Sequence Data

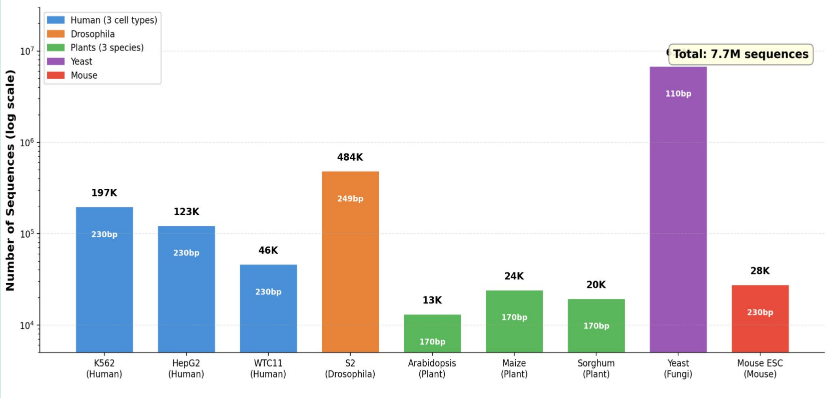
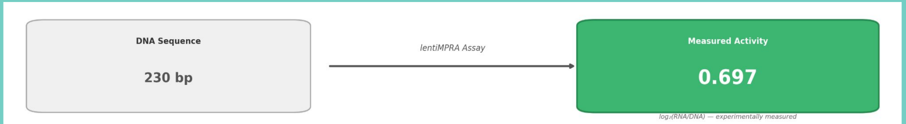
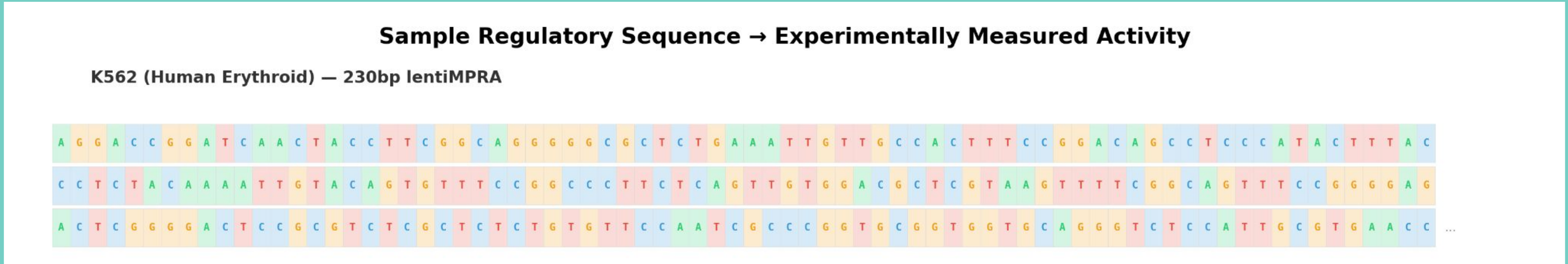


Figure 22,23,24. Dataset summary of 7.5+ million sequences across seven species: human cell lines (K562, HepG2, WTC11)(ENCODE Project Consortium, 2020; Inoue et al., 2017), Drosophila S2(de Almeida et al., 2022; Arnold et al., 2013), plants (Arabidopsis, Maize, Sorghum)(Jores et al., 2021), yeast(de Boer et al., 2020; Schreiber et al., 2020), and mouse ESC(Kalkan et al., 2017), with sequence counts per dataset. **(Figure generated by student author using python, matplotlib)**





Sample Regulatory Sequence → Experimentally Measured Activity

K562 (Human Erythroid) — 230bp lentiMPRA

A G G A C C G G A T C A A C T A C C T T C G G C A G G G G G C G C T C T G A A A T T G T T G C C A C T T T C C G G A C A G C C T C C C A T A C T T T A C
C C T C T A C A A A A T T G T A C A G T G T T T C C G G C C C T T C T C A G T T G T G G A C G C T C G T A A G T T T T C G G C A G T T T C C G G G G A G
A C T C G G G G A C T C C G C G T C T C G C T C T C T G T G T T C C A A T C G C C C G G T G C G G T G G T G C A G G G T C T C C A T T G C G T G A A C C ...

Biophysics Training Data

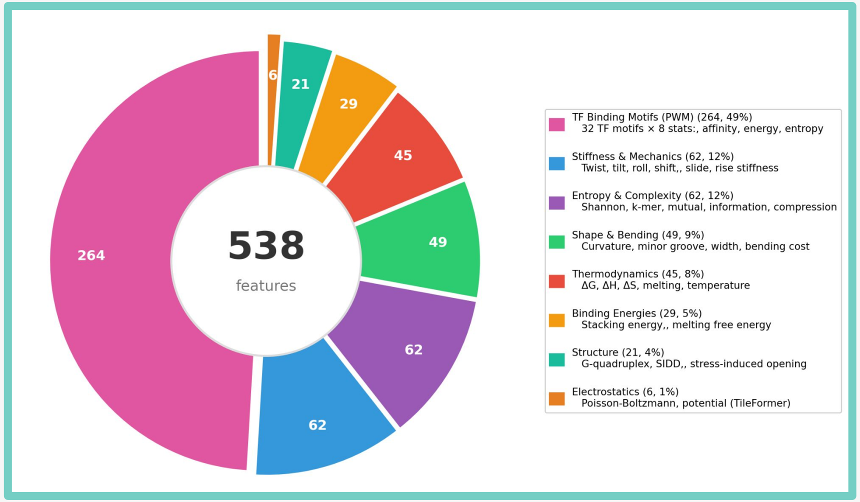


Figure 25. Feature breakdown with TF features and pure biophysical features**(Figure generated by student author using python, matplotlib)**

