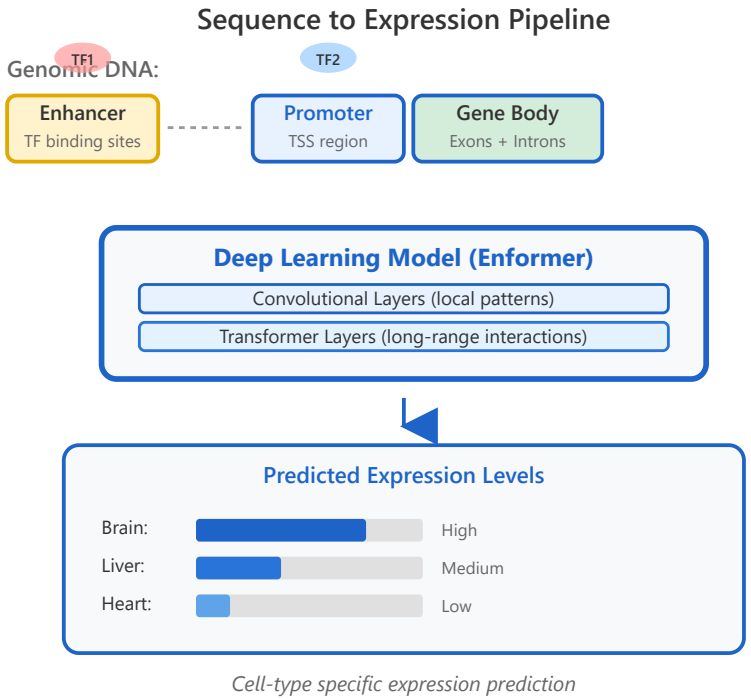
Gene Expression Prediction



Cell-type specific expression prediction

Sequence to expression
DNA → RNA abundance mapping

Promoter models
TSS region activity prediction

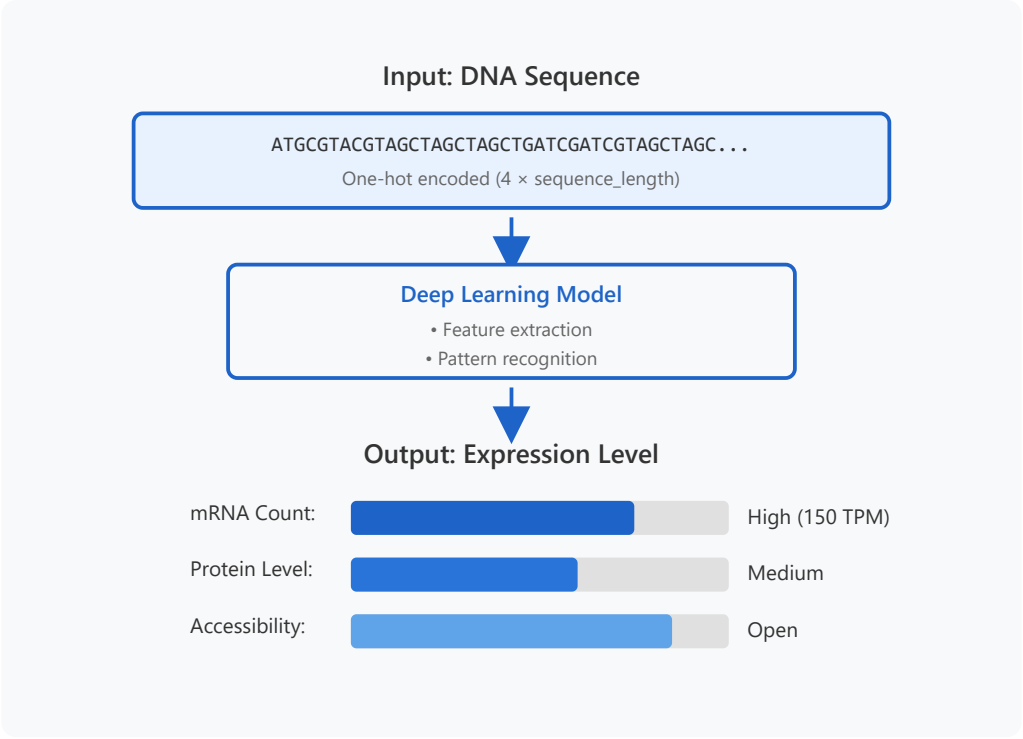
Enhancer grammar
TF binding syntax learning

