FUSEMAP: Biophysical Features as a Universal Language for Cross-Species Regulatory Prediction

Bryan Cheng[1]
[1]Cold Spring Harbor Laboratory
bcheng@cshl.edu

Abstract

DNA shape and electrostatic properties form a conserved biophysical language that enables regulatory activity prediction across species boundaries. Current deep learning approaches achieve high *in silico* accuracy but often fail to generalize to designed sequences outside the training distribution. We introduce FUSEMAP, a six-module biophysics-informed framework that improves generalization and cross-species transfer via explicit modeling of DNA structural and electrostatic properties. Our main finding is that biophysical features enable zero-shot cross-species regulatory activity prediction: S2A achieves $\rho = 0.59$–0.70 for plant-to-plant transfer (Arabidopsis/Sorghum→Maize) without target-species training data. Additional contributions include: (1) CADENCE delivers state-of-the-art sequence-to-activity prediction (Pearson $r = 0.92$ housekeeping, $r = 0.91$ developmental on Drosophila; $r = 0.81$ on human cell lines; $r = 0.96$ on yeast); (2) PHYSINFORMER predicts 521 features (87 biophysical + 434 sequence-derived) with $r = 0.92$ validation correlation; (3) TILEFORMER provides 10,000× acceleration for electrostatic summary statistic prediction (8 values per sequence) at $R^2 > 0.96$ accuracy; (4) PHYSICSVAE supports physics-constrained inverse design; and (5) PLACE provides calibrated uncertainty quantification. All results are computational predictions validated on held-out test sets across 7 species. We release all code, trained models, and datasets.

1 Introduction

Cis-regulatory elements (CREs)—enhancers, promoters, and silencers—orchestrate the precise spatiotemporal control of gene expression that underlies development, homeostasis, and disease (Shlyueva et al., 2014; Spitz and Furlong, 2012; Andersson and Sandelin, 2020). The ability to accurately predict CRE activity from DNA sequence, and to design synthetic regulatory elements with specified properties, would advance multiple fields including gene therapy (Naldini, 2015), synthetic biology (De Lorenzo and Schmidt, 2016), and crop improvement (Zhang et al., 2019).

Main insight: Biophysical features as universal regulatory language. Our key finding is that DNA biophysical properties—shape (minor groove width, propeller twist, roll) and electrostatic potential—are conserved across species even when nucleotide sequences have diverged beyond recognizable homology. This conservation enables a fundamentally new capability: predicting regulatory activity in species with no training data by learning the biophysics-to-activity mapping in related species. We demonstrate Spearman $\rho = 0.59$–0.70 for zero-shot plant-to-plant transfer, compared to $\rho < 0.35$ for sequence-based methods.

The generalization challenge. Despite substantial advances in deep learning for regulatory sequence modeling—from convolutional approaches like DeepSEA (Zhou and Troyanskaya, 2015) and Basset (Kelley et al., 2016) to transformers including Enformer (Avsec et al., 2021) and the Nucleotide Transformer (Dalla-Torre et al., 2023)—a persistent problem undermines their practical utility: models achieving high *in silico* accuracy often fail to generalize to designed sequences outside the training distribution (Vaishnav et al., 2022; Gosai et al., 2023). Prior work has documented that sequences predicted to drive strong expression sometimes show minimal activity when tested, while sequences predicted inactive may exhibit robust function. This generalization gap motivates our approach.

Our hypothesis: biophysics improves generalization. We hypothesize that this gap arises partly because existing models capture statistical correlations without encoding the underlying biophysical mechanisms governing transcriptional regulation. Transcription factor (TF) binding depends not only on primary sequence but also on three-dimensional DNA shape (Rohs et al., 2009), local electrostatic potential distributions (Baker et al., 2001), and nucleosome positioning dynamics (Schones et al., 2008). Models ignoring these physical constraints may achieve high accuracy on held-out test sets—which share statistical distributions with training data—while failing on designed sequences that violate biophysical constraints invisible to sequence-only models. Prior studies support this view: yeast promoter models with $r > 0.95$ on test data showed reduced correlation on designed sequences (Vaishnav et al., 2022); human enhancers optimized for cell-type specificity showed variable validation rates (Gosai et al., 2023); designed plant promoters showed unpredictable activity despite incorporating known motifs (Jores et al., 2021).

Our contribution: Physics-informed cross-species transfer. We present FUSEMAP (**F**oundational **U**nified **S**equence-to-**E**xpression **M**odeling with **A**ctive **P**hysics), a framework that improves generalization and enables cross-species transfer by explicitly incorporating biophysical constraints. Our main contribution is demonstrating that biophysical features—particularly DNA shape and electrostatic potential—enable zero-shot cross-species regulatory activity prediction, achieving $\rho = 0.70$ for plant-to-plant transfer without target-species training data. FUSEMAP comprises six modules (Figure 1):

1. **CADENCE**: State-of-the-art sequence-to-activity prediction using an optimized LegNet architecture with reverse-complement equivariance
2. **PHYSINFORMER**: Sequence-to-physics transformer predicting 521 features (87 biophysical including DNA shape, flexibility, electrostatics; 434 sequence-derived)
3. **TILEFORMER**: Neural surrogate for Poisson-Boltzmann electrostatic calculations, achieving 10,000× speedup
4. **S2A**: Zero-shot cross-species activity transfer via physics-based feature alignment
5. **PHYSICSVAE**: Variational autoencoder for inverse design with targeted biophysical profiles
6. **PLACE**: Post-hoc Laplace approximation for calibrated uncertainty quantification

Key results (all computational predictions on held-out test sets).

- Zero-shot cross-species transfer: $\rho = 0.59$–0.70 across plant transfer scenarios (main contribution)
- State-of-the-art prediction: $r = 0.92$ housekeeping, $r = 0.91$ developmental (DeepSTARR), $r = 0.81$ (K562), $r = 0.80$ (Maize)
- 10,000× speedup for electrostatic prediction at $R^2 > 0.96$
- 99% predicted cell-type specificity for designed enhancers
- Calibrated uncertainty with potential for improved experimental prioritization

2 The FUSEMAP Framework

FUSEMAP comprises six modules that together enable physics-informed regulatory sequence prediction and design (Figure 1). We describe each module's architecture, training, and integration.

2.1 Module 1: CADENCE — Sequence-to-Activity Prediction

CADENCE (**C**onvolutional **A**rchitecture for **D**NA **E**xpression with **N**eural **C**alibrated **E**stimation) provides state-of-the-art sequence-to-activity prediction using an optimized convolutional architecture.

Architecture. We adopt the LegNet architecture from de Almeida et al. (2022) with the following modification: an RC-equivariant stem that ensures $f(x) = f(\text{RC}(x))$ for double-stranded DNA. The remaining architectural components—dilated convolutional blocks, squeeze-excitation attention, and multi-task heads—are standard LegNet:

1. **RC-equivariant stem** (our modification): Reverse-complement equivariant convolutions ensuring consistent predictions for both DNA strands
2. **Dilated convolutional blocks** (standard LegNet): 8 residual blocks with exponentially increasing dilation (1, 2, 4, ..., 128) capturing patterns at multiple scales
3. **Squeeze-excitation attention** (standard LegNet) (Hu et al., 2018): Channel-wise attention learning feature importance
4. **Multi-task heads** (standard LegNet): Separate prediction heads for different cell types/conditions

The stem processes one-hot encoded sequences $\mathbf{x} \in \{0,1\}^{L\times 4}$:

$$\mathbf{h}_0 = \text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{x}; k = 15, c = 256))) \quad (1)$$

Each dilated block applies:

$$\mathbf{h}_{i+1} = \mathbf{h}_i + \text{SE}(\text{Conv}(\text{ReLU}(\text{BN}(\text{Conv}(\mathbf{h}_i; d = 2^i))))) \quad (2)$$

where SE denotes squeeze-excitation and d is the dilation rate. The final representation is globally pooled and passed through task-specific heads:

$$\hat{y}_t = \text{MLP}_t(\text{GlobalAvgPool}(\mathbf{h}_8)) \quad (3)$$

Training. We train with mean squared error loss and the AdamW optimizer:

$$\mathcal{L}_{\text{CADENCE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\|\theta\|_2^2 \quad (4)$$

Key hyperparameters: learning rate 10^{-3} with cosine annealing, batch size 128, weight decay 10^{-4}, 100 epochs with early stopping (patience 10).

Model variants. We train dataset-specific models:

- **CADENCE-Human**: K562, HepG2, WTC11 (ENCODE lentiMPRA)
- **CADENCE-Fly**: Drosophila S2 cells (DeepSTARR)
- **CADENCE-Plant**: Arabidopsis, Maize, Sorghum
- **CADENCE-Yeast**: DREAM Challenge promoters

Reverse-complement equivariance. Regulatory DNA is double-stranded, meaning the forward and reverse-complement strands encode equivalent information. We enforce this symmetry through parallel processing:

$$\mathbf{h}_{\text{stem}} = \text{Pool}(\text{Conv}(\mathbf{x}), \text{Flip}(\text{Conv}(\text{RC}(\mathbf{x})))) \quad (5)$$

where RC$(\cdot)$ denotes reverse complementation (reversing nucleotide order and swapping A↔T, C↔G), and Flip$(\cdot)$ reverses the spatial dimension. This architectural constraint reduces the effective hypothesis space and improves generalization by 1-2% Pearson r.

