Calico

Predicting dynamic expression patterns in budding yeast with a fungal DNA language model

Kuan-Hao Chao

03/25/2026

khchao.com

@kuanhaochao.bsky.social

@KuanHaoChao

Kuanhao-Chao

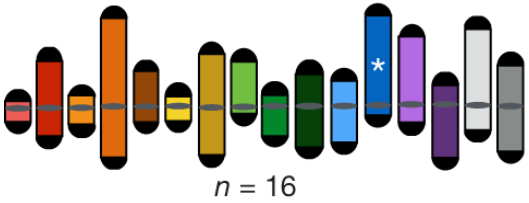
DISCLOSURES & ACKNOWLEDGEMENTS

This research was funded by **Calico Life Sciences LLC**, with additional support provided by the **U.S. National Institutes of Health (NIH)** under grants R01-HG006677 and R35-GM156470, and the **U.S. National Science Foundation (NSF)** under grant DBI-2412449. Computational analyses were performed using resources provided by the Advanced Research Computing at Hopkins (ARCH) core facility, which is supported in part by NSF grant OAC-1920103.

Introduction

Why Yeast Is an Ideal Model for Regulatory Grammar

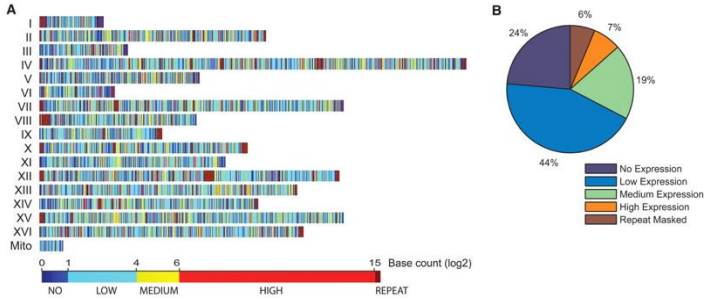
- *Saccharomyces cerevisiae* is the premier eukaryotic model organism to understand eukaryotic regulatory grammar
- With ~7,000 genes controlled by hundreds of transcription factors.



Luo et al, Nature 2018

SGD Saccharomyces GENOME DATABASE

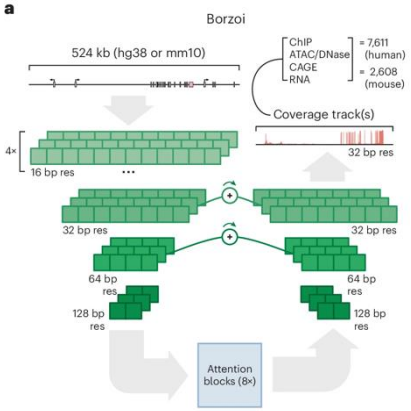
Engel et al, Genetics 2025

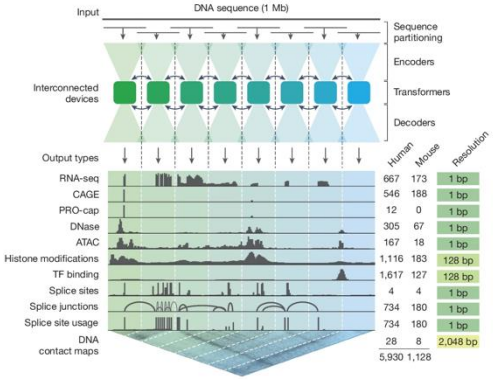


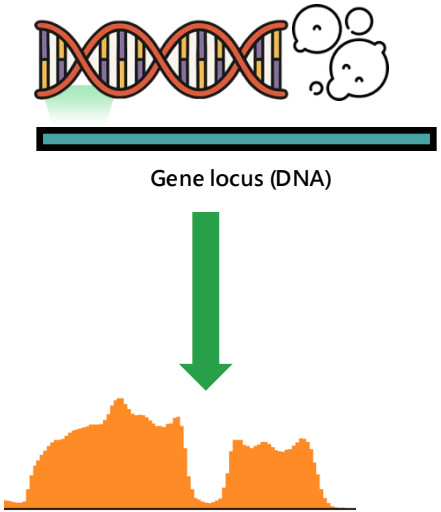
Nagalakshmi et al, Science 2008

Understanding Regulatory Grammar via Seq2func models

- Supervised deep learning can, in principle, learn regulatory features directly from DNA sequence.