Cell type specificity
Context-dependent prediction

Enformer architecture
Transformer + CNN hybrid model

1. Sequence to Expression

Mapping DNA sequences to RNA abundance levels



Overview

Sequence-to-expression models predict gene activity directly from DNA sequence. These models learn the complex regulatory code that determines when and where genes are expressed.

Input Features

Models take raw DNA sequences as input, typically encoded as one-hot vectors representing the four nucleotides (A, T, G, C). The sequence context can span from hundreds to hundreds of thousands of base pairs.

Output Predictions

Models output quantitative predictions of gene expression levels, which can include mRNA abundance (e.g., TPM, FPKM), chromatin accessibility (e.g., ATAC-seq), histone modifications (e.g., ChIP-seq), or protein levels.

Key Applications & Challenges

- **Variant Effect Prediction:** Predict how genetic variants affect gene expression (eQTLs)
- **Therapeutic Design:** Design synthetic regulatory elements for gene therapy
- **Disease Mechanisms:** Understand regulatory disruptions in disease states
- **Challenge:** Long-range interactions can span megabases, requiring large context windows
- **Challenge:** Cell-type specific regulation requires integrated models with epigenetic features

2. Promoter Models

Predicting transcription start site (TSS) region activity

Promoter Architecture

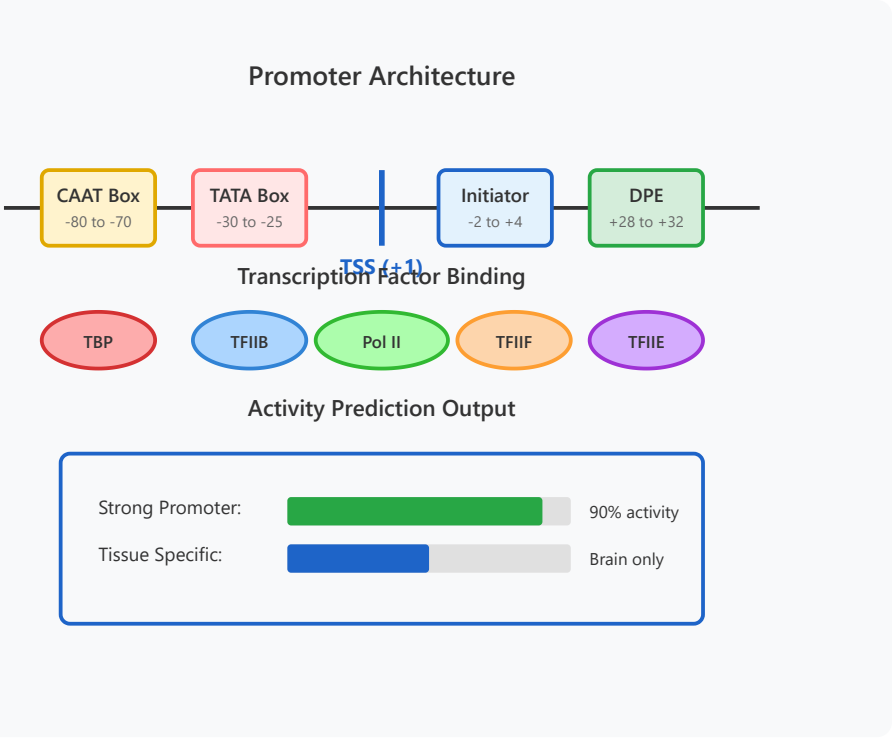
Promoters are regulatory DNA regions located upstream of genes, typically spanning ~1kb around the transcription start site (TSS). They contain core elements (TATA box, Initiator, DPE) and binding sites for general transcription factors and RNA polymerase II.

Model Approaches

Early models focused on position weight matrices (PWMs) for transcription factor binding sites. Modern deep learning approaches use CNNs to automatically learn motifs and their combinations, capturing complex syntax rules.

