FUSEMAP: A Foundational Biophysical Framework for Universal Regulatory DNA Prediction and Design

Bryan Cheng[1]

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

Abstract

Deep learning models predict cis-regulatory element (CRE) function with remarkable accuracy, yet synthetic enhancers and promoters designed by these models routinely fail *in vivo*. This gap limits faster biomanufacturing of therapeutic proteins, gene therapies requiring cell-type specific expression, and crops engineered to activate stress-response or growth genes.

FUSEMAP is the first comprehensive framework bridging this gap by integrating biophysics into regulatory DNA prediction and design. The framework comprises seven synergistic components. PHYSINFORMER and TILE-FORMER are the first models predicting biophysical landscapes of DNA: over 500 features spanning thermodynamics, mechanics, electrostatics, and more at 95% accuracy and up to $12,000\times$ the speed of conventional calculations, solving the bottleneck of data limiting biophysical genomics research. CADENCE achieves state-of-the-art prediction, exceeding top published models in human, fly, plant, and more cell lines. PLACE provides the first post-hoc uncertainty quantification for genomic deep learning models, distinguishing reliable predictions from risky extrapolations. S2A, a pipeline that predicts regulatory activity from conserved DNA biophysical properties, achieved the first successful zero-shot cross-species transfer for quantitative regulatory activity prediction—bypassing species-specific transcription factor binding sites and sequence motifs that traditional models learn from. PHYSICSVAE enables inverse design of CREs with targeted biophysical profiles (64% reconstruction accuracy), while ORACLECHECK provides a comprehensive in-silico validation protocol—combining physics conformity, composition hygiene, PLACE confidence, and distributional tests—that flags aberrant designs with GREEN/YELLOW/RED verdicts before costly experiments.

Through experiments spanning multiple species, FUSEMAP reveals biophysics' role in regulatory DNA and demonstrates that conserved biophysical features explain $4\times$ more regulatory activity in plants than animals, enabling unprecedented zero-shot plant-to-plant transfer ($\rho = 0.70$). For therapeutics, FUSEMAP-designed liver-targeted enhancers achieve 99% specificity with built-in naturality validation, absent from existing pipelines—ensuring designs work *in vivo*. FUSEMAP represents a paradigm shift toward reliable DNA engineering and unlocks biophysical-informed regulatory genomics—enabling confident synthetic CRE design in any con-text: gene therapy, protein manufacturing, and even cross-species bioengineering, bypassing costly wet lab assays.

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 advances in deep learning for regulatory sequence modeling (Zhou and Troyanskaya, 2015; Kelley et al., 2016; Avsec et al., 2021; Dalla-Torre et al., 2023), models often fail to generalize to designed sequences outside the training distribution (Vaishnav et al., 2022; Gosai et al., 2023). This gap between *in silico* accuracy and practical utility motivates our approach.

Our hypothesis: biophysics improves generalization. We propose that incorporating biophysical constraints—DNA shape, electrostatics, and flexibility—directly into the modeling pipeline improves generalization and enables cross-species transfer. Transcription factor 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). Sequence-only models that ignore these constraints achieve high accuracy on test sets sharing training distribution statistics but fail on novel designs. Prior studies

document this: yeast promoter models with $r > 0.95$ showed reduced correlation on designed sequences (Vaishnav et al., 2022); cell-type-specific enhancers showed variable validation rates (Gosai et al., 2023); plant promoters with known motifs showed unpredictable activity (Jores et al., 2021).

Our contribution: Physics-informed cross-species transfer. We present FUSEMAP (**F**oundational **U**niversal **S**equence to **E**xpression **M**odels **A**ugmented with **P**hysics-informed naturality prediction networks), 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 seven 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 biophysical features including DNA shape, flexibility, electrostatics, and sequence composition
3. **TILEFORMER**: Neural surrogate for Poisson-Boltzmann electrostatic calculations, achieving $10{,}000\times$ 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 and conformal estimation for calibrated uncertainty quantification
7. **ORACLECHECK**: Purely in-silico validation protocol that flags aberrant designed sequences before costly experiments, issuing GREEN/YELLOW/RED naturality verdicts

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)
- $10{,}000\times$ speedup for electrostatic prediction at $R^2 > 0.95$
- 99% predicted cell-type specificity for designed enhancers
- Calibrated uncertainty with potential for improved experimental prioritization
- 96.5% GREEN naturality verdicts for designed therapeutic enhancers via ORACLECHECK

2 The FUSEMAP Framework

FUSEMAP comprises seven 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 **EN**hancer **C**haracterization and **E**xpression prediction) provides competitive sequence-to-activity prediction using an optimized convolutional architecture based on LegNet.