Avsec et al. Nat. Methods (2021) Linder et al. Nature Genetics (2025) Žiga et al. Nature (2026)

The Yeast Data Bottleneck Motivates DNA Language Models

- **The Bottleneck:** The compact 12 Mb genome provides too few independent training examples to support large supervised models without overfitting
- **The Paradigm Shift: DNA Language Models**
  - Self-supervised LMs overcome this by training on diverse, unlabeled genomes.
- **The Unresolved Gap (The "Catch"):** The Unresolved Gap: Despite these advances, the mechanistic basis for how Masked Language Modeling (MLM) pretraining improves downstream Seq2func performance remains unclear.

Shorkie LM

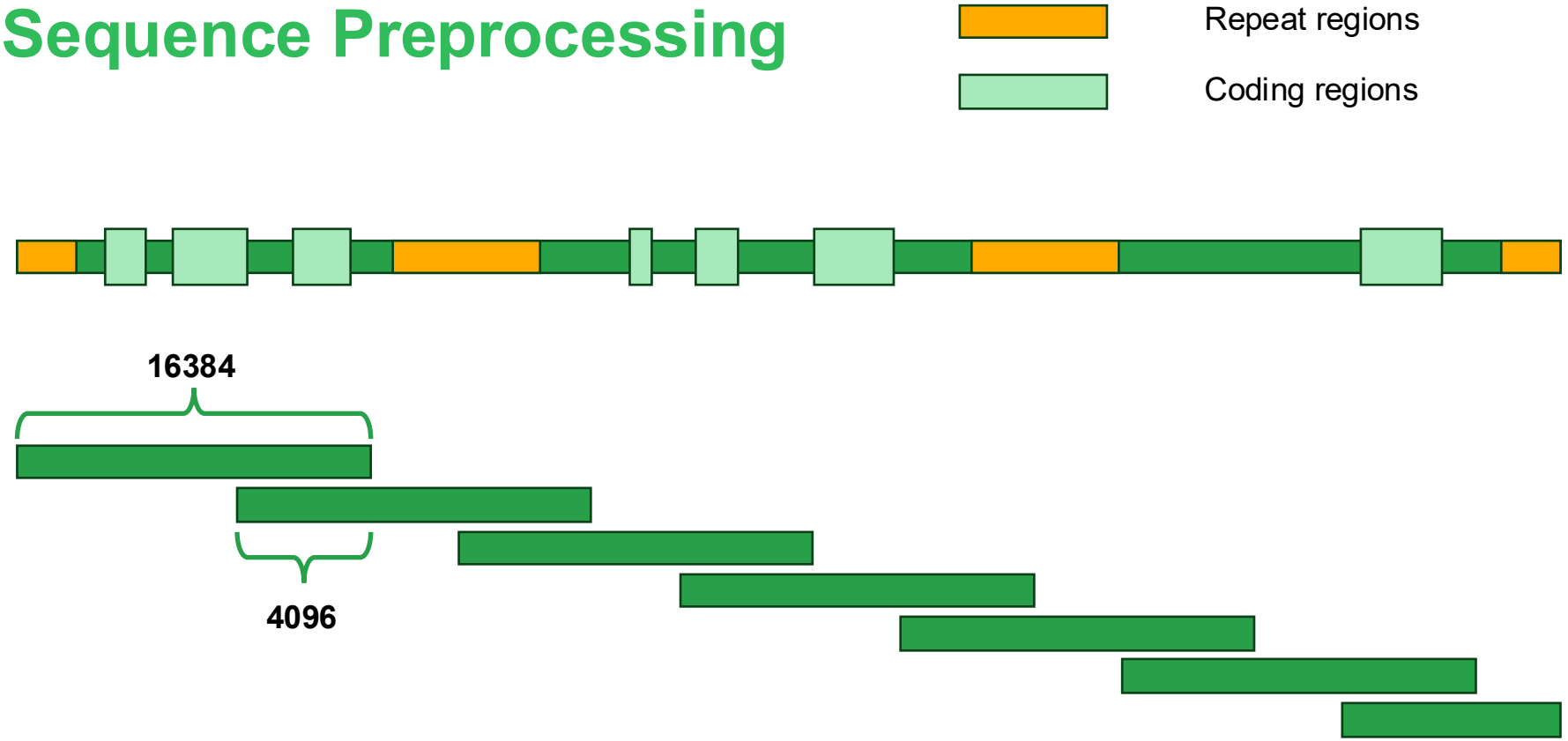
A Fungal DNA Language Model for Yeast

The Language Model Solution