Prediction Tasks

Models predict TSS activity strength, directionality, tissue-specificity, and response to transcription factors. Advanced models can predict effects of promoter mutations and design synthetic promoters with desired properties.

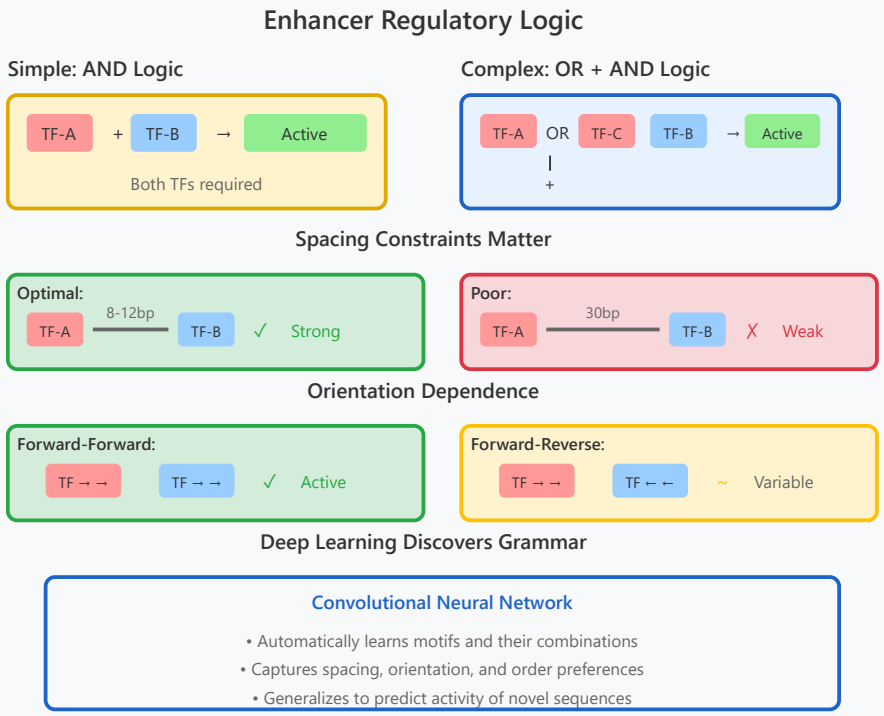


Key Insights & Applications

- **Core Promoter Elements:** TATA box, Initiator (Inr), and Downstream Promoter Element (DPE) determine basal transcription
- **Proximal Elements:** CAAT box and GC box enhance promoter activity
- **Synthetic Biology:** Design optimized promoters for gene expression systems
- **Disease Variants:** Predict how mutations in promoter regions affect gene expression
- **Recent Models:** ProCapNet, Xpresso, and ExPecto achieve high accuracy on human promoters

3. Enhancer Grammar

Learning the syntax of transcription factor binding



What is Enhancer Grammar?

Enhancer grammar refers to the rules governing how transcription factor binding sites combine to produce regulatory activity. Like linguistic grammar, it involves syntax (arrangement), semantics (meaning), and context-dependence.

Combinatorial Logic

Enhancers integrate signals from multiple transcription factors through boolean-like logic gates. Common patterns include AND gates (both TFs required), OR gates (either TF sufficient), and NOT gates (repressive interactions).

Spatial Constraints

The spacing and orientation between TF binding sites critically affects function. Optimal spacing allows protein-protein interactions, while poor spacing disrupts cooperative binding. Different TF pairs have characteristic preferred spacings.

Learning Grammar Rules

Deep learning models, particularly CNNs, automatically discover enhancer grammar from sequence and activity data. They learn motifs, their combinations, and higher-order syntax without explicit feature engineering.