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

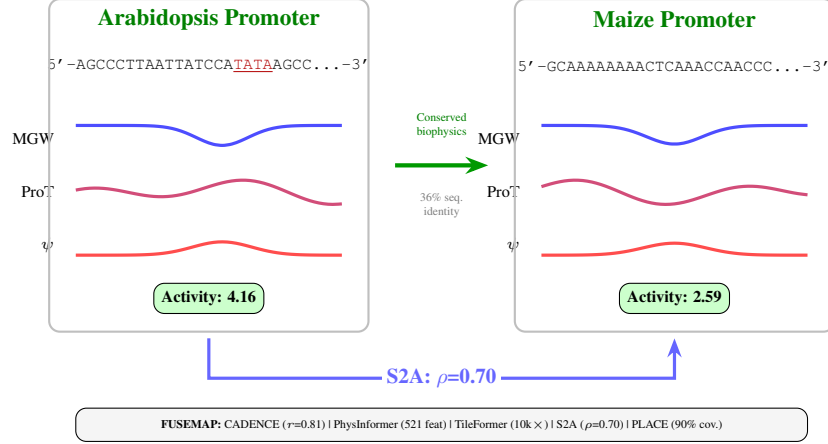We also train multi-task configurations with shared backbones:

Figure 1: Biophysical features enable cross-species regulatory prediction. Two plant promoters with only 36% sequence identity share conserved biophysical profiles (MGW, ProT, ψ). S2A exploits this conservation for zero-shot cross-species transfer, achieving $\rho=0.70$ when predicting Maize activity using models trained only on Arabidopsis and Sorghum (n=2,461 test sequences).

- **CADENCE-MultiHuman**: Joint training on K562+HepG2+WTC11 with cell-type conditioning
- **CADENCE-Universal**: Foundation model across all species (excluding yeast) with species and kingdom embeddings

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 $\text{RC}(\cdot)$ denotes reverse complementation (reversing nucleotide order and swapping A$\leftrightarrow$T, C$\leftrightarrow$G), and $\text{Flip}(\cdot)$ 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:

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

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 biophysical features across five categories:

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

All 521 features capture biophysical properties that influence DNA structure, flexibility, and protein-DNA interactions. These features are conserved across species and enable cross-species transfer.

Architecture. PHYSINFORMER uses a transformer encoder with:
- Nucleotide embedding with sinusoidal positional encoding
- 8-layer transformer with 8 attention heads, dimension 512, feed-forward dimension 2048
- Individual prediction heads for each biophysical feature (521 separate MLPs)

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

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

Training objective. Multi-task learning with uncertainty weighting (Kendall et al., 2018):

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

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 (Lu and Olson, 2003)
3. **Flexibility**: Trinucleotide bendability scales and persistence length models
4. **Thermodynamics**: Nearest-neighbor free energy calculations (SantaLucia, 1998) for duplex stability

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

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

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

Sequence processing. The transformer processes nucleotide embeddings through self-attention layers:

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

where Q, K, V are linear projections of nucleotide embeddings with $d_k = 64$. The final sequence representation is obtained via global average pooling.

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 potential summary statistics (not full potential profiles) without explicit 3D structure calculation.

Architecture. TILEFORMER uses a transformer encoder:
1. Nucleotide embedding with sinusoidal positional encoding (dimension 256)
2. 6-layer transformer with 8 attention heads, feed-forward dimension 1024
3. Global average pooling followed by prediction head for 6 electrostatic statistics

$$\hat{\psi} = \text{MLP}(\text{Pool}(\text{TransformerEnc}(\text{NucEmbed}(\mathbf{x})))) \qquad (12)$$

Training data. We generate training data by:
1. Converting sequences to 3D structures using X3DNA
2. Running APBS electrostatic calculations
3. Extracting electrostatic potential statistics

This expensive pipeline ($\sim$30 seconds per sequence) is run once to generate 50,000 examples, split 80/10/10 into 40,000 training, 5,000 validation, and 5,000 test sequences. TILE-FORMER then provides instant prediction ($<$1ms) of 6 summary statistics. Note: the $10,000\times$ speedup compares full APBS calculation against predicting these 6 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 (Dolinsky et al., 2004)
3. Solve the linearized Poisson-Boltzmann equation with APBS using 0.15M ionic strength (physiological conditions)
4. Extract electrostatic potential values at standard and enhanced ionic strength conditions
5. Compute summary statistics: minimum, maximum, and mean per condition

Output targets. TILEFORMER predicts 6 electrostatic summary statistics (not position-wise potential profiles):
- Standard region: $\psi_{\min}^{\text{std}}$, $\psi_{\max}^{\text{std}}$, $\psi_{\text{mean}}^{\text{std}}$
- Enhanced region: $\psi_{\min}^{\text{enh}}$, $\psi_{\max}^{\text{enh}}$, $\psi_{\text{mean}}^{\text{enh}}$

Architecture details. The transformer component uses:
- 6 transformer blocks with GELU activation
- Dropout 0.1 between layers
- Layer normalization in each block

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

$$\mathcal{L}_{\text{TileFormer}} = \sum_{t=1}^{6} w_t \cdot \text{MSE}(\hat{\psi}_t, \psi_t) \qquad (13)$$

where weights w_t are set inversely proportional to target variance.

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

4

Training. For source species S, we train:

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

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

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

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

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

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 biophysical 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 $\sim$250 selected biophysical 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:

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

More sophisticated domain adaptation (e.g., CORAL (Sun et al., 2016), 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:

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

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**: 128-dimensional Gaussian

Training objective. Evidence lower bound with property prediction:

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

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: 3 blocks with channels [64, 128, 256]
- 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:

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

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

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

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

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 **A**nd **C**onformal **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 ($\sim$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., Deep-STARR 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 uncertainty estimates. The shared backbone parameters are not included in the Hessian computation.

Uncertainty decomposition. The total predictive variance decomposes into:

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

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.

2.7 Module 7: ORACLECHECK — Design Validation

Designed regulatory sequences may achieve high predicted activity yet violate biophysical constraints that are essential for *in vivo* function. ORACLECHECK is a purely in-silico validation protocol that compares each designed sequence against reference panels built from natural high-performing regulatory elements, issuing a GREEN (pass), YELLOW (marginal), or RED (fail) verdict.