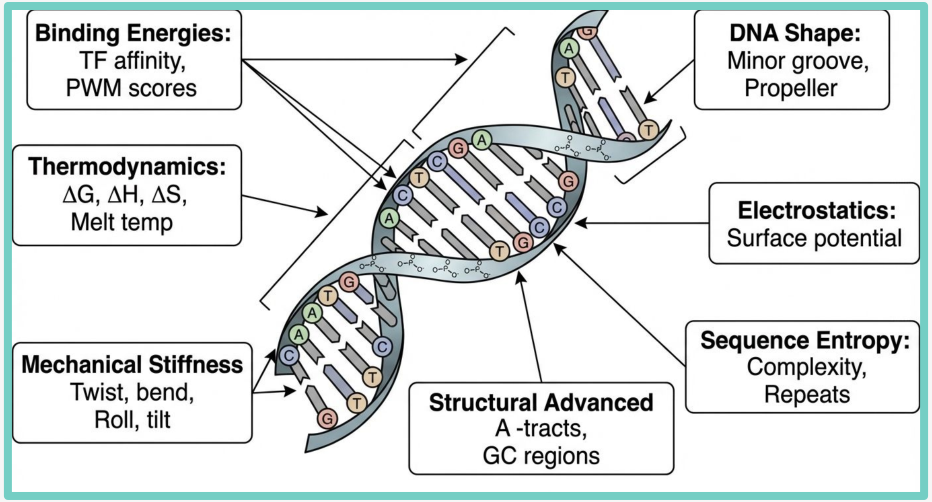
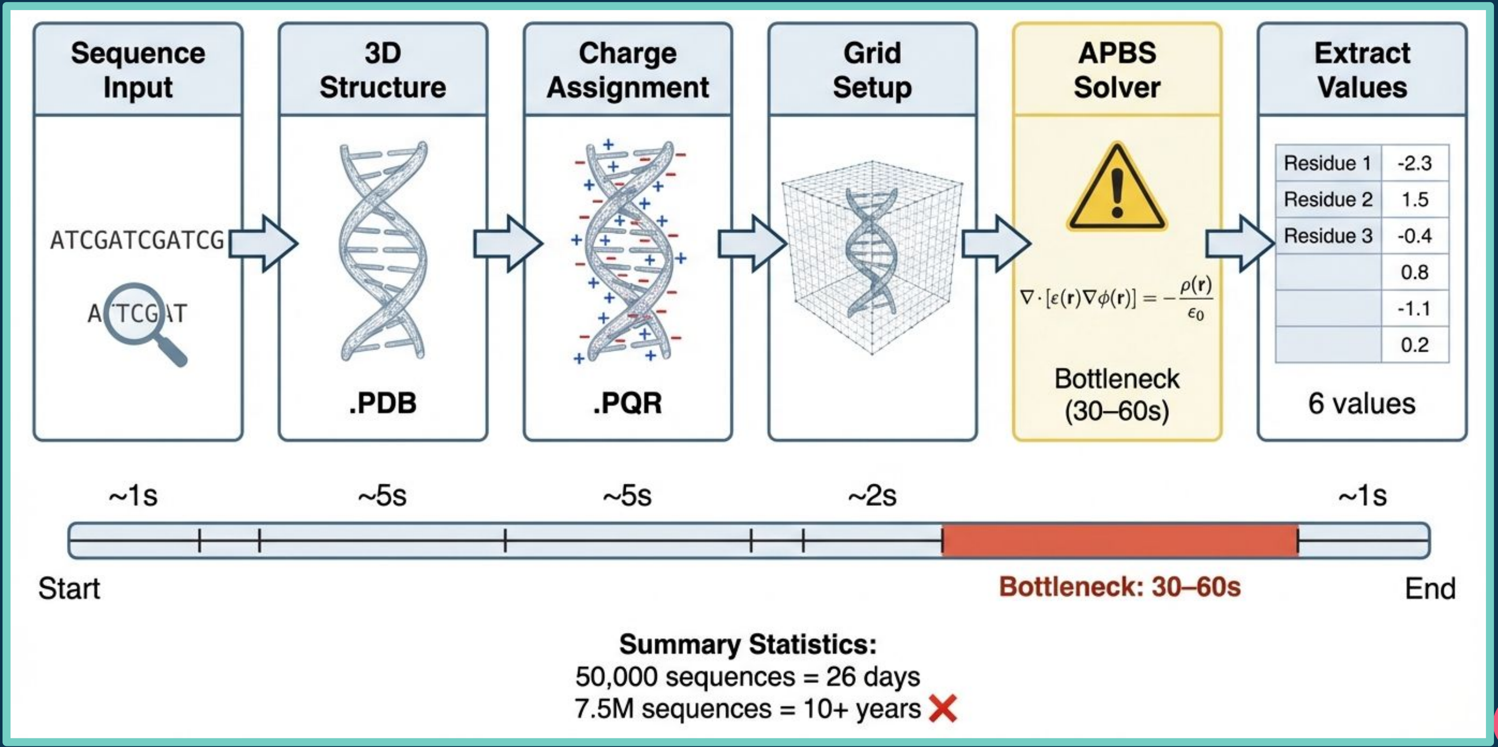


