TEXAS A&M UNIVERSITY
Engineering

Learning to Discover Regulatory Elements for Gene Expression Prediction

Xingyu Su[1]*, Haiyang Yu[1]*, Degui Zhi[2], and Shuiwang Ji[1]†
[1]Texas A&M University
[2]The University of Texas Health Science Center at Houston

Presenter: Cong Fu, Texas A&M University

* denotes equal contribution
† denotes corresponding author

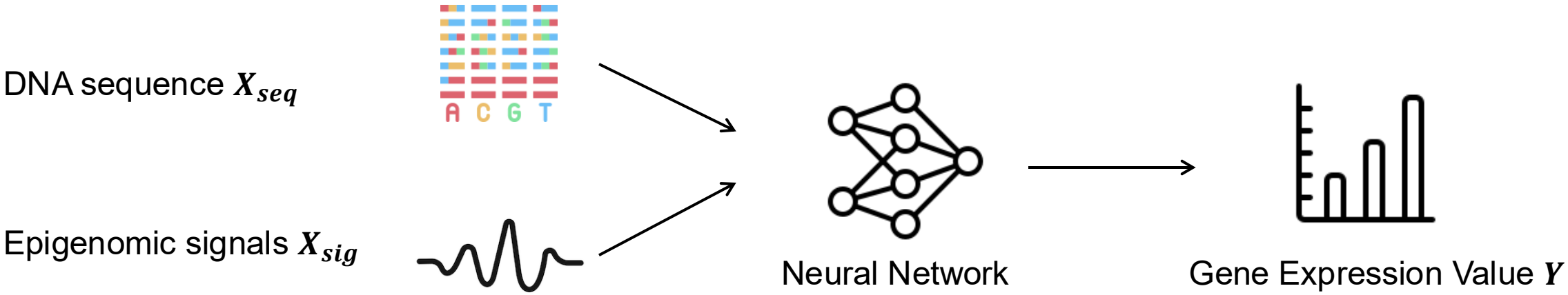
TEXAS A&M UNIVERSITY
Engineering

Background

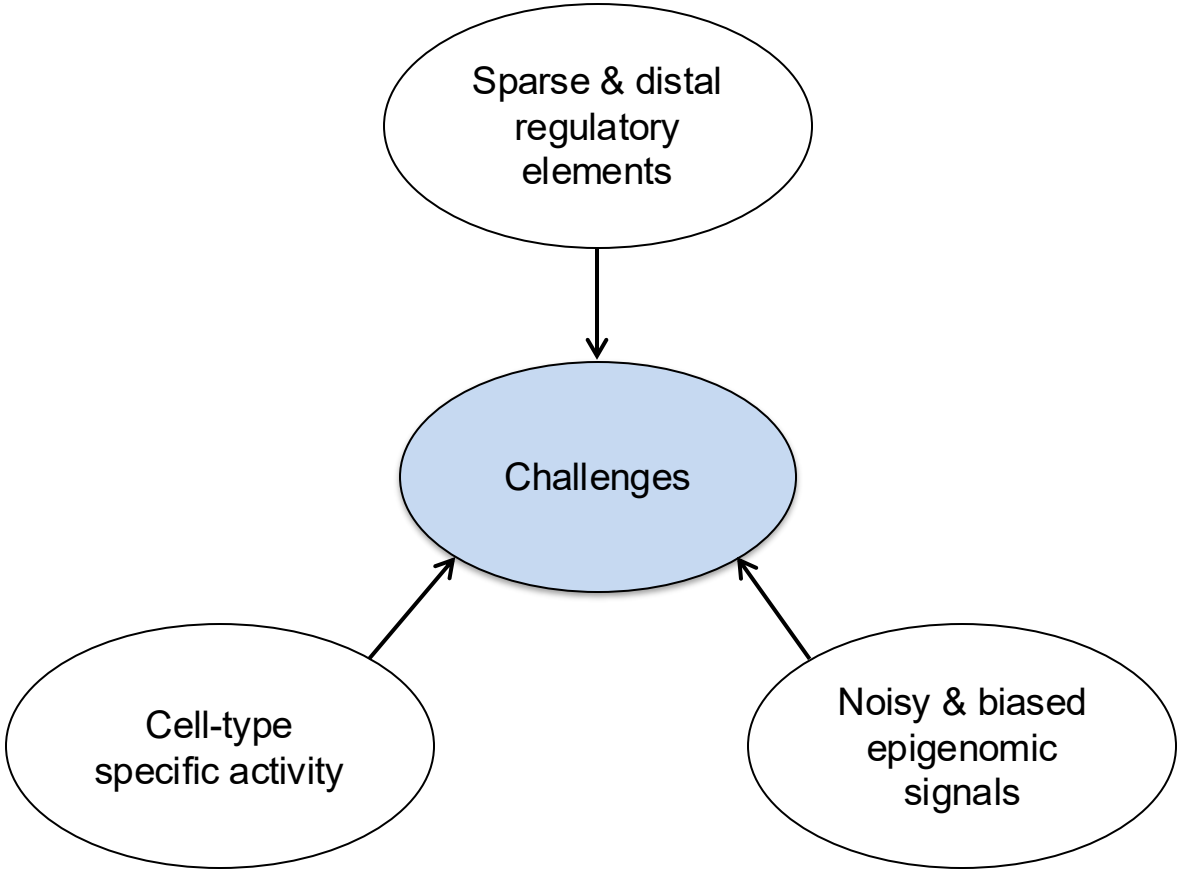
Gene Expression Prediction

TEXAS A&M UNIVERSITY Engineering

- Gene expression prediction is important because it helps us understand how genetic and regulatory variations influence cellular behavior and disease outcome
- Predicting target gene expression value $\boldsymbol{Y}$ for various cell types
  - Given DNA sequence $\boldsymbol{X_{seq}}$ and epigenomic signals $\boldsymbol{X_{sig}}$
- DNA sequence $\boldsymbol{X_{seq}}$
  - Same static genomic context for all the cells
- Epigenomic signals $\boldsymbol{X_{sig}}$
  - Dynamic, cell-type-specific markers



Biological Challenges



- Regulatory elements are sparse and distal
  - Only a few functional elements
  - Long-range interactions
- Functional activity is cell-type specific
  - The same promoters/enhancers can be active in one cell type and inactive in another
  - Requires context-aware modeling to reflect epigenetic variation
- Epigenomic signals are noisy and biased
  - E.g., DNase-seq, ChIP-seq, and Hi-C each have different biases
  - Signal quality varies across genomic regions and cell types