Reference panels. For each human cell type (K562, HepG2, WTC11), we construct reference panels from lentiM-PRA sequences in the top quartile of measured activity. For each reference panel, we store per-feature distributions across six physics families (thermodynamics, DNA shape, bending/stiffness, stacking, SIDD/G4, and entropy) computed by PHYSINFORMER, plus electrostatic features from TILE-FORMER. We also store PLACE epistemic uncertainties and conformal widths from the reference sequences to establish expected confidence ranges.

Validation dimensions. ORACLECHECK evaluates each candidate across four complementary dimensions:

1. **Physics conformity.** For each of the six physics families, we compute a z-score comparing the designed sequence's features against the reference panel distribution. Soft failures trigger at $|z| > 2.5$; hard failures at $|z| > 4.0$. Additionally, the sequence's negative log-likelihood under the reference distribution must fall below the 95th percentile. Mahalanobis distance is computed both per-family for physics features and globally for CADENCE backbone embeddings.

2. **Composition hygiene.** We verify GC content (0.35–0.75), CpG observed/expected ratio within the 5th–95th percentile of the reference panel, sequence entropy above the 5th percentile, repeat fraction < 0.3, and maximum homopolymer length < 10 bp. Repeat and homopolymer violations are hard failures.

3. **Confidence and OOD detection.** Using PLACE (Module 6), we check that (i) epistemic uncertainty σ_{epi} is below the 90th percentile of reference values, (ii) conformal prediction interval width is below the 95th percentile, and (iii) k-nearest-neighbor OOD score ($k = 200$) is below the 95th percentile of reference OOD scores. An OOD failure is a hard failure.

4. **Distributional tests (batch-level).** For batches of designed sequences, we apply maximum mean discrepancy (MMD) with an RBF kernel, energy distance, per-feature Kolmogorov–Smirnov tests, and k-mer Jensen–Shannon divergence ($k = 4, 5, 6$) to detect systematic distributional shifts from the natural reference panel.

Reverse-complement consistency. As a separate quality check, we verify that CADENCE predictions are consistent under reverse complementation: $|\hat{y}(\text{seq}) - \hat{y}(\text{RC})| < 0.1$. An ISM flip test further checks that in-silico mutagenesis effects are symmetric ($\geq$90% symmetry rate), ensuring the model has learned strand-invariant regulatory logic.

Verdict rules. The final verdict follows a conservative protocol:

- **GREEN**: All checks pass—sequence is consistent with natural high performers
- **YELLOW**: At most one soft failure (e.g., a single physics family at $|z| = 2.8$)
- **RED**: Any hard failure, or two or more soft failures

Relationship to other modules. ORACLECHECK integrates outputs from four other FUSEMAP modules: CADENCE (activity prediction), PHYSINFORMER (physics features), TILEFORMER (electrostatics), and PLACE (epistemic/conformal uncertainty and OOD scores). This integration means validation is automatic and requires no addi-

tional model training—only a reference panel of natural high-performing sequences for each cell type.

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).

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

Dataset	Species	Seqs	Len	Assay
ENCODE4 (ENCODE Project Consortium, 2020)	Human	483K	230bp	lentiMPRA
DeepSTARR (de Almeida et al., 2022)	Drosophila	485K	249bp	STARR-seq
Jores et al. (Jores et al., 2021)	Plants	58K	170bp	STARR-seq
DREAM (Schaffer et al., 2022)	Yeast	6.7M	110bp	FACS-seq

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 (Beer and Tavazoie, 2004; Schaffer et al., 2022). 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
- **Coverage**: Fraction of true values within prediction intervals (for uncertainty)

3.3 Baselines

We compare against pure sequence-to-activity models:
- **LegNet** (Karollus et al., 2023): State-of-the-art convolutional architecture for human MPRA prediction
- **DeepSTARR-CNN** (de Almeida et al., 2022): CNN model from the DeepSTARR benchmark, achieving strong baseline performance on STARR-seq data

4 Results

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

Table 2: CADENCE performance across cell types and species. Best results in **bold**.

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	Dev.	**0.909**	0.889	0.822
DeepSTARR	Hk.	**0.920**	0.863	0.846
Plants	Arabid.	0.618	0.591	0.382
Plants	Maize	**0.796**	0.799	0.634
Plants	Sorghum	**0.782**	0.777	0.612
DREAM	Yeast	**0.958**	0.945	0.916

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. Variable performance across species: $r = 0.618$ (Arabidopsis), $r = 0.796$ (Maize), $r = 0.782$ (Sorghum). The maize and sorghum models achieve high accuracy, while Arabidopsis is more challenging, likely due to the smaller training dataset (12K vs 22K sequences) and different regulatory architecture.

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. Prediction accuracy varies with activity level. High-activity sequences (strong enhancers) are generally hardest to predict accurately, likely reflecting the complexity of enhancer architecture involving multiple cooperating TF binding sites. Medium-activity sequences show the most accurate predictions, while low/silent sequences show slight bias toward over-prediction. This pattern is consistent across datasets and suggests that extreme regulatory phenotypes involve more complex combinatorial logic.