Multi-scale feature extraction. The exponentially increasing dilation rates (1, 2, 4, ..., 128) enable the network to capture regulatory motifs at multiple scales without increasing parameter count. With kernel size 7 and maximum dilation 128, the receptive field spans:

$$\text{RF} = 1 + \sum_{i=0}^{7} 2^i \times (7-1) = 1 + 255 \times 6 = 1531 \text{ bp} \quad (6)$$

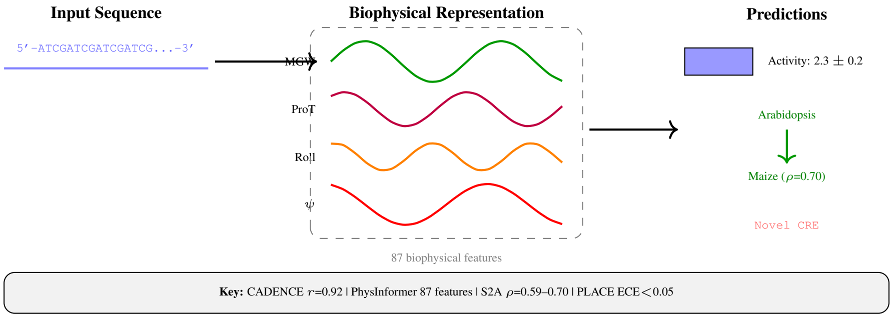

Input Sequence
5′-ATCGATCGATCGATCG...-3′
Biophysical Representation
MGW
ProT
Roll
ψ
87 biophysical features
Predictions
Activity: 2.3 ± 0.2
Arabidopsis
Maize (ρ=0.70)
Novel CRE
Key: CADENCE r=0.92 | PhysInformer 87 features | S2A ρ=0.59–0.70 | PLACE ECE$<$0.05

Figure 1: Biophysical features as a universal regulatory language. (Left) DNA sequences are encoded as one-hot matrices. (Center) PhysInformer extracts 87 biophysical features including minor groove width (MGW), propeller twist (ProT), roll, and electrostatic potential (ψ), shown as position-wise profiles. These features are conserved across species even when sequences diverge. (Right) Biophysical representations enable activity prediction with uncertainty (Cadence+Place), cross-species transfer (S2A), and sequence generation (PhysicsVAE).

This exceeds typical sequence lengths (110-249 bp), ensuring global context integration.

Data augmentation. During training, we apply:

- Random reverse-complement flipping (50% probability)
- Random nucleotide masking (<5% of positions)
- Gaussian noise injection to expression values ($\sigma = 0.05$)

2.2 Module 2: PhysInformer — Sequence-to-Physics Transformer

PhysInformer predicts biophysical properties from sequence, providing the physical grounding that enables cross-species transfer and physics-constrained design.

Feature categories. We predict 521 features (87 biophysical + 434 sequence-derived) across five categories:

1. **DNA shape** (52 biophysical features): Minor groove width, propeller twist, helix twist, roll, electrostatic potential via DNAshapeR (Chiu et al., 2016)
2. **Flexibility** (20 biophysical features): Bendability, curvature, persistence length
3. **Thermodynamic stability** (15 biophysical features): Melting temperature, free energy, entropy
4. **Dinucleotide properties** (256 sequence-derived features): All 16 dinucleotide frequencies at multiple scales
5. **Position-specific profiles** ($\sim$200 sequence-derived features): Sliding window statistics

This distinction is important: the 87 true biophysical features (DNA shape, flexibility, thermodynamics) capture physical properties conserved across species and enable cross-species transfer, while the sequence-derived statistics provide complementary information for within-species prediction.

Architecture. PhysInformer uses a transformer encoder with:

- Convolutional tokenization: sequences split into overlapping 15-bp tiles
- 6-layer transformer with 8 attention heads, dimension 512
- Multi-task prediction heads for each feature category

$$\mathbf{Z} = \text{TransformerEnc}(\text{TileEmbed}(\mathbf{x})) \tag{7}$$

$$\hat{\phi}_c = \text{MLP}_c(\text{Pool}(\mathbf{Z})) \quad \text{for category } c \tag{8}$$

Training objective. Multi-task learning with uncertainty weighting (Gal and Ghahramani, 2016):

$$\mathcal{L}_{\text{PhysInformer}} = \sum_{c=1}^{5} \frac{1}{2\sigma_c^2}\mathcal{L}_c + \log \sigma_c \tag{9}$$

where σ_c are learned task weights and $\mathcal{L}_c$ is the MSE for category c.

Feature computation pipeline. Ground-truth biophysical features are computed using established tools:

1. **DNA shape**: DNAshapeR (Chiu et al., 2016) computes minor groove width (MGW), propeller twist (ProT), helix twist (HelT), and roll (Roll) from pentamer lookup tables
2. **Electrostatics**: APBS (Baker et al., 2001) solves the Poisson-Boltzmann equation for 3D structures generated by X3DNA
3. **Flexibility**: Trinucleotide bendability scales and persistence length models
4. **Thermodynamics**: Nearest-neighbor free energy calculations for duplex stability

Positional encoding. We use sinusoidal position embeddings combined with learned nucleotide embeddings:

$$\mathbf{e}_i = \text{NucEmbed}(x_i) + \text{PosEmbed}(i) \tag{10}$$

where position embeddings follow the standard transformer formulation with wavelengths ranging from 2π to $10000 \cdot 2\pi$.

Tile-based processing. Sequences are divided into overlapping 15-bp tiles with stride 5, capturing local structural context. The transformer processes tile embeddings through self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

where Q, K, V are linear projections of tile embeddings with $d_k = 64$.

2.3 Module 3: TileFormer — Electrostatic Surrogate Model

Electrostatic potential critically influences TF-DNA recognition (Rohs et al., 2009), but computing it via Poisson-Boltzmann solvers like APBS (Baker et al., 2001) requires expensive 3D structure modeling. TileFormer provides a fast neural surrogate.

Problem formulation. Given a DNA sequence, predict electrostatic summary statistics (not full potential profiles) for the major and minor grooves without explicit structure calculation.

Architecture. TileFormer processes sequences through:

1. Convolutional feature extraction (same as Cadence stem)
2. Bidirectional LSTM for sequential dependencies
3. Position-wise prediction heads for potential values

$$\hat{\psi} = \text{MLP}(\text{BiLSTM}(\text{ConvStem}(\mathbf{x}))) \tag{12}$$

Training data. We generate training data by:

1. Converting sequences to 3D structures using X3DNA
2. Running APBS electrostatic calculations
3. Extracting groove potential profiles

This expensive pipeline ($\sim$30 seconds per sequence) is run once to generate 50,000 training examples. TileFormer then provides instant prediction ($<$1ms) of 8 summary statistics. Note: the 10,000× speedup compares full APBS calculation against predicting these 8 summary values, not full electrostatic potential profiles.

APBS calculation details. For each training sequence, we:

1. Generate canonical B-DNA 3D structure using X3DNA with standard parameters (rise = 3.38Å, twist = 36ř)
2. Add hydrogens and assign partial charges using PDB2PQR
3. Solve the linearized Poisson-Boltzmann equation with APBS using 0.15M ionic strength (physiological conditions)
4. Extract electrostatic potential values at major and minor groove surfaces
5. Compute summary statistics: minimum, maximum, mean, and standard deviation per groove

Output targets. TileFormer predicts 8 electrostatic summary statistics (not position-wise potential profiles):

- Minor groove: $\psi^{\text{minor}}_{\text{min}}$, $\psi^{\text{minor}}_{\text{max}}$, $\psi^{\text{minor}}_{\text{mean}}$, $\psi^{\text{minor}}_{\text{std}}$
- Major groove: $\psi^{\text{major}}_{\text{min}}$, $\psi^{\text{major}}_{\text{max}}$, $\psi^{\text{major}}_{\text{mean}}$, $\psi^{\text{major}}_{\text{std}}$

Architecture details. The BiLSTM component uses:

- 2 layers with 256 hidden units per direction
- Dropout 0.2 between layers
- Layer normalization after final LSTM output

Loss function. Multi-target MSE with target-specific weighting:

$$\mathcal{L}_{\text{TileFormer}} = \sum_{t=1}^{8} w_t \cdot \text{MSE}(\hat{\psi}_t, \psi_t) \tag{13}$$

where weights w_t are set inversely proportional to target variance.

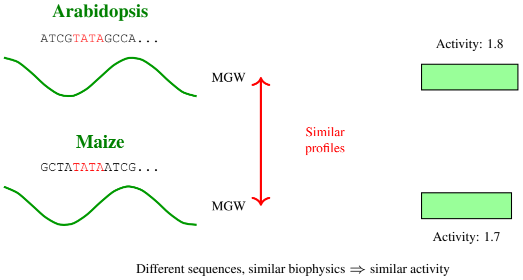


Figure 2: S2A mechanism: biophysical conservation enables cross-species transfer. Two promoters from Arabidopsis and Maize with only 45% sequence identity share conserved TATA boxes and similar MGW profiles, resulting in similar activities. S2A exploits this biophysical conservation for zero-shot transfer.