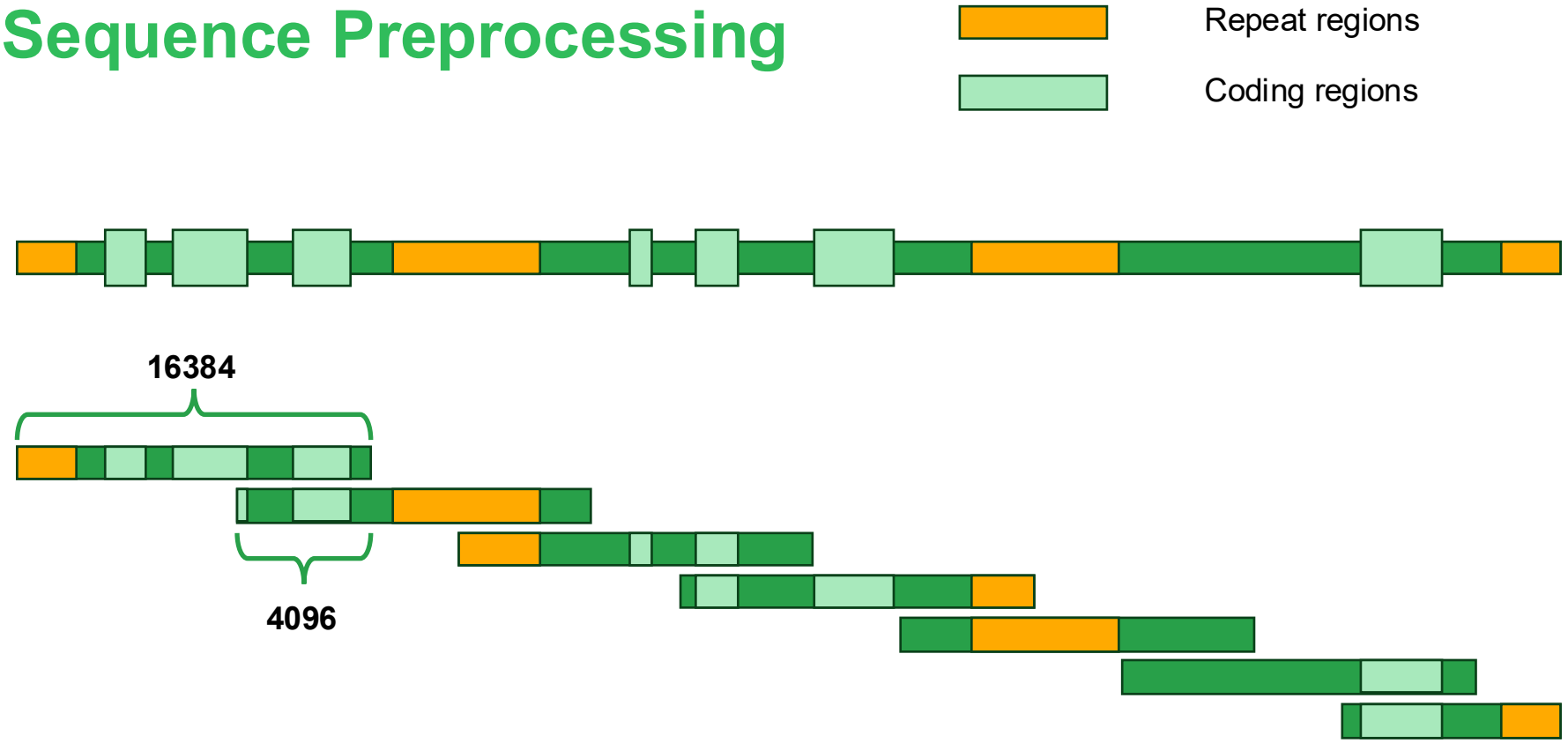
Repeat regions

Coding regions

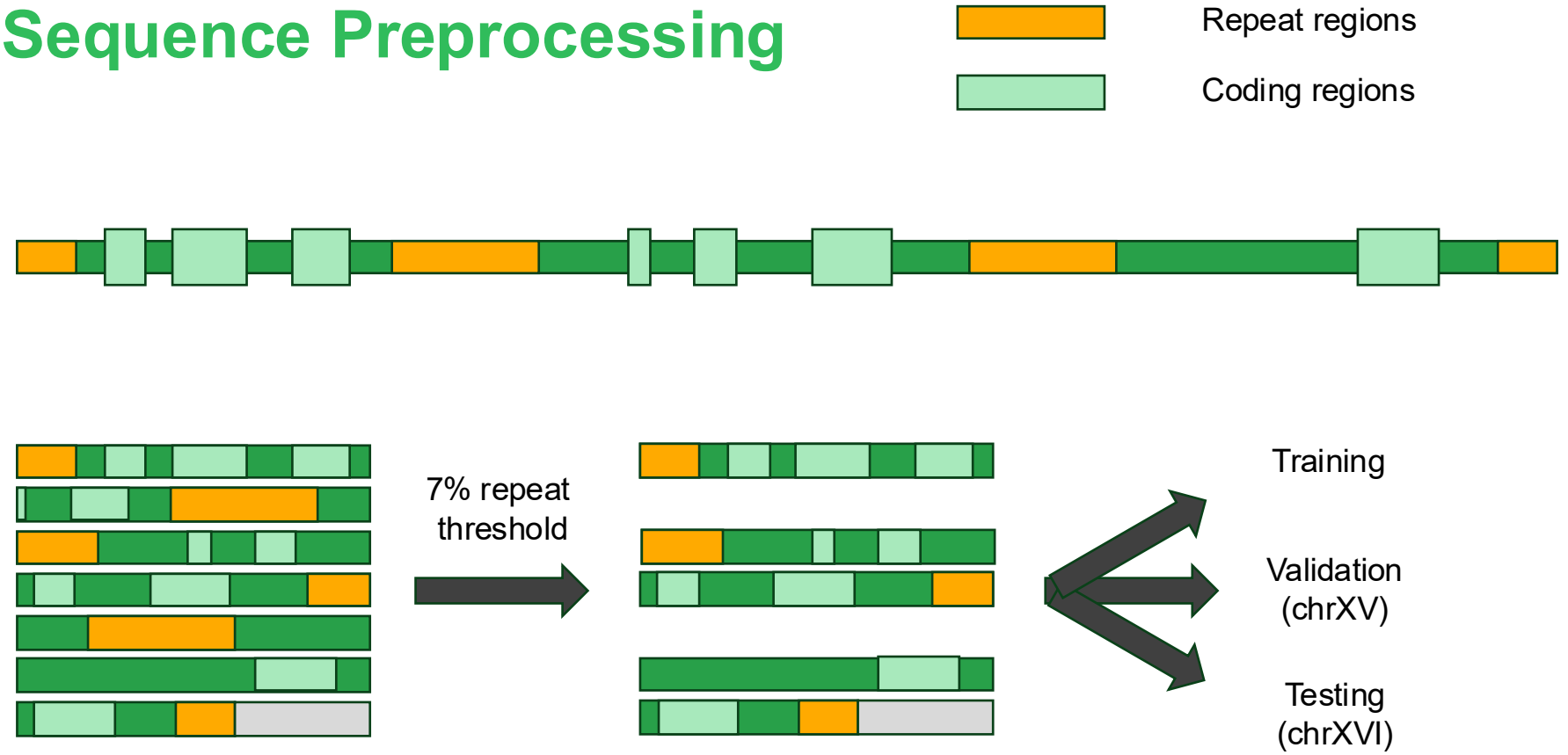
Sequence Preprocessing

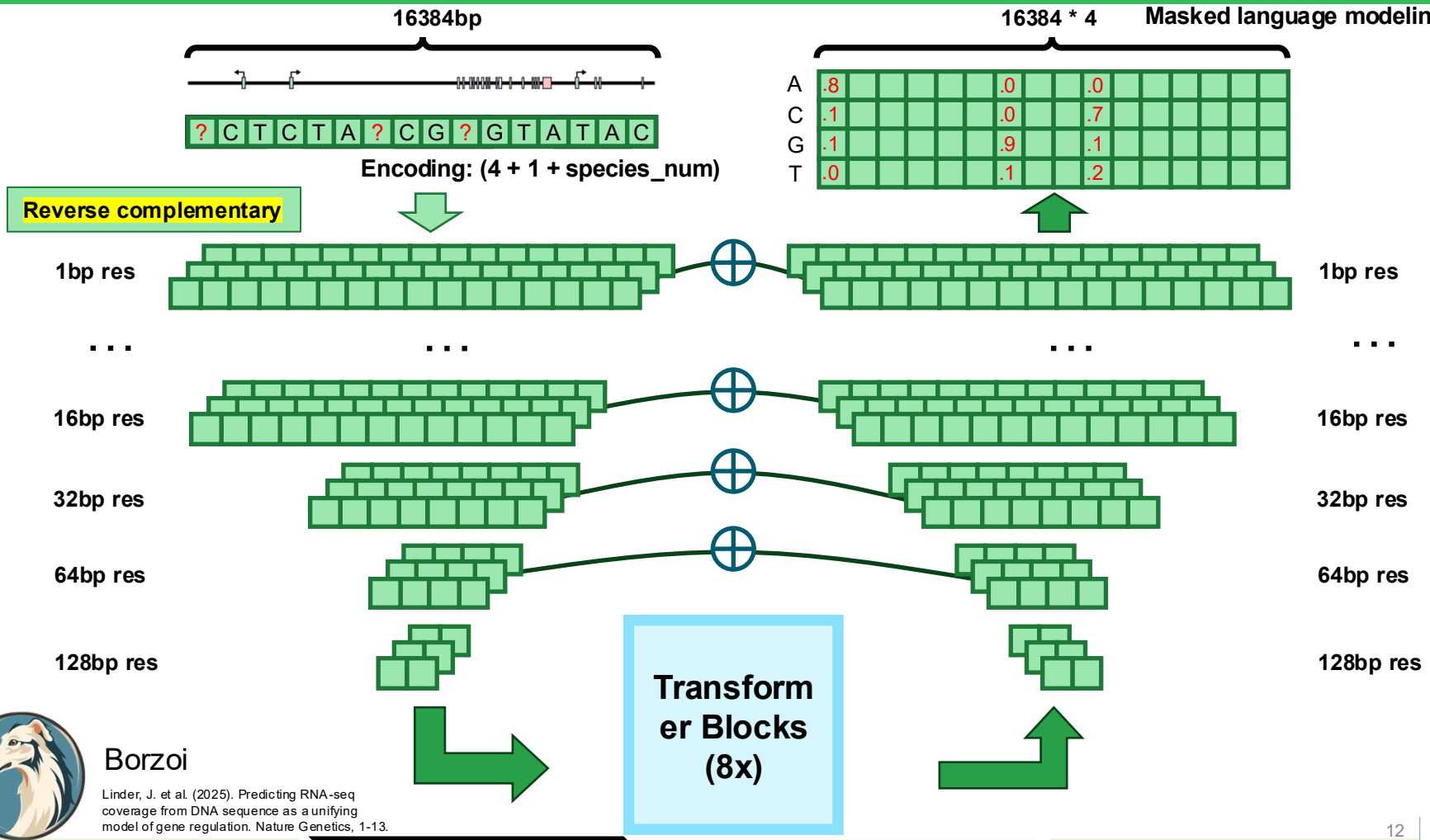


Sequence Preprocessing



Sequence Preprocessing



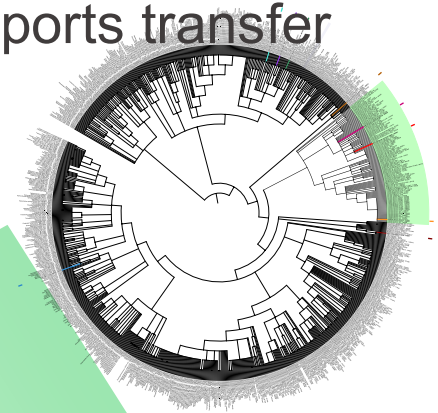


Borzoi

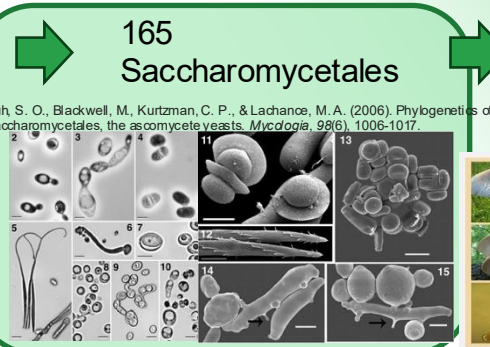
Linder, J. et al. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics, 1-13.

Finding the Evolutionary "Sweet Spot"

- **The Core Question:** What evolutionary scope best supports transfer learning to *S. cerevisiae*?
- The Four Training Corpora
  - Species-level: Single *S. cerevisiae* reference (R64).
  - Strain-level: 80 *S. cerevisiae* strains.
  - Order-level: 165 Saccharomycetales genomes.
  - Kingdom-level: 1,341 fungal genomes spanning the entire kingdom.