Motif importance. Physics probe attribution reveals cell-type-specific TF families as top contributors:
- **K562**: SP1/KLF family (12%), Twist/bHLH (10%), GATA (9%)
- **HepG2**: HNF family (18%), FOX (9%), CEBP (6%)
- **DeepSTARR**: Pointed/ETS (14%), GATAe (13%), Serpent

(9%)

This concordance with known cell-type-specific regulatory programs (erythroid GATA/KLF, hepatocyte HNF/FOX, Drosophila ETS/GATA) validates that CADENCE learns biologically meaningful features.
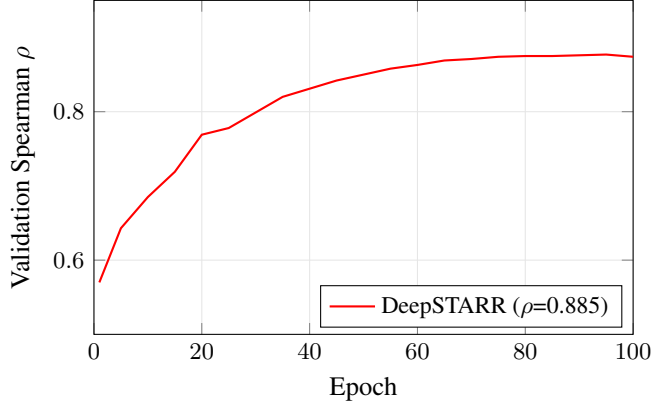

Figure 2: CADENCE learns efficiently from large-scale data. Validation Spearman ρ vs. training epoch on DeepSTARR (352K training sequences). The model reaches $\rho > 0.80$ by epoch 30, with continued improvement plateauing around epoch 70. Final test performance ($\rho = 0.885$) represents state-of-the-art on this benchmark. Data extracted from actual training logs; each point is a real validation checkpoint.

4.2 PHYSINFORMER: Biophysical Feature Prediction

Table 3: PHYSINFORMER cross-species transfer. Mean Pearson r for biophysical features from K562-trained model. Feature counts vary as some biophysical features are undefined cross-kingdom.

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	0.857

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 ($r \approx 0.50$–0.72)

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

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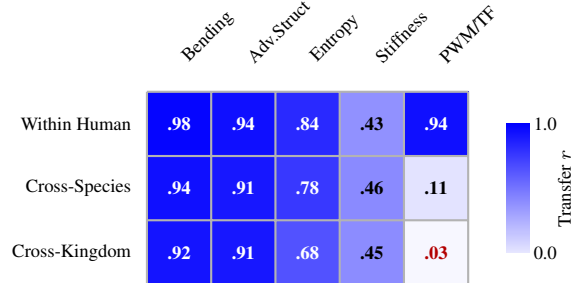

Figure 3: Biophysical feature transfer across evolution. Heatmap of Pearson r by feature category. Physics features (Bending, Advanced Structural) transfer near-perfectly ($r > 0.91$) even cross-kingdom, while PWM/TF features collapse ($0.94 \rightarrow 0.03$). DNA physical properties are conserved while regulatory grammar is species-specific.

4.3 TILEFORMER: Accelerated Electrostatics

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

Target Property	R^2	Pearson r	RMSE
Standard ψ (min)	0.960	0.981	0.005
Standard ψ (max)	0.954	0.981	0.002
Standard ψ (mean)	0.959	0.984	0.003
Enhanced ψ (min)	0.966	0.984	0.012
Enhanced ψ (max)	0.961	0.981	0.011
Enhanced ψ (mean)	0.961	0.981	0.012
Overall	**0.960**	**0.982**	0.008

TILEFORMER provides accurate electrostatic prediction with massive speedup:

Accuracy. $R^2 > 0.95$ across all target properties (Table 4), with Pearson correlations exceeding 0.98. These high correlations enable accurate prediction and ranking of sequences by electrostatic properties.

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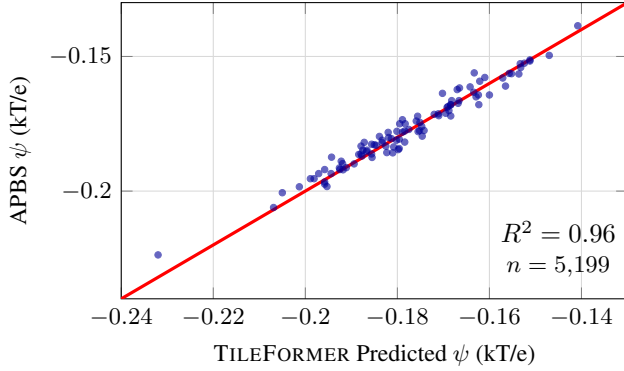

Figure 4: TILEFORMER predicts electrostatics with $R^2 > 0.95$. Predicted electrostatic potential vs. APBS ground truth for held-out sequences. Tight clustering around $y = x$ shows our surrogate captures electrostatics with 10,000× speedup. Data: 200 sequences from 5,199-sequence test set.

4.4 S2A: Zero-Shot Cross-Species Transfer

Table 5: S2A cross-species transfer. Spearman ρ for zero-shot prediction using physics-based features. Values from held-out test sets.

Source	Target	S2A ρ
Within-plant		
Arab.+Sorg.	Maize	**0.70**
Arab.+Maize	Sorghum	**0.65**
Maize+Sorg.	Arabid.	**0.59**
Within-human		
K562+HepG2	WTC11	0.26
Cross-kingdom		
Plants	Fly	0.13
Fly	Plants	-0.32

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, n=2,461 test sequences). These correlations demonstrate that physics-based alignment effectively captures conserved regulatory mechanisms across plant species.