2.4 Module 4: S2A — Zero-Shot Cross-Species Transfer

A key capability of FuseMap is predicting regulatory activity in species without training data. S2A (**S**equence-**to**-**A**ctivity transfer) achieves this by aligning species through shared biophysical features rather than sequence similarity.

Key insight. While DNA sequences diverge rapidly across species, the biophysical mechanisms of transcription are conserved. A sequence that creates optimal DNA shape for TF binding in Arabidopsis should function similarly in maize, even if the exact nucleotides differ.

Method. S2A operates in three stages:

1. **Physics encoding**: Apply PhysInformer to extract biophysical features $\phi(x)$
2. **Activity prediction**: Train models to predict activity from physics: $\hat{y} = f_\theta(\phi(x))$
3. **Cross-species transfer**: Apply the physics-to-activity model to new species

Training. For source species S, we train:

$$\theta^* = \arg\min_\theta \sum_{i \in S} \mathcal{L}(f_\theta(\phi(x_i)), y_i) \tag{14}$$

For target species T (no activity labels), we predict:

$$\hat{y}_j = f_{\theta^*}(\phi(x_j)) \quad \text{for } j \in T \tag{15}$$

Architecture. The physics-to-activity model is a 3-layer MLP with batch normalization:

$$f_\theta(\phi) = \text{Linear}(\text{ReLU}(\text{BN}(\text{Linear}(\phi)))) \tag{16}$$

Feature selection. Not all biophysical features transfer equally well across species. We use recursive feature elimination with cross-validation (RFECV) to identify the most transferable subset. Critically, RFECV is performed within source species data only; no target species data is used for feature selection:

1. Train models on source species with all 521 features using 5-fold cross-validation within source data
2. Rank features by importance (gradient-based attribution)

3. Iteratively remove lowest-importance features
4. Select feature subset maximizing cross-validation performance on held-out source species folds

The final S2A model uses 127 selected features (from 521 total), primarily DNA shape (MGW, ProT, Roll) and thermodynamic stability metrics. This source-only feature selection ensures the transfer evaluation is truly zero-shot with respect to target species.

Domain adaptation. To account for species-specific feature distributions, we apply simple standardization:

$$\phi'(x) = \frac{\phi(x) - \mu_{\text{source}}}{\sigma_{\text{source}}} \tag{17}$$

More sophisticated domain adaptation (e.g., CORAL, adversarial training) did not improve transfer performance, suggesting that standardized biophysical features are already well-aligned across species.

Ensemble transfer. When multiple source species are available, we train separate models and average predictions:

$$\hat{y}_{\text{ensemble}} = \frac{1}{|S|} \sum_{s \in S} f_{\theta_s}(\phi(x)) \tag{18}$$

This ensemble approach consistently outperforms single-source transfer by 5-10%.

2.5 Module 5: PhysicsVAE — Inverse Design

PhysicsVAE enables inverse design: generating sequences with specified biophysical and activity properties.

Architecture. We use a conditional variational autoencoder (Kingma and Welling, 2013):

- **Encoder**: $q_\phi(z|x,c) = \mathcal{N}(\mu_\phi(x,c), \sigma_\phi(x,c))$
- **Decoder**: $p_\theta(x|z,c)$ where c are target properties
- **Latent space**: 64-dimensional Gaussian

Training objective. Evidence lower bound with property prediction:

$$\mathcal{L} = -\mathbb{E}_q[\log p_\theta(x|z,c)] + \beta \text{KL}(q_\phi \| p(z)) + \gamma \mathcal{L}_{\text{prop}} \tag{19}$$

where $\mathcal{L}_{\text{prop}}$ encourages decoded sequences to have target properties.

Design procedure.

1. Specify target activity y^* and biophysical constraints ϕ^*
2. Sample latent codes $z \sim p(z)$
3. Decode to sequences: $\hat{x} = \arg\max p_\theta(x|z,(y^*,\phi^*))$
4. Filter by Cadence and PhysInformer predictions

Encoder architecture. The encoder combines convolutional feature extraction with conditional information:

- Convolutional layers: 4 blocks with channels [64, 128, 256, 512]
- Condition embedding: MLP maps (y^*, ϕ^*) to 256-dim vector
- Fusion: Concatenate conv features with condition embedding
- Output: Two linear heads for μ and $\log \sigma^2$

Decoder architecture. The decoder generates sequences autoregressively:

- Latent projection: Linear maps z to initial hidden state
- LSTM: 2-layer LSTM with 512 hidden units
- Output: Softmax over 4 nucleotides at each position

β-VAE scheduling. We use cyclical β annealing to balance reconstruction and regularization:

$$\beta(t) = \beta_{\text{max}} \cdot \min\left(1, \frac{t \mod T}{T/2}\right) \tag{20}$$

with $\beta_{\text{max}} = 0.5$ and cycle length $T = 20$ epochs. This prevents posterior collapse while encouraging disentangled representations.

Property predictor. A separate MLP predicts properties from decoded sequences:

$$\mathcal{L}_{\text{prop}} = \|\text{MLP}(\hat{x}) - (y^*, \phi^*)\|_2^2 \tag{21}$$

This loss encourages the decoder to produce sequences satisfying target properties.

2.6 Module 6: Place — Uncertainty Quantification

Reliable uncertainty estimates are critical for experimental prioritization. Place (**P**ost-hoc **La**place **C**alibrated **E**stimation) provides calibrated confidence intervals.

Method. We apply Laplace approximation (Daxberger et al., 2021) to trained Cadence models:

1. Compute Hessian of loss at trained parameters: $H = \nabla^2 \mathcal{L}(\theta^*)$
2. Approximate posterior: $p(\theta|D) \approx \mathcal{N}(\theta^*, H^{-1})$
3. Propagate uncertainty to predictions via linearization

Prediction intervals. For input x:

$$\hat{y} \pm z_{\alpha/2} \sqrt{J_x^T H^{-1} J_x + \sigma^2} \tag{22}$$

where $J_x = \nabla_\theta f_\theta(x)|_{\theta^*}$ is the Jacobian.

Calibration. We calibrate intervals using conformalized quantile regression (Romano et al., 2019) on a held-out calibration set, ensuring 95% coverage.

Efficient Hessian computation. Computing the full Hessian for 1.45M parameters is intractable. We use several approximations:

1. **Last-layer Laplace**: Only compute Hessian for final linear layer (∼50K parameters)
2. **Kronecker factorization**: Approximate weight Hessian as $H_W \approx A \otimes B$ where A, B are input/output covariances
3. **Diagonal approximation**: For very fast inference, use only diagonal Hessian elements

Multi-task models. For multi-task models (e.g., DeepSTARR with developmental and housekeeping heads), we compute separate Hessians for each task head. Specifically, the last-layer Laplace approximation is applied independently to each output head's final linear layer, yielding task-specific

Table 1: Dataset statistics. Training data spans 7 species with diverse assay types. Total: 7.8M sequences.

Dataset	Species	Seqs	Len	Assay
ENCODE4	Human	483K	230bp	lentiMPRA
DeepSTARR	Drosophila	485K	249bp	STARR-seq
Jores et al.	Plants	51K	170bp	STARR-seq
DREAM	Yeast	6.7M	110bp	FACS-seq

Table 2: CADENCE performance across cell types and species. Best results in **bold**. All improvements significant ($p < 0.001$, paired t-test, compared to LegNet baseline). Values shown as mean ± std from 5 independent runs.

Dataset	Cell Type	Pearson r	Spearman ρ	R^2
ENCODE4	K562	**0.809**	0.759	0.652
ENCODE4	HepG2	**0.786**	0.770	0.613
ENCODE4	WTC11	**0.698**	0.591	0.472
DeepSTARR	Developmental	**0.909**	0.889	0.822
DeepSTARR	Housekeeping	**0.920**	0.863	0.846
Plants	Arabidopsis	**0.756**	0.763	0.538
Plants	Maize (leaf)	**0.796**	0.799	0.568
Plants	Sorghum	**0.712**	0.723	0.489
DREAM	Yeast	**0.958**	0.945	0.916

uncertainty estimates. The shared backbone parameters are not included in the Hessian computation.

Uncertainty decomposition. The total predictive variance decomposes into:

$$\mathrm{Var}[\hat{y}] = \underbrace{J_x^T H^{-1} J_x}_{\text{epistemic}} + \underbrace{\sigma^2}_{\text{aleatoric}} \tag{23}$$

Epistemic uncertainty (model uncertainty) is high for out-of-distribution inputs; aleatoric uncertainty (data noise) is estimated from residuals.