Key Concepts & Applications

- **Motif Interactions:** TF binding sites work cooperatively or antagonistically based on their arrangement
- **Flexible Grammar:** Same TFs can produce different outputs depending on context and arrangement
- **Evolutionary Conservation:** Grammar rules are often conserved across species, indicating functional importance
- **Therapeutic Applications:** Design synthetic enhancers with predictable, cell-type specific activity
- **Key Models:** DeepSEA, Basset, and ChromBPNet excel at learning enhancer grammar

4. Cell Type Specificity

Context-dependent gene expression prediction

The Challenge

All cells in an organism share the same genome, yet express vastly different sets of genes. A neuron expresses different genes than a liver cell, despite having identical DNA sequences. This cell-type specificity arises from epigenetic regulation.

Epigenetic Context

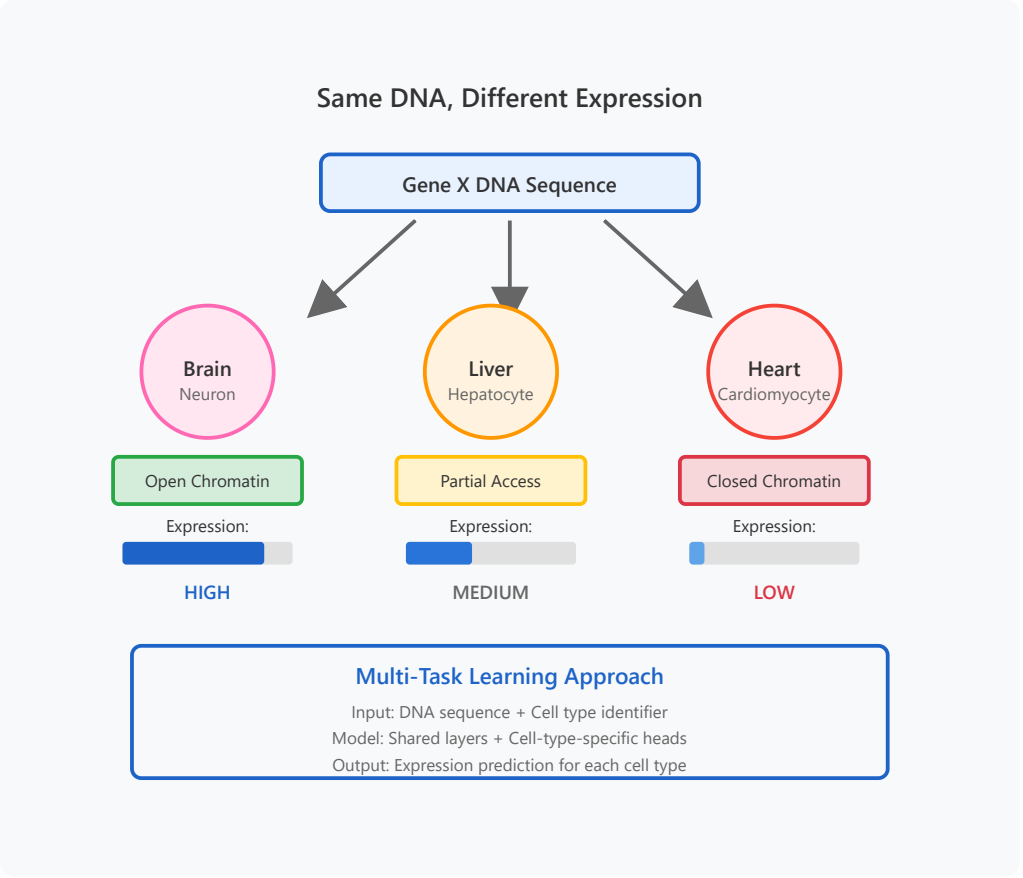
Cell type identity is encoded through chromatin accessibility, DNA methylation, and histone modifications. These epigenetic marks determine which regulatory elements are active in each cell type, creating a unique regulatory landscape.

Modeling Approaches

Modern models incorporate cell-type information through multi-task learning (predicting across cell types simultaneously), conditional models (cell type as input), or through learned cell-type embeddings that capture regulatory state.

Practical Impact

Understanding cell-type specificity enables prediction of how genetic variants affect different tissues, design of cell-type-specific gene therapies, and identification of regulatory elements driving cell fate decisions.