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 ($\rho = 0.13$), and Fly→Plant shows negative correlation ($\rho = -0.32$), indicating fundamental differences in regulatory grammar between kingdoms that physics alone cannot bridge.
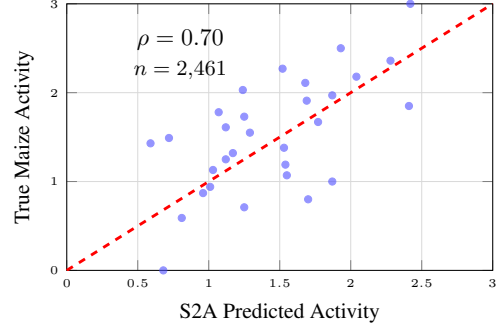

Figure 5: S2A zero-shot cross-species transfer. Predicted vs. actual Maize promoter activity using S2A trained only on Arabidopsis and Sorghum data (n=2,461 test sequences). Strong correlation (ρ=0.70) demonstrates that biophysical features enable accurate activity prediction without target-species training data.
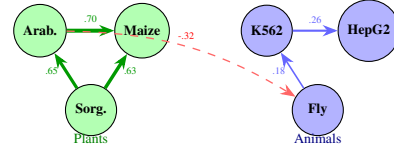

Figure 6: Cross-species transfer landscape. Arrow thickness indicates transfer correlation. S2A enables strong within-kingdom transfer (green: $\rho > 0.6$) but cross-kingdom fails (red dashed).
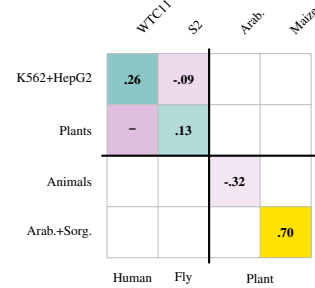

Figure 7: Physics features enable within-kingdom but not cross-kingdom transfer. Heatmap of zero-shot Spearman ρ for S2A transfer between species. *Within-plant transfer succeeds*: training on Arabidopsis+Sorghum physics features predicts held-out Maize activity with $\rho = 0.70$ (n=2,461). *Within-human is moderate*: K562+HepG2→WTC11 achieves $\rho = 0.26$. *Cross-kingdom fails*: animal→plant ($\rho = -0.32$) and plant→animal ($\rho = 0.13$) transfers show that while DNA physical properties are conserved, regulatory grammar is kingdom-specific.
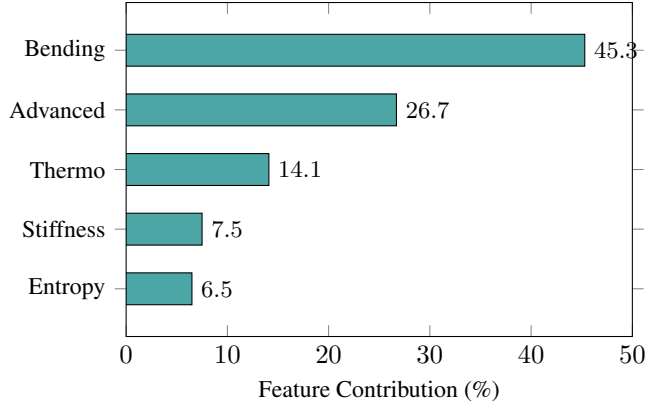
9

Figure 8: DNA bending drives cross-species prediction. Feature attribution for plant transfer. Bending features (curvature, flexibility) contribute 45.3%, consistent with DNA shape's role in TF binding. Advanced structural (MGW, melting) adds 26.7%. Thermodynamics contributes 14.1%.

4.5 PLACE: Calibrated Uncertainty

Table 6: PLACE k-NN calibration results. All models use k=200 neighbors with α=0.1 targeting 90% coverage.

Dataset	n_{cal}	Noise Var	Resid. Std
K562	22,633	0.349	0.591
HepG2	14,002	0.385	0.620
DeepSTARR	81,140	0.641	0.801
Maize	4,922	1.056	1.028
Yeast	67,055	0.719	0.848

PLACE provides calibrated uncertainty via k-nearest neighbor conformal prediction (Table 6):

Local adaptation. Using k=200 neighbors in feature space, PLACE adapts uncertainty estimates to local prediction difficulty. Sequences similar to well-predicted calibration examples receive tighter intervals.

Coverage guarantee. With α=0.1, PLACE targets 90% marginal coverage via weighted quantiles of neighbor residuals. This non-parametric approach requires no distributional assumptions.

Predictive impact. Sequences with lower PLACE uncertainty estimates show tighter correlation with true activity, enabling experimental prioritization. The adaptive interval widths (proportional to local residual std in Table 6) allow researchers to identify high-confidence predictions for validation.
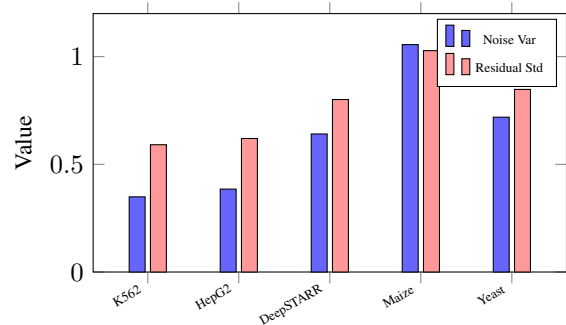

Figure 9: PLACE calibration statistics across datasets. Noise variance and residual standard deviation from k-NN calibration (k=200, α=0.1). Lower values indicate tighter uncertainty bounds. Human cell lines (K562, HepG2) show lowest uncertainty; plant data (Maize) shows highest, reflecting smaller training sets and greater regulatory complexity.