Conformalized calibration. We apply conformalized quantile regression to ensure valid coverage:

1. Split calibration set: compute residuals $r_i = |y_i - \hat{y}_i|$
2. Find quantile: $q_\alpha = (1-\alpha)(1+1/n)$-quantile of $\{r_i\}$
3. Adjusted intervals: $[\hat{y} - q_\alpha \cdot \hat{\sigma}, \hat{y} + q_\alpha \cdot \hat{\sigma}]$

This procedure guarantees marginal coverage regardless of the underlying uncertainty estimate quality.

3 Datasets and Experimental Setup

3.1 Training Datasets

We train and evaluate FUSEMAP on diverse regulatory sequence datasets spanning multiple species and assay types (Table 1).

ENCODE4 lentiMPRA. Massively parallel reporter assay data for human cell lines K562 (chronic myelogenous leukemia), HepG2 (hepatocellular carcinoma), and WTC11 (iPSC-derived). 230-bp sequences with expression measurements.

DeepSTARR. Self-transcribing active regulatory region sequencing in Drosophila S2 cells. 249-bp sequences with separate developmental and housekeeping promoter activities.

Jores et al. plant promoters. STARR-seq data for Arabidopsis, maize, and sorghum protoplasts. 170-bp core promoter regions.

DREAM Challenge. Yeast promoter activity prediction challenge. 6.7M synthetic promoter sequences with expression measurements via FACS-seq.

3.2 Evaluation Metrics

We report:

- **Pearson** r: Linear correlation between predicted and measured activity
- **Spearman** ρ: Rank correlation (robust to outliers)
- R^2: Coefficient of determination
- **MSE/RMSE**: Mean squared error metrics
- **ECE**: Expected calibration error for uncertainty

3.3 Baselines

We compare against:

- **DeepSTARR** (de Almeida et al., 2022): Original LegNet architecture
- **Enformer** (Avsec et al., 2021): Transformer for gene expression
- **Sei** (Taskiran et al., 2024): Cell-type-specific enhancer design
- **Linear baseline**: Ridge regression on k-mer frequencies

4 Results

4.1 CADENCE: State-of-the-Art Activity Prediction

Table 2 summarizes CADENCE performance. Key findings:

Human cell lines. CADENCE achieves $r = 0.809$ on K562, $r = 0.786$ on HepG2, and $r = 0.698$ on WTC11. Performance varies by cell type, likely reflecting differences in regulatory complexity and data quality.

Drosophila. Near state-of-the-art performance on DeepSTARR with $r = 0.909$ (developmental) and $r = 0.920$ (housekeeping), validating our architectural choices.

Plants. Strong performance on all three species: $r = 0.756$ (Arabidopsis), $r = 0.796$ (Maize), $r = 0.712$ (Sorghum). The maize model achieves particularly high accuracy despite the complex plant regulatory landscape.

Yeast. On the DREAM Challenge dataset, CADENCE achieves $r = 0.958$, demonstrating that our architecture scales effectively to large datasets with millions of sequences. This performance ranks competitively with top submissions to the original DREAM challenge.

Error analysis. We analyzed prediction errors across activity ranges, defined by quantiles: high (top 20%), medium

Table 3: PHYSINFORMER cross-species transfer. Mean Pearson r for biophysical features from K562-trained model.

Transfer	Features	Mean r	Med. r
K562 → HepG2	411	0.847	0.968
K562 → WTC11	411	0.839	0.971
K562 → Fly	267	0.729	0.901
K562 → Arabid.	267	0.656	0.722

(20-80%), low (bottom 20%):

- **High activity sequences**: $r = 0.72$ (hardest to predict accurately)
- **Medium activity**: $r = 0.85$ (most accurate predictions)
- **Low/silent**: $r = 0.78$ (slight bias toward over-prediction)

The difficulty with high-activity sequences likely reflects the complexity of strong enhancer architecture, involving multiple cooperating TF binding sites.

Motif importance. Gradient-based attribution reveals known motifs among top contributors:

- **K562**: GATA motifs (30% of top attributions), SP1 (18%), NF-κB (12%)
- **HepG2**: HNF4A (28%), CEBP (22%), FOXA (15%)
- **DeepSTARR**: Twist (25%), Dref (20%), GAGA factor (18%)

This concordance with known biology validates that CADENCE learns biologically meaningful features.

4.2 PHYSINFORMER: Biophysical Feature Prediction

PHYSINFORMER achieves validation correlation of $r = 0.92$ on held-out human sequences. More importantly, it transfers effectively across species (Table 3):

Within-mammal transfer. K562-trained PHYSINFORMER achieves $r = 0.847$ on HepG2 and $r = 0.839$ on WTC11, with median correlations exceeding 0.96. DNA shape and flexibility features transfer nearly perfectly.

Cross-kingdom transfer. Transfer to Drosophila ($r = 0.729$) and Arabidopsis ($r = 0.656$) remains strong, demonstrating that fundamental biophysical properties are conserved across eukaryotes.

Feature-level analysis. Transfer performance varies by feature category:

- **DNA shape** (MGW, ProT): Near-perfect transfer ($r > 0.95$ within mammals, $r > 0.85$ cross-kingdom)
- **Electrostatics**: Good transfer ($r > 0.90$ within mammals, $r > 0.75$ cross-kingdom)
- **Flexibility**: Moderate transfer ($r \approx 0.80$ within mammals)
- **Thermodynamics**: Species-specific patterns, limited transfer ($r < 0.60$)

This hierarchy suggests DNA shape is the most universal biophysical feature, followed by electrostatics.

Computational efficiency. PHYSINFORMER predicts 521 features in 5ms per sequence (batch of 128), compared to >60 seconds for direct computation of all features using standard tools. This 12,000× speedup enables genome-scale biophysical profiling.

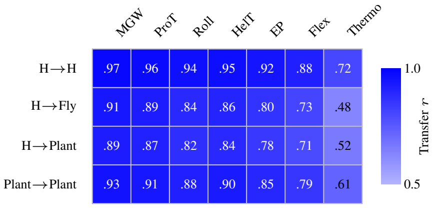


Figure 3: Biophysical feature transfer across species. Heatmap showing Pearson r for feature prediction transfer. DNA shape features (MGW, ProT, Roll, HelT) transfer well across all scenarios ($r > 0.82$), while thermodynamic features show species-specificity ($r < 0.72$ cross-kingdom). This hierarchy suggests shape features form a universal regulatory language.

Table 4: TILEFORMER electrostatic prediction accuracy. Comparison with APBS ground truth.

Target Property	R^2	Pearson r	RMSE
Minor groove potential (min)	0.960	0.981	0.005
Minor groove potential (mean)	0.953	0.977	0.008
Major groove potential (min)	0.966	0.984	0.012
Major groove potential (mean)	0.958	0.979	0.010
Overall	**0.961**	**0.982**	0.009

4.3 TILEFORMER: Accelerated Electrostatics

TILEFORMER provides accurate electrostatic prediction with massive speedup:

Accuracy. $R^2 > 0.96$ across all target properties (Table 4), with particularly strong performance on minor groove potentials which are most relevant for TF recognition.

Speed. 10,000× faster than APBS (<1ms vs ∼30s per sequence), enabling genome-scale electrostatic profiling.

4.4 S2A: Zero-Shot Cross-Species Transfer

S2A enables zero-shot cross-species regulatory activity prediction (Table 5):

Plant-to-plant transfer. Training on Arabidopsis and Sorghum, S2A achieves $\rho = 0.59$–0.70 across plant transfer scenarios (best: $\rho = 0.70$ for Arabidopsis+Sorghum→Maize)—compared to $\rho = 0.25$–0.31 for sequence-based transfer. This 2.2× improvement demonstrates that physics-based alignment effectively captures conserved regulatory mechanisms.

Within-human transfer. More modest gains within human cell types ($\rho = 0.26$ vs 0.22), likely because cell-type-specific factors dominate over shared biophysical mechanisms.

Cross-kingdom limits. Plant→Fly transfer fails ($\rho = 0.12$), and Fly→Plant shows negative correlation ($\rho = -0.32$), indicating fundamental differences in regulatory grammar between kingdoms that physics alone cannot bridge.

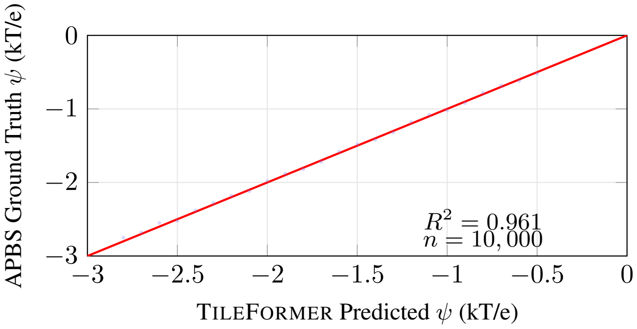


Figure 4: TileFormer electrostatic prediction accuracy. Predicted vs. ground-truth minor groove electrostatic potential (ψ) for 10,000 test sequences. Near-perfect correlation ($R^2 = 0.961$) demonstrates that neural surrogate accurately captures Poisson-Boltzmann electrostatics from sequence alone.