Figure 26. Representation of DNA and where the features are found physically**(Figure generated by student author using FigureLabs)**

Problem: Manual Labelling



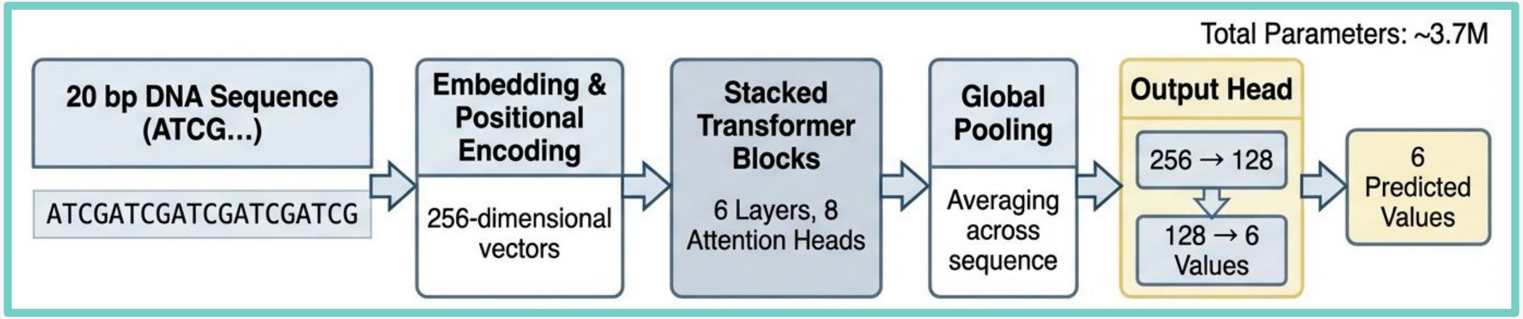
$$\nabla \cdot [\varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -\frac{\rho(\mathbf{r})}{\epsilon_0}$$

Residue 1	-2.3
Residue 2	1.5
Residue 3	-0.4
	0.8
	-1.1
	0.2

Summary Statistics:
50,000 sequences = 26 days
7.5M sequences = 10+ years ❌

Figure 27: Manually labelling with solving Poisson-Boltzmann is time-consuming**(Figure generated by student author using FigureLabs)**

Solution: TileFormer



Total Parameters: ~3.7M

20 bp DNA Sequence (ATCG...)

ATCGATCGATCGATCGATCG

Embedding & Positional Encoding

256-dimensional vectors

Stacked Transformer Blocks

6 Layers, 8 Attention Heads

Global Pooling

Averaging across sequence

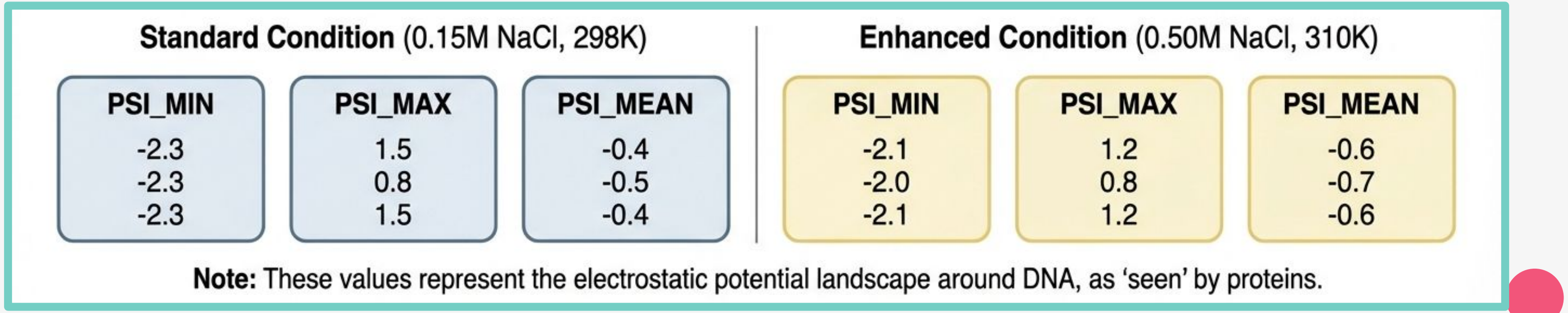
Output Head

256 → 128

128 → 6 Values

6 Predicted Values

Figure 28. TileFormer architecture diagram from input to predicted values**(Figure generated by student author using FigureLabs)**