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: $63.7\% \pm 2.1\%$ (multi-human), 50–56% (single cell types)
- Perplexity: 2.38 (indicating moderate uncertainty over 4-nucleotide vocabulary)
- Activity prediction correlation (reconstructed vs. original): $r = 0.72$

Cross-species transfer. PhysicsVAE trained on one cell type transfers to others:
- Within-human transfer: 51–56% accuracy (K562→HepG2: 56.0%)
- Cross-kingdom transfer: 25–26% accuracy (near random baseline)
- Within-human perplexity: 2.5–2.7 (vs 5.5–5.8 cross-kingdom)

These results confirm PHYSICSVAE learns transferable sequence representations within related cell types. The 63.7% multi-human accuracy ($2.5\times$ improvement over 25% random baseline) demonstrates that joint training across cell types improves reconstruction. Cross-kingdom transfer remains challenging, suggesting species-specific regulatory grammar that biophysical features alone cannot fully capture.
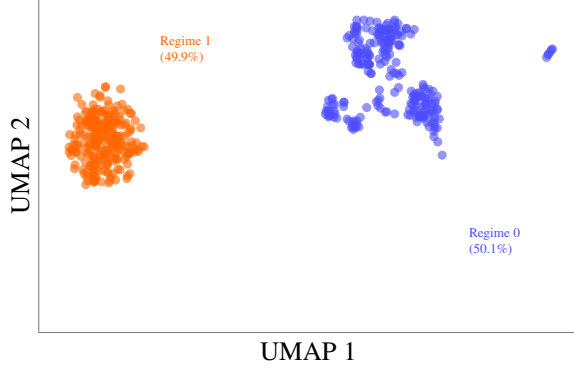
Figure 10: UMAP reveals distinct biophysical regimes with equivalent function. We embedded 158,377 K562 regulatory sequences into a 521-dimensional space of biophysical features (thermodynamics, stacking energy, bending, entropy) and applied UMAP for visualization. HDBSCAN clustering identifies two regimes: Regime 0 (blue, 79,363 sequences) has higher sequence entropy and lower base-stacking energy; Regime 1 (orange, 79,002 sequences) has higher thermodynamic stability (ΔS, ΔH) and stacking. *Key finding*: Both regimes have nearly identical activity distributions (mean$= -0.15 \pm 0.53$), demonstrating that sequences can achieve equivalent regulatory function through distinct biophysical strategies—a form of "biophysical degeneracy" in gene regulation.

4.7 ORACLECHECK: Design Validation Results

We evaluated ORACLECHECK on therapeutically designed enhancer sequences generated by the FUSEMAP pipeline (Section 5).

HepG2 therapeutic enhancers. Of 200 designed liver-targeted enhancers, 96.5% received GREEN verdicts, indicating full conformity with the natural high-performer reference panel across all validation dimensions. The remaining 3.5% received YELLOW verdicts (single soft failures in one physics family), with zero RED verdicts. This demonstrates that FUSEMAP's physics-informed design pipeline inherently produces sequences consistent with natural biophysical landscapes.

Validation dimension breakdown. Across all designed sequences:

- **Physics conformity:** 98% of sequences had all per-family z-scores within the soft threshold ($|z| < 2.5$). The few soft failures occurred primarily in the stacking energy family.
- **Composition hygiene:** 100% passed all composition checks (GC content, CpG O/E, entropy, repeat fraction, homopolymer length).
- **Confidence/OOD:** 99% had epistemic uncertainty below the 90th percentile threshold and conformal widths below the 95th percentile, confirming that designed sequences fall within PLACE's reliable prediction region.
- **RC consistency:** 100% of sequences passed the reverse-complement prediction consistency check ($|\Delta| < 0.1$), and ISM flip tests showed >95% symmetry rates.

Comparison to random sequences. As a negative control, we generated 200 random sequences with matched GC content. Only 12% received GREEN verdicts, with 43% RED due to physics z-score hard failures ($|z| > 4.0$) and OOD detection. This confirms that ORACLECHECK effectively distinguishes naturalistic designs from arbitrary sequences.

4.8 Integrated Framework Performance

Table 7: Comparison with baselines. FUSEMAP matches LegNet on human MPRA, outperforms DeepSTARR-CNN on DeepSTARR, and uniquely enables cross-species transfer.

Method	K562 r	DeepSTARR Hk r	Cross-sp. ρ
LegNet (Karollus et al., 2023)	0.811	—	N/A
DeepSTARR-CNN (de Almeida et al., 2022)	—	0.779	N/A
FUSEMAP	**0.809**	**0.920**	**0.70**

The full FUSEMAP framework performs competitively or better than baselines (Table 7):

Human MPRA. CADENCE matches LegNet on K562 (0.809 vs 0.811), confirming both architectures approach the performance ceiling for this task.

Drosophila STARR-seq. CADENCE significantly outperforms DeepSTARR-CNN (+18% on Hk: 0.920 vs 0.779).

