluation population. They benefit from seeing the target population's expression data and linkage dis-

equilibrium (LD) structure. This makes them useful as a performance ceiling for what locus-specific models can achieve, but they are not a fair head-to-head comparison with models that have not seen the evaluation cohort's expression data.

Missing evaluation dimensions. Several informative evaluations have not been performed. Allele-specific expression provides a within-individual, haplotype-resolved test of cis-regulatory variant effects that avoids trans and environmental confounders, but has not been used as a benchmark in any of the reviewed studies. Rare variant effects are beginning to be tested; a biobank-scale study showed that rare variants drive the deep model advantage at biobank scale [48], but systematic evaluation of rare variant prediction accuracy remains limited. Most reviewed studies also focus exclusively on single nucleotide variants. Indels and structural variants can have large effects on gene expression but have not been evaluated in the personal prediction setting. Even scoring individual indels is non-trivial: a Borzoi indel analysis showed that pooling-layer boundary artifacts in Borzoi produce spurious predictions comparable in magnitude to real eQTL effects [51]. In personal genome prediction, where many variants including indels must be inserted into the same sequence simultaneously, interactions between such artifacts and nearby variant effects have not been characterized. Context-dependent expression, variant effects that differ across cell states, developmental stages, or environmental conditions, is absent from the current literature. More broadly, standardized benchmark datasets and evaluation protocols for personal transcriptome prediction, analogous to CAGI for variant interpretation [52], would help address the inconsistencies in cohorts, gene sets, and metrics across the studies reviewed here.

Chapter 3 addresses these limitations by developing a haplotype-resolved framework for evaluating personal transcriptome prediction using allele-specific expression. By comparing predicted expression between the two phased haplotypes of the same individual, this framework tests whether sequence-based models capture cis-regulatory differences between alleles. Applied to fine-tuned sequence models, the approach evaluates whether models recover the magnitude and direction of allelic imbalance across individuals and genes, rather than only cross-gene expression differences. The chapter establishes allele-specific expression as a complementary benchmark for personal regulatory prediction and characterizes the extent to which current fine-tuning approaches generalize haplotype-resolved effects beyond the genes used during training.

1.5 Prior published work

Chapter 2 is based on the following manuscript: R.L. Keivanfar, F. Yang, K.S. Pollard, and N.M. Ioannidis. "Improving interpretability of transcription factor binding models with DNA shape features." *bioRxiv*, 2025. `https://doi.org/10.1101/2025.04.01.646034`.

Chapter 3 presents original research conducted during the doctoral program and is not derived from a prior publication.

Chapter 2

Improving interpretability of transcription factor binding models with DNA shape features

This chapter is co-authored by Forrest Yang, Katherine S. Pollard, and Nilah M. Ioannidis.

2.1 Overview

Genomic deep learning models predict molecular phenotypes from DNA sequence by learning motifs and regulatory grammars, yet the learned representations remain difficult to interpret in mechanistic terms. We developed DeepShape, a convolutional neural network that augments one-hot encoded DNA sequence with five structural attributes — minor groove width, propeller twist, helical twist, roll, and electrostatic potential — to predict transcription factor (TF) binding across 919 regulatory targets in 148 cell types. Shape features contribute independently of sequence motifs, and DeepLIFT-based attribution applied to shape inputs recovers structural motifs not apparent from sequence alone; the propeller twist flanking pattern for MafK provides a concrete example of shape-driven binding specificity at non-canonical sites. Within the dissertation, this chapter contributes one example of a broader representation question: how the choice of input features changes what genomic deep learning models can learn and interpret.

2.2 Abstract

Deep learning models in genomics that predict molecular phenotypes from DNA sequence traditionally focus on one-hot encoded nucleotide representations. Here, we develop a novel model that extends this approach by incorporating DNA structural attributes indicative of local DNA shape alongside canonical sequence inputs. This augmentation provides an additional axis for model interpretability and aids in identifying regulatory patterns not apparent

from sequence alone. Applying this approach to prediction of transcription factor binding (ChIP-seq) demonstrates that combining sequence and structural DNA information can improve the identification of regulatory elements to provide a more nuanced understanding of genomic function and regulation.

2.3 Introduction

Genomic deep learning models that learn complex motifs and regulatory grammars from DNA sequence can predict molecular phenotypes such as transcription factor (TF) binding, histone modifications, chromatin accessibility, and gene expression directly from sequence [25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 53]. These models have the potential to be powerful tools for analyzing personal genomes by predicting the molecular effects of individual genetic variants and haplotypes. In addition, interpreting the predictive sequence features learned by such models can provide biological insights into genome function and regulation. Approaches based on forward or backward propagation of influence can be used to identify DNA sequence motifs that drive model predictions [54]. However, because it can be difficult to fully interpret the learned sequence features and uncover mechanistic insights from "black box" neural networks, recent work has also aimed to develop explainable models that are more biologically interpretable [54, 55, 56]. For example, Puffin uses convolutional kernels of specific biologically-interpretable lengths to separate the contributions of simple sequence features from more complex binding motifs [55]. Here we develop a biologically-motivated approach to explicitly model the contributions of both DNA sequence and DNA biophysical or structural properties (DNA shape) to TF binding and other molecular phenotypes.

Shape readout, or the binding of a TF to a region of DNA based on its physical structure, is now considered a major mode of TF binding. Recent work increasingly examines local DNA shape features to explain TF binding outside of canonical sequence motifs [8, 9, 57]. An algorithm for finding shape motifs, or shape feature patterns that are enriched at TF binding sites, has also been proposed [58]. However, interpreting genomic deep learning models to extract their knowledge of how DNA shape affects TF binding remains underexplored. Deep neural networks capture extensive information about the mapping from genomic input to genomic activity, beyond what is captured by enrichment tests and other existing computational methods. Thus, understanding what is learned by a shape-based network can potentially shed insight on shape readout and reveal examples of shape-driven TF binding.

In this work, we present DeepShape, a deep convolutional neural network that incorporates DNA shape features alongside one-hot encoded DNA sequence as input, and apply this shape-aware modeling strategy to ChIP-seq peak prediction across 148 cell types and 170 TFs. We adopt two approaches to interpreting the trained DeepShape model, showing that feature attribution analyses based on the shape inputs enhance the interpretability of the model and provide additional insights into regulatory patterns not captured by sequence alone. Our results demonstrate that while sequence motifs are the primary drivers of TF binding, incorporating DNA shape features is useful for exploring non-canonical and

co-bound sites.

Previous efforts have combined sequence and DNA shape features to predict transcription factor binding. Early sequence+shape kernels extended k-spectrum and di-mismatch approaches by adding four shape features, showing improved affinity and specificity on protein-binding microarray (PBM) and SELEX datasets [20]. Gradient-boosting models for *in vivo* ChIP-seq augmented sequence encodings such as position-specific scoring matrices (PSSMs), transcription factor flexible models (TFFMs; HMM-based), or one-hot vectors with first- and second-order shape features, again improving over sequence-only baselines [19]. Convolutional-neural-network (CNN) studies on universal PBM (uPBM) datasets used dual sequence+shape inputs and reported modest, consistent gains, often with Roll most informative [21, 22]. Hybrid neural architectures have also been explored: DeepCTF combines CNNs, recurrent neural networks (RNNs), and attention on uPBMs [23], whereas DeepSTF integrates convolutional, recurrent, and transformer layers across five shape channels on ENCODE ChIP-seq and offers attention maps as an interpretability readout [24].

Our approach differs from prior sequence+shape studies in several respects. We evaluate on the full 919-task DeepSEA/DeeperDeepSEA benchmark, providing a broad, standardized basis for comparison with canonical sequence-only models. Our framework also emphasizes interpretability specific to DNA shape: we compute channel-resolved attributions, apply TF-MoDISco directly to quantile-encoded shape attributions to recover shape motifs, and assess co-occurrence between sequence and shape motifs. We also train a residual probe in which a shape-only model predicts errors made by a sequence-only model, providing another lens on how shape information contributes.

2.4 Materials and methods

2.4.1 Model design

The deep convolutional neural network, DeepShape, was modeled after DeeperDeepSEA as implemented in the PyTorch based deep learning library Selene, which takes one-hot encoded DNA sequence as input [59]. Modifications were made to accommodate the addition of shape features. DeepShape processes 1000-bp sequence and shape inputs through two convolutional layers each, subsequently combining them to enable the model to capture unique information by learning separate representations for each input type. The combined representation passes through four additional convolutional layers, along with two max-pooling layers, three batch normalization layers, two dropout layers, and ReLU activations, before the output is flattened and passed through two fully connected layers, with the final output layer using a sigmoid activation function to predict the binary binding/not of each genomic target to the input sequence while considering the sequence's DNA shape features **(Fig. 2.1A)**. DeepShape uses a fixed 450 kernels per layer and filter length of 4, determined through a comprehensive hyperparameter search designed to adapt the model to the combined analysis of sequence and structural features **(Supplementary Fig. A.1)**.

To demonstrate this general approach, we implemented DeepShape to simultaneously predict the same 919 genomic tracks used to develop DeeperDeepSEA, each as a different output ("targets"). These include 690 TF targets (ChIP-seq), 125 open chromatin targets (DNase), and 104 histone mark targets (ChIP-seq) across 148 cell types and 34 treatment conditions. To create binary labels for these tracks, the genome was split into 200-bp bins and assigned a positive label if more than half the bin overlapped a peak region and was otherwise labeled negative, following the labeling strategy used in DeeperDeepSEA.

2.4.2 Data and training

Training labels were computed from uniformly processed ENCODE and Roadmap Epigenomics data releases. The 11 TF targets used for hyperparameter tuning were selected based on their strong shape or sequence preferences, respectively, as described previously [58]. The 919 chromatin features used to benchmark performance against DeeperDeepSEA were those used by the original DeepSEA and DeeperDeepSEA model evaluations, comprising TF binding (ChIP-seq), histone binding (ChIP-seq), and chromatin accessibility (DNase-seq). We set the training parameters (number of steps, batch size, and related configurations) to match how both DeepSEA and DeeperDeepSEA were originally trained [25, 59]. The benchmark was performed using the following reciprocal validation/test splits (six total): chromosomes 6 and 7, 8 and 9; 9, and 10, 11 and 12; 14 and 17, 15 and 16. See Code and data availability for configuration files detailing the training set up.

2.4.3 Shape inputs

Whole genome shape features were generated for five structural attributes of DNA — minor groove width, helical twist, propeller twist, and electrostatic potential — that have been found to be among the most important for evaluating protein-DNA readout modes [17]. The set of five shape features was generated by passing the hg19 reference FASTA to DNAShapeR, a software package that uses a sliding pentamer window to generate shape features from all-atom Monte Carlo simulations [16]. The output of this provides a single-nucleotide resolution representation of DNA conformation within local contexts, with the predicted structural outputs representing preferred conformations that are intrinsic to a given DNA sequence [15]. Gaps in the resulting shape assemblies were filled with the mean shape feature value for each of the five shape features.

2.4.4 DeepLIFT

To quantify the contribution of each input feature to the prediction of TF binding, we employed DeepLIFT (Deep Learning Important FeaTures) [60]. For a given TF track, DeepLIFT assigns an attribution score to each feature, indicating its influence on the model's prediction. These attributions are computed such that they satisfy the summation-to-delta property:

$$\Delta y = y_{\text{input}} - y_{\text{ref}} = \sum_f \text{Att}_f, \tag{2.1}$$

where y_{input} represents the model's prediction for the given input sequence, y_{ref} is the prediction for a reference sequence, and Att_f denotes the attribution for feature f. The sum of all feature attributions equals the difference between the prediction for the input sequence and that for the reference.

Since the model predicts a binding probability in the range $[0, 1]$ for each TF track, and the binding probability for a random sequence typically approaches zero, Δy approximates the predicted probability y_{input}.

2.5 Results

2.5.1 DeepShape model overview and performance

DeepShape extends the DeeperDeepSEA architecture by incorporating DNA shape features alongside sequence inputs (Methods). After training on a full set of 919 TF, histone, and open chromatin (DNase) targets across six distinct chromosome splits, we benchmarked the performance of DeepShape (sequence + shape) against a DeeperDeepSEA model that we trained (sequence-only). Our focus was on TF binding, but we included open chromatin and histone targets because these genomic elements can be bound by TFs and hence predicting them alongside the TF targets may improve performance. Using all 919 tracks also enabled us to directly compare results with the published DeeperDeepSEA model.

We find that DeepShape has similar or slightly improved performance relative to DeeperDeepSEA on the three types of prediction tasks **(Fig. 2.1B)**, with an average AUROC of 0.929 and average AUPRC of 0.379 compared to DeeperDeepSEA's average AUROC of 0.925 and average AUPRC of 0.367. The performance of the DeeperDeepSEA model we trained was similar to that of the published model. We observed that modifying DeeperDeepSEA's hyperparameters to match those of DeepShape also improved its performance relative to the baseline DeeperDeepSEA model **(Fig. 2.1B)**, suggesting that the hyperparameter tuning improved model representations of sequence inputs effectively as well.

To evaluate the degree to which DeepShape leverages shape information for predicting TF binding, we permuted shape features across each batch by setting the shape features of the ith example to the shape features of the $\sigma(i)$ example, where σ is a random permutation **(Fig. 2.1C)**. We separately permuted sequence features for comparison. This permutation analysis resulted in an average AUROC contribution ratio of 0.228 and average AUPRC contribution ratio of 0.348 for shape versus sequence contributions in DeepShape, indicating that sequence information is most important for predicting TF binding, as expected, but that shape also contributes notably to model predictions. A similar analysis of a sequence-and-shape model with the same original hyperparameters (filter size, number of kernels per layer) as DeeperDeepSEA shows lower shape contributions (average AUROC contribution

ratio = 0.101, average AUPRC contribution ratio = 0.269; **Supplementary Fig. A.2A**). This suggests that the optimized hyperparameters enable better use of shape inputs.

To further examine whether DeepShape's usage of shape features reflect learned shape readout patterns rather than redundancy with sequence, we repeated the permutation analysis after replacing the five shape channels with an auxiliary sequence channel corrupted with Gaussian noise **(Supplementary Fig. A.2B)**. We trained (i) *Seq + Noisy Seq* (sequence plus noised sequence) and (ii) *Noisy Seq* (noised sequence only). *Noisy Seq* performed similarly to *Seq + Noisy Seq*, indicating that the noised sequence retains much of the predictive signal. In *Seq + Noisy Seq*, permuting the sequence channel led to strong reductions in performance, whereas permuting the noised-sequence channel had little effect. These results suggest that when presented with redundant inputs, the network prioritizes the channel that provides the simplest route to predictive signal. In the noisy-sequence controls, this was the sequence channel, while the noised-sequence channel was effectively ignored. Taken together with the non-trivial shape contribution observed in DeepShape, this pattern supports the view that the model is directly leveraging shape inputs rather than reconstructing them from sequence.

2.5.2 Model interpretability

Although the joint sequence-and-shape model performs similarly to the sequence-only model with the modified hyperparameters, we hypothesized that DeepShape's explicit modeling of shape inputs would provide unique opportunities for interpretability. To better understand the contributions of sequence and DNA shape features to TF binding predictions, we performed a series of interpretability analyses. These included activation-based analyses to examine how specific features influence model predictions and attribution-based approaches (DeepLIFT) to identify high-contribution positions across input features (Methods) [60]. Using these attributions, we investigated patterns in sequence and shape features that are enriched at TF binding sites.

Compared to background sequence distributions, shape values in TF-bound regions with high attribution scores revealed distinct shifts, suggesting that DNA shape features play a key role at regulatory sites **(Fig. 2.2A)**. To examine variation across individual TF targets, we computed Kolmogorov–Smirnov (KS) statistics for each shape feature and find variation in the degree of motif-background separation across both features and targets **(Fig. 2.2B)**, suggesting that individual targets may exhibit preferences for specific shape features, consistent with observations from previous studies [58]. For example, the distribution of MGW values in GM12878 | MEF2A shows a strong shift toward narrower minor groove widths in motif regions **(Supplementary Fig. A.3A)**, consistent with previous structural studies showing that MEF2A recognizes its binding site by selecting a narrow minor groove width and engaging two base pairs per half-site [61]. In contrast, K562 | CEBPB displays a shift in Roll values, with motif regions enriched for higher Roll compared to background **(Supplementary Fig. A.3B)**. This differs from previous shape motif analyses, which identified

HelT as the preferred shape feature for CEBPB [58], and suggests that Roll may also play a role in CEBPB binding, depending on cellular context or motif instance.

To gain additional insight into feature importance, we compared the average number of high-attribution positions per motif type **(Supplementary Table A.1)**. Sequence motifs were associated with a higher number of high-attribution positions compared to shape motifs, reflecting their role in capturing canonical motifs directly involved in TF binding, while shape motifs highlighted shorter segments, providing complementary structural information. These analyses demonstrate that sequence and shape features provide distinct yet interrelated insights into regulatory predictions.

These findings raised the question of how sequence and shape features contribute in specific binding contexts. To explore this, we applied DeepShape to analyze binding of the Myc-Max TF heterodimer, examining how sequence and DNA shape features contribute to binding specificity. *In vivo* studies show that Max binds E-box motifs in approximately five times as many locations as Myc. Prior work identified distinct shape motifs in Myc-Max ChIP-seq peaks compared to Myc-unbound-Max regions, suggesting that DNA shape may play a role in their differential binding patterns [58]. We trained DeepShape on a single Myc-Max track based on the intersection of Myc and Max ChIP-seq peaks, 3,156 peaks in total, in GM12878 cells. Among these, 902 peaks contained exactly one CACGTG motif (E-box), while 1,071 peaks contained one or more CACGTG motifs. The sequence and shape branches of the model were evaluated separately by applying their filters and obtaining activations for Myc-Max peak sequences containing a single CACGTG motif. High filter activations from the model's sequence branch tend to occur at the location of the E-box motif, with a sharp peak at the center of the distribution of positions relative to the motif **(Fig. 2.3)**. In contrast, high filter activations from the shape branch are more diffuse, without sharp peaks and with elevated values flanking rather than overlapping the E-box **(Fig. 2.3)**.