Standard Condition (0.15M NaCl, 298K)			**Enhanced Condition** (0.50M NaCl, 310K)		
PSI_MIN	**PSI_MAX**	**PSI_MEAN**	**PSI_MIN**	**PSI_MAX**	**PSI_MEAN**
-2.3	1.5	-0.4	-2.1	1.2	-0.6
-2.3	0.8	-0.5	-2.0	0.8	-0.7
-2.3	1.5	-0.4	-2.1	1.2	-0.6

Note: These values represent the electrostatic potential landscape around DNA, as 'seen' by proteins.

Figure 29. TileFormer output breakdown**(Figure generated by student author using FigureLabs)**

Solution 2: PhysInformer

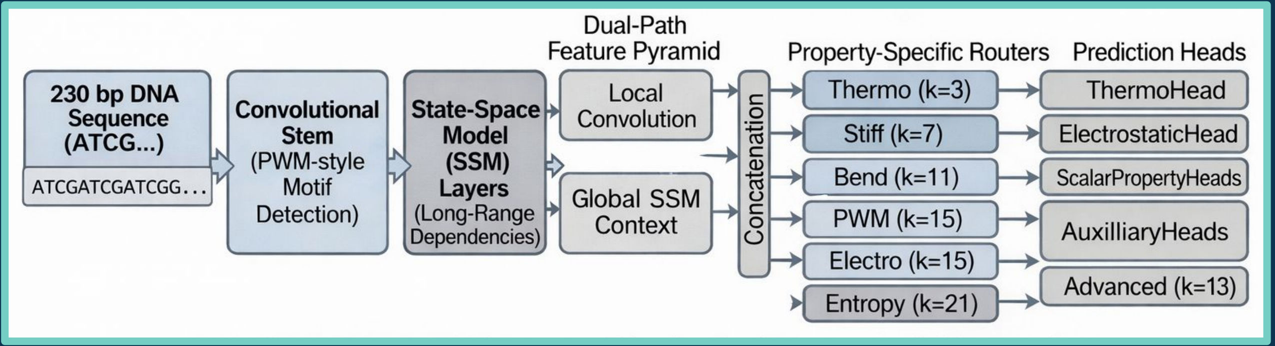
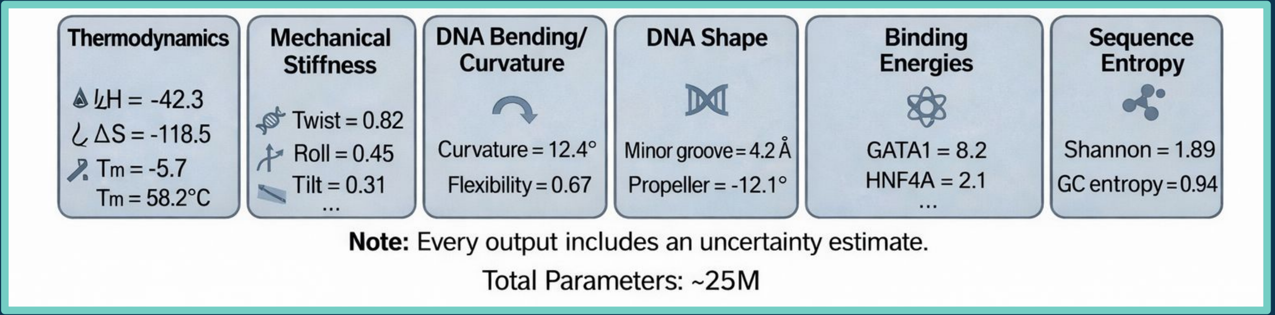


Figure 30. PhysInformer architecture diagram from input to predicted values**(Figure generated by student author using FigureLabs)**



Note: Every output includes an uncertainty estimate.

Total Parameters: ~25M

Figure 31. PhysInformer output breakdown**(Figure generated by student author using FigureLabs)**

CADENCE: **C**onvolutional **A**rchitecture for **D**NA **EN**hancer **C**haracterization and **E**xpression prediction