Limitations of Existing Methods

TEXAS A&M UNIVERSITY Engineering

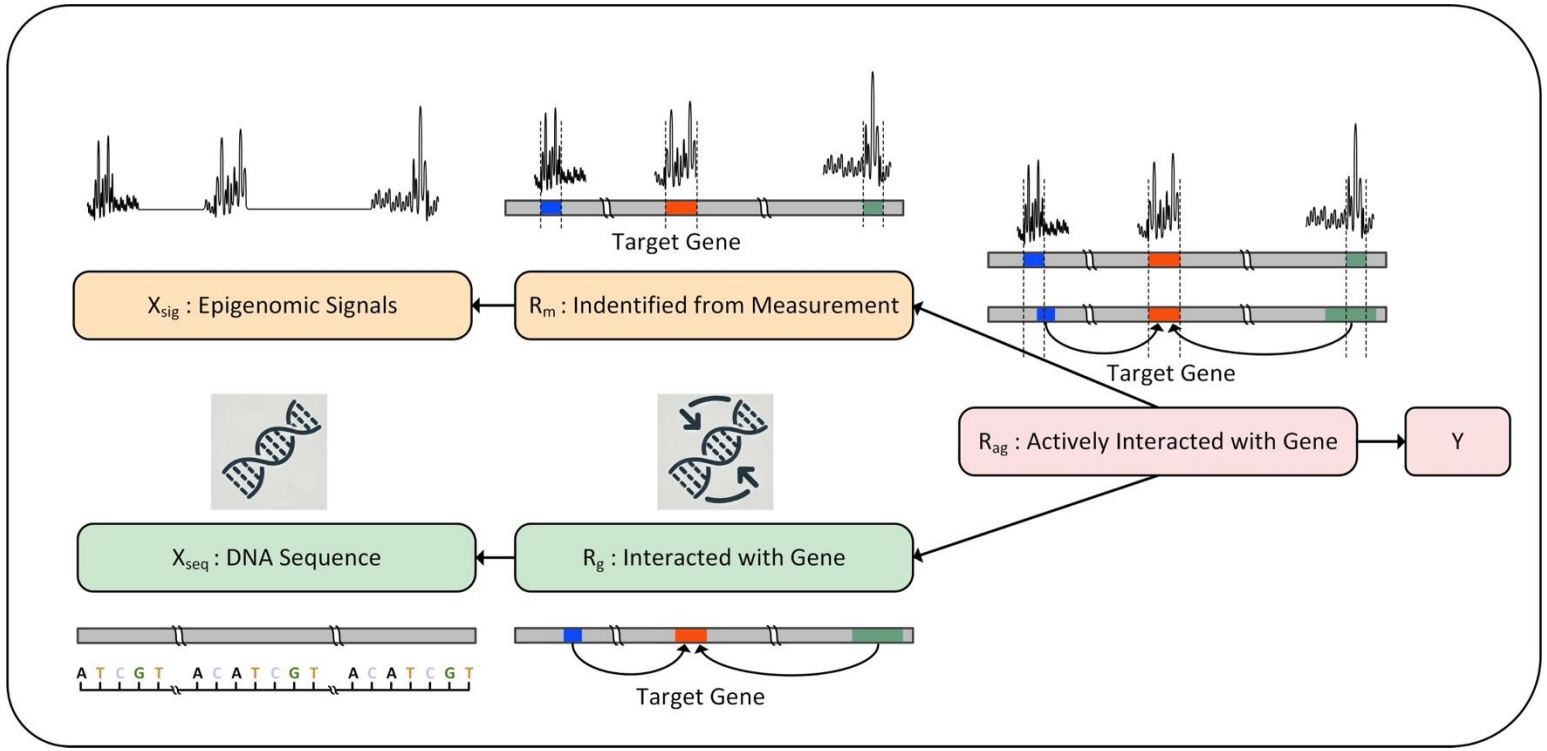
Existing Methods	General Approach	Limitations
DNA-only Methods	• Encode full DNA sequence with deep learning models • Use DNA alone to predict expression	• Cannot model dynamic, cell-type-specific regulatory elements • Fail to capture functional differences from the same static sequence
Peak-based Methods	• Identify regulatory regions using epigenomic peak-calling methods • Use extracted signal features from peaks for prediction	• Heuristic region selection is not task-optimized • Miss low-signal but important regions

- Other general limitations
  - ➢ No explicit modeling of causal relationships
  - ➢ Poor generalization to unseen cell types