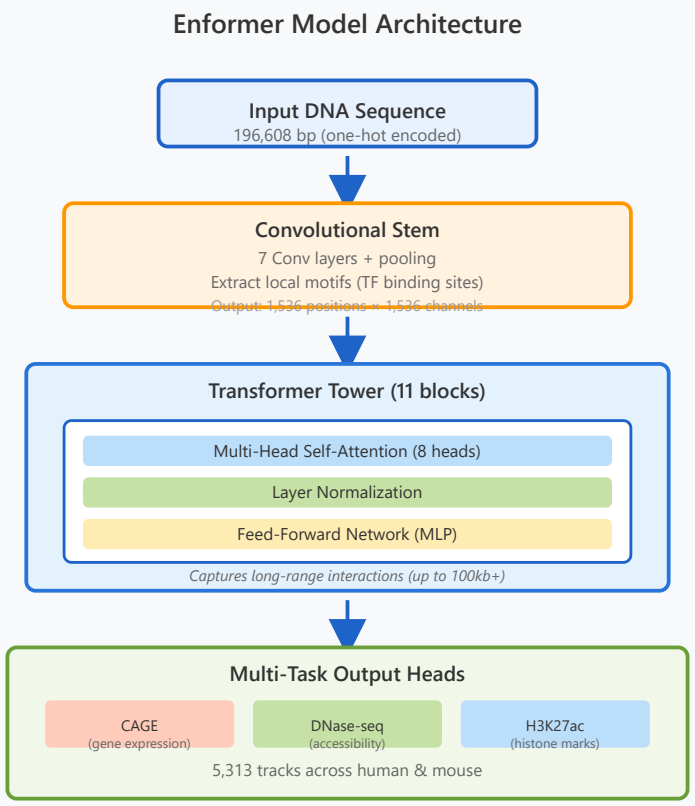
Key Mechanisms & Applications

- **Master Regulators:** Cell-type-specific transcription factors (e.g., MyoD in muscle, GATA1 in blood) drive expression programs
- **Chromatin Accessibility:** DNase-seq and ATAC-seq data reveal which regulatory elements are accessible in each cell type
- **Histone Marks:** H3K4me3 (promoters), H3K27ac (active enhancers), H3K27me3 (repression) mark regulatory states
- **Clinical Applications:** Predict tissue-specific effects of disease variants (e.g., heart vs brain)

- **Notable Models:** Basenji predicts cell-type-specific chromatin and expression across 200+ cell types

5. Enformer Architecture

Transformer + CNN hybrid for long-range regulatory prediction



Why Enformer?

Previous models (like Basenji) used only CNNs and were limited to ~40kb context windows. Enformer uses transformers to capture interactions across 200kb, dramatically improving predictions by modeling distal enhancers and TAD structures.

CNN Stem

The convolutional stem processes raw DNA sequence to extract local features like transcription factor binding motifs. This reduces the sequence length while enriching the representation with biologically meaningful patterns.

Transformer Tower

The transformer blocks use self-attention to model long-range interactions between regulatory elements. Unlike CNNs with limited receptive fields, attention can directly connect distant positions, capturing enhancer-promoter loops and chromatin interactions.

Multi-Task Learning

Enformer simultaneously predicts thousands of genomic tracks (CAGE, ChIP-seq, DNase-seq, etc.) across cell types. This multi-task approach helps the model learn generalizable regulatory principles and improves performance through shared representations.

Performance Gains

Enformer achieves state-of-the-art accuracy, explaining ~60% of variance in human expression data. It substantially outperforms previous models on variant effect prediction and can identify regulatory variants missed by GWAS.

Key Innovations in Enformer

Technical Details & Impact

- **Parameters:** ~250M parameters, trained on TPUs for several weeks
- **Training Data:** Thousands of genomic assays from ENCODE, Roadmap Epigenomics, and GTEx
- **Architecture Benefit:** Attention mechanism provides computational efficiency and better gradient flow than deep CNNs
- **Interpretability:** Attention weights reveal which genomic regions interact to regulate expression
- **Applications:** Variant prioritization, synthetic biology, understanding disease mechanisms, drug target identification
- **Future Directions:** Single-cell predictions, 3D genome structure integration, protein sequence co-modeling