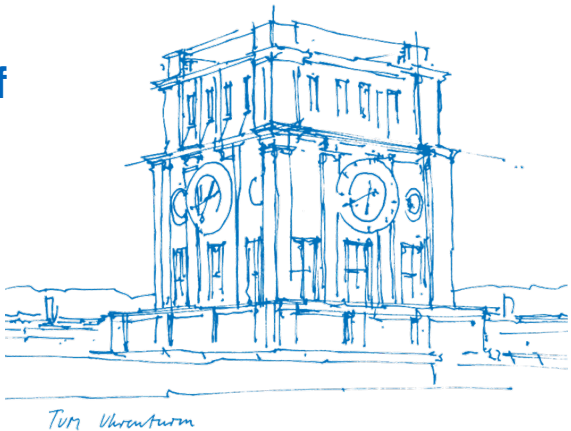
Binding-Bench Challenge: Unsupervised prediction of TF binding sites

Jannis-Alexander Arnold, Katharina Miller, Ana Cunha, Arjun Bommadevara

Technical University of Munich

July 2026



Outline

1 Introduction

2 Methodology

3 Results

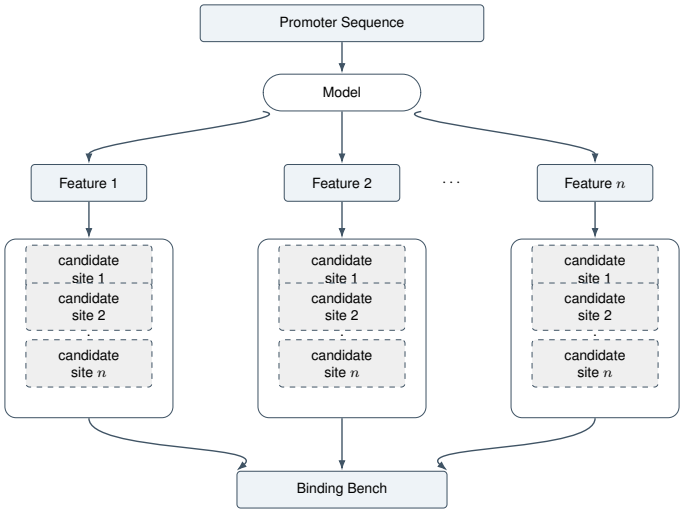
4 Limitations & Future Work

5 References

Introduction
Motivation

- **Biology**: Transcription Factors (TFs) regulate gene expression by binding specific promoter DNA sites.
- **Bottleneck**: binding-site annotations are **sparse** for most species and TFs (ChIP is costly and condition-specific).
- **Need**: methods should **generalize beyond available ChIP data**.
- **Goal**: develop generalizable methods that leverage information encoded in the genome to predict individual TF binding sites.

Introduction
BindingBench



Outline

1 Introduction

2 Methodology

3 Results

4 Limitations & Future Work

5 References

Motif Baselines
STREME and JASPAR+FIMO

STREME

- Discovers motifs from the benchmark sequences.
- Convert motif-supported sites to BindingBench predictions.

JASPAR+FIMO

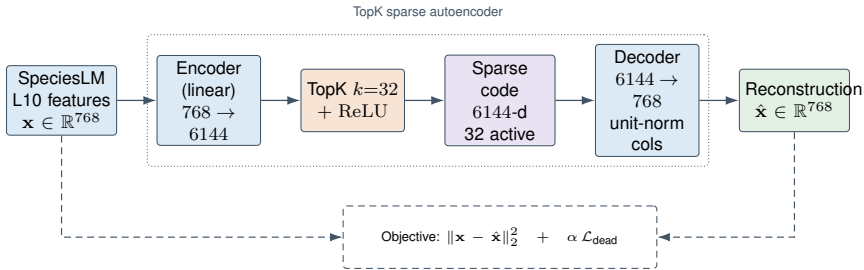
- Scan promoters with curated PWMs (JASPAR).

Takeaway

Both approaches predict binding from short local sequence motifs and provide interpretable baselines.