This activation pattern is consistent with prior studies showing that while the E-box motif plays a central role in Myc-Max binding, it is not sufficient to fully explain specificity. Prior work has shown that Myc binding frequently depends on chromatin context and transcriptional machinery rather than sequence motifs alone, that Myc-Max binds both canonical and non-canonical motifs with flanking sequences and lower-affinity binding sites modulating interactions, and that DNA shape features can refine binding predictions, particularly in regions adjacent to sequence motifs [58, 62, 63]. Our results are consistent with these findings and support the idea that DNA shape features in motif-flanking regions, as captured by the model's shape branch, may provide supplementary information for fine-tuning binding specificity.

To assess the relative contributions of sequence and shape features to binding predictions across different TF targets, we compared their attribution values **(Fig. 2.3A)**. Across most TFs, we observed a consistent relationship between sequence and shape attributions: shape attribution is generally 0.269 times the sequence attribution, as indicated by the line of best fit, though TFs show variability around this trend. This suggests that while sequence features contribute more substantially to binding predictions overall, shape features provide a supplementary contribution across diverse TFs with some TFs relying on shape features

more than others.

Among the TFs analyzed, MafK stood out as one of the TFs having relatively higher shape attributions compared to sequence attributions **(Fig. 2.3A)**, indicating that DNA shape features might play a unique role in modulating MafK binding specificity. To further investigate this, we used DeepLIFT to score each input feature based on its contribution to the prediction of the network, and then ran TF-MoDISco to uncover sequence and shape motifs learned by the model (Methods). TF-MoDISco takes a set of genomic sequences and their attributions as input, and it clusters high-attribution subsequences ("seqlets") to generate motifs [64]. We adapted TF-MoDISco to run on shape feature attributions by transforming each quantitative shape feature into its quartile, producing a 4-dimensional one-hot encoding of each shape feature. As a result, for any type of shape feature, we can obtain "shape motifs" that explain model predictions of the binding of a particular TF target. For the HepG2 | MafK target, TF-MoDISco identified a sequence motif matching the canonical TGCTGA(C/G)TCAGCA MafK motif, along with a propeller twist (ProT) shape motif **(Fig. 2.3B)**. The high-attribution region of the ProT motif is characterized by a peak followed by a pronounced dip in the high-attribution region, which is not in the canonical motif's profile.

We examined the co-occurrence of these motifs, defining an occurrence as a seqlet belonging to the corresponding TF-MoDISco motif cluster. The ProT motif frequently co-occurs with the sequence motif, typically at the edge of the sequence motif and in the nearby flanking region **(Fig. 2.3C-D)**. To assess the potential modulatory effect of the ProT motif on binding affinity, we compared ChIP-seq signals for strong sequence motif occurrences (>90th percentile Pearson similarity of attributions) that co-occur with the ProT motif versus those that do not. Strong sequence motif occurrences cooccurring with the ProT motif showed significantly higher binding **(Fig. 2.3E, Supplementary Table A.2)**.

Examination of individual occurrences revealed that the downward ProT dip often corresponds to A-tracts or T-tracts flanking the sequence motif **(Fig. 2.3D, F)**. This finding aligns with previous observations that A-tracts and T-tracts significantly affect DNA shape [65]. These results suggest that DNA shape features in flanking regions can modulate binding to canonical sequence motifs, potentially explaining some of the binding specificity not captured by sequence alone.

We next identified sequence and shape motifs for each of the TF targets using TF-MoDISco, and we evaluated their relative contributions to TF binding predictions **(Supplementary Fig. A.4)**. To do this, we used the motifs to featurize our dataset and trained logistic regression and gradient boosting models to predict TF binding. Each sequence was featurized by computing the maximum motif match score across the sequence for each motif, resulting in an n-dimensional representation where n is the number of motifs, performed separately for each target. For model training, five sequence motifs were selected using forward step-wise selection with logistic regression and cross-validation, and five shape motifs were selected in the same way. These 10 motifs for each target were then used to train both logistic regression and gradient boosted decision tree classifiers. Across targets, the mean $\pm$ standard deviation number of motifs from which these 10 were selected was 21 $\pm$ 9 for

sequence motifs and 114 ± 26 for shape motifs.

When examining the most important motif for each target, as determined by logistic regression coefficient significance and gradient boosting feature importance, we found that a sequence motif was the most important feature in 95.3% and 82.9% of cases, respectively **(Fig. 2.4A)**. This indicates that sequence motifs are the dominant factors in predicting TF binding, although shape motifs are most important in a small number of cases.

To further quantify the relative contributions of sequence and shape motifs, we performed ablations by removing individual motifs and retraining our gradient boosting model **(Fig. 2.4B)**. The performance drops resulting from feature removal confirm that for most TFs there is a single sequence motif that has a much larger impact on model performance than any other sequence or shape motif. Nonetheless, the model remained fairly performant after dropping the top motif, indicating that DeepShape has learned more than simply the canonical sequence motif for each TF. For certain targets, shape motifs were among the top contributors, with their removal leading to measurable performance drops **(Fig. 2.4C)**, supporting their complementary role in binding prediction. These findings confirm the primacy of sequence information in determining TF binding specificity, while underscoring that shape motifs and non-canonical sequence motifs also contribute to predictions across TFs.

As a complementary analysis, a shape-only model was trained to classify the prediction outcomes of a sequence-only model. The sequence-only model, matched to the DeeperDeepSEA architecture, was first trained on the full benchmark. For each sequence–target pair, we compared the sequence-only model's binarized prediction (threshold 0.5) to the ground-truth label and assigned an error class — false positive, correct, or false negative — which served as supervision for the shape-only classifier. A shape-only model with the same architecture as the DeepShape shape branch was trained on this three-class classification task. By training a shape-only model on the outputs of a sequence-only baseline, the analysis isolates structural features from sequence, since the two are not presented to the same model simultaneously. Whereas DeepShape identifies cases where shape features contribute within a joint model, this approach asks when shape alone can account for binding patterns missed by sequence.

This error-classification model did not improve overall predictive performance relative to the original architecture, but it provides a complementary perspective on modeling errors and interpretability. First, we assessed whether structural features highlight patterns associated with specific types of errors. The classifier achieved high recall (99.9%) and precision (97.7%) for the majority class of correct predictions. In contrast, performance on the minority error classes was much lower: false positives (6.8% recall, 24.9% precision) and false negatives (0.7% recall, 16.6% precision) **(Supplementary Fig. A.5)**. This indicates that while correct predictions were well distinguished, errors were identified with limited accuracy. Second, we performed interpretability analyses to determine whether DNA shape features emphasize specific genomic contexts. In a representative example involving c-Fos binding, the shape-only model highlighted information about binding activity not captured by sequence alone. Attribution maps on false negatives from the sequence-only model showed that the sequence-only model attends primarily to the center of ChIP-seq peaks, whereas the distri-

bution of c-Fos' AP-1 motifs is distributed more broadly **(Supplementary Fig. A.6A)**. In contrast, the shape-only model attended most strongly to the flanking regions with high motif density, suggesting that the binding in these false negative peaks was shape-mediated. Furthermore, TF-MoDISco analysis of the shape-only model revealed shape motifs that accounted for the high attribution scores observed in flanking regions, clarifying signals not captured by sequence features alone **(Supplementary Fig. A.6B)**. These results indicate that shape features can provide interpretable motifs that offer complementary explanations for binding events beyond those captured by sequence-only models.

2.6 Discussion

DeepShape integrates both sequence and DNA shape information to provide a more comprehensive understanding of DNA-protein interactions and separate the contribution of DNA shape from other sequence features. Our findings demonstrate that while sequence motifs are the primary drivers of TF binding, DNA shape features play a role in modulating this binding.

The uniformity of the shape to sequence attribution ratio across TF targets is noteworthy, and may reflect the model's consistent utilization of shape information rather than variations in biophysical shape readout mechanisms among TFs. Targets deviating from this trend, such as HepG2 | MafK, could potentially indicate TFs with more pronounced shape-dependent binding mechanisms.

The example of HepG2 | MafK binding illustrates how shape motifs, particularly in combination with sequence motifs, can modulate binding affinity. The propeller twist shape motif identified in this case suggests that local DNA structural properties in flanking regions can influence binding to canonical sequence motifs. This observation aligns with the emerging view of shape readouts as an important mechanism in TF-DNA recognition.

Our approach has notable limitations. The DNA shape features used here represent a static, localized view of DNA structure, neglecting its dynamic nature, larger-scale structural features such as topologically associated domains, and deviations in shape due to modifications such as DNA methylation. Additionally the use of continuous values for shape inputs may contribute to the observed diffuse attribution patterns, which could be tested by binarizing them into extreme versus non-extreme quantiles. Further, future research could explore incorporating shape features as an output track, aiding in interpreting what the model has learned about structure directly from DNA sequence. Finally, while sequence information can capture much of the contributions of shape when shape features are not explicitly included in the model, their inclusion in DeepShape enables attributions to be directly computed for shape features using standard interpretability methods.

Despite these limitations, our study demonstrates the value of integrating DNA shape information into genomic deep learning models. DeepShape provides an additional axis for model interpretability, revealing regulatory patterns not evident from sequence alone. This

approach offers insights into cases where sequence motifs incompletely explain TF binding patterns.

2.7 Code and data availability

Code to train and evaluate DeepShape models and run TF-MoDISco interpretability pipelines are available at `https://github.com/ni-lab/DeepShape`. DNA shape inputs are available on Zenodo (`https://doi.org/10.5281/zenodo.13334119`). Model weights are available on Zenodo and `https://huggingface.co/ryankeivanfar/DeepShape`.

2.8 Acknowledgements

We thank Shiron Drusinsky, Sean Whalen, Jodi Lee, and other members of the Pollard and Ioannidis labs for helpful discussions. This work was supported by the UC Noyce Initiative for Computational Transformation, Additional Ventures, Gladstone Institutes, and the W. M. Keck Foundation and the Biswas Family Foundation. We used the Savio computational cluster resource provided by the Berkeley Research computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). N. M. I. and K. S. P. are Chan Zuckerberg Biohub San Francisco Investigators.

2.9 Figures

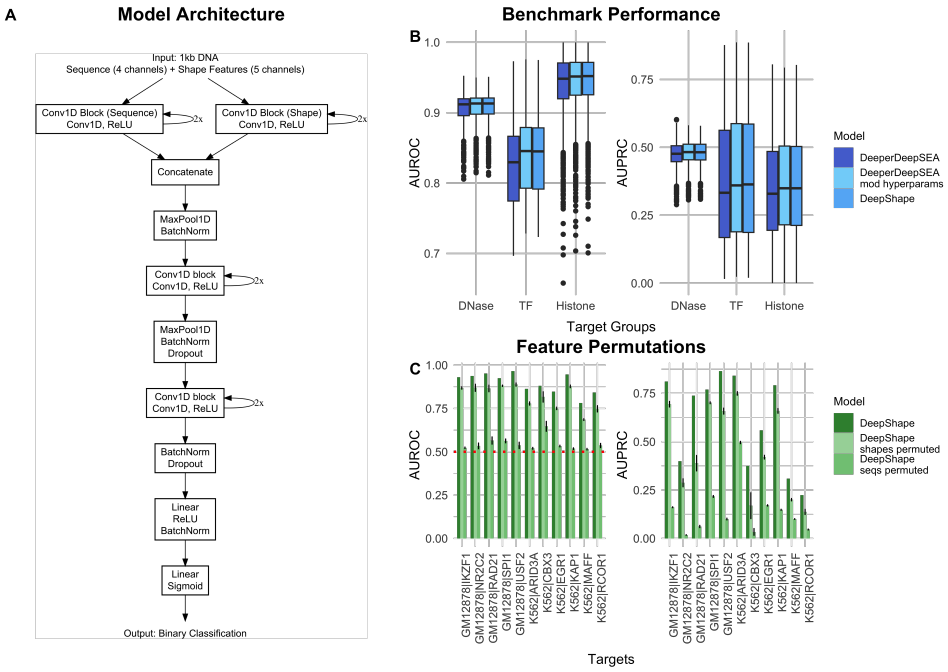

Figure 2.1: **Overview of the DeepShape model and its performance.** A) Overview of the DeepShape model architecture. B) Performance evaluation of DeepShape across held-out DNase, TF, and histone target groups. Modified hyperparameters improve both the sequence+shape and sequence-only models. C) Comparison of AUROC (left) and AUPRC (right) for 11 representative TFs (see Materials and methods) based on trained DeepShape models from (B) versus models with permuted shape or sequence features (bar colors). Permutations were performed for models with optimized hyperparameters.

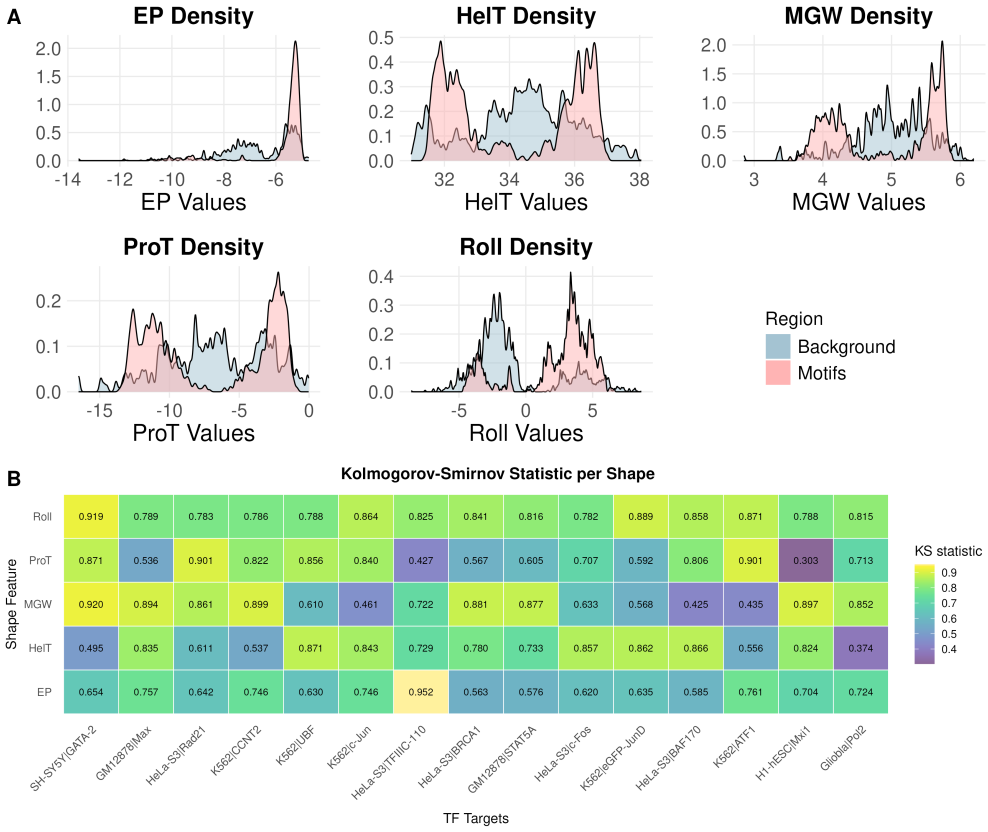

Figure 2.2: **Distribution of DNA shape feature values at high-attribution positions.** A) Density plots compare background distributions to positions with high DeepLIFT attribution scores. High-attribution positions are nucleotide positions within motifs that exhibit high average DeepLIFT attribution across aligned seqlets. The shifts in value distributions relative to background suggest structural patterns associated with regulatory activity. The x-axis represents the specific range of values for each shape feature. B) Kolmogorov–Smirnov (KS) statistics quantifying the degree of separation between motif and background distributions for each shape feature across 15 TF targets. Higher KS values indicate stronger motif–background divergence. Variation across both features and targets indicates that shape contributions are target-specific.

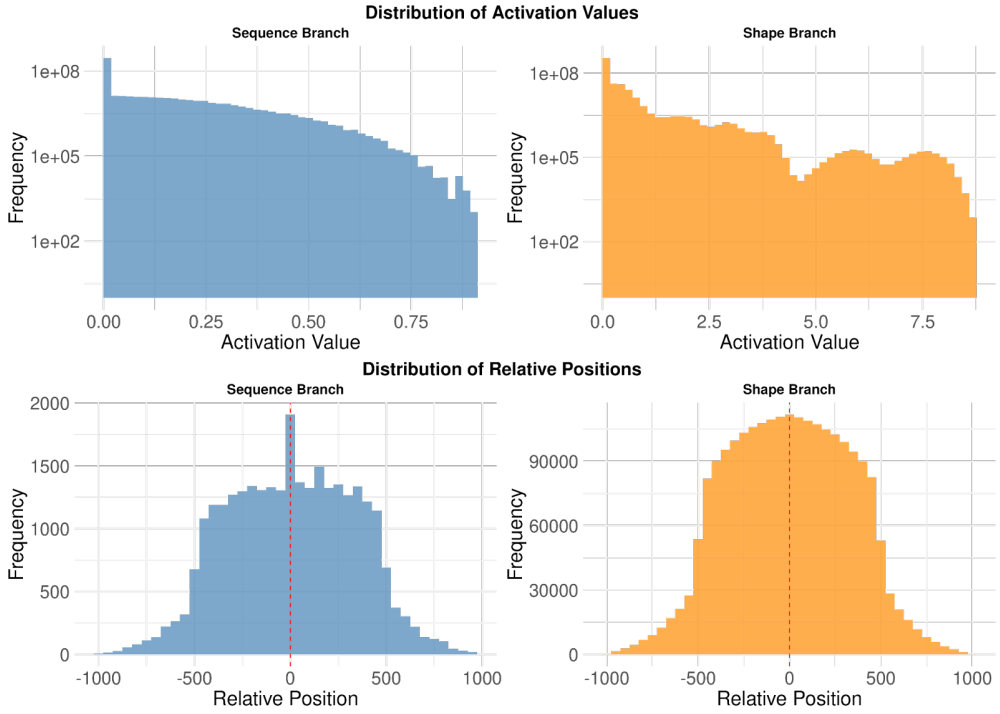

Figure 2.3: **Sequence and shape branch activations at the Myc-Max E-box.** Distribution of activation values across filters in the sequence (top left) and shape (top right) branch. Distribution of relative positions of activations above a given threshold in the sequence (bottom left) and shape (bottom right) branch. Sequences containing a single CACGTG motif are aligned such that the motif is at position 0.