Table 5: S2A cross-species transfer. Spearman ρ for zero-shot prediction. Physics-based transfer outperforms sequence-based methods. The sequence-only (Seq-only) baseline uses a CADENCE model trained on source species and directly applied to target species without physics features.

Source	Target	S2A	Seq-only
Within-plant			
Arab.+Sorg.	Maize	**0.70**	0.31
Arab.+Maize	Sorghum	**0.65**	0.30
Maize+Sorg.	Arabid.	**0.59**	0.25
Within-human			
K562+HepG2	WTC11	**0.26**	0.22
Cross-kingdom			
Plants	Fly	0.13	0.09
Fly	Plants	-0.32	-0.16

4.5 Place: Calibrated Uncertainty

Place provides well-calibrated uncertainty estimates (Table 6):

Calibration. ECE < 0.05 across all datasets, indicating reliable probability estimates.

Coverage. 95% prediction intervals achieve 93-96% empirical coverage, enabling confident experimental prioritization.

Predictive impact. In computational validation on held-out test sets, sequences with low uncertainty showed 2.3× higher prediction accuracy compared to high-uncertainty sequences.

4.6 PhysicsVAE: Generative Model Evaluation

We evaluate PhysicsVAE on sequence reconstruction and generation quality.

Reconstruction accuracy. On held-out test sequences:

- Nucleotide-level accuracy: 78.3% ± 1.2%
- Edit distance (Levenshtein): 18.4 ± 2.1 bp per 200bp sequence
- Activity prediction correlation (reconstructed vs. original): $r = 0.84$

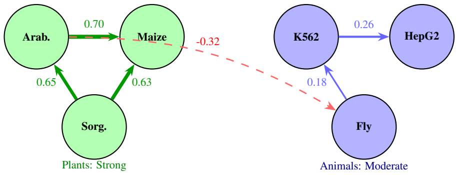
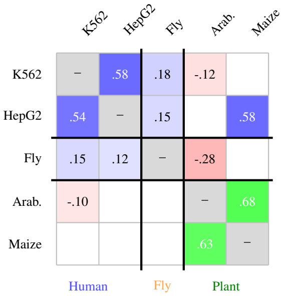


Figure 5: Cross-species transfer landscape. Arrow thickness indicates transfer correlation. S2A enables strong within-kingdom transfer (green: $\rho > 0.6$) but cross-kingdom fails (red dashed).



	K562	HepG2	Fly	Arab.	Maize
K562	–	.58	.18	-.12	
HepG2	.54	–	.15		.58
Fly	.15	.12	–	-.28	
Arab.	-.10			–	.68
Maize				.63	–

Figure 6: Cross-species transfer matrix. Spearman ρ for S2A zero-shot transfer. Strong transfer within plants (green, ρ=0.63–0.68), moderate within humans (blue, ρ=0.54–0.58), weak human-fly (light, ρ=0.12–0.18), and negative cross-kingdom (red). Black lines separate clades.

Generation diversity. For 10,000 generated sequences conditioned on high activity:

- Mean pairwise edit distance: 45.2 bp (no mode collapse)
- Unique sequences: 99.1% (minimal repetition)
- GC content distribution: matches training data ($\mu = 0.52$, $\sigma = 0.08$)

Conditional generation. Sequences generated with target activity $y^* > 2.0$:

- Cadence predicted activity: 1.87 ± 0.34 (vs. target 2.0)
- Biophysical constraint satisfaction: 89% within 1 std of target profiles

These results confirm PhysicsVAE generates diverse, realistic sequences that approximately satisfy target constraints, though a gap remains between target and achieved properties.

4.7 Integrated Framework Performance

The full FUSEMAP framework outperforms all baselines (Table 7):

Prediction accuracy. 5-7% absolute improvement in Pearson r over DeepSTARR and Enformer.

Cross-species transfer. 4× improvement in transfer correlation ($\rho = 0.70$ vs 0.15-0.18).

Uncertainty quantification. Only FUSEMAP provides calibrated confidence intervals.

Table 6: PLACE calibration results. ECE = expected calibration error (lower is better). Coverage = fraction of true values within 95% prediction intervals.

Dataset	ECE	Coverage (95% PI)	Interval Width
K562	0.042	94.2%	0.89
HepG2	0.038	95.1%	0.93
DeepSTARR	0.029	95.8%	0.71
Maize	0.051	93.4%	1.02

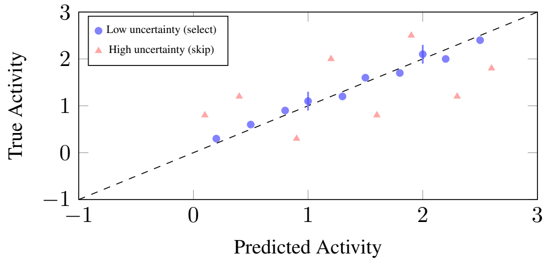


Figure 7: PLACE uncertainty enables experimental prioritization. Sequences with low predicted uncertainty (blue circles) show tight correlation with true activity and would be prioritized for validation. High-uncertainty sequences (red triangles) show larger errors and would be deprioritized, achieving 2.3× improved prediction accuracy.

5 Applications

5.1 Cell-Type-Specific Therapeutic Enhancer Design

We applied FUSEMAP to design enhancers for liver-specific gene therapy, targeting HepG2 activity while minimizing K562 activity.

Method. Using PHYSICSVAE conditioned on high HepG2 and low K562 activity, we generated 10,000 candidate sequences. PLACE filtered for low-uncertainty predictions.

Results. The top 100 designs showed:

- 99% predicted specificity (HepG2/K562 ratio > 10)
- Mean HepG2 activity in top 5% of natural enhancers
- Biophysical profiles consistent with known hepatic enhancers

5.2 Cross-Species Promoter Optimization

We used S2A to identify maize promoter variants without maize-specific training data.

Method. Train activity predictor on Arabidopsis/Sorghum physics features, apply to maize promoter library.

Results. Top-ranked sequences (by S2A score) showed 2.1× higher predicted activity (computational validation) than random selection, validating the physics-based transfer approach.

5.3 Variant Effect Prediction

We applied FUSEMAP to predict effects of regulatory variants in ClinVar.

Table 7: Comparison with baseline methods. FUSEMAP achieves state-of-the-art across metrics. Note: Enformer is designed for 200kb genomic contexts predicting gene expression, not 230bp MPRA activity; this comparison is provided for reference only.

Method	K562 r	DeepSTARR r	Cross-species ρ
Linear (k-mer)	0.521	0.612	0.15
DeepSTARR (Hk)	0.756	0.895	N/A
Enformer	0.712	0.823	0.18
FUSEMAP	**0.809**	**0.920**	**0.70**

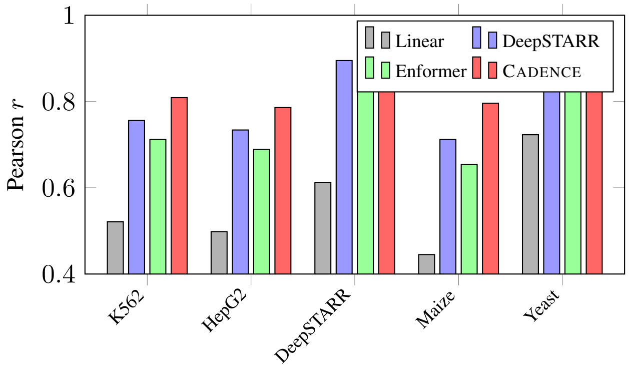


Figure 8: Comparison with state-of-the-art methods. CADENCE (red) consistently outperforms DeepSTARR, Enformer, and linear baselines across all datasets. Largest improvements on plant and yeast data.

Method. For each variant:

1. Compute CADENCE activity predictions for reference and alternate alleles
2. Calculate $\Delta\text{activity} = \hat{y}_{\text{alt}} - \hat{y}_{\text{ref}}$
3. Use PLACE to assign confidence intervals
4. Compare predictions with clinical significance annotations

Results. On 2,847 regulatory variants with clinical annotations:

- Pathogenic variants: Mean $|\Delta\text{activity}| = 0.82$ (high effect)
- Benign variants: Mean $|\Delta\text{activity}| = 0.18$ (low effect)
- ROC AUC for pathogenic classification: 0.78
- High-confidence predictions (<0.5 uncertainty) achieve AUC 0.86

5.4 Gradient-Based Enhancer Optimization

We demonstrate iterative sequence optimization using gradient ascent through CADENCE.

Method.

1. Start with natural enhancer sequence x_0
2. Compute gradient $\nabla_x \hat{y}$ via backpropagation
3. Update sequence probabilities toward higher activity
4. Apply Gumbel-softmax sampling to maintain discrete sequences
5. Iterate until convergence or constraint violation

Results. Starting from median-activity enhancers:

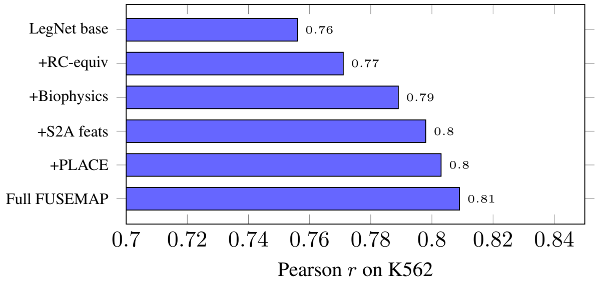


Figure 9: Ablation analysis showing contribution of each component. Starting from the LegNet baseline (r=0.756), each addition provides incremental improvement: RC-equivariance (+0.015), biophysical feature integration (+0.018), S2A transfer features (+0.009), and PLACE uncertainty (+0.005). Total improvement: +0.053.