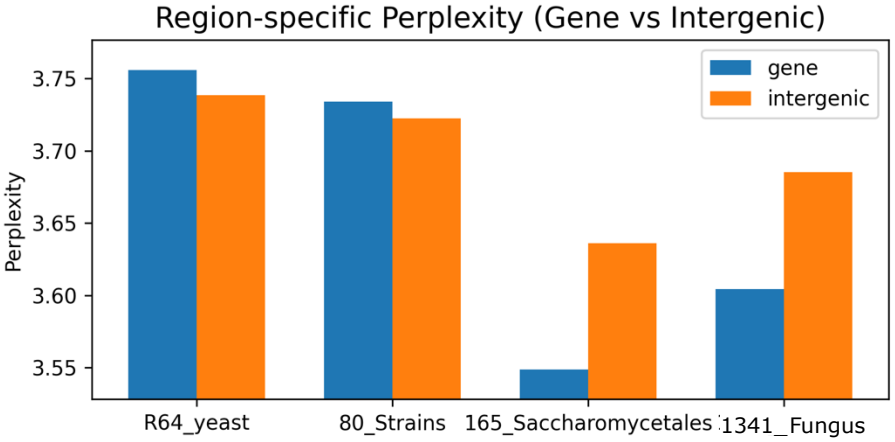


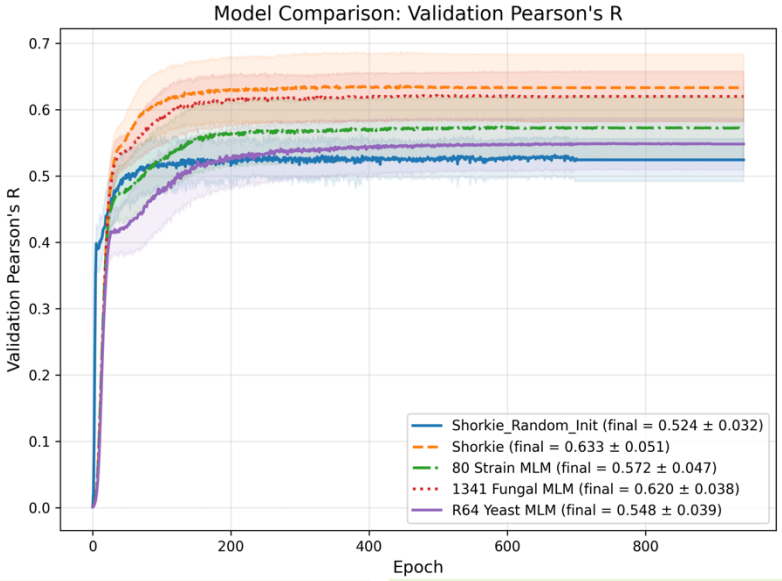
R64 reference yeast → 80 strains of yeasts → 165 