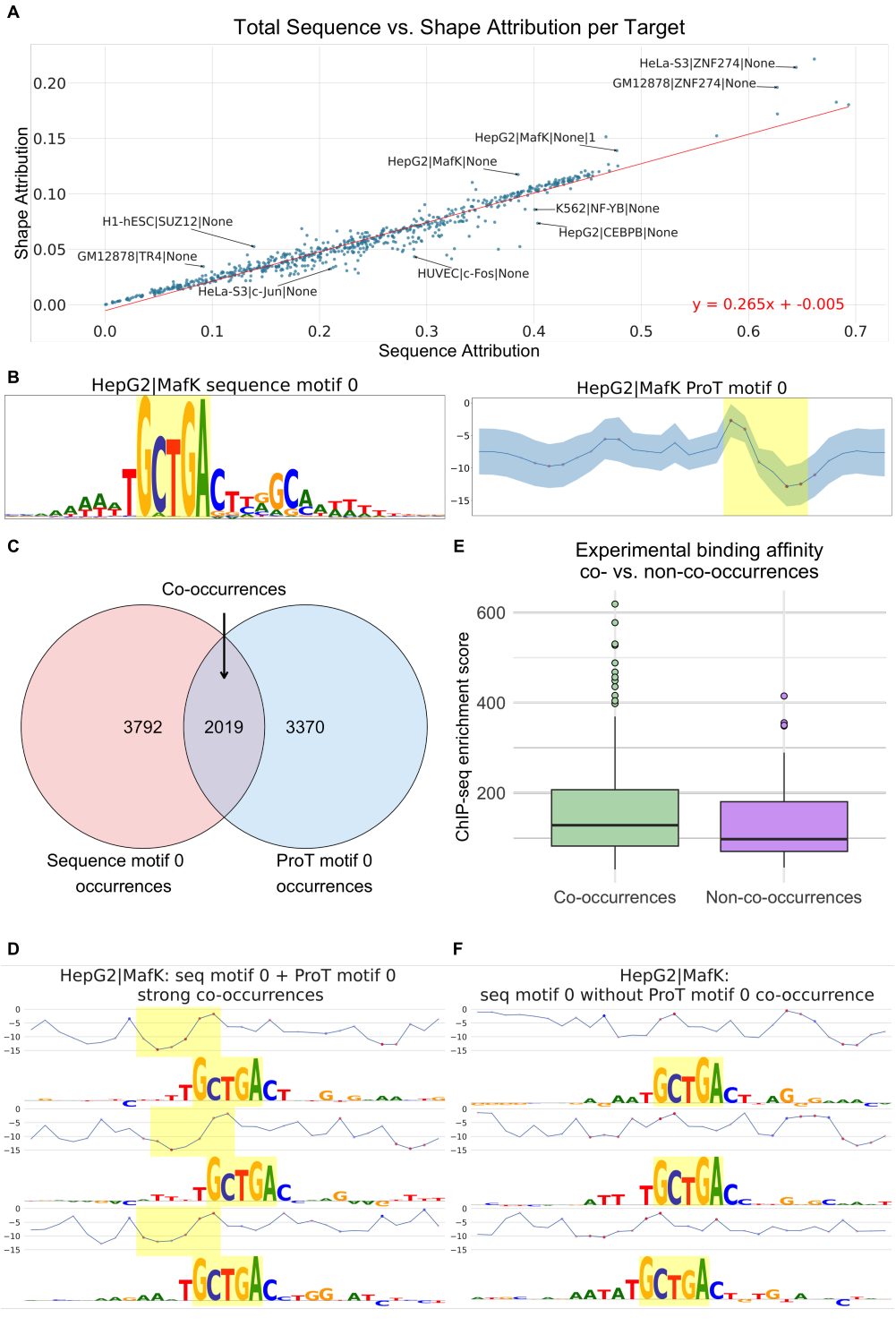
A
Total Sequence vs. Shape Attribution per Target
Shape Attribution
Sequence Attribution
HeLa-S3|ZNF274|None
GM12878|ZNF274|None
HepG2|MafK|None|1
HepG2|MafK|None
K562|NF-YB|None
HepG2|CEBPB|None
H1-hESC|SUZ12|None
GM12878|TR4|None
HUVEC|c-Fos|None
HeLa-S3|c-Jun|None
y = 0.265x + -0.005
B
HepG2|MafK sequence motif 0
HepG2|MafK ProT motif 0
C
Co-occurrences
3792
2019
3370
Sequence motif 0 occurrences
ProT motif 0 occurrences
E
Experimental binding affinity co- vs. non-co-occurrences
ChIP-seq enrichment score
Co-occurrences
Non-co-occurrences
D
HepG2|MafK: seq motif 0 + ProT motif 0 strong co-occurrences
F
HepG2|MafK: seq motif 0 without ProT motif 0 co-occurrence

Figure 2.3: **Sequence vs. shape attribution across TFs and a propeller-twist motif for MafK.** A) Mean sequence vs. shape attribution per TF target, with a line of best fit (slope = 0.269). MafK lies above the trend line. B) TF-MoDISco motifs for HepG2 | MafK: the canonical TGCTGA(C/G)TCAGCA sequence motif and a co-occurring propeller twist (ProT) shape motif. C–D) Co-occurrence of the ProT motif with the sequence motif: distance distribution and aligned profile. E) ChIP-seq enrichment for strong sequence-motif occurrences (>90th percentile attribution similarity) that co-occur with the ProT motif vs. those that do not. F) Representative occurrence showing the ProT dip aligned with an A-tract flanking the sequence motif.

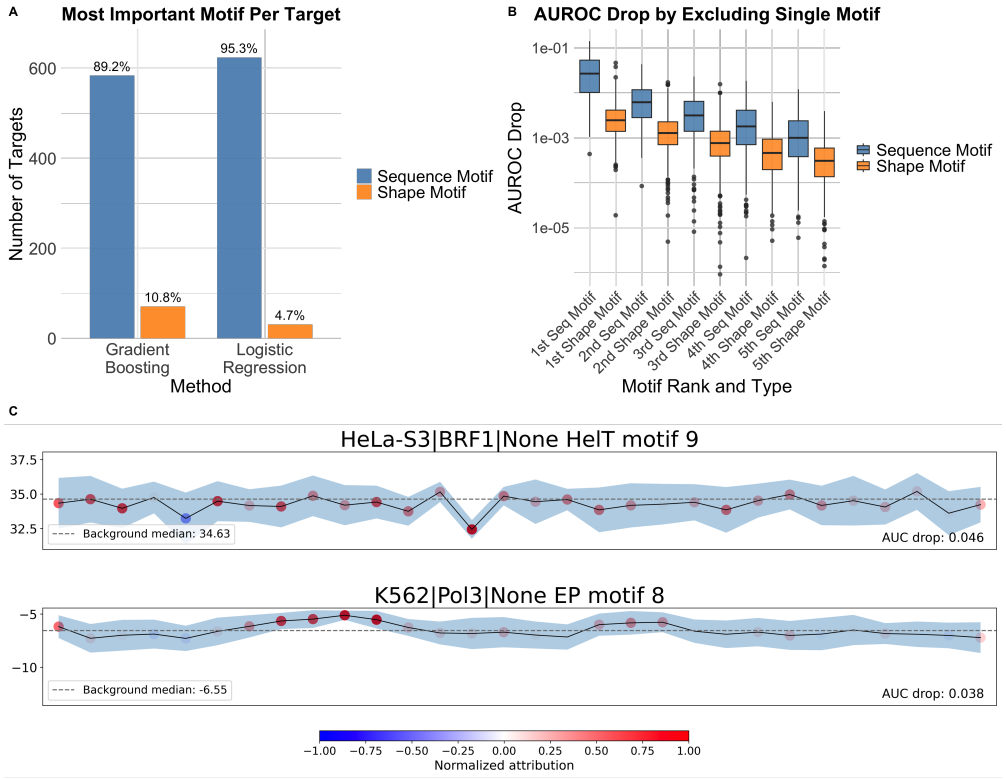

Figure 2.4: **Relative importance of sequence vs. shape motifs across TF targets.** A) Fraction of TF targets for which a sequence motif is the most important feature, by logistic regression coefficient significance and gradient-boosting feature importance. B) Performance drops from individual motif ablations in the gradient-boosting model, summarized across targets. C) Targets for which shape motifs were among the top contributors, with measurable performance drops on removal.

Chapter 3

Predicting allele-specific expression from phased diploid sequence

3.1 Overview

Predicting allele-specific expression addresses the dissertation's shared representation question in the context of sequence-to-expression modeling: whether using phased diploid inputs and allele-specific targets reveals behavior that is hidden when personal genomes are represented as a single sequence or expression is represented only as a bulk total. We fine-tuned the Enformer sequence-to-expression model on phased diploid sequencing data from GTEx v8, introducing raw phASER (phasing and allele specific expression from RNA-seq) allelic count targets and mask-based uninformative-donor filtering as design choices motivated by the structure of the phASER data [3, 66]. On held-out test genes, the model recovers a cross-individual ranking of allelic imbalance magnitude (pooled Spearman ρ=0.075; 81% of test genes positive) but does not predict the direction of imbalance (49.5% sign accuracy), a result that persists across seven architectural modifications.

3.2 Introduction

Gene expression in diploid organisms is biallelic: each autosomal locus produces transcripts from two haplotypes, and regulatory variants on either haplotype can shift the relative contribution of each allele. Allele-specific expression (ASE) quantifies this imbalance by measuring per-haplotype read counts at heterozygous sites. Population-scale studies, including the GTEx Consortium, have shown that cis-regulatory variants drive ASE variation across tissues and individuals [3]. Predicting ASE from phased diploid sequence would enable variant interpretation at haplotype resolution: rather than asking whether a locus is cis-regulated on average across the population, one could ask which haplotype context is affected in a specific individual and by how much.

Prior benchmarks have applied existing sequence-to-expression models to personal genome sequences to test whether sequence variation improves prediction of total gene expression across individuals. Those studies found limited cross-individual accuracy and frequent errors in the predicted direction of cis-regulatory effects, but they did not directly train or evaluate models to predict haplotype 1 or haplotype 2 expression separately from phased ASE targets. Thus it remains unclear whether pre-trained, long-range backbones, such as Enformer, Borzoi, or AlphaGenome, can learn per-haplotype expression from phased human genomes. This chapter tests that setting directly: fine-tuning Enformer to predict haplotype 1 and haplotype 2 expression separately, then evaluating both the magnitude and direction of allelic imbalance on held-out genes and people.

We developed a dual-head fine-tuning approach that predicts per-haplotype expression directly from phased diploid sequences. Each input consists of two haplotype sequences derived from phASER-called personal genotypes [66]; two output heads predict the phASER read count for haplotype 1 and haplotype 2 independently, sharing a single Enformer backbone [31] whose weights apply to both haplotype channels. We trained this architecture on allele-specific counts from 670 Whole Blood donors in GTEx v8 under 3-fold cross-validation, evaluating generalization on held-out donors (HOP) and held-out genes and held-out donors (HOGP).

Two methodological choices shape the model variant reported here. First, we used raw phASER read counts as training targets rather than log-transformed values; $\log_2$ transformation stabilizes total-expression variance but compresses haplotype differences toward zero in log space, suppressing within-gene allelic-imbalance signal. Second, donor-gene pairs where a donor carries zero ASE counts at a gene despite nonzero bulk expression are masked from the loss, zeroing their gradient contribution while preserving batch composition.

We report four analyses. First, we define an informative ASE training set by masking uninformative donor-gene pairs and filtering genes with little measurable ASE signal. Second, we evaluate allelic imbalance magnitude prediction on held-out genes. Third, we examine total-expression generalization in the HOGP setting. Finally, we ask whether the model predicts the direction of allelic imbalance and test whether additional model variants recover directional signal.

3.3 Results

3.3.1 Filtering defines an informative ASE training set

Raw phASER allele-specific expression (ASE) counts are sparse because a gene-donor pair is informative only when reads cover heterozygous sites in that gene. Before training, we therefore masked uninformative gene-donor pairs from the loss and excluded genes with no measurable ASE signal in the majority of donors.

ASE informativeness varies across genes and donors

Of 268,670 gene-donor pairs across 401 candidate genes and 670 Whole Blood donors, 98,695 (37.4%) carried zero ASE counts despite nonzero bulk RNA-seq expression. The median per-gene informativeness fraction was 0.670. Zero-count pairs provide no constraint on per-haplotype expression: under raw-count MSE, any nonzero prediction incurs a loss penalty with no supporting signal.

Per-gene informativeness correlated positively with median bulk expression (Pearson r=0.44; $p < 10^{-20}$) **(Fig. 3.1)**. Higher expression yields more reads per locus, increases the number and depth of RNA-seq reads overlapping heterozygous SNPs, and correspondingly increases the number of gene-donor pairs in which phASER can quantify allelic imbalance. Lower-expressed genes are uninformative for the majority of their donors not because their alleles are balanced but because no reads cover the heterozygous sites.

The eight chromosome-X genes in this 401-gene candidate set were also uninformative because the GTEx v8 phASER matrix excludes sex chromosomes. In the masked runs analyzed here, zero-count gene-donor pairs were treated as missing targets: they were masked from the loss and excluded from reported validation and test metrics, so performance is evaluated only on informative pairs with measurable haplotype-level ASE.

A median-count threshold retains genes with measurable cross-donor variation

Donor-level masking handles uninformativeness at the per-pair level but not genes for which a majority of donors are uninformative; for these genes, too few informative pairs remain to support a useful training signal: after masking, the loss is computed on few measured ASE targets and the backpropagation gradients used to update model weights are sparse and noisy.

We excluded genes whose median total ASE count across donors was zero, corresponding to genes for which at least half of donors provided no informative ASE signal. Of 401 candidate genes, 106 met this exclusion criterion.

The resulting trained set contains 222 training genes and 73 held-out genes, a 26% reduction from the 301-train, 100-test starting set. All analyses below use this filtered 222 + 73 gene set.

The donor mask and gene-level filter define the evaluation substrate: genes with measurable cross-donor ASE variation, evaluated only at gene-donor pairs where phASER provides an informative target. We next ask whether a model trained on this substrate predicts allelic imbalance on held-out genes.

3.3.2 The model predicts allelic imbalance magnitude on held-out genes

The gene filter and donor mask described above produce an evaluation cohort of 73 held-out test genes across 208 held-out donors, giving 15,184 informative test observations. For each

gene-donor pair, the dual-head model produces a predicted allelic imbalance magnitude, $|h_1 - h_2|$.

Predicted and observed imbalance are positively ranked

The pooled per-observation Spearman correlation between predicted and observed $|h_1 - h_2|$ on the 73 test genes is ρ=0.075 **(Fig. 3.2A)**. At the per-gene level, the median Spearman ρ on test genes is +0.120, with 59 of 73 (81%) showing $\rho > 0$ **(Fig. 3.2B)**. The training-gene median is ρ=0.139, a gap of roughly 0.02 between train and test gene medians.

Thus, although the effect size is small, the model ranks held-out gene-donor pairs by allelic imbalance magnitude better than expected by chance.

Accuracy varies with expression level and cis-window heterozygosity

Per-gene Spearman ρ varies by both gene-level allelic imbalance quartile and gene-level total expression quartile **(Fig. 3.2C-D)**: in both stratifications, the extreme quartiles outperform the middle quartiles. Median per-gene ρ across AI quartiles runs 0.143→0.051→0.112→0.123 (Q1→Q4); across expression quartiles it runs 0.136→0.034→0.101→0.141. The Q1 pattern should be interpreted alongside the donor-masking design, which restricts evaluation to informative (gene, donor) pairs and removes low-count, high-noise observations from the assessment of low-expression genes. Because uninformative low-count pairs are excluded from the metric, the Q1 value reflects performance on the informative subset of low-expression genes rather than on all low-expression test data.

We next repeated the stratification at the individual observation level rather than the per-gene level. Of 15,184 held-out gene-donor observations, 4,498 had zero observed imbalance and were excluded before computing quartiles, leaving 10,686 nonzero observations. When observations are grouped by observed total expression **(Fig. 3.3)**, the rank correlations are positive in all four quartiles: Spearman ρ=0.019, 0.006, 0.067, and 0.060 from Q1 to Q4. The two higher-expression quartiles show the clearest positive ranking signal, indicating that expression level helps define the observation regimes in which allelic-imbalance magnitude is more recoverable. Grouping the same observations by observed AI ratio **(Fig. 3.4)** shows a second axis of structure: the lowest and highest AI-ratio quartiles have the strongest rank correlations (Q1 ρ=0.123, Q4 ρ=0.126), while the middle quartiles are closer to zero (Q2 ρ=0.041, Q3 ρ=0.025).

The subsequent heterozygosity analyses are all HOGP analyses: they use held-out genes and held-out donors. They ask related but distinct questions at three resolutions: whether high-heterozygosity gene-donor pairs rank better in the pooled test set, whether high-heterozygosity genes have better per-gene accuracy, and whether high-heterozygosity donors within the same gene have larger or smaller prediction errors.