Method 1: Sparse Autoencoder (SAE)

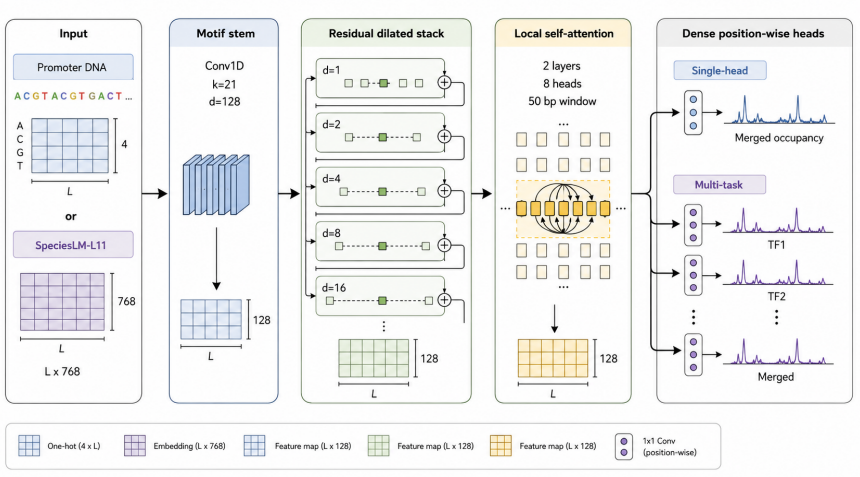
- Train a TopK SAE on mean-centered SpeciesLM L10 embeddings (no binding labels).
- Expansion factor 8: $768 \rightarrow 6144$ latent features, keep TopK $k = 32$ activations per position.
- Evaluate: per TF pick best feature on a training split, report AUROC on held-out genes.



Method 2: DNA-only Models

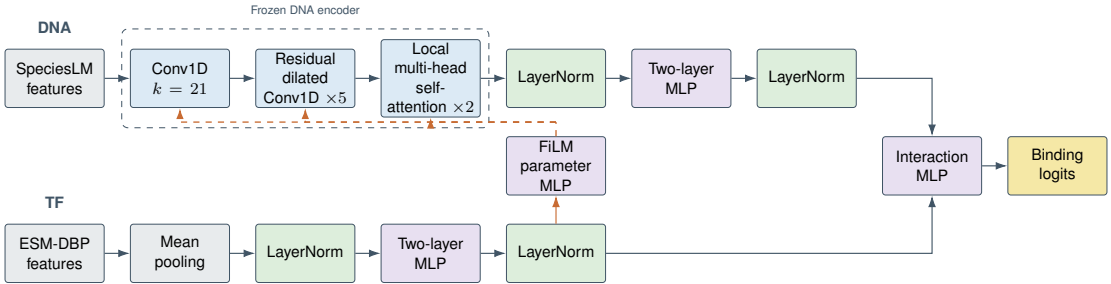
- Input: one-hot DNA or SpeciesLM L11 embeddings.
- Shared backbone: motif stem + dilated conv residual stack + local self-attention.
- **Single-head**: only merged occupancy target.
- **Multi-head**: TF-specific auxiliary heads + merged occupancy head (evaluated on held-out TF benchmark).

Method 2: DNA-only Architecture



Method 3: DNA + Protein Embeddings

- Shared scorer combines DNA features with protein embedding representation.
- Protein representation generates FiLM parameters that modulate DNA encoder.