Saccharomycetales → 1341 Fungus genomes



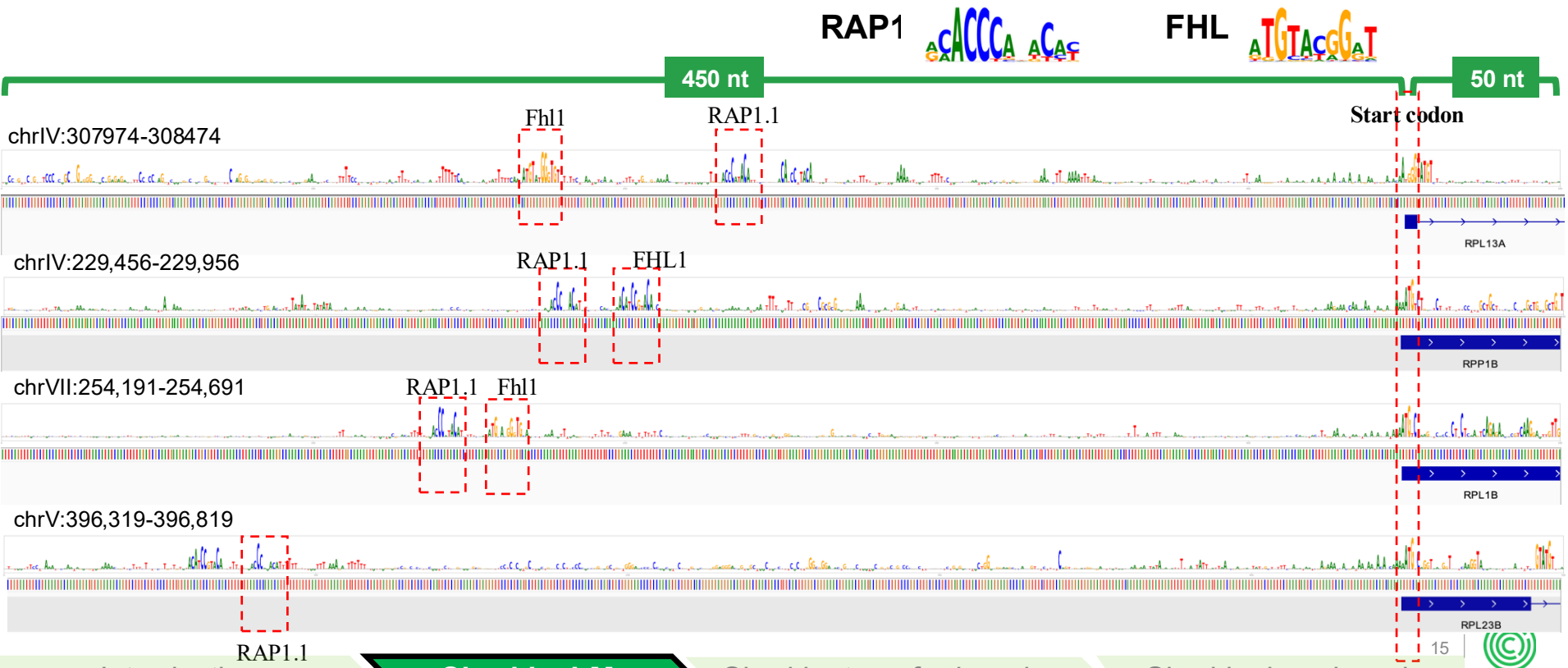
Order-Level Pretraining Gives the Best Transfer to Yeast