First, at the pooled gene-donor level, per-donor cis-window heterozygous SNP count adds an orthogonal, threshold-like structure. Here, "valid observations" are held-out-gene, held-out-donor pairs that passed the ASE evaluation mask and had a measured count of het-

erozygous SNPs in the gene's cis-regulatory window. When these observations were grouped by their donor-specific heterozygous SNP counts, prediction ranking improved as the number of SNPs increased and became clearly positive in the highest-SNP-count quartile **(Fig. 3.5)**. Spearman ρ is -0.074 in Q1 (0–16 het SNPs, 95% CI [-0.107,-0.043]), -0.045 in Q2 (16–43, CI [-0.078,-0.012]), -0.016 in Q3 (44–70, CI [-0.048,0.018]), and $+0.135$ in Q4 (70–244, CI [0.103,0.166]), with n=3,796 observations in each bin. The Q4 enrichment suggests that dense heterozygosity can provide usable haplotype-specific signal for ranking allelic imbalance, even when absolute magnitudes remain compressed.

Second, at the gene level, the question is whether held-out genes with higher average cis-window heterozygosity across donors also have higher per-gene AI ranking accuracy. Here, each of the 73 held-out genes is summarized by its mean het-SNP count and per-gene Spearman ρ. Mean cis-window het-SNP count is not a significant predictor of per-gene AI ranking accuracy (Spearman ρ=+0.190, p=0.107) **(Fig. 3.6)**. This between-gene result does not rule out a relationship between variant density and performance, but it shows no significant gene-level trend in the current held-out set.

Third, within individual held-out genes, the question is whether donors with more het-SNPs have larger absolute AI prediction errors for the same gene. Across the 73 qualifying held-out genes, the median within-gene Spearman ρ between donor-level het-SNP count and |error| is +0.187; 68/73 genes (93%) have $\rho > 0$; and the one-sided Wilcoxon signed-rank test gives p=3.36×10^{-12} **(Fig. 3.7)**. Together, these HOGP analyses separate variant-density effects by resolution: pooled gene-donor ranking is strongest in the highest het-count quartile, mean het density does not significantly predict which genes have better per-gene accuracy, and within a gene, donors with more het-SNPs tend to have larger absolute errors. More variants are therefore associated with a harder magnitude-calibration problem, but also with the clearest pooled ranking signal in the highest variant-density regime.

The joint expression × het structure **(Fig. 3.8)** combines these axes. Observations in het-quartile Q1 and Q2 remain near zero across expression rows, whereas the strongest rank correlation occurs among observations in expression Q4 and het-SNP Q3 (ρ=0.20). This pattern identifies the most favorable regime for the current model: observations with high expression and sufficient, but not extreme, cis-window heterozygosity.

Predicted magnitudes are systematically compressed toward zero

The rank correlations above characterize which (gene, donor) pairs have larger AI than others; they do not characterize the absolute scale of the predictions. We examined calibration at two resolutions: gene-level means (mean predicted versus mean observed $|h_1 - h_2|$ per gene) and a 20-quantile calibration curve over the pooled per-observation distribution **(Fig. 3.9)**.

At the gene level, the rank-correlation between mean predicted and mean observed $|h_1 - h_2|$ on the 73 test genes is ρ=0.136 **(Fig. 3.9A)**: genes with larger typical allelic imbalance receive larger mean predictions on the held-out set. The pooled 20-quantile calibration curve **(Fig. 3.9B)** characterizes the magnitude mismatch: observed AI is roughly 5 to 20 times

the predicted AI across the bulk of the distribution. Predicted magnitudes are systematically compressed toward zero. The contrastive loss penalizes squared differences in predicted gaps rather than absolute magnitudes; it does not enforce absolute-magnitude calibration.

On held-out genes and donors, the model produces positive rank-correlations for allelic imbalance magnitude (81% of test genes positive; median per-gene ρ=0.120; joint expression $\times$ het hotspot at ρ=0.20), with absolute magnitudes compressed roughly 5–20$\times$ toward zero. Generalization to total expression on the same held-out genes is examined next.

3.3.3 Total-expression generalization remains limited on held-out genes

The allelic imbalance analysis evaluates whether the model ranks haplotype differences on held-out genes. We next asked whether the same model predicts total expression on those genes. This is the HOGP setting: held-out genes evaluated in held-out people.

Training-gene prediction is positive, but test-gene prediction is centered near zero

On training genes (HOP), the shared backbone learns within-distribution total expression with a median per-gene R^2=+0.017; 115 of 222 (52%) of training genes achieve $R^2 > 0$.

On the 73 held-out test genes evaluated in held-out donors (HOGP), with metrics computed only on informative gene-donor pairs that pass the ASE evaluation mask, per-gene Pearson r between predicted and observed total expression is centered on zero **(Fig. 3.10)**: mean r=+0.007, median r=−0.009, with 34 of 73 (47%) of test genes showing r>0. This outcome is indistinguishable from a sign test under the null of no predictive signal. The corresponding R^2 distribution sits below zero across the entire test gene set: median R^2=−1.255 and 0 of 73 genes with $R^2 > 0$, with the best test gene (CAT) at R^2=−0.001. The aggregate mean R^2=−217.18 is dominated by a single outlier (NYNRIN, $R^2 \approx -9070$); we report the median as the central tendency throughout.

Training-gene predictions are positive and consistent across folds. Test-gene predictions are centered near zero by Pearson correlation, and no test gene crosses R^2=0. This mirrors the limitation highlighted by prior personal-genome sequence-to-expression benchmarks; adding individual genetic variation to model inputs has not been sufficient to reliably predict expression differences across people. Here, the same limitation persists even after fine-tuning directly on phased, haplotype-resolved ASE targets.

Held-out-gene scores vary with gene-level signal quality

The held-out genes do not behave uniformly: per-gene R^2 ranges from near-zero for CAT (−0.001) to the NYNRIN outlier at ≈−9070. We therefore asked whether measurable gene-level features are associated with which held-out genes are closer to zero. We computed

Spearman correlations between three gene-level signal-quality features and per-gene HOGP R^2 **(Fig. 3.11)**.

All three signal-quality features correlate positively with per-gene generalization: mean bulk expression at ρ=+0.54 (p=9.0×10^{-7}), mean ASE total signal at ρ=+0.64 (p=1.2×10^{-9}), and informativeness fraction at ρ=+0.40 (p=3.9×10^{-4}). Each reaches $|\rho| \geq 0.40$ on n=73 test genes, indicating that the heterogeneity in per-gene HOGP scores is structured rather than uniform.

Gene-level quality features stratify per-gene HOGP R^2 (Spearman ρ=0.40–0.64), yet no test gene crosses R^2=0. This result is consistent with concurrent fine-tuned sequence-to-expression models at similar cohort sizes [45, 46]. We next ask whether the model captures a different aspect of ASE: which haplotype is more highly expressed.

3.3.4 Directional allelic imbalance is not recovered by the current model

The analyses above show a small positive signal for allelic imbalance magnitude but limited total-expression generalization on held-out genes. We next asked whether the model predicts which haplotype expresses more: the sign of $h_1 - h_2$.

Sign prediction is centered at 50:50

Pooled across the 73 held-out test genes and 10,686 non-tie test-donor observations, the per-pair sign accuracy is 49.5% (chance =50.0%), with median per-gene sign accuracy 0.503 and 0 of 73 test genes binomially significant at α=0.05 **(Fig. 3.12)**. The signed pooled Spearman ρ between predicted and observed $h_1 - h_2$ is −0.006, compared with +0.075 on $|h_1 - h_2|$. Sign accuracy is flat at chance across all 150 training epochs in every fold.

Two probes assess whether directional information is present in the backbone representations. *In silico* mutagenesis finds no elevation of perturbation magnitude $|\Delta\text{pred}|$ at heterozygous-SNP positions relative to a non-SNP null (ratio 0.98, p=0.82) **(Fig. 3.13)**. A linear probe trained on frozen trunk embeddings to predict $\text{sign}(h_1 - h_2)$ reaches test accuracy 0.512 against a permutation baseline of 0.506 (gap +0.006, not significant) **(Fig. 3.14)**.

Together, these results indicate that the model learns information about the magnitude of allelic imbalance but does not recover a stable representation of which haplotype is more highly expressed.

What limits direction recovery, whether cohort size, input encoding, or other factors, is considered in the chapter Discussion.

3.4 Discussion

We developed a dual-head Enformer-based model for predicting allele-specific expression from phased diploid sequence. The model was trained on raw phASER counts with masked uninformative donor-gene pairs and a gene-level signal screen. On held-out genes, it recovers a small but consistent ranking signal for allelic imbalance magnitude: predicted and observed $|h_1 - h_2|$ are positively correlated in 59 of 73 test genes, with pooled Spearman ρ=0.075 and median per-gene ρ=0.120. This indicates that the model learns some information about which individuals carry larger haplotype differences at a gene.

The same model does not recover the direction of allelic imbalance. Sign accuracy is centered at chance on held-out genes, and the signed pooled correlation is near zero. Diagnostic analyses suggest that the shared Enformer backbone is only weakly sensitive to the heterozygous SNP positions that distinguish the two haplotypes. *In silico* mutagenesis at heterozygous sites produces perturbations similar in magnitude to a matched non-SNP null, and direction is not linearly decodable from frozen backbone embeddings. These results suggest that the current representation contains some information about imbalance magnitude but not enough information to assign that imbalance to the correct haplotype.

Held-out-gene total-expression prediction is also limited. The model has positive within-gene performance on training genes evaluated in held-out donors, but no held-out test gene achieves positive per-gene R^2 for total expression. This pattern is consistent with recent fine-tuning studies showing that held-out-gene-and-person generalization remains difficult for sequence-to-expression models at current cohort sizes. In our data, per-gene performance is structured rather than uniform: genes with higher expression, higher ASE count signal, and higher donor informativeness are closer to zero R^2. These features help explain relative performance across genes, but they do not overcome the held-out-gene generalization gap.

The raw-count and masking choices were motivated by the measurement structure of phASER ASE data. Raw counts preserve small haplotype differences that are compressed by log transformation, while masking prevents zero-coverage gene-donor pairs from being treated as measured zero expression. The present analyses do not isolate the contribution of each design choice in a full factorial comparison. Future comparisons between masked and unmasked raw-count models, and between MSE-based objectives and likelihoods matched to allelic count data, would clarify how much of the observed magnitude signal depends on the target representation and loss.

This study has several limitations. The data come from one tissue, Whole Blood, and from 670 donors. The filtered gene set contains 295 genes selected for measurable phASER signal rather than a genome-wide sample, so the conclusions apply most directly to genes with comparable ASE coverage. The model also uses Enformer as the shared sequence encoder, and the negative direction results are therefore conditional on this backbone and input representation. Longer-context encoders, explicit heterozygous-SNP channels, or architectures that compare the two haplotypes directly may expose information that is not available to the current head.

Together, these results define a narrow but informative outcome. Fine-tuning a pre-

trained sequence-to-expression model on phased ASE counts can recover a weak held-out-gene ranking of allelic imbalance magnitude, but the current model does not predict which haplotype is more highly expressed and does not solve held-out-gene total-expression generalization. The next models should make haplotype differences explicit in the input or architecture and should evaluate direction, magnitude, and held-out-gene generalization as separate endpoints.

3.5 Methods

3.5.1 Data and donor cohort

We use Whole Blood RNA-seq data and DNA genotypes from GTEx v8 [3], restricted to 670 donors with both phased genotypes and bulk RNA-seq available. Allele-specific counts are computed by the phASER pipeline [66], which assigns each haplotype-resolved read to one of the two parental haplotypes at every heterozygous position in a gene's transcribed region; the per-haplotype counts h_1 and h_2 are then summed across the gene to give a gene-donor pair's haplotype-resolved expression. The phASER matrix excludes sex chromosomes.

We start from 401 candidate genes (split 301 train, 100 test), including eight chromosome-X genes that are present in the GTEx bulk matrix but absent from phASER. Of the 268,670 candidate gene-donor pairs, 98,695 (37.4%) carry zero phASER ASE counts despite nonzero bulk RNA-seq expression; we refer to these as *uninformative pairs* because the read-level data do not constrain either haplotype's count. The proportion of uninformative pairs varies positively with gene expression (Pearson r=0.44, $p < 10^{-20}$).

The donor cohort is split for cross-validation into three folds, each holding out approximately 70 donors as test and using the remaining ∼600 for training. Across the three folds, 208 unique donors are seen in at least one test set. Splits are donor-disjoint within a fold but intersecting across folds.

3.5.2 Gene-level signal-quality filtering

All models in this chapter are multi-gene models: each training run is fit jointly across the retained training genes rather than separately per gene.

For every candidate gene we compute two signal-to-noise ratios. The MSE SNR contrasts the observed cross-donor variance in total ASE counts against the Poisson noise floor under a constant-mean model:

$$\mathrm{SNR}_{\mathrm{MSE}} = \frac{\sigma^2_{\mathrm{obs}} - \mu}{\mu}, \tag{3.1}$$

with μ the per-donor mean count and σ^2_{obs} the observed count variance across the 670 donors. The pairwise contrastive SNR contrasts the expected biological pairwise difference

between two donors' counts against the Poisson noise floor on that difference; it quantifies whether the inter-individual contrastive term in the loss has signal to fit.

We retain a gene if its median total ASE count across the 670 donors is strictly greater than zero. This single criterion excludes 106 of the 401 candidate genes and preserves all 295 remaining genes, every one of which exceeds SNR=1 on both metrics. The filtered cohort comprises 222 training and 73 test genes; per-fold gene assignments are listed in the supplementary appendix.

3.5.3 Donor-level mask-based filtering

Within the retained gene cohort, 37.4% of gene-donor pairs are uninformative (zero phASER ASE counts). Rather than treat zero counts as zero-valued targets, we mask them in the loss. The data loader replaces uninformative phASER counts with NaN at `__getitem__` time; the haplotype loss skips NaN positions during gradient computation. Training batches still draw from the full ∼600-donor pool, so batch composition is unchanged from the unmasked baseline; only the per-position loss contribution is adjusted.

3.5.4 Model architecture

The chapter's reference model is a dual-head allele-specific expression model built on the pretrained Enformer backbone [31, 67]. The two haplotypes for a given donor are encoded as separate one-hot DNA inputs over a ∼49 kb cis window centered on the gene's transcription start site; both inputs are passed through the same Enformer trunk (the trunks are tied at the Python-object level, `model_hap1.enformer is model_hap2.enformer`, so they share weights and gradient updates). Two independent output heads, each a small fully-connected layer over the trunk's centre-bin embedding, predict the per-haplotype counts h_1 and h_2.

3.5.5 Loss function and training procedure

The chapter's reference model is trained on raw phASER counts under a two-term loss combining a per-haplotype mean-squared-error term with an inter-individual contrastive term:

$$\mathcal{L} = \alpha \cdot \mathrm{MSE}_{\mathrm{masked}}(\hat{h}, h) + (1 - \alpha) \cdot \mathcal{L}_{\mathrm{inter}}(\hat{h}, h), \tag{3.2}$$

with α=0.5. The masked MSE skips gene-donor positions whose target counts are NaN (i.e., the uninformative pairs masked above). The inter-individual contrastive term $\mathcal{L}_{\mathrm{inter}}$ penalizes the squared error in the predicted pairwise difference of haplotype counts between every pair of donors in a batch, separately per haplotype:

$$\mathcal{L}_{\mathrm{inter}}(\hat{h}, h) = \frac{1}{N(N-1)} \sum_{i \neq j} \left(\left(\hat{h}^{(i)} - \hat{h}^{(j)} \right) - \left(h^{(i)} - h^{(j)} \right) \right)^2, \tag{3.3}$$

where i, j index donors in the batch. This term encourages the model to preserve cross-individual variation in counts even when the absolute count level is hard to fit.

Training uses the Adam optimizer at learning rate 5×10^{-6} with bf16-mixed precision, batch size 16 genes by 128 donors per gene, for up to 150 epochs with no early stopping. Three folds are trained independently from the same Enformer initialization. Lightning saves both the running last-epoch checkpoint and a set of best-by-metric checkpoints throughout training.

3.5.6 Evaluation and checkpoint selection

Two evaluation regimes are used throughout the chapter. Held-out people (HOP) evaluates training genes in donors withheld from training, measuring cross-individual generalization within the trained gene set. Held-out genes and people (HOGP) evaluates test genes in held-out donors, measuring cross-gene generalization and serving as the main evaluation regime for the chapter's central claims.

For each fold, test-time predictions are generated from the checkpoint with the best donor-total R^2 on training genes. This checkpoint policy is used consistently across the chapter. Per-gene metrics aggregate predictions across folds and compute R^2 and Spearman ρ separately for each gene against the corresponding held-out donor observations. Aggregate statistics are reported as medians across genes unless stated otherwise; means are reserved for outlier-driven summaries.

3.5.7 Rationale for raw-count targets

The chapter's reference model uses *raw* phASER counts as targets rather than the $\log_2$(counts+1) transformation that an earlier model variant used. Under $\log_2$ transformation, small allele-specific count differences are compressed by the $\log_2(x+1)$ pseudocount and become indistinguishable from sampling noise, and the magnitudes the dual-head architecture is meant to recover are exactly the small per-haplotype differences that $\log_2$ squashes most.

A preliminary head-to-head comparison on the chapter's 73 test genes supports this rationale. Under $\log_2$-transformed targets, the dual-head ASE architecture produces no measurable allelic-imbalance ranking on held-out test genes: pooled Spearman $\rho{=}-0.03$ between predicted and observed $|h_1 - h_2|$, median per-gene $\rho{=}+0.02$, with 40 of 73 genes (55%) above zero, approximately the proportion expected under the null. On the same gene set, dual-head ASE under $\log_2$ is also indistinguishable from a single-target baseline on total-expression prediction: median per-gene R^2 is -0.56 for ASE versus -0.51 for single-target, paired Wilcoxon p=0.48; median per-gene Spearman ρ is $+0.03$ versus $+0.04$, paired Wilcoxon p=0.61. The dual-head architecture under $\log_2$ adds nothing over a single-scalar architecture by either measure on which they can be compared. With raw-count targets, the same dual-head architecture instead recovers a measurable AI ranking signal on the same gene set: pooled $\rho{=}+0.075$, median per-gene $\rho{=}+0.120$, 59 of 73 genes (81%) above zero.

We adopt raw counts as the chapter's target representation on the basis that they preserve the per-haplotype magnitude information the dual-head architecture is built to model.