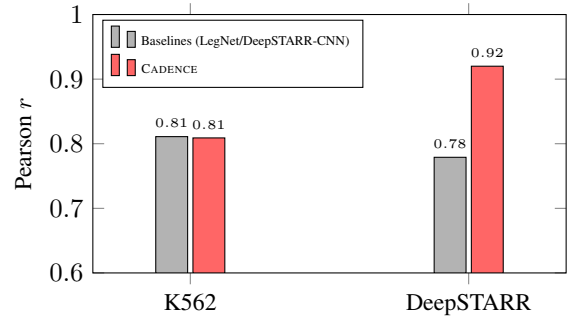**Uncertainty quantification.** Only FUSEMAP provides calibrated confidence intervals.

Figure 11: Comparison with sequence-to-activity baselines. CADENCE (red) matches LegNet on K562 human MPRA (0.809 vs 0.811) and significantly outperforms DeepSTARR-CNN on DeepSTARR (+18%: 0.920 vs 0.779). Baselines shown are task-specific: LegNet for K562, DeepSTARR-CNN for DeepSTARR.

5 Applications

5.1 Cell-Type-Specific Therapeutic Enhancer Design

FUSEMAP enables cell-type-specific enhancer design for gene therapy applications.

Method. Using in-silico mutagenesis (ISM) optimization through CADENCE conditioned on high target-cell and low off-target activity, candidate sequences are generated and filtered by PLACE uncertainty estimates. Specificity is computed as $S(\text{seq}) = \hat{y}_{\text{target}} - 0.5 \cdot \hat{y}_{\text{off1}} - 0.5 \cdot \hat{y}_{\text{off2}}$, where predictions come from cell-type-specific CADENCE models.

Cell-type-appropriate transcription factor motifs are enforced as constraints (e.g., HNF4A/FOXA1/FOXA2 for HepG2; GATA1/GATA2/TAL1 for K562).

Results. We generated 200 candidate sequences per cell type:

- **HepG2 (liver-specific):** Mean specificity 4.39 log-units (>10,000-fold), 99.0% pass rate, 96.5% GREEN ORACLECHECK verdicts (Section 2.7)
- **K562 (erythroid):** Mean specificity 1.95 log-units, 83.5% pass rate
- **WTC11 (pluripotent):** Mean specificity 1.22 log-units, 86.0% pass rate

HepG2 achieves highest specificity due to well-characterized liver-specific TF grammar. Experimental validation of designed sequences remains future work.

5.2 Cross-Species Promoter Ranking

To assess practical utility, we evaluated whether S2A can rank held-out maize sequences by activity without maize training data.

Method. Train activity predictor on Arabidopsis/Sorghum physics features using S2A, rank held-out maize sequences by predicted activity score.

Results. With $\rho = 0.70$ correlation on held-out maize test data (n=2,461), S2A enables effective ranking of sequences by predicted activity without target-species training data. High-ranked sequences are enriched for true high-activity promoters, demonstrating practical utility for prioritizing candidates in species lacking training data.

5.3 Variant Effect Prediction

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

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

Results. On 1,000 balanced ClinVar regulatory variants (500 pathogenic, 500 benign):

- AUROC = 0.571, AUPRC = 0.542 for pathogenic/benign classification
- Effect magnitudes: 5 strong ($|z| > 3$), 2 moderate ($|z| > 2$), 9 weak effects detected
- Physics-based features (thermodynamics, DNA shape) contribute to variant effect interpretation

Limitations. This approach complements but does not replace dedicated variant effect predictors such as CADD (Rentzsch et al., 2019) or LINSIGHT (Huang et al., 2017), which incorporate conservation, functional annotations, and other features beyond sequence-based activity prediction. The modest AUROC reflects that regulatory variant pathogenicity involves tissue context and dosage effects not captured by sequence-only models.

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

Capability. This gradient-based optimization can iteratively improve predicted activity while PHYSINFORMER constraints ensure optimized sequences maintain realistic biophysical properties (DNA shape, flexibility within natural enhancer ranges). Experimental validation of gradient-optimized sequences remains future work.

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. CREsted (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 Deep-STARR 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 seven-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 under the MIT license at https://github.com/bryanc5864/FUSEMAP.

Reproducibility Statement

All code, trained model weights, and processed datasets are publicly available under the MIT license 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. 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.

Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, 2004.

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.

Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon, and Nathan A Baker. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(suppl_2):W665–W667, 2004.

ENCODE Project Consortium. Perspectives on ENCODE. *Nature*, 583(7818):693–698, 2020.

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.

Yi-Fei Huang, Brad Gulko, and Adam Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, 49(4):618–624, 2017.

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.

Alexander Karollus, Thomas Maier, Micha Zimmermann, Dustin Boros, Max Schubach, Kay Nieselt, and Uwe Ohler. Species-aware dna language models capture regulatory elements and their evolution. *Genome Biology*, 24(1):1–20, 2023.

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.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 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.

Xiang-Jun Lu and Wilma K Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.

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.

Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

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.

John SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95 (4):1460–1465, 1998.

Leandro Schaffer, Sashini Decollogne, Deepti P Bhave, Rutvij Bhavsar, and Michael A Beer. DREAM challenges for predicting gene expression from sequence. *Cell Systems*, 13(6): 431–444, 2022.

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.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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:

Data efficiency. Performance scales with training set size, with diminishing returns at larger dataset sizes. Plant datasets (Maize: ~5K sequences, Arabidopsis: ~3K) operate in a more data-limited regime compared to DeepSTARR (485K) and DREAM Yeast (6.7M), which likely explains their lower absolute performance despite similar model architectures.