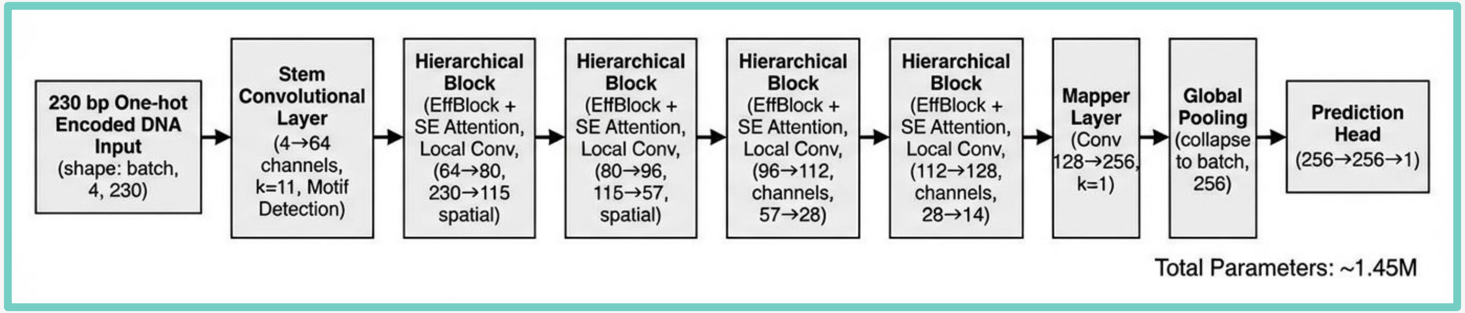
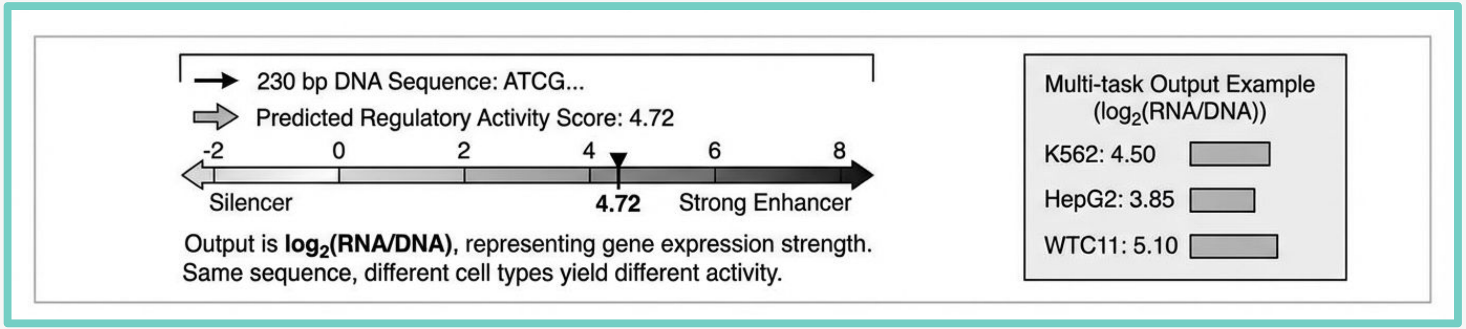

230 bp One-hot Encoded DNA Input (shape: batch, 4, 230)
Stem Convolutional Layer (4→64 channels, k=11, Motif Detection)
Hierarchical Block (EffBlock + SE Attention, Local Conv, (64→80, 230→115 spatial)
Hierarchical Block (EffBlock + SE Attention, Local Conv, (80→96, 115→57, spatial)
Hierarchical Block (EffBlock + SE Attention, Local Conv, (96→112, channels, 57→28)
Hierarchical Block (EffBlock + SE Attention, Local Conv, (112→128, channels, 28→14)
Mapper Layer (Conv 128→256, k=1)
Global Pooling (collapse to batch, 256)
Prediction Head (256→256→1)
Total Parameters: ~1.45M

Figure 32. CADENCE architecture diagram from input to predicted values**(Figure generated by student author using FigureLabs)**


230 bp DNA Sequence: ATCG...
Predicted Regulatory Activity Score: 4.72
-2
0
2
4
6
8
Silencer
4.72
Strong Enhancer
Multi-task Output Example ($\log_2$(RNA/DNA))
K562: 4.50
HepG2: 3.85
WTC11: 5.10

Output is **$\log_2$(RNA/DNA)**, representing gene expression strength. Same sequence, different cell types yield different activity.

Figure 33. CADENCE output breakdown**(Figure generated by student author using FigureLabs)**