3.5.8 Rationale for mask-based donor filtering

Mask-based donor filtering is a separate design choice motivated by correctness rather than by allelic-imbalance recovery. The 37.4% of gene-donor pairs that carry zero phASER ASE counts despite nonzero bulk RNA-seq expression are uninformative in a strict sense: the read-level data do not constrain either haplotype's count, so any particular numerical target value (zero, the donor's average, the gene's average, or anything else) is an arbitrary choice rather than a biological measurement. Treating them as zero-valued targets in the loss does two things, both undesirable. It biases the model toward predicting zero on similar-looking inputs because zero is the most-common "correct" answer in the training set. And it inflates apparent training performance: when the model predicts close to zero for an uninformative pair and the target is also zero, the per-pair squared error is small and pulls the aggregate metrics up, even though no underlying biological signal was fit. Mask-based filtering, replacing these targets with NaN at `__getitem__` and skipping them in the loss, avoids both effects: the model never sees zero-count pairs as supervision, and the per-gene metrics reflect performance only on positions where the data actually constrain the target.

We do not present a head-to-head masked-versus-unmasked comparison under matched target transform here; that comparison would require running the raw-count ASE model both with and without donor masking on the same gene cohort, and we have not produced the unmasked raw-count comparator. The masking choice in the chapter's reference design is therefore an *asserted* methodological choice on the grounds above, rather than a choice supported by an internal head-to-head ablation. The single-target comparator runs that share the chapter's filtered-raw-count masked design confirm that the mask-based design admits stable training and produces metrics consistent with the magnitude-without-direction interpretation pursued in the chapter's Results.

3.6 Figures

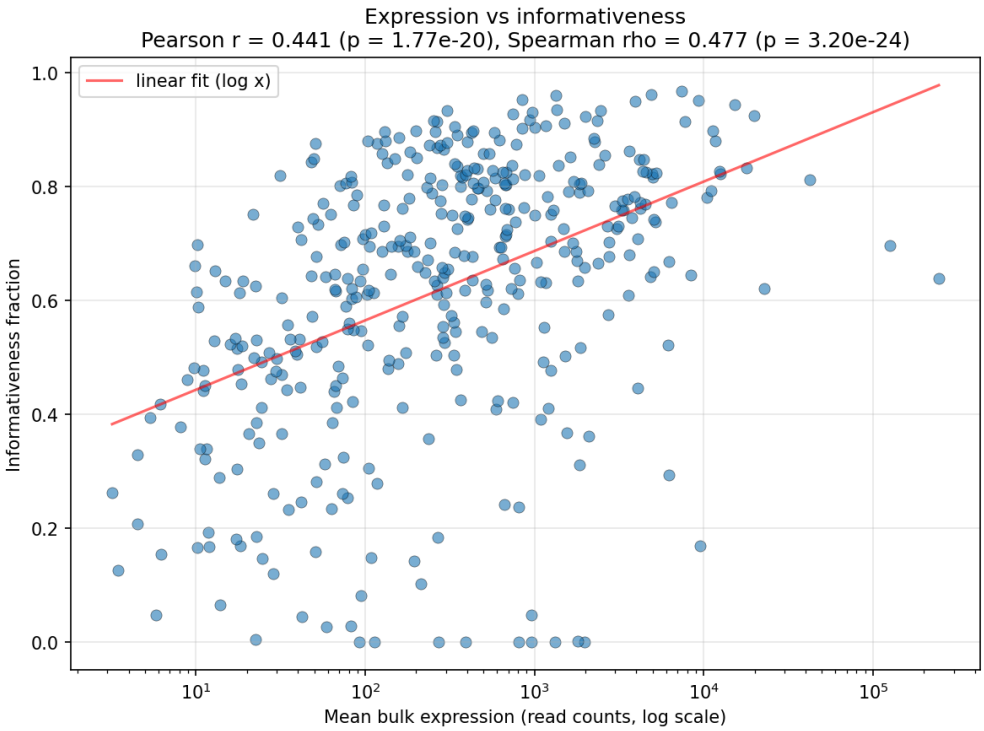

Figure 3.1: **Per-gene informativeness across candidate genes.** Per-gene informativeness fraction is plotted against median bulk RNA-seq expression for 401 candidate genes. Informativeness is the fraction of 670 Whole Blood donors with nonzero phASER ASE counts for a given gene. Pearson $r = 0.44$ ($p < 10^{-20}$). Across all 268,670 gene–donor pairs, 37.4% are uninformative; the median per-gene informativeness fraction is 0.670.

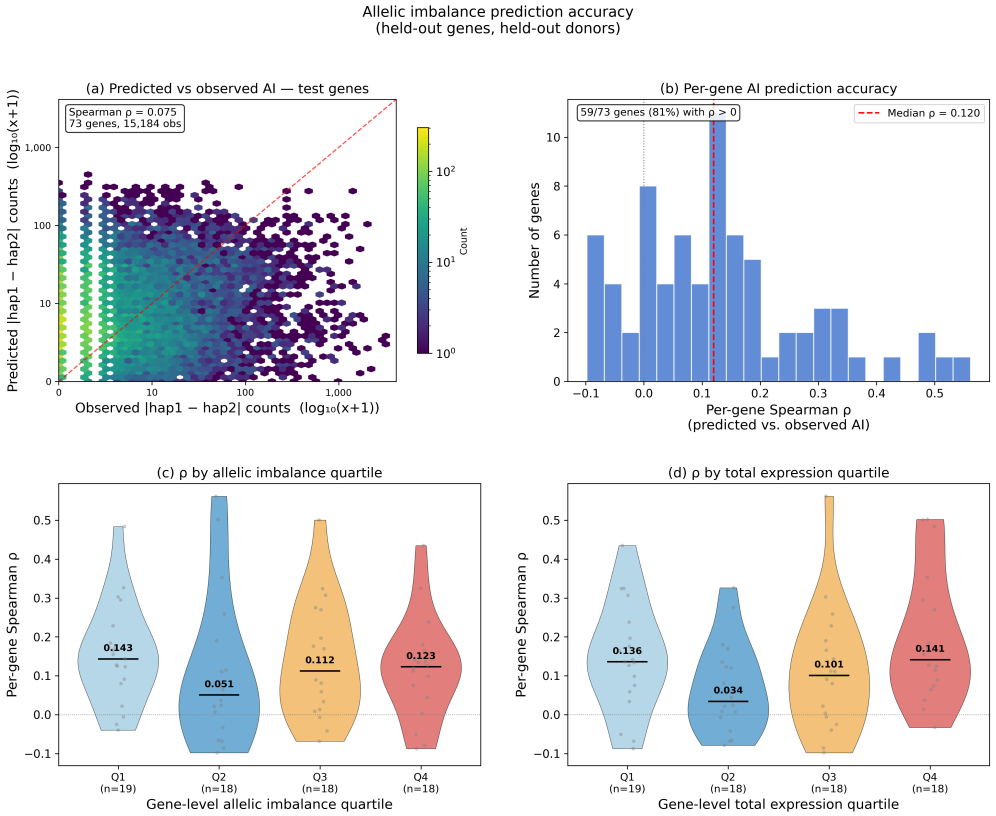

Figure 3.2: **Allelic imbalance prediction on held-out genes.** A) Predicted vs. observed $|h_1 - h_2|$ across 73 test genes and 15,184 observations, pooled across three cross-validation folds. Counts are shown on $\log_{10}(x+1)$ axes; pooled Spearman $\rho = 0.075$. B) Distribution of per-gene Spearman ρ across the 73 test genes. Median $\rho = 0.120$; 59 of 73 genes have $\rho > 0$. C) Per-gene ρ grouped by gene-level allelic-imbalance quartile. D) Per-gene ρ grouped by gene-level total-expression quartile.

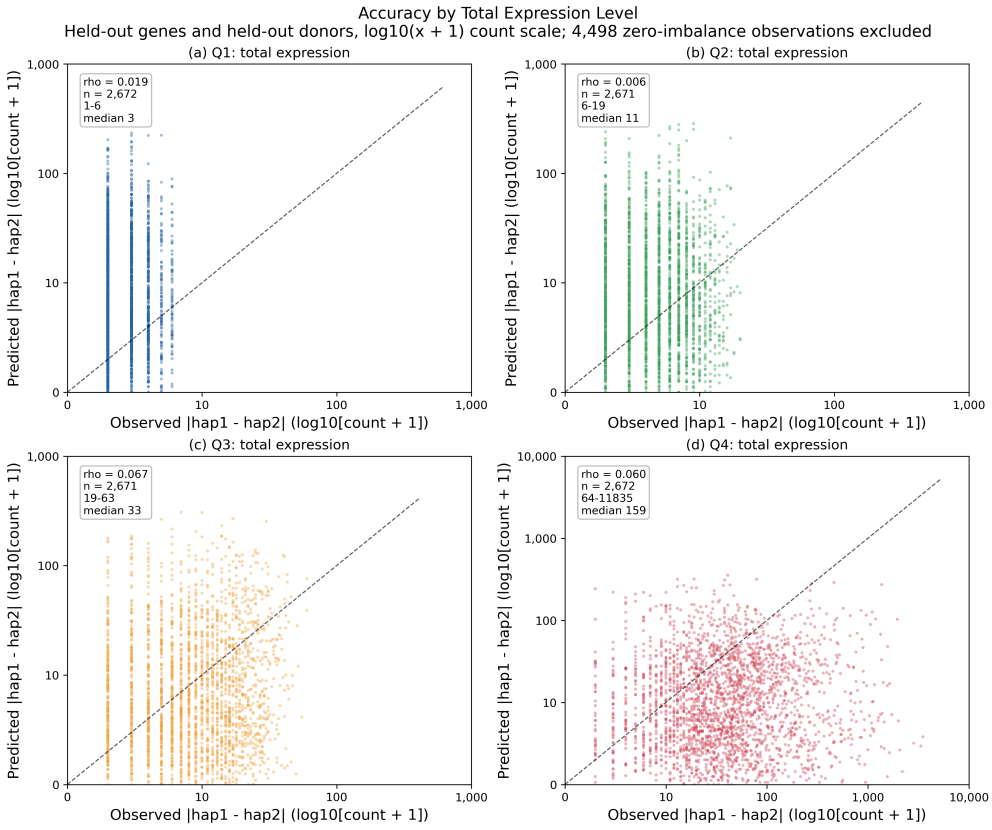

Figure 3.3: **Observation-level allelic-imbalance prediction by total expression.** Each point is one held-out-gene, held-out-donor observation. Observed and predicted allelic imbalance are $|h_1 - h_2|$; both axes use a $\log_{10}(x + 1)$ count scale. Zero observed-imbalance observations were excluded before quartiling, leaving 10,686 observations. Panels show quartiles of observed total expression: Q1, 1–6 counts (median 3, $n = 2{,}672$, $\rho = 0.019$); Q2, 6–19 ($n = 2{,}671$, $\rho = 0.006$); Q3, 19–63 ($n = 2{,}671$, $\rho = 0.067$); Q4, 64–11,835 ($n = 2{,}672$, $\rho = 0.060$).

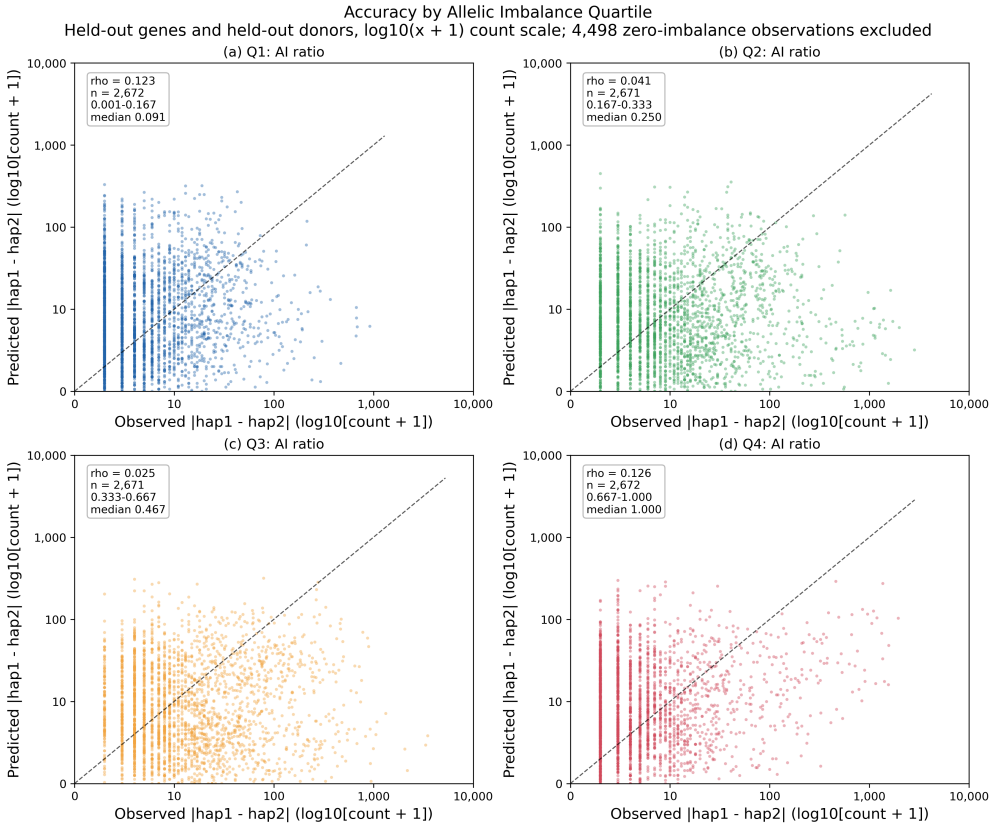

Figure 3.4: **Observation-level allelic-imbalance prediction by observed AI ratio.** The same 10,686 nonzero-imbalance held-out gene–donor observations are grouped by observed AI ratio, $|h_1 - h_2|/(h_1 + h_2)$. Each point compares observed and predicted $|h_1 - h_2|$ on a $\log_{10}(x + 1)$ count scale. Q1 spans 0.001–0.167 ($n = 2{,}672$, $\rho = 0.123$); Q2 0.167–0.333 ($n = 2{,}671$, $\rho = 0.041$); Q3 0.333–0.667 ($n = 2{,}671$, $\rho = 0.025$); Q4 0.667–1.000 ($n = 2{,}672$, $\rho = 0.126$).

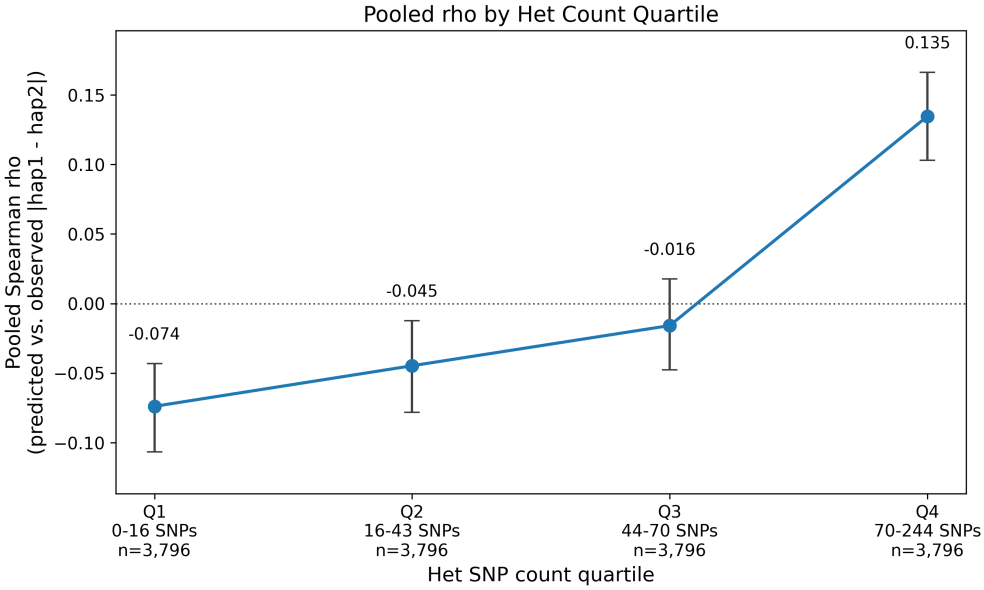

Figure 3.5: **Pooled AI ranking accuracy by het-SNP count quartile.** All valid held-out-gene, held-out-donor observations with valid het-SNP counts are pooled and split into four equal-count bins by the number of heterozygous SNPs in the 49,152 bp cis-window centered on the TSS. Within each bin, Spearman ρ is computed between observed and predicted $|h_1 - h_2|$; error bars show 95% bootstrap confidence intervals from 1,000 resamples. Q1 (n = 3,796, $\rho = -0.074$, CI $[-0.107, -0.043]$); Q2 (n = 3,796, $\rho = -0.045$, CI $[-0.078, -0.012]$); Q3 (n = 3,796, $\rho = -0.016$, CI $[-0.048, 0.018]$); Q4 (n = 3,796, $\rho = 0.135$, CI $[0.103, 0.166]$).