Outline

1 Introduction

2 Methodology

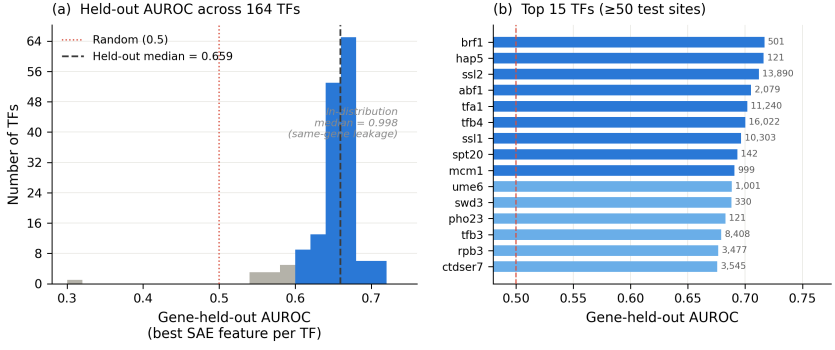
3 Results

4 Limitations & Future Work

5 References

SAE Results

- Gene-held-out evaluation avoids same-gene leakage.
- Median best-feature AUROC across 164 held-out TFs: **0.659**.
- Strongest recovery for high-occupancy TFs (e.g. *abf1*, *hap5*).



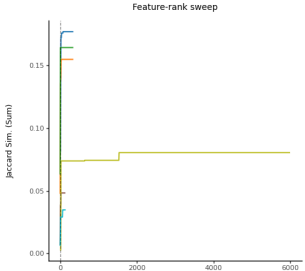
Protein-aware Results

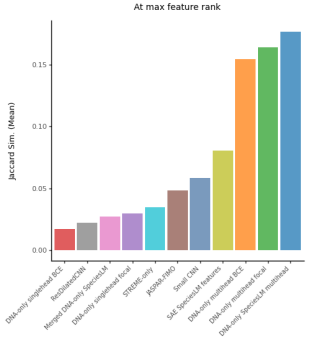
Model	AP	AUROC
Late fusion	0.0106	0.8351
FiLM, shuffled proteins	0.0059	0.8318
FiLM conditioning	**0.0107**	**0.8581**

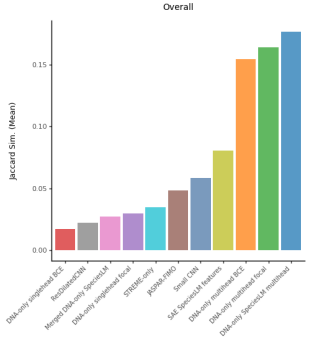
Takeaways

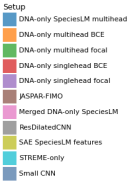
- AP almost the same between FiLM and late fusion
- shuffling proteins hurts AP strongly $\rightarrow$ TF identity in protein embeddings matters
- Absolute AP still low!

BindingBench Comparison: Merged held-out TFs

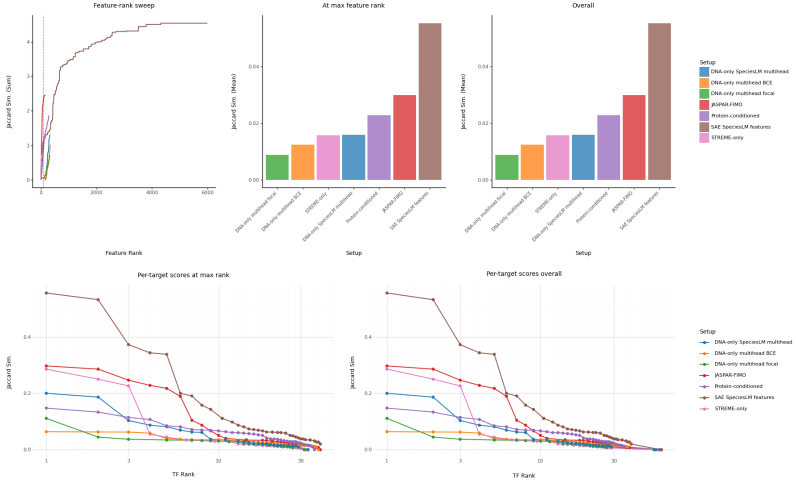








BindingBench Comparison: TF-specific



Outline

1 Introduction

2 Methodology

3 Results

4 Limitations & Future Work

5 References

Limitations

- **Scope**: only *S. cerevisiae* promoters (1kb upstream of ATG) and ChIP-exo derived sites.
- **Matched evaluation**: SAE export used lower top-k peak caps than supervised models (runtime constraints).
- **TF-specific setting**: DNA-only/SAE rely on post hoc best-assignment (no TF identity at test time).
- **Protein embedding**: mean pooling may dilute DNA-binding-domain signal; embedding anisotropy can obscure specificity.