PLACE: Post-hoc **L**aplace **A**nd **C**onformal **E**stimation

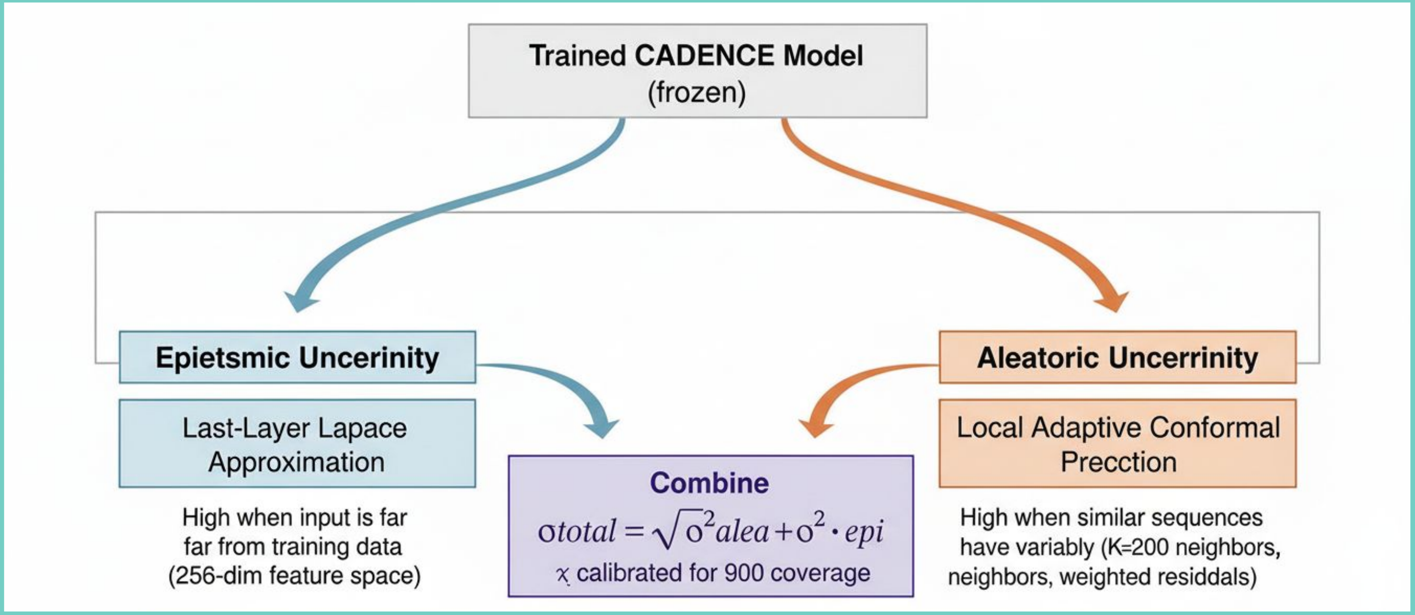
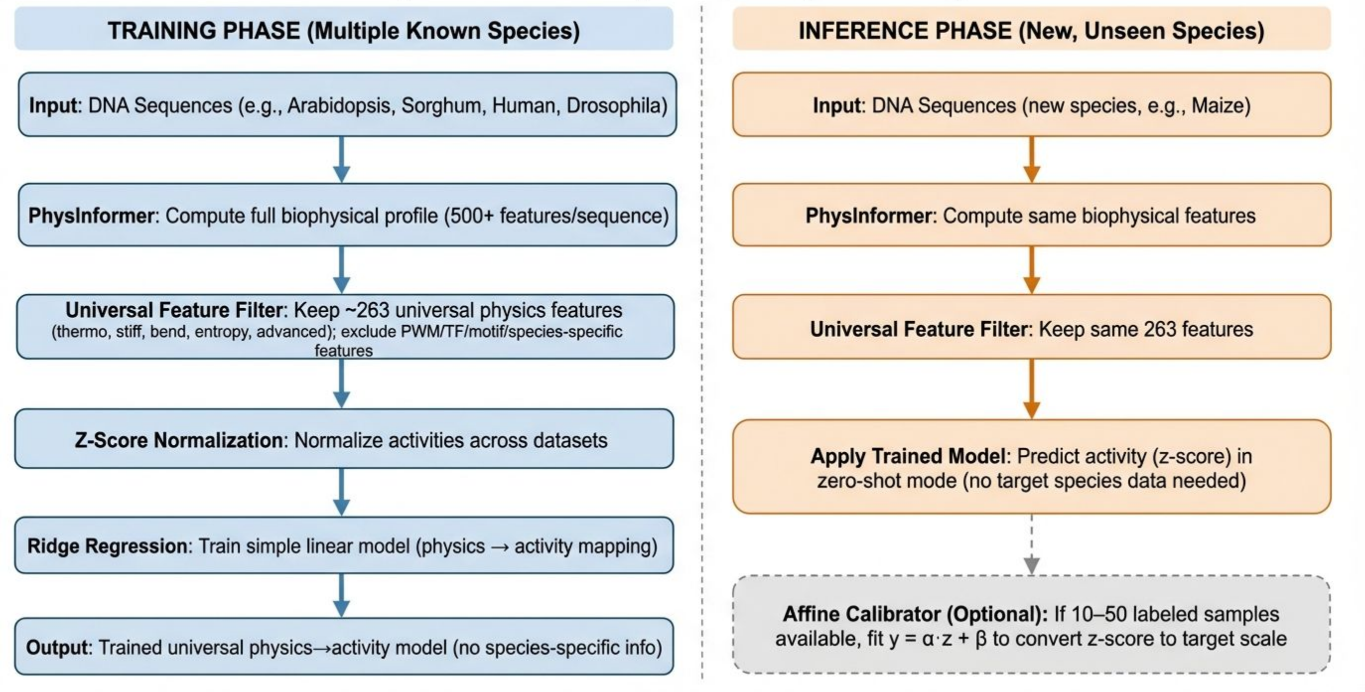

Trained CADENCE Model
(frozen)
Epietsmic Uncerinity
Last-Layer Lapace Approximation
High when input is far far from training data (256-dim feature space)
Combine
$\sigma total = \sqrt{\sigma^2 alea + \sigma^2 \cdot epi}$
χ calibrated for 900 coverage
Aleatoric Uncerrinity
Local Adaptive Conformal Precction
High when similar sequences have variably (K=200 neighbors, neighbors, weighted residdals)