TEXAS A&M UNIVERSITY
Engineering

Methods

Causal Relationship



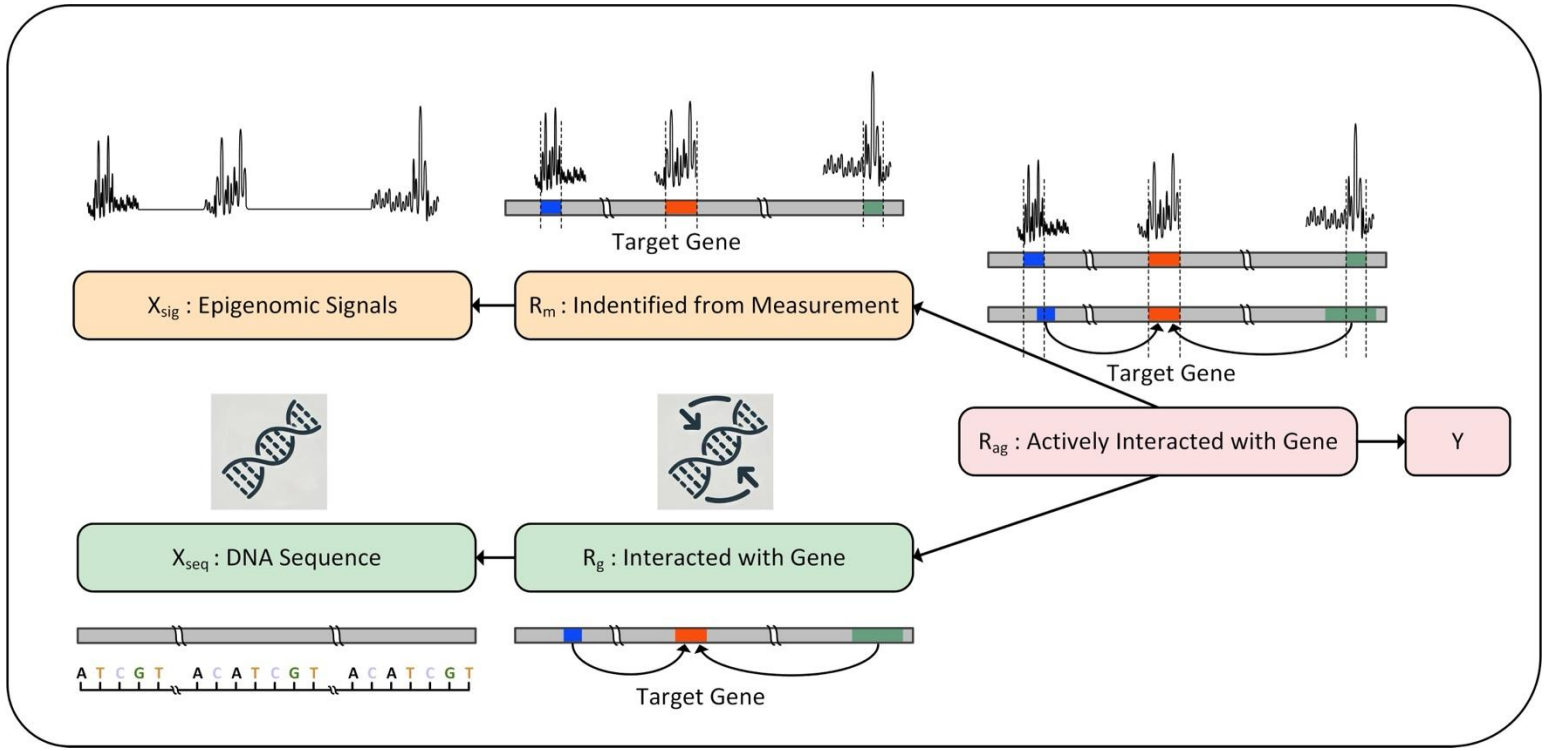
Three Key Components

R_g - ***Genomic candidates***: Regulatory elements that could interact with the target gene; but may be inactive in certain cell types

R_m - ***Measured regions***: Detected via epigenomic signals may correlate with expression, but their true target gene is often unknown

R_{ag} - ***Active and causal regulatory elements***: Regulatory elements that directly influence target gene expression. These are the true causal drivers we aim to discover

Causal Relationship



$\boldsymbol{X_{seq}} \leftarrow \boldsymbol{R_g}$: DNA sequence containing of all regulatory elements R_g interacted with target gene and other non-causal regions

$\boldsymbol{R_{ag}} \rightarrow \boldsymbol{Y}$: Causal active regulatory region that directly affects the gene expression

$\boldsymbol{R_g} \leftarrow \boldsymbol{R_{ag}} \rightarrow \boldsymbol{R_m}$: Key causal region shared by both DNA and epigenomic signals

$\boldsymbol{R_m} \rightarrow \boldsymbol{X_{sig}}$: Regulatory elements R_m by measurement reflecting active regions in the cell, often showing as peaks in experiments

Task Objective

TEXAS A&M UNIVERSITY Engineering

- Based on information bottleneck, we learn a token level soft/hard mask M, to extract regulatory elements
- Assume each token mask selection is independent given the input sequences $p(M|X) = \prod_i p(m_i|X)$
- Objective loss function

$$L \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p_\theta(m_i|x_i)}[-log q_\phi(y_i|m_i \odot x_i)] + \beta KL[p_\theta(m_i|x_i), r(m_i)]$$

  - Task specific loss + constraint on the learned mask m

Decomposition of Sequences & Signals

TEXAS A&M UNIVERSITY Engineering

- Assumption 1
  - Conditioned on the selection of regulatory elements M, the DNA sequences and epigenomic signals are condition independent

$$p(X_{sig}, X_{seq}|M) = p(X_{sig}|M)p(X_{seq}|M)$$

  - Based on the previous causal relationships
- Proposition 1
  - The estimation of $p_\theta(M|X)$ can be divided into

$$p_\theta(M|X) \propto p_{\theta_1}(M|X_{seq})p_{\theta_2}(M|X_{sig})$$

  - Divide the learning of mask into contributions from DNA sequence and epigenomic signals