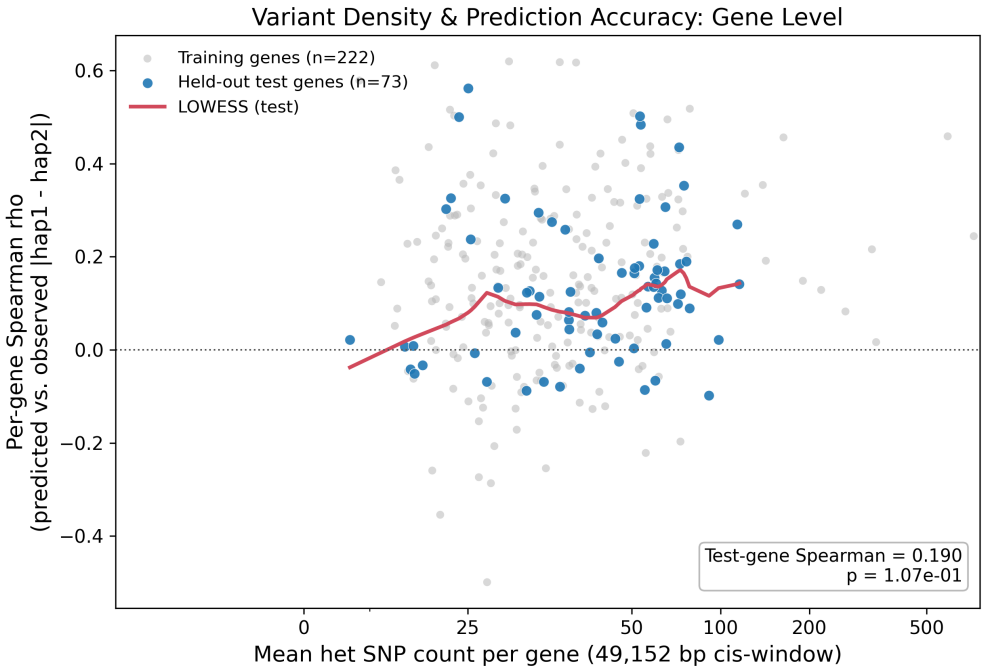

Figure 3.6: **Gene-level variant density and AI prediction accuracy.** Each point is one gene, with training genes in gray and held-out test genes in blue. The x-axis shows the mean number of heterozygous SNPs per gene across donors in the 49,152 bp cis-window centered on the TSS; the y-axis shows per-gene Spearman ρ between predicted and observed $|h_1 - h_2|$. The LOWESS curve is fit only to held-out test genes. Among 73 held-out genes, mean het-SNP count is not significantly correlated with per-gene accuracy (Spearman $\rho = 0.190$, $p = 0.107$).

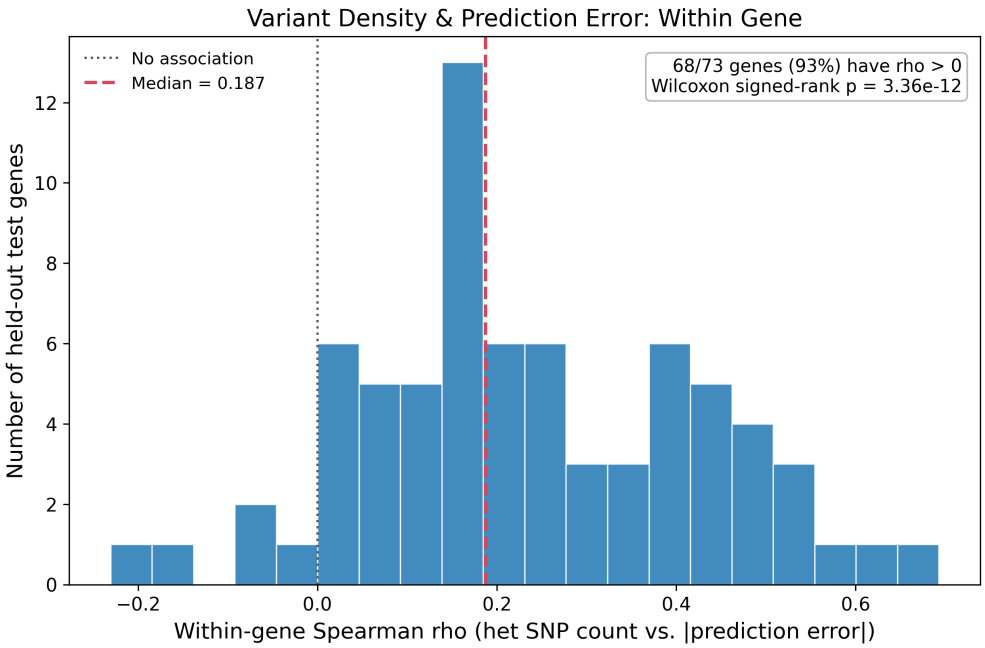

Figure 3.7: **Within-gene association between donor het-SNP count and absolute AI error.** For each held-out test gene, Spearman ρ is computed between donor-level het-SNP count and absolute error, $\big||h_1 - h_2| - |\hat{h}_1 - \hat{h}_2|\big|$. Genes are included if at least 10 donors have nonzero het-count and error variance; all 73 held-out genes qualify. Median within-gene $\rho = 0.187$; 68 of 73 genes have $\rho > 0$; Wilcoxon signed-rank $p = 3.36 \times 10^{-12}$.

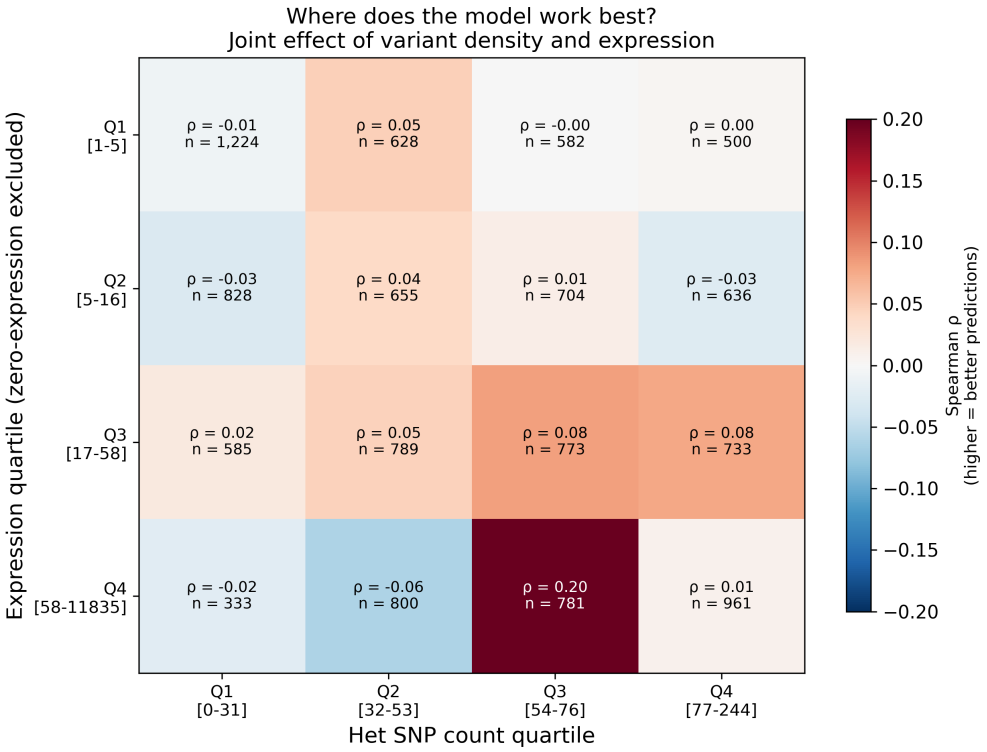

Figure 3.8: **Joint expression × heterozygosity stratification.** Pooled Spearman ρ between predicted and observed $|h_1 - h_2|$ for each cell of a 4×4 grid defined by gene-level expression quartile and per-pair cis-window het-SNP count quartile. Cell color indicates ρ; each label gives the number of observations in the cell. The highest cell is expression Q4 and het-SNP count Q3 ($\rho = 0.20$, $n = 781$).

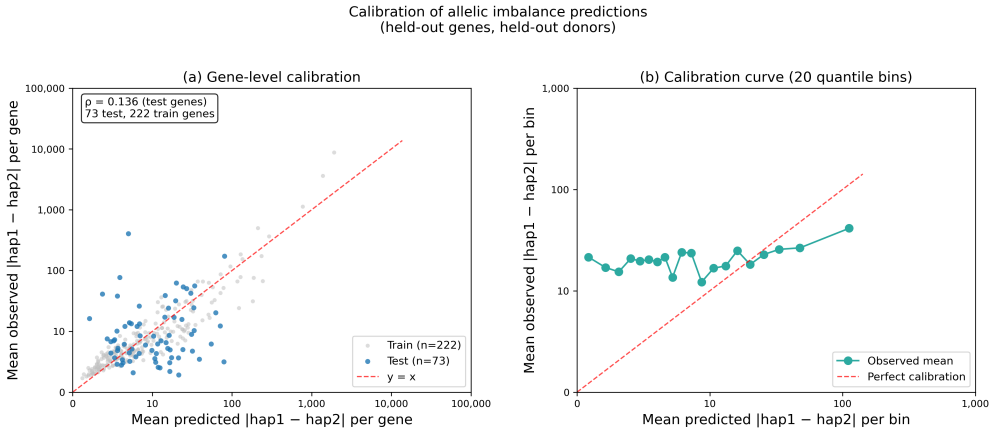

Figure 3.9: **Calibration of allelic-imbalance predictions.** A) Mean predicted vs. mean observed $|h_1 - h_2|$ per gene, with train genes in gray ($n = 222$) and test genes in blue ($n = 73$). Axes are shown on a $\log_{10}$ scale; the dashed line marks $y = x$. Test-gene Spearman $\rho = 0.136$. B) Mean observed $|h_1 - h_2|$ in 20 bins of predicted magnitude, pooled across held-out genes, donors, and folds.

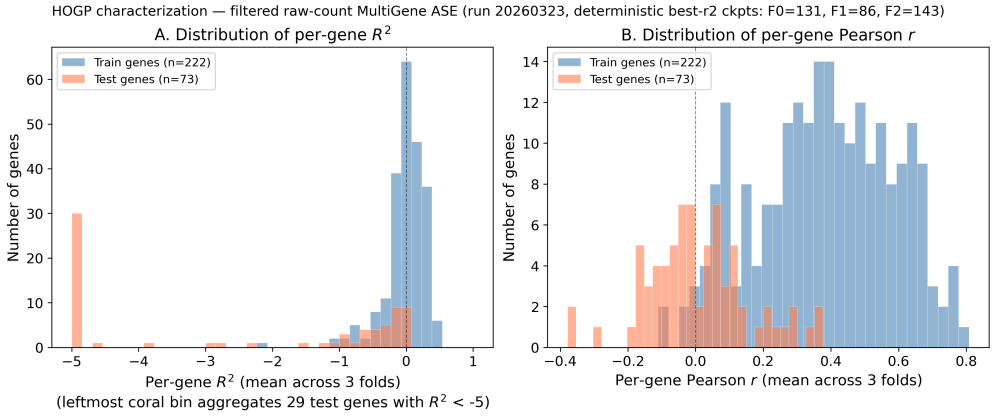

Figure 3.10: **Held-out-gene total-expression performance.** Per-gene R^2 (left) and Pearson r (right) for 73 held-out test genes, with train-gene distributions overlaid for reference. Values are averaged across three folds. Test-gene R^2 is below zero for all genes (median -1.255; best gene CAT, $R^2 = -0.001$); test-gene Pearson r is centered near zero (median -0.009; 34/73 genes above zero). The R^2 axis uses a symlog scale to show the full range, including NYNRIN ($R^2 \approx -9070$).

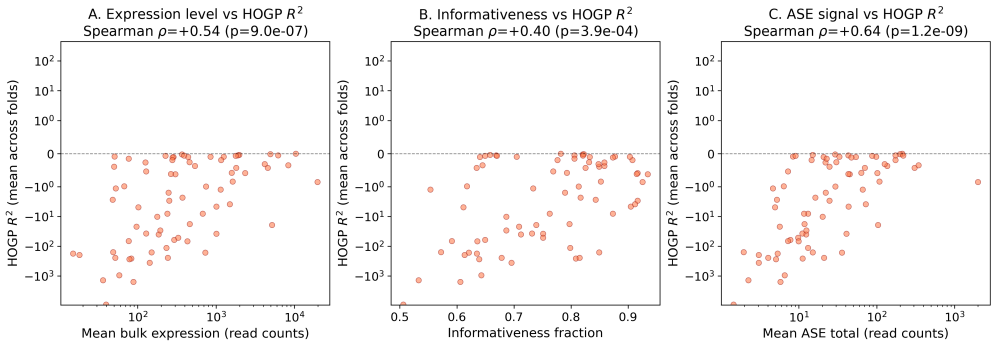

Figure 3.11: **Per-gene HOGP R^2 stratified by gene-level signal features.** Per-gene HOGP R^2 for 73 held-out test genes against A) mean bulk expression, B) informativeness fraction, and C) mean ASE total signal. Features computed from the training donor pool. Spearman $\rho = +0.54$ for mean bulk expression ($p = 9.0 \times 10^{-7}$); $\rho = +0.40$ for informativeness fraction ($p = 3.9 \times 10^{-4}$); $\rho = +0.64$ for mean ASE total signal ($p = 1.2 \times 10^{-9}$). The y-axis uses a symlog scale in all panels.

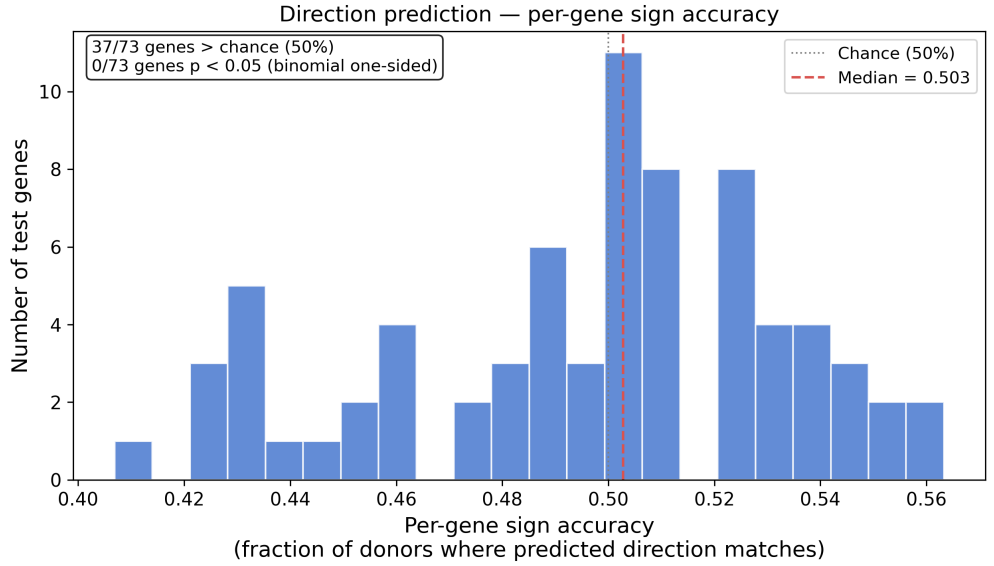

Figure 3.12: **Sign accuracy on held-out genes.** Distribution of per-gene sign accuracy across 73 held-out test genes. The dotted vertical line marks chance accuracy (0.5); the dashed red line marks the median per-gene accuracy (0.503). Per-gene significance was assessed using a one-sided binomial test against chance; no gene reached nominal $\alpha = 0.05$.

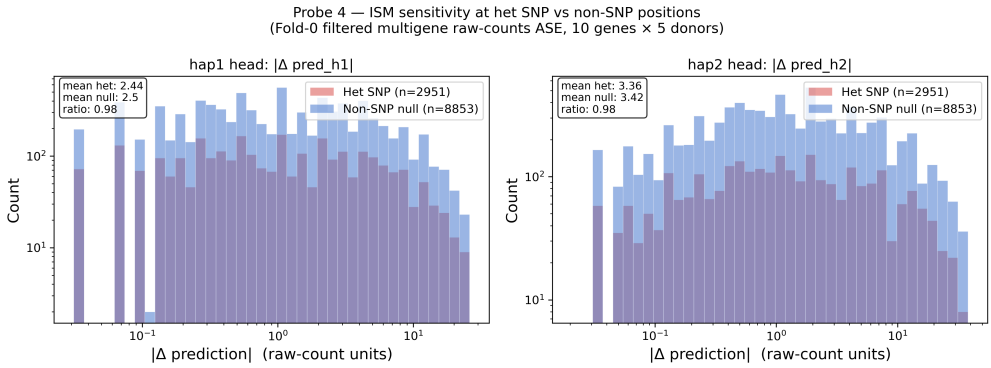

Figure 3.13: **In-silico mutagenesis at heterozygous-SNP positions.** Per-position |Δpred| is compared between heterozygous-SNP positions and matched non-SNP positions across 670 donors and 295 genes. The het-SNP and matched non-SNP distributions overlap closely; the |Δ| ratio is 0.98 ($p = 0.82$).

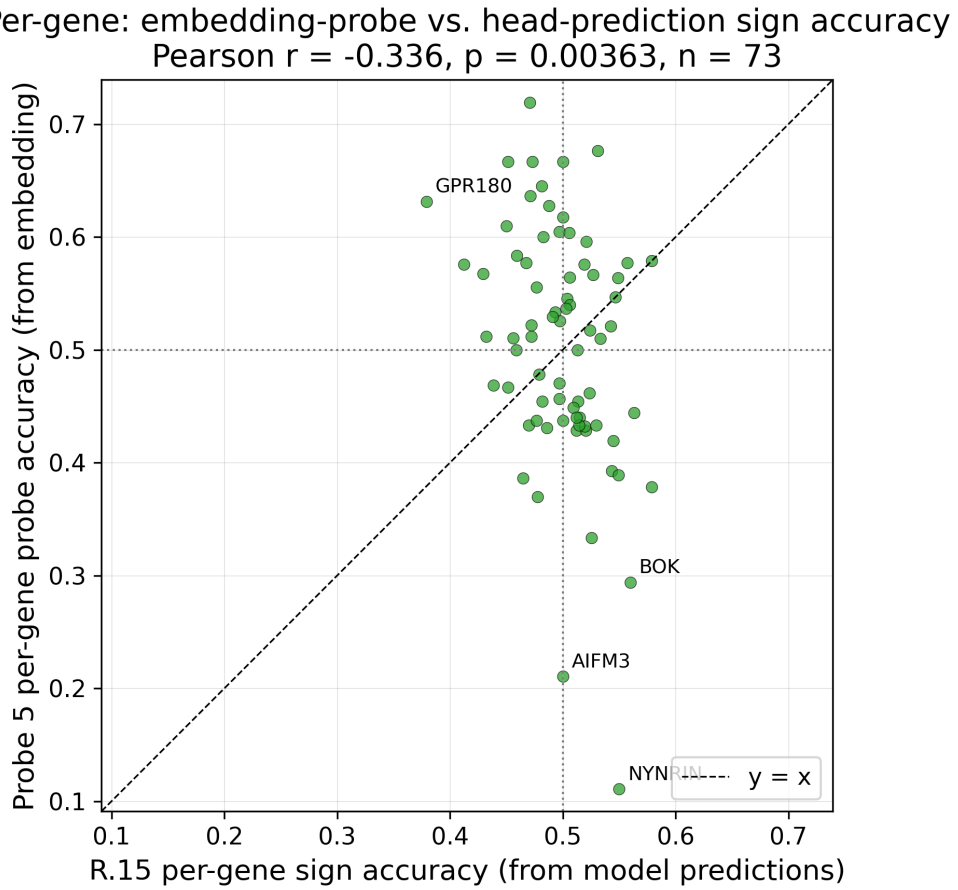

Figure 3.14: **Linear probe for directional information in backbone embeddings.** Per-gene test accuracy for a linear probe trained to predict $\text{sign}(h_1 - h_2)$ from frozen backbone embeddings, alongside a within-gene permutation null. Mean test accuracy is 0.512, vs. 0.506 for the permutation baseline.