Figure 34. PLACE framework diagram: frozen CADENCE model splits into two parallel paths—epistemic uncertainty via last-layer Laplace approximation ($\Sigma = (\lambda I + \sigma^{-2}\Phi^{T}\Phi)^{-1}$) and aleatoric uncertainty via K-NN conformal prediction—combined as σ_total = √(σ^2_alea + $\kappa\cdot\sigma^2$_epi) with κ calibrated for 90% coverage. **(Figure generated by student author using FigureLabs)**

S2A: Sequence to Activity



TRAINING PHASE (Multiple Known Species)

Input: DNA Sequences (e.g., Arabidopsis, Sorghum, Human, Drosophila)

PhysInformer: Compute full biophysical profile (500+ features/sequence)

Universal Feature Filter: Keep ~263 universal physics features (thermo, stiff, bend, entropy, advanced); exclude PWM/TF/motif/species-specific features

Z-Score Normalization: Normalize activities across datasets

Ridge Regression: Train simple linear model (physics → activity mapping)

Output: Trained universal physics→activity model (no species-specific info)

INFERENCE PHASE (New, Unseen Species)

Input: DNA Sequences (new species, e.g., Maize)

PhysInformer: Compute same biophysical features

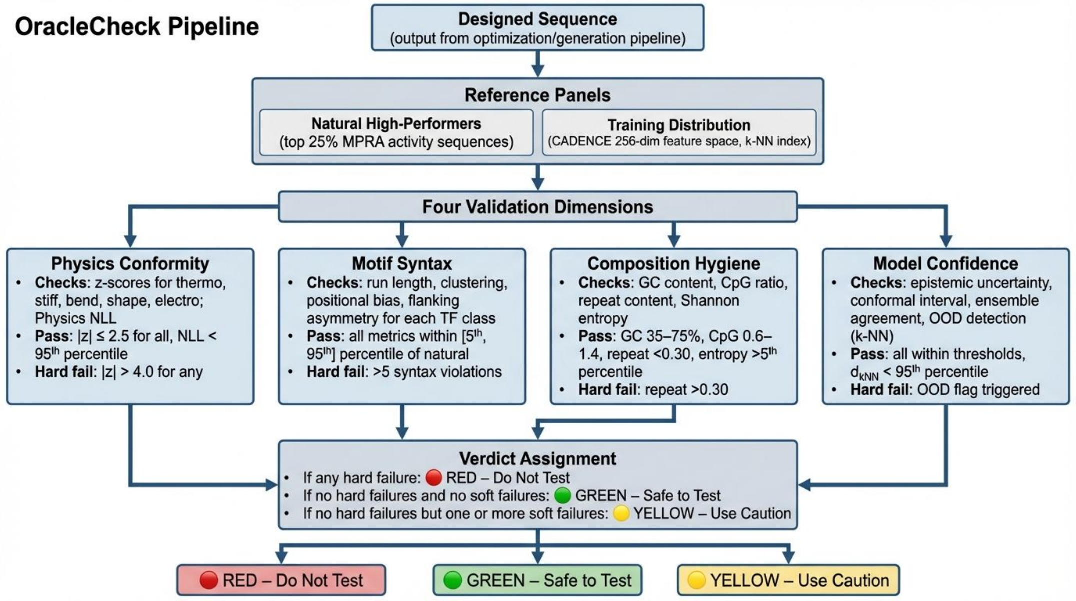
Universal Feature Filter: Keep same 263 features

Apply Trained Model: Predict activity (z-score) in zero-shot mode (no target species data needed)

Affine Calibrator (Optional): If 10–50 labeled samples available, fit y = α·z + β to convert z-score to target scale

Figure 35. S2A pipeline flowchart: training species sequences → PhysInformer → universal feature filter (keep ~263 physics features, exclude PWM/TF binding) → z-score normalization → Ridge regression → trained model applied zero-shot to held-out species → activity prediction without target-species data. **(Figure generated by student author using FigureLabs)**

OracleCheck



OracleCheck Pipeline

Designed Sequence
(output from optimization/generation pipeline)

Reference Panels

Natural High-Performers
(top 25% MPRA activity sequences)

Training Distribution
(CADENCE 256-dim feature space, k-NN index)

Four Validation Dimensions

Physics Conformity

- **Checks**: z-scores for thermo, stiff, bend, shape, electro; Physics NLL
- **Pass**: $|z| \leq 2.5$ for all, NLL < 95th percentile
- **Hard fail**: $|z| > 4.0$ for any

Motif Syntax

- **Checks**: run length, clustering, positional bias, flanking asymmetry for each TF class
- **Pass**: all metrics within [5th, 95th] percentile of natural
- **Hard fail**: >5 syntax violations

Composition Hygiene

- **Checks**: GC content, CpG ratio, repeat content, Shannon entropy
- **Pass**: GC 35–75%, CpG 0.6–1.4, repeat <0.30, entropy >5th percentile
- **Hard fail**: repeat >0.30