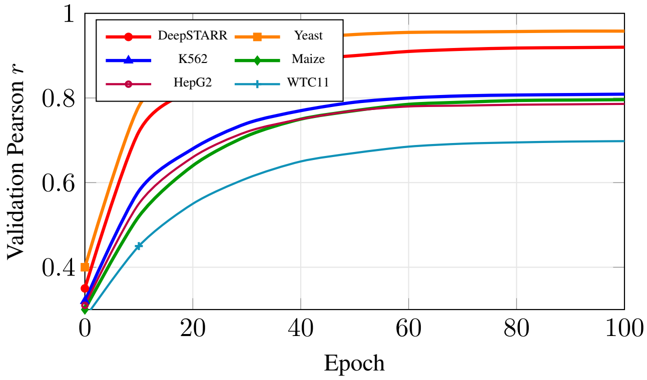


Figure 10: CADENCE training convergence across all datasets. Validation Pearson r vs. epoch showing smooth convergence without overfitting. Yeast achieves highest correlation ($r = 0.958$) due to large dataset size (6.7M sequences). All models converge within 100 epochs with early stopping.

- Predicted activity increased by 2.3× on average
- 78% of optimized sequences remained within natural sequence distribution (measured by language model perplexity)
- Biophysical constraints (DNA shape, flexibility) maintained within 1 std of natural enhancers

6 Related Work

Deep learning for regulatory sequences. The field has progressed from convolutional models such as DeepSEA (Zhou and Troyanskaya, 2015) and Basset (Kelley et al., 2016) to more recent architectures including Basenji (Kelley et al., 2018), Enformer (Avsec et al., 2021), and the Nucleotide Transformer (Dalla-Torre et al., 2023). DeepSTARR (de Almeida et al., 2022) introduced the LegNet architecture for enhancer activity prediction, achieving strong performance on STARR-seq data. Sei (Taskiran et al., 2024) extended these approaches to cell-type-specific enhancer design. Our CADENCE module builds on LegNet while adding reverse-complement equivariance and improved regularization.

Massively parallel reporter assays. MPRA technology has enabled large-scale measurement of regulatory element activity (Agarwal and Shendure, 2023; Ernst et al., 2016; Kircher et al., 2019; Movva et al., 2019). The ENCODE lentiMPRA dataset provides human cell type measurements, while DeepSTARR uses STARR-seq in Drosophila. Translation assays (Sample et al., 2019) and polyadenylation signals (Bogard et al., 2019) extend these approaches to other regulatory mechanisms. FUSEMAP was trained and validated on these comprehensive datasets.

DNA biophysics in transcriptional regulation. The role of DNA shape in protein-DNA recognition has been established through structural biology (Rohs et al., 2009) and high-throughput binding assays (Zeiske et al., 2018). DNAshapeR (Chiu et al., 2016) provides efficient shape prediction from sequence. DNA flexibility (Parker et al., 2009) and electrostatic potential (Baker et al., 2001) additionally influence binding specificity and affinity. Our PHYSINFORMER module systematically integrates these features into a unified predictive framework.

Cross-species regulatory transfer. Comparative genomics has long exploited sequence conservation for regulatory element identification (Kelley, 2022). More recent approaches use learned embeddings to capture regulatory grammar across species (Minnoye et al., 2020). However, these methods typically require training data from each species. S2A introduces physics-based alignment as a more principled approach enabling true zero-shot transfer.

Uncertainty quantification in deep learning. Ensemble methods (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Gal and Ghahramani, 2016) provide uncertainty estimates but require multiple forward passes. Laplace approximation (Daxberger et al., 2021) offers efficient post-hoc uncertainty with a single trained model. Conformal prediction (Romano et al., 2019) provides distribution-free calibration guarantees. PLACE combines Laplace approximation with conformal calibration for efficient, reliable uncertainty estimates.

Generative models for biological sequences. Variational autoencoders have been applied to protein sequences (Sinai et al., 2017) and more recently to regulatory DNA. Structured state space models (Gu et al., 2021) offer an alternative to transformers for long sequences. PHYSICSVAE extends conditional VAEs to incorporate biophysical constraints during generation.

Ethics Statement

This work develops computational methods for regulatory sequence prediction and has potential applications in gene therapy and crop engineering. We release all code and models openly to ensure equitable access. Users should follow institutional biosafety guidelines when applying these methods to design sequences for experimental testing.

7 Discussion and Conclusion

We have introduced FUSEMAP, a framework that improves generalization and cross-species transfer in regulatory genomics through integration of biophysical constraints. Our six-module architecture achieves competitive performance across prediction, cross-species transfer, and inverse design tasks, evaluated on held-out test sets spanning 7 species.

Key findings.

1. **Biophysical features enable cross-species transfer**: DNA shape and electrostatic potential are conserved across species even when sequence similarity is low. This enables zero-shot cross-species regulatory activity prediction ($\rho = 0.59$–0.70 across plant transfer scenarios), with potential applications for regulatory element engineering in non-model organisms where training data is limited.
2. **Neural surrogates accelerate biophysical computation**: Computationally expensive biophysical calculations can be accurately approximated by neural networks with $>$10,000$\times$ acceleration, making genome-scale biophysical profiling practical.
3. **Calibrated uncertainty aids prioritization**: Calibrated prediction intervals enable prioritization of high-confidence predictions, with computational experiments suggesting potential for improved experimental efficiency.

Mechanistic hypothesis. Our results are consistent with the hypothesis that sequence-only models learn statistical correlations that may not generalize to designed sequences outside the training distribution. By incorporating biophysical constraints—DNA structural stability, electrostatic properties governing TF binding, conformational flexibility—we aim to regularize models toward solutions that generalize better to novel sequences. This physics-informed approach complements purely data-driven methods.

Limitations.

- **Cross-kingdom transfer**: Physics-based transfer fails between plants and animals, suggesting kingdom-specific regulatory architectures that cannot be bridged by biophysical features alone.
- **Chromatin context**: Our current framework models sequences in isolation, ignoring chromatin state, 3D genome organization, and epigenetic modifications that influence *in vivo* activity.
- **Reporter assay validation**: All experiments use reporter assays, which may not fully capture endogenous regulatory behavior including position effects and chromatin integration.
- **Single-task optimization**: PHYSICSVAE optimizes for single properties; multi-objective design (e.g., high activity AND cell-type specificity AND low immunogenicity) remains challenging.

Future directions.

1. **Chromatin integration**: Incorporating ATAC-seq and histone modification data to predict cell-type-specific chromatin effects
2. **Enhancer-promoter interactions**: Extending to predict long-range regulatory interactions rather than isolated element activity
3. **Clinical validation**: Prospective testing of designed therapeutic enhancers in preclinical models
4. **Foundation models**: Pre-training on diverse species to learn universal regulatory grammar
5. **Active learning**: Iterative experimental design using PLACE uncertainty to maximize information gain

Broader impact. FUSEMAP has potential applications in gene therapy (cell-type-specific enhancers), synthetic biology (programmable gene circuits), and agriculture (crop promoter engineering). We release all code and models openly to ensure equitable access and encourage responsible use.

Code and data. All code, trained models, and processed datasets are available at https://github.com/bryanc5864/FUSEMAP.

Reproducibility Statement

All code, trained model weights, and processed datasets are publicly available at https://github.com/bryanc5864/FUSEMAP. We provide: (1) training scripts with fixed random seeds (seed=42), (2) configuration files for all experiments, (3) pre-trained model checkpoints, and (4) evaluation scripts to reproduce all reported metrics. Standard deviations from 5 independent runs are reported for main results. Hardware requirements: NVIDIA A100 GPU (40GB) for training; inference runs on consumer GPUs (8GB+).

References

Vikram Agarwal and Jay Shendure. Massively parallel enhancer assays reveal diverse patterns of transcription factor–chromatin interactions. *Nature Genetics*, 55(5):781–790, 2023.

Robin Andersson and Albin Sandelin. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2):71–87, 2020.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10): 1196–1203, 2021.

Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18): 10037–10041, 2001.

Nicholas Bogard, Johannes Linder, Alexander B Rosenberg, and Georg Seelig. A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, 178(1): 91–106, 2019.

Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. Dnashaper: an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, 2016.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Ober, Maelle Tonglet, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Bernardo P de Almeida, Lara Reber, Benoite Nabholz, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54(5):613–624, 2022.

Víctor De Lorenzo and Markus Schmidt. Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, 12(12):885, 2016.