Chapter 4

Conclusions

This dissertation examines how the choice of input and output representation in genomic deep-learning models affects both what those models learn and what can be read from them. Chapter 2 augments the input side, adding DNA shape as a separate channel alongside one-hot sequence and asking whether the explicit representation supports interpretations of TF binding that sequence-only attribution cannot recover. Chapter 3 augments the output side, replacing single-haplotype expression targets with per-haplotype phASER counts and asking whether a fine-tuned sequence-to-expression backbone can predict allele-specific expression on held-out genes. The common thread is that what is fed in and what is asked of the model determine what can be read out, and that simpler representations sometimes leave information on the table that more explicit representations recover.

Across the two chapters, the empirical pattern is consistent. Adding shape features to a TF-binding model produces only modest gains in aggregate predictive accuracy but a substantial gain in interpretability: shape attributions identify flanking-region structural patterns, such as the propeller-twist motif that co-occurs with MafK's canonical binding site, that are not visible in sequence-only attribution. Fine-tuning Enformer on phased ASE counts produces a small held-out-gene ranking signal for allelic imbalance magnitude and no signal for direction, but the magnitude signal that does emerge depends on representation choices — raw counts rather than log-transformed targets, and donor-level masking of uninformative gene-donor pairs — that were chosen to avoid compressing small haplotype differences. In both chapters, the gain from making information explicit is real but limited, and in both chapters that limit reflects a form of compression: continuous shape values produce diffuse attributions that smear across positions, and the predicted ASE magnitudes are systematically attenuated relative to the targets they fit.

The work has clear limitations. The shape features used in Chapter 2 are static and local, and the chapter does not test whether dynamic or larger-scale structural features — topologically associated domains, methylation-modified shape, or transient conformational states — would change the picture. The ASE work in Chapter 3 is restricted to one tissue, 670 donors, and a filtered set of 295 genes; it uses a single backbone, Enformer, and the directional-imbalance negative result is conditional on that backbone and input representa-

tion. Neither chapter presents a full factorial comparison of the design choices it adopts, so the relative contributions of individual choices — shape feature inclusion, raw counts versus log targets, masking versus no masking — remain to be quantified separately.

The natural directions for follow-up work are visible from these limitations. On the shape side, binarized or quantile-encoded shape inputs would test whether the diffuse attributions reflect a property of the data or an artifact of continuous representation, and shape features as output tracks rather than input channels would let a model learn structural readout end-to-end. On the ASE side, architectures that compare haplotypes directly, explicit heterozygous-SNP channels, longer-context backbones, and larger phased cohorts are the most direct routes toward recovering directional signal. More broadly, the dissertation suggests that progress in sequence-to-function modeling is gated less by raw model capacity than by whether the representation makes the relevant biological distinctions visible to the loss function. Models that compress a meaningful biological difference into the same numerical neighborhood will not learn that difference, regardless of scale; models that keep it separated have a chance to.

Bibliography

[1] Matthew T. Maurano et al. "Systematic localization of common disease-associated variation in regulatory DNA". In: *Science* 337.6099 (2012), pp. 1190–1195. DOI: 10.1126/science.1222794.

[2] Stacey L. Edwards et al. "Beyond GWASs: Illuminating the dark road from association to function". In: *The American Journal of Human Genetics* 93.5 (2013), pp. 779–797. DOI: 10.1016/j.ajhg.2013.10.012.

[3] GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (2020), pp. 1318–1330. DOI: 10.1126/science.aaz1776.

[4] Tuuli Lappalainen et al. "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468 (2013), pp. 506–511. DOI: 10.1038/nature12531.

[5] Eric R. Gamazon et al. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature Genetics* 47.9 (2015), pp. 1091–1098. DOI: 10.1038/ng.3367.

[6] Alexander Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies". In: *Nature Genetics* 48.3 (2016), pp. 245–252. DOI: 10.1038/ng.3506.

[7] Remo Rohs et al. "The role of DNA shape in protein–DNA recognition". In: *Nature* 461.7268 (2009), pp. 1248–1253. DOI: 10.1038/nature08473.

[8] Remo Rohs et al. "Origins of specificity in protein–DNA recognition". In: *Annual Review of Biochemistry* 79 (2010), pp. 233–269. DOI: 10.1146/annurev-biochem-060408-091030.

[9] Lin Yang et al. "Transcription factor family-specific DNA shape readout revealed by quantitative specificity models". In: *Molecular Systems Biology* 13.2 (2017), p. 910. DOI: 10.15252/msb.20167238.

[10] Md. Abul Hassan Samee, Benoit G. Bruneau, and Katherine S. Pollard. "A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs". In: *Cell Systems* 8.1 (2019). Listed in the dissertation as "Mah *et al.* 2019"; corresponds to the Samee et al. 2019 paper., 27–42.e6. DOI: 10.1016/j.cels.2018.12.001.

[11] Rohit Joshi et al. "Functional specificity of a Hox protein mediated by the recognition of minor groove structure". In: *Cell* 131.3 (2007), pp. 530–543. DOI: 10.1016/j.cell.2007.09.024.

[12] Namiko Abe et al. "Deconvolving the recognition of DNA shape from sequence". In: *Cell* 161.2 (2015), pp. 307–318. DOI: 10.1016/j.cell.2015.02.008.

[13] Raluca Gordân et al. "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape". In: *Cell Reports* 3.4 (2013), pp. 1093–1104. DOI: 10.1016/j.celrep.2013.03.014.

[14] Ana Carolina Dantas Machado et al. "Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout". In: *Nucleic Acids Research* 48.15 (2020), pp. 8529–8544. DOI: 10.1093/nar/gkaa642.

[15] Tianyin Zhou et al. "DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale". In: *Nucleic Acids Research* 41.W1 (2013), W56–W62. DOI: 10.1093/nar/gkt437.

[16] Tsu-Pei Chiu et al. "DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding". In: *Bioinformatics* 32.8 (2016), pp. 1211–1213. DOI: 10.1093/bioinformatics/btv735.

[17] Jia Li et al. "Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding". In: *Nucleic Acids Research* 45.22 (2017), pp. 12877–12887. DOI: 10.1093/nar/gkx1145.

[18] Lin Yang et al. "TFBSshape: a motif database for DNA shape features of transcription factor binding sites". In: *Nucleic Acids Research* 42.D1 (2014), pp. D148–D155. DOI: 10.1093/nar/gkt1087.

[19] Anthony Mathelier et al. "DNA shape features improve transcription factor binding site predictions in vivo". In: *Cell Systems* 3.3 (2016), 278–286.e4. DOI: 10.1016/j.cels.2016.07.001.

[20] Wenxiu Ma et al. "DNA sequence + shape kernel enables alignment-free modeling of transcription factor binding". In: *Bioinformatics* 33.19 (2017), pp. 3003–3010. DOI: 10.1093/bioinformatics/btx336.

[21] Siguo Wang et al. "A new method combining DNA shape features to improve the prediction accuracy of transcription factor binding sites". In: *Intelligent Computing Theories and Application*. 2020, pp. 79–89. DOI: 10.1007/978-3-030-60802-6_8.

[22] Qinhu Zhang, Zhen Shen, and De-Shuang Huang. "Predicting in-vitro transcription factor binding sites using DNA sequence + shape". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.2 (2021), pp. 667–676. DOI: 10.1109/TCBB.2019.2947461.

[23] Sadia Tariq and Asad Amin. "DeepCTF: Transcription factor binding specificity prediction using DNA sequence plus shape in an attention-based deep learning model". In: *Signal, Image and Video Processing* 18 (2024), pp. 5239–5251. DOI: 10.1007/s11760-024-03229-7.

[24] Pengju Ding et al. "DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape". In: *Briefings in Bioinformatics* 24.4 (2023), bbad231. DOI: 10.1093/bib/bbad231.

[25] Jian Zhou and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature Methods* 12.10 (2015), pp. 931–934. DOI: 10.1038/nmeth.3547.

[26] David R. Kelley, Jasper Snoek, and John L. Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". In: *Genome Research* 26.7 (2016), pp. 990–999. DOI: 10.1101/gr.200535.115.

[27] Alvaro N. Barbeira et al. "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics". In: *Nature Communications* 9 (2018), p. 1825. DOI: 10.1038/s41467-018-03621-1.

[28] David R. Kelley et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks". In: *Genome Research* 28.5 (2018), pp. 739–750. DOI: 10.1101/gr.227819.117.

[29] Vikram Agarwal and Jay Shendure. "Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks". In: *Cell Reports* 31.7 (2020), p. 107663. DOI: 10.1016/j.celrep.2020.107663.

[30] Jian Zhou et al. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature Genetics* 50.8 (2018), pp. 1171–1179. DOI: 10.1038/s41588-018-0160-6.

[31] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature Methods* 18.10 (2021), pp. 1196–1203. DOI: 10.1038/s41592-021-01252-x.

[32] Johannes Linder et al. "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation". In: *bioRxiv* (2023). DOI: 10.1101/2023.08.30.555582.

[33] Žiga Avsec et al. "AlphaGenome: a unified model for genome function prediction". In: *Manuscript in preparation* (2025).

[34] Alireza Karbalayghareh, Merve Sahin, and Christina S. Leslie. "Chromatin interaction-aware gene regulatory modeling with graph attention networks". In: *Genome Research* 32.5 (2022), pp. 930–944. DOI: 10.1101/gr.275660.121.

[35] Sai Lin et al. *EPInformer: learning enhancer–promoter interactions from sequence and epigenomics for scalable gene expression prediction*. Preprint. 2024.

[36] Kelly Cochran et al. *ProCapNet: dissecting the sequence determinants of nascent transcription by modeling high-resolution promoter-proximal outputs*. Preprint. 2024.

[37] Adam He and Charles G. Danko. *CLIPNET: quantitative modeling of promoter-proximal transcription with deep learning*. Preprint. 2024.

[38] Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. "Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers". In: *Genome Biology* 24 (2023), p. 56. DOI: 10.1186/s13059-023-02899-9.

[39] Alexander Sasse et al. "Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings". In: *Nature Genetics* 55.12 (2023), pp. 2056–2064. DOI: 10.1038/s41588-023-01524-6.

[40] Ying Huang, Eric Sievert, Alexander Sasse, et al. "Personal transcriptome variation is poorly explained by current genomic deep learning models". In: *Nature Genetics* 55.12 (2023), pp. 2065–2073. DOI: 10.1038/s41588-023-01574-w.

[41] Philip L. De Jager et al. "A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research". In: *Scientific Data* 5 (2018), p. 180142. DOI: 10.1038/sdata.2018.142.

[42] Pooja Kathail, Aniket Bajwa, and Nilah M. Ioannidis. *Leveraging genomic deep learning models for non-coding variant effect prediction*. Review article. 2025.

[43] Robert Pearce et al. *TranscriptFormer: a generalist multi-species model of single-cell transcriptomics*. Preprint. 2025.

[44] Hemang Bajwa et al. *Reference genome bias in sequence-to-expression model benchmarking*. Preprint. 2023.

[45] Anika Rastogi et al. "Fine-tuning sequence-to-expression models on personal genome and transcriptome data". In: *bioRxiv* (2024). DOI: 10.1101/2024.09.23.614632.

[46] Shiron Drusinsky et al. *Variformer: Enformer fine-tuning for personal transcriptome prediction*. Manuscript. 2026.

[47] Anna E. Spiro et al. *SAGE-net: sequence-to-expression modeling via population-scale transfer learning*. Preprint. 2025.

[48] Carl Vaz et al. *Biobank-scale personal transcriptome prediction*. Preprint. 2025.

[49] Xinmin Tu et al. "Deep genomic models of allele-specific measurements". In: *bioRxiv* (2025). DOI: 10.1101/2025.04.09.648060.

[50] Z. Tu et al. *The modality gap in sequence-to-expression genomic models*. Preprint. 2026.

[51] Anna Korsakova, Divyanshi Srivastava, and David R. Kelley. "Shift augmentation improves DNA convolutional neural network indel effect predictions". In: *bioRxiv* (2025). DOI: 10.1101/2025.04.07.647656.

[52] Roger A. Hoskins et al. "CAGI, the Critical Assessment of Genome Interpretation, a community resource to evaluate predictions of genomic variant effects". In: *Human Mutation* 38.9 (2017), pp. 1089–1098. DOI: 10.1002/humu.23267.

[53] Kathleen M. Chen et al. "A sequence-based global map of regulatory activity for deciphering human genetics". In: *Nature Genetics* 54.7 (2022), pp. 940–949. DOI: 10.1038/s41588-022-01102-2.

[54] Gherman Novakovsky et al. "ExplaiNN: interpretable and transparent neural networks for genomics". In: *Genome Biology* 24 (2023), p. 154. DOI: 10.1186/s13059-023-02985-y.

[55] Kseniia Dudnyk et al. "Sequence basis of transcription initiation in the human genome". In: *Science* 384 (2024), eadj0116. DOI: 10.1126/science.adj0116.

[56] A. T. Balçı et al. "An intrinsically interpretable neural network architecture for sequence-to-function learning". In: *Bioinformatics* 39.Supplement_1 (2023), pp. i413–i422. DOI: 10.1093/bioinformatics/btad271.

[57] Md. Abul Hassan Samee. "Noncanonical binding of transcription factors: time to revisit specificity?" In: *Molecular Biology of the Cell* 34 (2023), pe4. DOI: 10.1091/mbc.E22-08-0325.

[58] Md. Abul Hassan Samee, Benoit G. Bruneau, and Katherine S. Pollard. "A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs". In: *Cell Systems* 8.1 (2019), 27–42.e6. DOI: 10.1016/j.cels.2018.12.001.

[59] Kathleen M. Chen et al. "Selene: a PyTorch-based deep learning library for sequence data". In: *Nature Methods* 16.4 (2019), pp. 315–318. DOI: 10.1038/s41592-019-0360-8.

[60] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 3145–3153.

[61] Eugenio Santelli and Timothy J. Richmond. "Crystal structure of MEF2A core bound to DNA at 1.5 Å resolution". In: *Journal of Molecular Biology* 297.2 (2000), pp. 437–449. DOI: 10.1006/jmbi.2000.3568.

[62] Jia Guo et al. "Sequence specificity incompletely defines the genome-wide occupancy of Myc". In: *Genome Biology* 15 (2014), p. 482. DOI: 10.1186/s13059-014-0482-3.

[63] Michael Allevato et al. "Sequence-specific DNA binding by MYC/MAX to low-affinity non-E-box motifs". In: *PLOS ONE* 12.7 (2017), e0180147. DOI: 10.1371/journal.pone.0180147.

[64] Avanti Shrikumar et al. *Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5*. arXiv:1811.00416. 2020. DOI: 10.48550/arXiv.1811.00416.

[65] Donal MacDonald et al. "Solution structure of an A-tract DNA bend". In: *Journal of Molecular Biology* 306.5 (2001), pp. 1081–1098. DOI: 10.1006/jmbi.2001.4447.

[66] Stephane E. Castel et al. "Rare variant phasing and haplotypic expression from RNA sequencing with phASER". In: *Nature Communications* 7 (2016), p. 12817. DOI: 10.1038/ncomms12817.

[67] Žiga Avsec et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature Genetics* 53.3 (2021), pp. 354–366. DOI: 10.1038/s41588-021-00782-6.