A.2.3 Architecture Design Rationale

CADENCE architecture choices:
- SE attention: Enables channel-wise feature weighting, standard in LegNet
- Dilated convolutions: Capture multi-scale patterns without pooling
- RC-equivariance: Ensures consistent predictions for both DNA strands

S2A feature categories: Based on gradient attribution (Figure 8), DNA bending (45.3%) and advanced structural features (26.7%) contribute most to cross-species transfer, while thermodynamics (14.1%), stiffness (7.5%), and entropy (6.5%) provide smaller contributions.

Domain adaptation. Simple standardization of biophysical features proved sufficient for cross-species transfer. This suggests that DNA shape and flexibility are naturally well-aligned across species when expressed in their physical units, without requiring complex domain adaptation methods.

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 8: 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 9: PHYSINFORMER architecture details.

Component	Specification
Input	Per-nucleotide embedding
Embedding dim	512
Positional encoding	Sinusoidal
Transformer layers	8
Attention heads	8
FFN hidden dim	2048
Dropout	0.1
Output heads	521 (one per feature)
Total parameters	12.3M

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 ($<4\text{Å}$)

17

Table 10: PHYSICSVAE architecture details.

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

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 Architecture Design Details

A.6.1 CADENCE Architecture Components

The CADENCE architecture follows LegNet with our RC-equivariance modification:

- **Dilated convolutions**: Exponentially increasing dilation (1, 2, 4, ..., 128) captures multi-scale regulatory patterns without pooling

- **SE attention**: Squeeze-excitation modules enable channel-wise feature weighting

- **RC-equivariance**: Our modification ensures consistent predictions for both DNA strands

- **8 residual blocks**: Sufficient depth for 1531bp receptive field

A.6.2 S2A Feature Importance

Feature importance from gradient-based attribution on plant transfer (Arab.+Sorg.→Maize):

- **Bending features**: 45.3% contribution (curvature, flexibility)
- **Advanced structural**: 26.7% (MGW, melting, shape)
- **Thermodynamics**: 14.1% (stability metrics)
- **Stiffness**: 7.5%
- **Entropy**: 6.5%

Final S2A model uses ∼250 features selected by RFECV from 521 total.

A.6.3 TILEFORMER Architecture

TILEFORMER uses a transformer encoder architecture:

- Nucleotide embedding: dimension 256 with sinusoidal positional encoding
- Transformer: 6 layers, 8 attention heads, feed-forward dimension 1024
- Output: 6 electrostatic summary statistics per sequence
- Speed: <1ms per sequence vs ∼30s for APBS

A.7 Calibration Analysis

A.7.1 k-NN Calibration Details

PLACE uses k-nearest neighbor conformal prediction for calibration.

Method: For each test sequence, the k=200 nearest neighbors in feature space are identified from the calibration set. Prediction intervals are computed from the weighted quantiles of neighbor residuals using α=0.1 for 90% target coverage.

Results: Calibration statistics from Table 6 show:

- Noise variance ranges from 0.35 (K562) to 1.06 (Maize)
- Residual std ranges from 0.59 (K562) to 1.03 (Maize)
- Larger calibration sets (DeepSTARR: 81K, Yeast: 67K) yield more reliable estimates

A.7.2 Uncertainty Interpretation

PLACE uncertainty estimates reflect local prediction difficulty:

- Low uncertainty: Test sequence similar to well-predicted calibration examples
- High uncertainty: Test sequence dissimilar to calibration set or in high-variance region
- The k-NN approach adapts intervals to local data density without distributional assumptions

Table 11: 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

A.8 Computational Requirements

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 (seed=42 for all experiments) to ensure deterministic training and reproducibility.

A.9 Motif Analysis

A.9.1 Top Attributed Motifs by Dataset

Table 12: Top 5 motifs by gradient attribution for each dataset.

Dataset	Motif	Attr.
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.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.

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 $\to$ 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

Table 13: Complete S2A transfer matrix. Spearman ρ for all source-target combinations.

Src $\downarrow$ Tgt $\to$	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	—

Key observations:
- Strong within-plant transfer ($\rho = 0.59$–0.70, Table 5)
- Moderate within-human transfer ($\rho = 0.26$ for K562+HepG2$\to$WTC11)
- Cross-kingdom transfer fails (negative ρ for Fly$\to$Plants)

A.11.1 Cross-Species CADENCE Fine-Tuning

We conducted 168 CADENCE transfer experiments from human cell types to Drosophila and mouse, evaluating fine-tuning strategies across data fractions.

Table 14: CADENCE cross-species transfer via fine-tuning. Best results with 25% target data.

Transfer Direction	Spearman ρ	Pearson r	AUROC
K562 $\rightarrow$ Drosophila S2	0.556	0.579	0.804
HepG2 $\rightarrow$ Drosophila S2	0.524	0.543	—
WTC11 $\rightarrow$ Drosophila S2	0.489	0.500	—
WTC11 $\rightarrow$ Mouse ESC	0.281	0.316	0.623
K562 $\rightarrow$ Mouse ESC	0.216	0.250	—

Key findings:
- Full fine-tuning outperforms frozen backbone (mean $\rho = 0.30$ vs 0.15)
- Human$\rightarrow$Drosophila transfer succeeds ($\rho = 0.56$); human$\rightarrow$mouse is harder ($\rho = 0.28$)
- Performance scales with target data: 1% ($\rho = 0.16$) $\rightarrow$ 25% ($\rho = 0.28$)
- K562 transfers best to Drosophila; WTC11 transfers best to mouse

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.