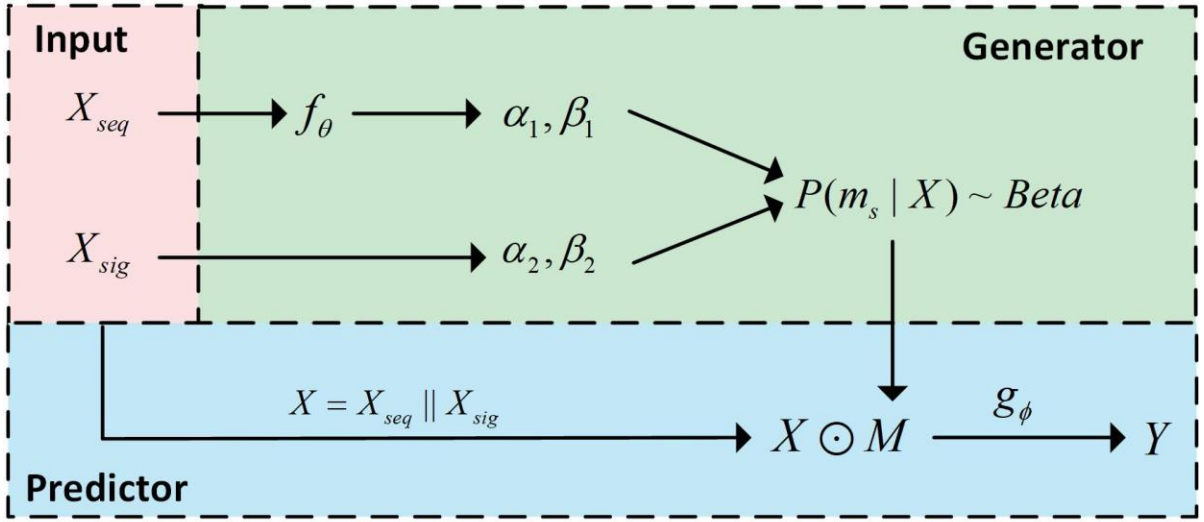
Mask Distribution

- Assumption 2
  - ➢ We assume the mask distribution follows Beta distribution, i.e., $m_s \sim Beta(\alpha, \beta)$
  - ➢ Reason for choosing Beta distribution
    - ❑ Beta distribution naturally quantifies the success rates.
    - ❑ The success rates is controlled by varying the value of α and β
    - ❑ Product of two Beta distributions can result in another Beta distribution
- Proposition 2
  - ➢ Given $p_{\theta_1}(m_s|X_{seq}) \sim Beta(\alpha_1, \beta_1)$ and $p_{\theta_2}(m_s|X_{sig}) \sim Beta(\alpha_2, \beta_2)$, the product follows: $Beta(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_2 - 1)$
  - ➢ We can directly model both distributions by predicting parameters of α and β
- We use straight-through estimator to ensure differentiability for hard mask
- Set prior $r(m_s)$ to control the sparsity

Seq2Exp Architecture

TEXAS A&M UNIVERSITY Engineering

- Model both mask distribution $p_{\theta_1}(m_s|X_{seq})$ and $p_{\theta_2}(m_s|X_{sig})$
- Inputs
  - DNA sequence X_{seq} and epigenomic signals X_{sig}
- Mask from DNA sequences
  - Use mask generator f_θ to learn Beta distribution parameters (α_1, β_1) from X_{seq}
- Mask from epigenomic signals
  - (α_2, β_2) is derived from X_{sig} in a non-parametric way
- Result mask distribution
  - Combine two Beta distribution to get $p(m_s|X)$



- Regulatory Element Selection
  - Sample a mask to select actively interacting regions
- Predictor Model (g_ϕ)
  - Use selected sub-sequences to predict gene expression

TEXAS A&M UNIVERSITY
Engineering

Experiments

Experiments - Settings

TEXAS A&M UNIVERSITY Engineering

- Predict CAGE (Cap Analysis of Gene Expression) values on cell K562 and GM12878
  - DNA sequences: HG38 human reference genome
  - Epigenomic signals: DNase-seq, H3K27ac, Hi-C
  - Consider 18,377 protein-coding target genes
- Baselines
  - DNA based models: Enformer, HyenaDNA, Mamba, Caduceus
  - DNA & signal model: EPInformer
- Metric
  - MSE, MAE, Pearson correlation between the predicted and real CAGE
- Implementation
  - Both generator p_θ and predictor q_ϕ are based on Caduceus architecture
  - Leverage cross-chromosome validation strategy
    - chr 3&21 as validation, chr 22&X as testing
  - Consider window size of 200k surrounded by target genes