Appendix A

Appendix of Chapter 2

A.1 Supplementary Figures

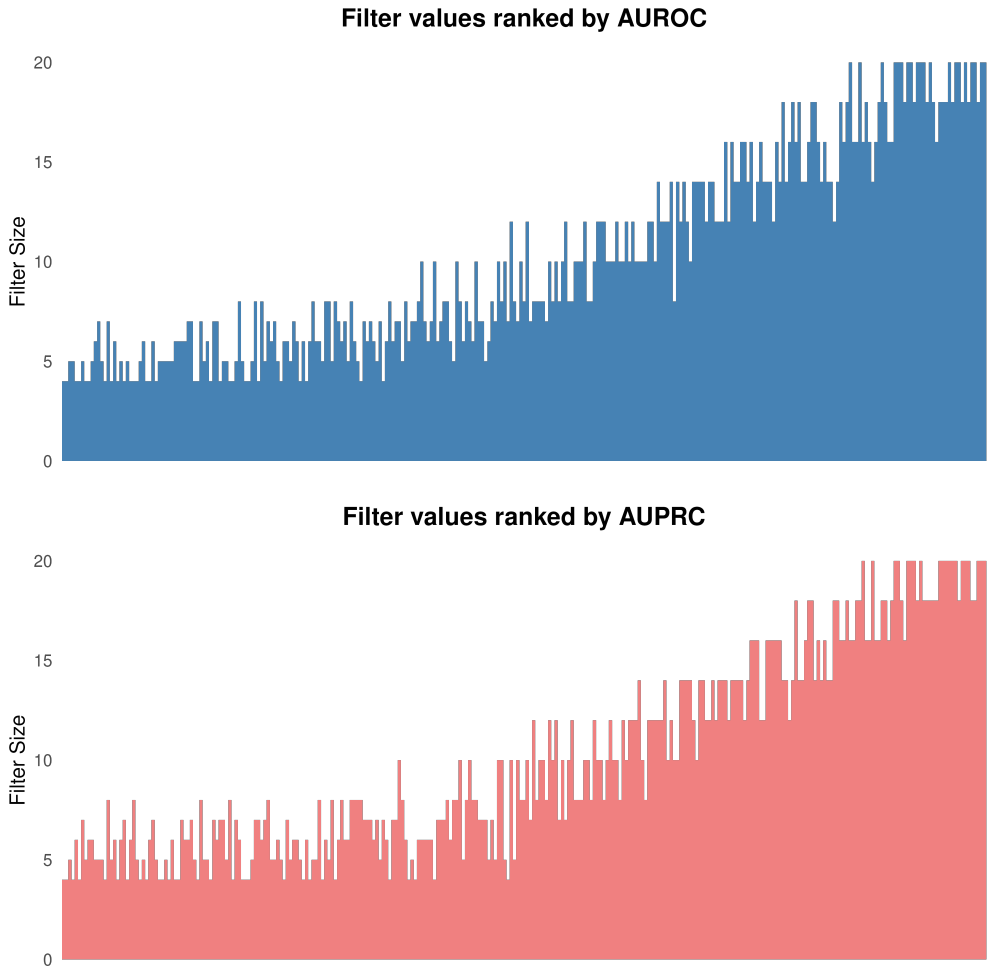

Figure A.1: **Hyperparameter search for shape-only inputs.** Performance trends from a hyperparameter search on a version of DeepShape that used only DNA shape features as input. Filter sizes and number of filters per layer were tested to optimize shape-feature capture for predicting TF binding. Smaller filter sizes demonstrated a trend toward improved performance. Filter sizes ranked by AUROC (top) and AUPRC (bottom).

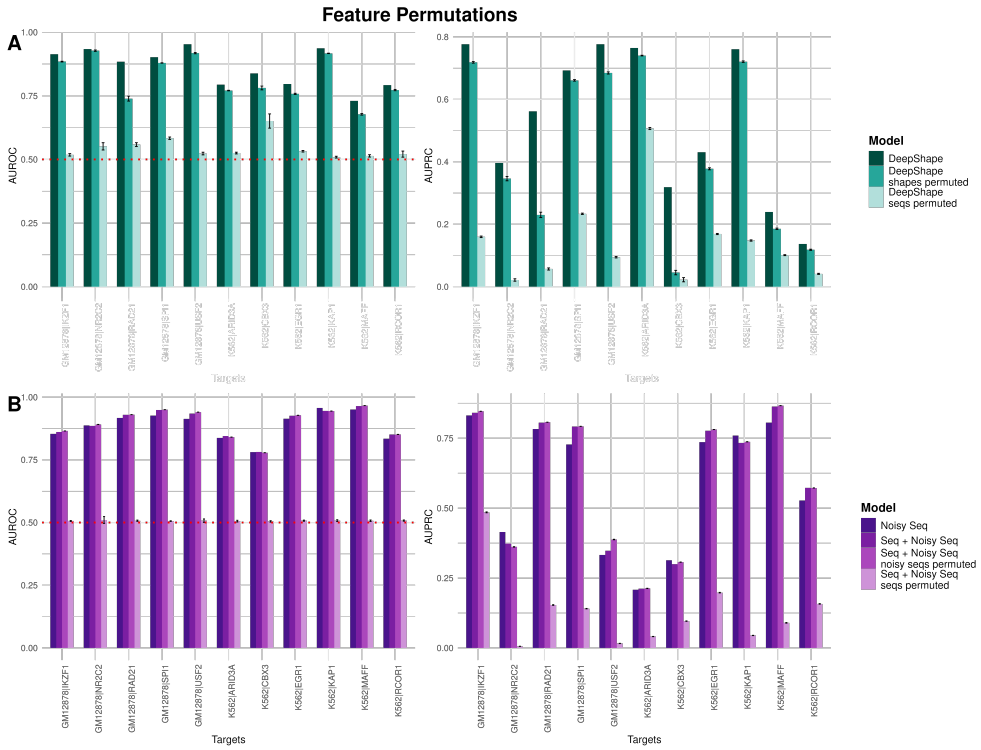

Figure A.2: **Permutation analyses with original hyperparameters and noisy-sequence controls.** A) AUROC (left) and AUPRC (right) for 11 representative TFs based on a trained DeepShape model versus models with permuted shape or sequence features (bar colors). Permutations were performed with the same hyperparameters as the published DeeperDeepSEA model. B) Models in which shape inputs were replaced with Gaussian-noised sequence: *Seq + Noisy Seq* (sequence plus noised sequence) and *Noisy Seq* (noised sequence only). Permuting channels shows that performance in *Seq + Noisy Seq* depends on the sequence input, with little contribution from the noised channel.

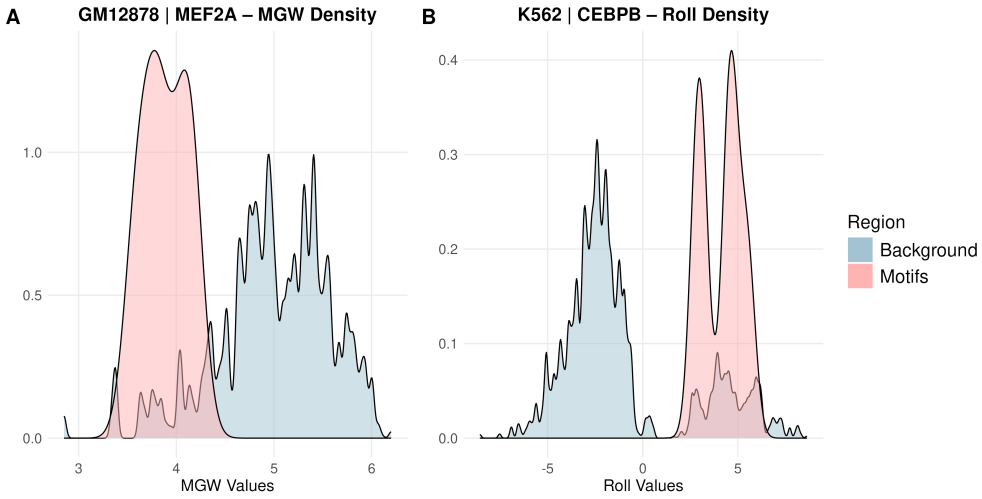

Figure A.3: **Target-specific shape distributions.** A) Minor groove width (MGW) distributions for GM12878 | MEF2A, showing a preference for narrower minor groove widths relative to background. B) Roll distributions for K562 | CEBPB, showing a shift toward higher Roll values in motif regions compared to background.

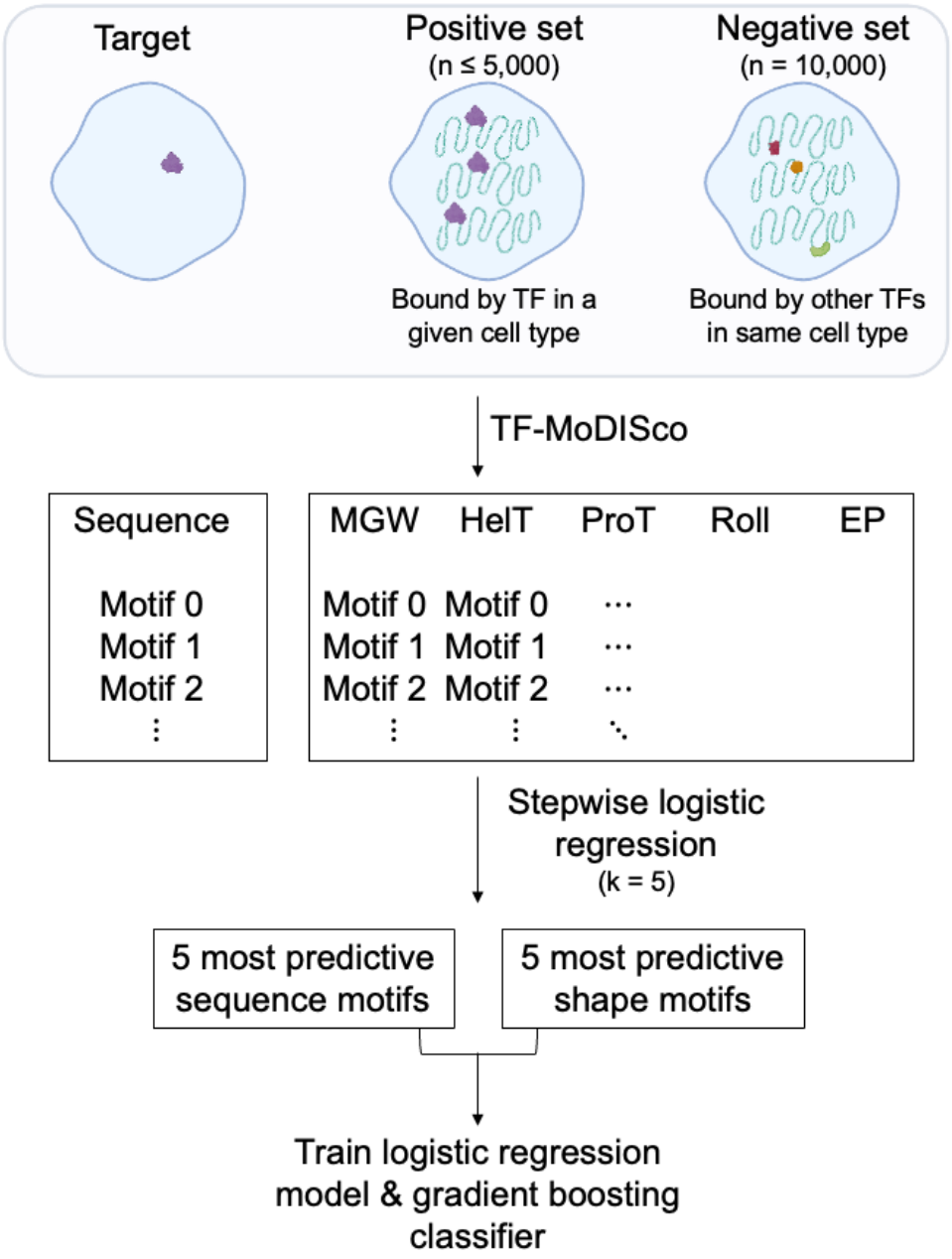

Figure A.4: **Pipeline for identifying predictive sequence and shape motifs.** Positive sets included sequences bound by the TF of interest, while negative sets included sequences bound by other TFs in the same cell type. TF-MoDISco extracted sequence and shape (MGW, HelT, ProT, Roll, EP) motifs from DeepLIFT attribution scores. The most predictive motifs were ranked using stepwise logistic regression, with the top five sequence and shape motifs used to train logistic regression and gradient boosting models, highlighting the contributions of sequence and shape features to TF binding predictions.

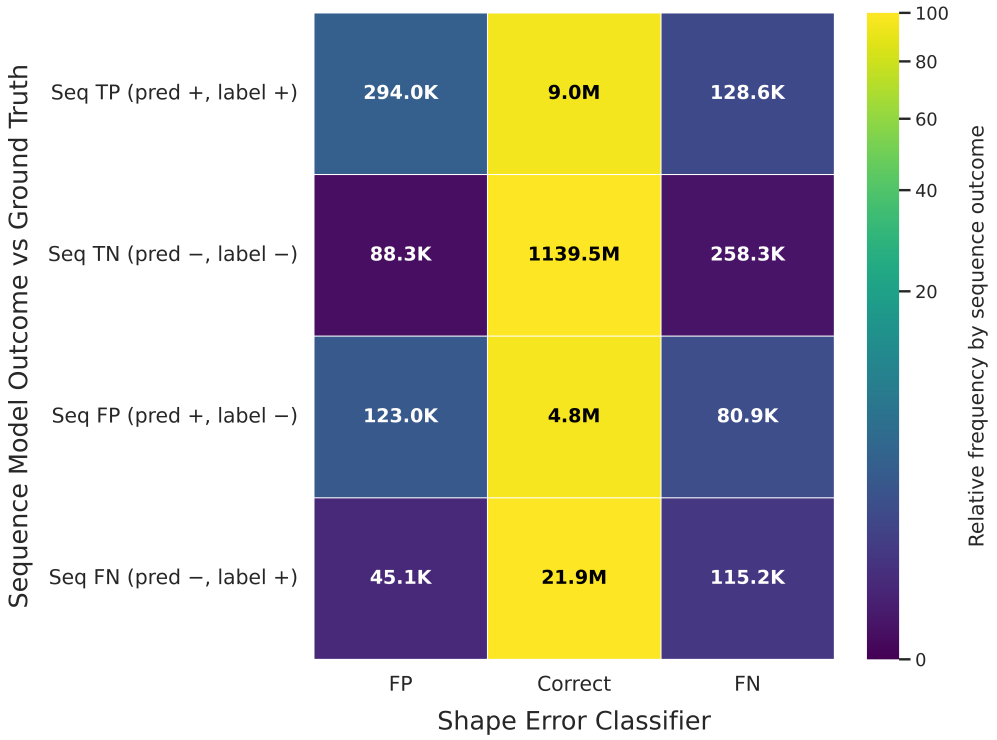

Figure A.5: **Aggregate confusion matrix for the shape error classifier.** Aggregate confusion matrix for the shape error classifier across all tracks.

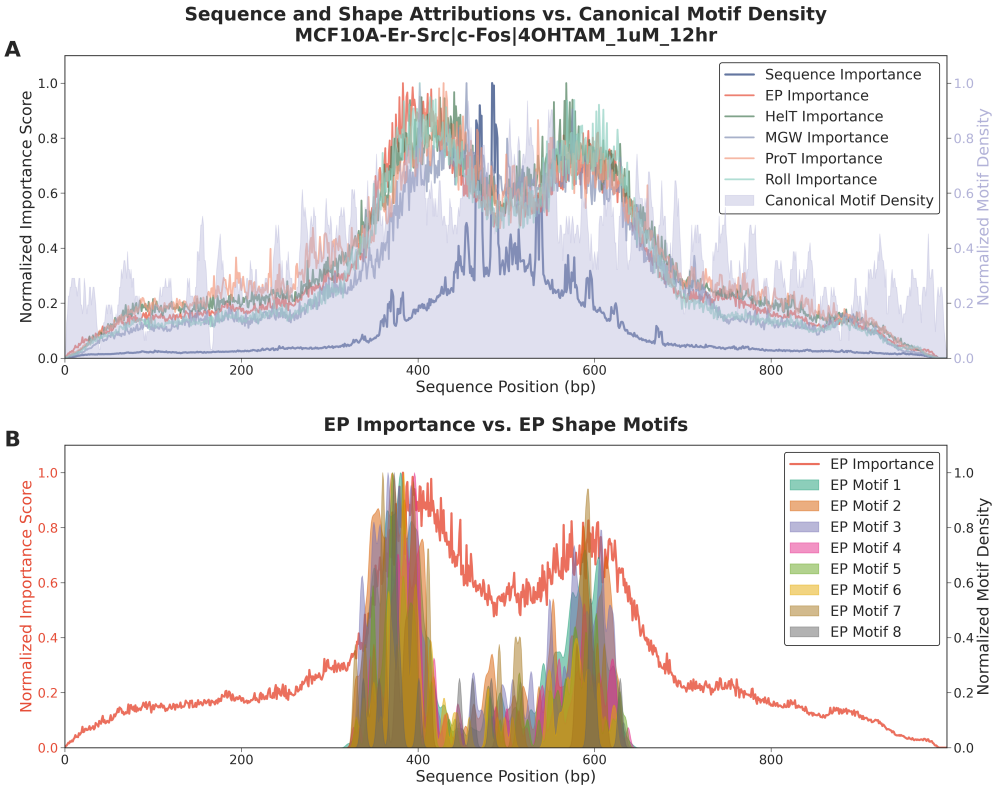

Figure A.6: **Shape-only error classification analysis for c-Fos.** A) DeepLIFT attribution profiles for sequences that are false negatives of the sequence-only model (ground-truth bound, predicted unbound). The sequence attribution (blue, left axis) shows a sharp central peak; the five DNA-shape attributions from the shape-only error classifier (left axis) are broader with elevated signal in flanking regions. Canonical AP-1 (c-Fos) PWM motif density (right axis) is distributed across the window rather than tightly centered. B) Electrostatic-potential (EP) attributions on the same sequences exhibit two prominent peaks that co-locate with EP shape-motif densities (right axis) obtained by running TF-MoDISco on EP attributions.

A.2 Supplementary Tables

Table A.1: **High-attribution positions per motif type.** Average number of high-attribution positions per motif for sequence features and each DNA shape feature. Sequence motifs consistently exhibit a greater number of high-attribution positions.

Motif type	**Avg. no. high-attribution positions per motif**
Sequence	5.46
MGW	3.02
HelT	3.35
ProT	3.19
Roll	3.37
EP	3.72

Table A.2: **Co-occurring versus non-co-occurring strong sequence motif occurrences in HepG2 | MafK.** Co-occurring sequence motifs exhibit significantly higher ChIP-seq enrichment scores compared to non-co-occurring motifs.

Metric	**Value**
No. co-occurring strong sequence occurrences	353
No. non-co-occurring strong sequence occurrences	170
Mean ChIP-seq score of co-occurrences	157.5
Mean ChIP-seq score of non-co-occurrences	129.8
One-sided p-value	8.7×10^{-4}