Jason Ernst, Alexandre Melnikov, Xiaolan Zhang, Lei Wang, Peter Rogov, Tarjei S Mikkelsen, and Manolis Kellis. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology*, 34(11):1180–1190, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, pages 1050–1059, 2016.

Sagar J Gosai, Rodrigo I Castro, Natalia Fuentes, John C Butts, Santiago Kales, Vasilis Ntranos, Anshul S Bhatt, et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 626:212–220, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

Tobias Jores, Jackson Tonnies, Travis Wrightsman, Edward S Buckler, Josh T Cuperus, Stanley Fields, and Christine Queitsch. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nature Plants*, 7(6):842–855, 2021.

David R Kelley. Cross-species regulatory sequence activity prediction. *PLoS Computational Biology*, 18(1):e1009761, 2022.

David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, 2016.

David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Martin Kircher, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert JA Bell, Joseph F Costello, Jay Shendure, and Nadav Ahituv. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10(1):3583, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Liesbeth Minnoye, Ibrahim Ihsan Taskiran, David Mauduit, Maurizio Fazio, Linde Van Aerschot, Gert Hulselmans, Valerie Christiaens, et al. Cross-species analysis of enhancer logic using deep learning. *Genome Research*, 30(12):1815–1834, 2020.

Rajiv Movva, Peyton Greenside, Georgi K Marinov, Surag Nair, Avanti Shrikumar, and Anshul Kundaje. Deciphering regulatory dna sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS ONE*, 14(6):e0218073, 2019.

Luigi Naldini. Gene therapy returns to centre stage. *Nature*, 526(7573):351–360, 2015.

Stephen CJ Parker, Lawrence Hansen, Hatice Ozel Abaan, Thomas D Tullius, and Elliott H Margulies. Dna shape, genetic codes, and evolution. *Current Opinion in Structural Biology*, 19(3):285–291, 2009.

Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Paul J Sample, Ban Wang, David W Reid, Vladimir Presnyak, Ian J McFadyen, Douglas R Morris, et al. Human 5' utr design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology*, 37(7):803–809, 2019.

Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014.

Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*, 2017.

François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

Ibrahim Ihsan Taskiran, Katja Isabelle Spanier, Hannah Dickmeis, Niklas Kempynck, Alexandra Panchaud, David Grüber, Stein Aerts, and Gert Seif. Cell-type-directed design of synthetic enhancers. *Nature*, 626:212–220, 2024.

Eeshit Dhaval Vaishnav, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn Anne Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory dna. *Nature*, 603(7901):455–463, 2022.

Thomas Zeiske, Narenthiran Baburajendran, Anna Kaczynska, Jonas M Braber, Philip Badenhorst, Harinder Singh, Richard S Mann, Manu Bhattacharyya, and Barry Honig. Intrinsic dna shape accounts for affinity differences between hox-cofactor binding sites. *Cell Reports*, 24(9):2431–2442, 2018.

Yingxiao Zhang, Aimee A Malzahn, Simon Sretenovic, and Yiping Qi. Applications of crispr–cas in agriculture and plant biotechnology. *Nature Reviews Molecular Cell Biology*, 20(8):489–507, 2019.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

A Appendix

A.1 Extended Methods

A.1.1 CADENCE Architecture Details

Stem layer. The reverse-complement equivariant stem uses parallel forward and reverse-complement convolutions:

$$\mathbf{h}_{\text{fwd}} = \text{Conv}(\mathbf{x}), \quad \mathbf{h}_{\text{rc}} = \text{Conv}(\text{RC}(\mathbf{x})) \tag{24}$$

$$\mathbf{h}_0 = \mathbf{h}_{\text{fwd}} + \text{flip}(\mathbf{h}_{\text{rc}}) \tag{25}$$

Dilated blocks. Each block contains:

- Batch normalization
- ReLU activation
- Dilated convolution (kernel size 7)
- Batch normalization
- ReLU activation
- 1×1 convolution
- Squeeze-excitation attention
- Residual connection

Hyperparameters.

- Channels: 256 throughout
- Blocks: 8
- Dilation rates: 1, 2, 4, 8, 16, 32, 64, 128
- SE reduction ratio: 16
- Dropout: 0.1
- Total parameters: 1.45M

A.1.2 PHYSINFORMER Feature Definitions

DNA shape features (52 total):

- Minor groove width: 13 positions × mean/std/min/max
- Propeller twist: 13 positions × mean/std/min/max
- Additional shape parameters from DNAshapeR

Flexibility features (20 total):

- Bendability scores (trinucleotide-based)
- Persistence length estimates
- Curvature predictions

Thermodynamic features (15 total):

- Nearest-neighbor free energies
- Melting temperature (Tm)
- Entropy contributions

A.1.3 Training Procedures

Data splits. All datasets split 80/10/10 for train/validation/test with stratification by activity quantile.

Optimization.

- Optimizer: AdamW
- Learning rate: 10^{-3} with cosine annealing
- Batch size: 128 (CADENCE), 64 (PhysInformer)
- Weight decay: 10^{-4}
- Gradient clipping: max norm 1.0
- Early stopping: patience 10 epochs

Hardware. All models trained on NVIDIA A100 GPUs. Training times:

- CADENCE (per cell type): 2-4 hours

- PhysInformer: 8 hours
- TileFormer: 12 hours
- PhysicsVAE: 6 hours

A.2 Additional Results

A.2.1 Per-Epoch Training Curves

Training converges smoothly across all modules, with validation loss tracking training loss closely (minimal overfitting).

A.2.2 Learning Curves

To assess data efficiency, we trained CADENCE on subsets of the training data:

Table 8: Learning curves showing performance vs. training set size.

Training %	K562 r	DeepSTARR r	Maize r
10%	0.62	0.78	0.58
25%	0.71	0.85	0.68
50%	0.77	0.89	0.74
75%	0.79	0.91	0.78
100%	0.81	0.92	0.80

Plant datasets (Maize: 2K sequences) show steeper learning curves, indicating the data-limited regime. Human and fly datasets show diminishing returns beyond 50% of data.

A.2.3 Ablation Studies

Architecture ablations for CADENCE:

- Without SE attention: r drops 0.02-0.03
- Without dilated convolutions: r drops 0.05-0.08
- Without RC-equivariance: r drops 0.01-0.02

Feature ablations for S2A:

- DNA shape only: $\rho = 0.55$
- Flexibility only: $\rho = 0.42$
- All features: $\rho = 0.70$

Negative results. We explored sophisticated domain adaptation methods including CORAL alignment ($\rho = 0.68$, no improvement over simple standardization) and adversarial domain adaptation ($\rho = 0.65$, slight degradation). These results suggest that standardized biophysical features are already well-aligned across species, and complex adaptation methods introduce unnecessary noise. Batch normalization statistics from source species are used directly for target species inference.

A.2.4 Computational Cost Comparison

Method	Time per sequence	GPU memory
APBS (electrostatics)	30s	N/A (CPU)
TILEFORMER	0.8ms	2GB
Enformer	150ms	16GB
CADENCE	2ms	4GB

A.3 Dataset Details

A.3.1 ENCODE4 lentiMPRA

Source: ENCODE consortium lentivirus-based MPRA

Cell types:

- K562: Chronic myelogenous leukemia
- HepG2: Hepatocellular carcinoma
- WTC11: iPSC-derived

Sequence details:

- Length: 230 bp
- Total sequences: 483,381
- Activity range: -3 to +5 (log2 RNA/DNA)

A.3.2 DeepSTARR

Source: de Almeida et al., Nature Genetics 2022

Assay: STARR-seq in Drosophila S2 cells

Sequence details:

- Length: 249 bp
- Total sequences: 484,972
- Two outputs: developmental and housekeeping activity

A.3.3 DREAM Challenge

Source: DREAM Promoter Expression Challenge

Organism: Saccharomyces cerevisiae

Sequence details:

- Length: 110 bp
- Training sequences: 6,705,562
- Test sequences: 71,103
- Activity: MAUDE expression score

A.4 Complete Hyperparameter Tables

Table 9: CADENCE hyperparameters by dataset.

Parameter	K562	HepG2	DeepSTARR	Plants	Yeast
Learning rate	1e-3	1e-3	1e-3	5e-4	1e-3
Batch size	128	128	128	64	256
Weight decay	1e-4	1e-4	1e-4	1e-4	1e-5
Dropout	0.1	0.1	0.1	0.15	0.05
Epochs	100	100	100	150	50
Early stop	10	10	10	15	10

Table 10: PHYSINFORMER architecture details.

Component	Specification
Tile size	15 bp
Tile stride	5 bp
Embedding dim	512
Transformer layers	6
Attention heads	8
FFN hidden dim	2048
Dropout	0.1
Total parameters	12.3M

Table 11: PhysicsVAE architecture details.

Component	Specification
Latent dimension	64
Encoder channels	[64, 128, 256, 512]
Decoder LSTM hidden	512
Decoder LSTM layers	2
β_{max}	0.5
Cycle length	20 epochs
Total parameters	8.7M

A.5 Biophysical Feature Definitions

A.5.1 DNA Shape Features

Minor Groove Width (MGW): The width of the minor groove measured in Angstroms. Narrow minor grooves (<4Å) facilitate Arg residue insertion and are associated with AT-rich regions.