Experiments

TEXAS A&M UNIVERSITY Engineering

Methods:

- Seq2Exp-hard: assign hard mask to DNA base pair
- Seq2Exp-soft: assign soft probability to DNA base pair

Cell types: K562

	MSE ↓	MAE ↓	Pearson ↑
Enformer	0.2889 ± 0.0009	0.4185 ± 0.0013	0.8327 ± 0.0025
HyenaDNA	0.2217 ± 0.0018	0.3562 ± 0.0012	0.8729 ± 0.0010
Mamba	0.2145 ± 0.0021	0.3446 ± 0.0022	0.8788 ± 0.0011
Caduceus	0.2124 ± 0.0037	0.3436 ± 0.0031	0.8819 ± 0.0009
Caduceus w/ signals	0.1942 ± 0.0058	0.3269 ± 0.0048	0.8928 ± 0.0017
EPInformer	0.1975 ± 0.0031	0.3246 ± 0.0025	0.8907 ± 0.0011
MACS3	0.2340 ± 0.0028	0.3654 ± 0.0017	0.8634 ± 0.0020
Seq2Exp-hard	0.1890 ± 0.0045	0.3199 ± 0.0040	0.8916 ± 0.0027
Seq2Exp-soft	**0.1873 ± 0.0044**	**0.3137 ± 0.0028**	**0.8951 ± 0.0038**

GM12878

	MSE ↓	MAE ↓	Pearson ↑
Enformer	0.2920 ± 0.0050	0.4056 ± 0.0040	0.7961 ± 0.0019
HyenaDNA	0.2265 ± 0.0013	0.3497 ± 0.0012	0.8425 ± 0.0008
Mamba	0.2241 ± 0.0027	0.3416 ± 0.0026	0.8412 ± 0.0021
Caduceus	0.2197 ± 0.0038	0.3327 ± 0.0070	0.8475 ± 0.0014
Caduceus w/ signals	0.1959 ± 0.0036	0.3187 ± 0.0036	0.8630 ± 0.0008
EPInformer	0.2140 ± 0.0042	0.3291 ± 0.0031	0.8473 ± 0.0017
MACS3	0.2195 ± 0.0023	0.3455 ± 0.0018	0.8435 ± 0.0013
Seq2Exp-hard	0.1863 ± 0.0051	0.3074 ± 0.0036	0.8682 ± 0.0045
Seq2Exp-soft	**0.1856 ± 0.0032**	**0.3054 ± 0.0024**	**0.8723 ± 0.0012**

Experiments

Evaluate the regulatory elements extracted by different methods

- Seq2Exp-hard: train both generator and predictor together to extract regulatory elements
- Seq2Exp-retrain: extract top 10% base pairs from soft mask and only train the predictor
- MACS3: statistical methods to extract regulatory elements region

	K562				GM12878			
	MSE ↓	MAE ↓	Pearson ↑	Mask Ratio	MSE ↓	MAE ↓	Pearson ↑	Mask Ratio
Seq2Exp-hard	0.1863	0.3074	0.8682	6.88%	0.1890	0.3199	0.8916	6.32%
Seq2Exp-retrain	0.1979	0.3168	0.8623	10.00%	0.1887	0.3177	0.8941	10.00%
MACS3	0.2195	0.3455	0.8435	13.61%	0.2340	0.3654	0.8634	15.95%

Conclusion

TEXAS A&M UNIVERSITY
Engineering

Summary

- Introduced Seq2Exp, a causal and interpretable framework for gene expression prediction
- Jointly models DNA sequences and epigenomic signals to learn regulatory element masks
- Achieves state-of-the-art performance on gene expression prediction in K562 and GM12878

Future works

- Extend to more diverse cell types and additional epigenomic signals
- Develop pretraining models for generalizable regulatory region extraction across cell types
- Explore causal interpretation in broader biological and clinical contexts

TEXAS A&M UNIVERSITY
Engineering

Thank you!