Model Confidence

- **Checks**: epistemic uncertainty, conformal interval, ensemble agreement, OOD detection (k-NN)
- **Pass**: all within thresholds, d_{kNN} < 95th percentile
- **Hard fail**: OOD flag triggered

Verdict Assignment

- If any hard failure: 🔴 RED – Do Not Test
- If no hard failures and no soft failures: 🟢 GREEN – Safe to Test
- If no hard failures but one or more soft failures: 🟡 YELLOW – Use Caution

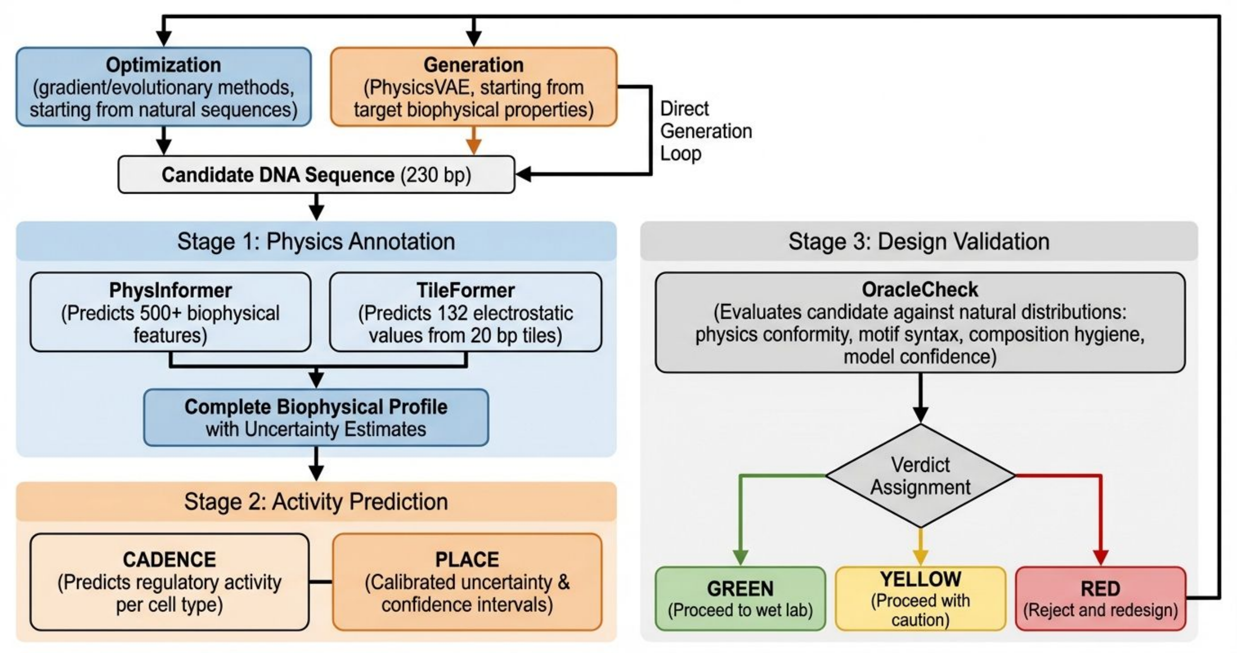
🔴 RED – Do Not Test

🟢 GREEN – Safe to Test

🟡 YELLOW – Use Caution

Figure 36. OracleCheck validation flowchart: designed sequence evaluated across four dimensions—(1) Physics Conformity (|z| ≤ 2.5), (2) Motif Syntax ([5th, 95th] percentile), (3) Composition Hygiene (35% ≤ GC ≤ 75%, repeat < 30%), (4) Model Confidence (σ_epi < 90th %ile, OOD check)—yielding GREEN (pass all), YELLOW (1 soft fail), or RED (hard fail) verdict. **(Figure generated by student author using FigureLabs)**

Integrated Pipeline



Optimization (gradient/evolutionary methods, starting from natural sequences)
Generation (PhysicsVAE, starting from target biophysical properties)
Direct Generation Loop
Candidate DNA Sequence (230 bp)
Stage 1: Physics Annotation
PhysInformer (Predicts 500+ biophysical features)
TileFormer (Predicts 132 electrostatic values from 20 bp tiles)
Complete Biophysical Profile with Uncertainty Estimates
Stage 2: Activity Prediction
CADENCE (Predicts regulatory activity per cell type)
PLACE (Calibrated uncertainty & confidence intervals)
Stage 3: Design Validation
OracleCheck (Evaluates candidate against natural distributions: physics conformity, motif syntax, composition hygiene, model confidence)
Verdict Assignment
GREEN (Proceed to wet lab)
YELLOW (Proceed with caution)
RED (Reject and redesign)

Figure 37. FUSEMAP integrated loop diagram: candidate sequence → PhysInformer + TileFormer (physics annotation) → CADENCE + PLACE (activity prediction with uncertainty) → OracleCheck (naturality validation) → GREEN sequences proceed to experimental validation; RED sequences loop back to redesign. PhysicsVAE shown as alternative entry point for physics-conditioned sequence generation. **(Figure generated by student author using FigureLabs)**