Future Work

- **More matched comparisons**: unified top-k and NMS parameters; rerun SAE exports at full scale.
- **Stronger generalization tests**: stricter gene/TF holdouts; similarity-aware splits.
- **Better protein representations**: focus on DNA-binding domains (e.g., InterPro-guided subsequences).
- **Hybrid models**: combine motif priors with dense predictors; explore cross-attention and fine-tuning.

Outline

1 Introduction

2 Methodology

3 Results

4 Limitations & Future Work

5 References

References I

[Bai+09] Timothy L. Bailey et al. “MEME SUITE: tools for motif discovery and searching”. In: *Nucleic Acids Research* 37.suppl$_2$ (2009), W202–W208. DOI: 10.1093/nar/gkp335.

[BC26] Shreya Basnet and Jianlin Cheng. “Integrating protein and DNA embeddings for improving genome-wide transcription factor binding site prediction”. In: *NAR Genomics and Bioinformatics* 8.2 (2026), lqag047. DOI: 10.1093/nargab/lqag047.

[BE95] Timothy L. Bailey and Charles Elkan. “Unsupervised learning of multiple motifs in biopolymers using expectation maximization”. In: *Machine Learning* 21 (1995), pp. 51–80. DOI: 10.1007/BF00993334.

[BLN24] Bart Bussmann, Patrick Leask, and Neel Nanda. “BatchTopK Sparse Autoencoders”. In: *arXiv preprint arXiv:2412.06410* (2024).

References II

[Blu+21] Matthias Blum et al. “The InterPro protein families and domains database: 20 years on”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D344–D354. DOI: 10.1093/nar/gkaa977. URL: https://doi.org/10.1093/nar/gkaa977.

[Eng+25] Stacia R. Engel et al. “Saccharomyces Genome Database: Advances in Genome Annotation, Expanded Biochemical Pathways, and Other Key Enhancements”. In: *Genetics* 229.3 (2025), iyae185. DOI: 10.1093/genetics/iyae185.

[GBN11] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. “FIMO: scanning for occurrences of a given motif”. In: *Bioinformatics* 27.7 (2011), pp. 1017–1018. DOI: 10.1093/bioinformatics/btr064.

References III

[Kar+24] Alexander Karollus et al. “Species-aware DNA language models capture regulatory elements and their evolution”. In: *Genome Biology* 25.1 (2024), p. 83. DOI: 10.1186/s13059-024-03221-x.

[MEM26] MEME Suite. *STREME: Sensitive, Thorough, Rapid, Enriched Motif Elicitation*. https://meme-suite.org/meme/doc/streme.html. Accessed: 2026-07-08. 2026.

[Ng] Andrew Ng. *Sparse Autoencoder*. CS294A Lecture Notes, Stanford University. URL: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf.

[Par09] Peter J. Park. “ChIP-seq: advantages and challenges of a maturing technology”. In: *Nature Reviews Genetics* 10 (2009), pp. 669–680. DOI: 10.1038/nrg2641.

References IV

[Per+17] Ethan Perez et al. *FiLM: Visual Reasoning with a General Conditioning Layer*. 2017. arXiv: 1709.07871 [cs.CV]. URL: https://arxiv.org/abs/1709.07871.

[Rau+24] Ieva Rauluseviciute et al. "JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* (2024). DOI: 10.1093/nar/gkad1059.

[Ros+21] Matthew J Rossi et al. "A high-resolution protein architecture of the budding yeast genome". In: *Nature* 592.7853 (2021), pp. 309–314.

[Zen+24] Wenwu Zeng et al. "Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein". In: *Nature Communications* 15.1 (2024), p. 7838.

References V

[Zho+23] Zhihan Zhou et al. “DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome”. In: *arXiv preprint arXiv:2306.15006* (2023).

[ZT15] Jian Zhou and Olga G. Troyanskaya. “Predicting effects of noncoding variants with deep learning-based sequence model”. In: *Nature Methods* 12 (2015), pp. 931–934. DOI: 10.1038/nmeth.3547.