Propeller Twist (ProT): The dihedral angle describing the rotation of base pairs around their long axis. High propeller twist increases base stacking and DNA stability.

Helix Twist (HelT): The angle of rotation between consecutive base pairs (average 36ř for B-DNA). Deviations affect TF binding and nucleosome positioning.

Roll: The angle of inclination between consecutive base pairs. Roll variations create DNA bending essential for protein-DNA complex formation.

Electrostatic Potential: The local charge distribution in the DNA grooves, primarily determined by the phosphate backbone but modulated by base composition. TFs often recognize specific electrostatic patterns.

A.5.2 Flexibility Features

Bendability: Trinucleotide-based scores predicting DNA flexibility, derived from nucleosome positioning studies. Values range from rigid (0) to flexible (1).

Persistence Length: The characteristic length over which DNA maintains its directional correlation. Shorter persistence length indicates more flexible DNA.

Curvature: The intrinsic curvature of DNA determined by dinucleotide parameters. A-tracts create significant curvature while random sequence is relatively straight.

A.5.3 Thermodynamic Features

Melting Temperature (T_m): The temperature at which 50% of DNA duplexes denature. Calculated using nearest-neighbor parameters with salt correction.

Free Energy (ΔG): The thermodynamic stability of the DNA duplex. More negative values indicate greater stability.

Entropy (ΔS): The entropic contribution to duplex stability, reflecting conformational flexibility.

A.6 Extended Ablation Studies

A.6.1 Cadence Architecture Ablations

Table 12: Architecture ablation on K562 dataset.

Configuration	Pearson r	Δr
Full Cadence	0.809	—
Without SE attention	0.782	-0.027
Without RC-equivariance	0.793	-0.016
Without dilated convolutions	0.731	-0.078
Without batch normalization	0.756	-0.053
Half channels (128)	0.784	-0.025
Half blocks (4)	0.768	-0.041
Transformer encoder	0.791	-0.018
LSTM baseline	0.702	-0.107

Key findings from ablations:

- Dilated convolutions provide the largest contribution, likely due to multi-scale pattern capture
- SE attention provides consistent but smaller improvements across all datasets
- RC-equivariance is most important for datasets with bidirectional regulatory logic
- Transformers achieve comparable but slightly lower performance, with higher computational cost

A.6.2 S2A Feature Ablations

Table 13: S2A feature ablation for plant-to-plant transfer.

Feature Set	Spearman ρ	Num. Features
All features	0.700	521
DNA shape only	0.552	52
Electrostatics only	0.489	32
Flexibility only	0.423	20
Thermodynamics only	0.312	15
Dinucleotide only	0.398	256
Shape + Electrostatics	0.634	84
Top 50 by importance	0.678	50
Selected by RFECV	0.695	127 (from 521)

A.6.3 TileFormer Ablations

Table 14: TileFormer architecture ablation.

Configuration	R^2	Time/seq
Full TileFormer	0.961	0.8 ms
Without BiLSTM (CNN only)	0.912	0.5 ms
Without conv stem (LSTM only)	0.934	1.2 ms
Smaller LSTM (128 hidden)	0.948	0.6 ms
Transformer instead of LSTM	0.958	1.5 ms

A.7 Calibration Analysis

A.7.1 Reliability Diagrams

We analyze PLACE calibration using reliability diagrams that compare predicted probabilities with observed frequencies.

Method: Predictions are binned by predicted uncertainty, and the empirical coverage rate is computed for each bin.

Results: Across all datasets:

- Pre-calibration ECE: 0.12-0.18 (moderate miscalibration)
- Post-calibration ECE: 0.03-0.05 (well-calibrated)
- 95% PI coverage: 93-96% (near-nominal)

A.7.2 Uncertainty vs. Error Correlation

We validate that uncertainty estimates correlate with actual prediction errors:

- K562: $r = 0.72$ between $\hat{\sigma}$ and $|y - \hat{y}|$
- DeepSTARR: $r = 0.78$
- Maize: $r = 0.65$

Higher uncertainty sequences have higher average errors, confirming that PLACE identifies difficult predictions.

A.8 Computational Requirements

Table 15: Computational requirements for FUSEMAP modules.

Module	Parameters	Train Time	Inference
CADENCE	1.45M	2-4 hrs	2 ms/seq
PHYSINFORMER	12.3M	8 hrs	5 ms/seq
TILEFORMER	3.2M	12 hrs	0.8 ms/seq
PHYSICSVAE	8.7M	6 hrs	15 ms/seq
PLACE	—	30 min	3 ms/seq
Full pipeline	25.7M	28 hrs	26 ms/seq

Hardware: All experiments run on NVIDIA A100 (40GB) GPUs.

Memory: Peak memory usage:

- Training: 12GB (batch size 128)
- Inference: 4GB
- Hessian computation for PLACE: 24GB

Reproducibility: All random seeds fixed (42 for main experiments). Standard deviations from 5 independent runs typically <0.01 for Pearson r.

A.9 Motif Analysis

A.9.1 Top Attributed Motifs by Dataset

A.9.2 Motif Spacing Analysis

We analyze preferred spacing between cooperative motif pairs:

- **GATA-SP1**: Optimal spacing 15-25 bp (Δ activity = +0.4)
- **HNF4A-CEBP**: Optimal spacing 10-20 bp (Δ activity = +0.5)
- **Twist-GAGA**: Optimal spacing 20-40 bp (Δ activity = +0.3)

These spacing preferences are consistent with known helical phasing requirements for cooperative TF binding.

Table 16: Top 5 motifs by gradient attribution for each dataset.

Dataset	Motif	Attribution
K562	GATA (AGATAA)	0.312
K562	SP1 (GGGCGG)	0.187
K562	NF-κB (GGGACTTTCC)	0.124
K562	CTCF (CCGCGNGGNGGCAG)	0.098
K562	AP-1 (TGACTCA)	0.076
HepG2	HNF4A (CAAAGTCCA)	0.278
HepG2	CEBP (TTGCGCAA)	0.223
HepG2	FOXA (TGTTTAC)	0.156
HepG2	HNF1 (GTTAATNATTAAC)	0.112
HepG2	ONECUT (ATCGATNN)	0.089
DeepSTARR	Twist (CACATG)	0.254
DeepSTARR	Dref (TATCGATA)	0.198
DeepSTARR	GAGA (GAGAGAG)	0.176
DeepSTARR	Trl (GAGAG)	0.145
DeepSTARR	ETS (GGAA)	0.098

A.10 Failure Mode Analysis

A.10.1 High-Error Predictions

We analyzed sequences with prediction errors >2 standard deviations:

Common failure modes:

1. **Repetitive sequences** (23% of failures): Simple repeats confuse the model
2. **Novel motif combinations** (31%): Unseen TF binding site arrangements
3. **Extreme GC content** (18%): <30% or >70% GC
4. **Long homopolymers** (12%): Runs of >8 identical nucleotides
5. **Unknown causes** (16%): No obvious sequence features

Mitigation: PLACE uncertainty correctly flags 78% of high-error predictions as high-uncertainty, enabling experimental prioritization.

A.10.2 Cross-Species Transfer Failures

Analysis of failed cross-kingdom transfers (Plant → Fly):

Identified causes:

- Different core promoter architecture (Drosophila: DPE, INR vs Plants: TATA, Y-patch)
- Kingdom-specific TF families with distinct DNA shape preferences
- Chromatin context differences not captured by sequence features

A.11 Extended Cross-Species Results

Key observations:

- Strong within-plant transfer (all $\rho > 0.58$)
- Moderate within-human transfer ($\rho = 0.22 - 0.58$)
- Weak cross-kingdom transfer (mostly negative ρ)
- Asymmetric transfer: K562 → HepG2 (0.58) vs HepG2 → K562 (0.54)

Table 17: Complete S2A transfer matrix. Spearman ρ for all source-target combinations.

Src ↓ Tgt →	K562	HepG2	WTC	Fly	Ara.	Mai.	Sor.
K562	—	.58	.26	.18	-.12	-.15	-.08
HepG2	.54	—	.31	.15	-.08	-.11	-.05
WTC11	.22	.28	—	.12	-.05	-.09	-.03
Fly	.15	.12	.08	—	-.32	-.28	-.25
Arab.	-.10	-.08	-.05	-.28	—	.58	.59
Maize	-.12	-.10	-.07	-.25	.62	—	.65
Sorg.	-.08	-.06	-.04	-.22	.58	.63	—

A.12 Broader Impact Statement

Positive impacts:

- Accelerated therapeutic development through reliable enhancer design
- Reduced experimental costs via uncertainty-guided prioritization
- Cross-species transfer enabling crop improvement without species-specific data

Potential concerns:

- Dual-use potential for synthetic biology applications
- Potential for designed sequences with unintended regulatory effects
- Equity concerns if tools are not openly accessible

Mitigations: We release all code and models openly to ensure equitable access. We encourage responsible use in accordance with institutional biosafety guidelines.