- The kingdom-level model suffered from extreme heterogeneity and noisier training dynamics, while the species/strain models overfit.

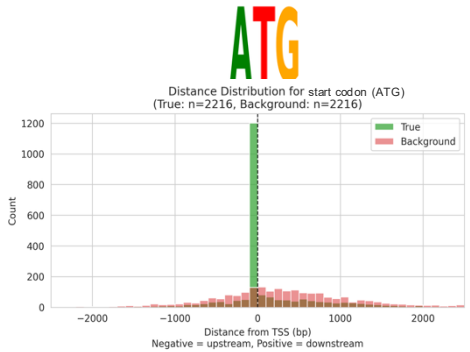


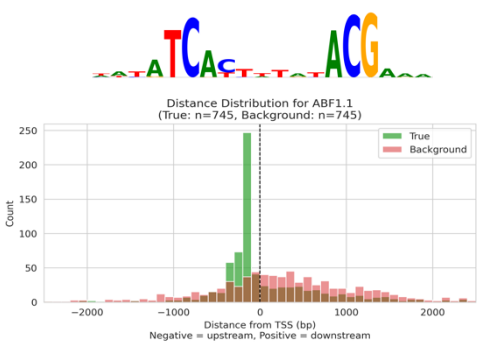


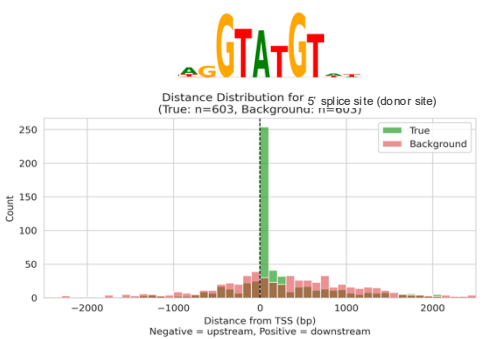
Zero-Shot Regulatory Grammar: Shorkie LM predicts motifs in RP gene promoter regions

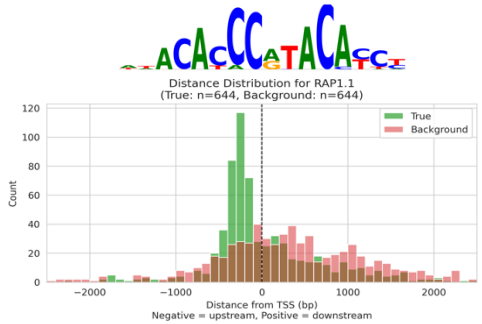


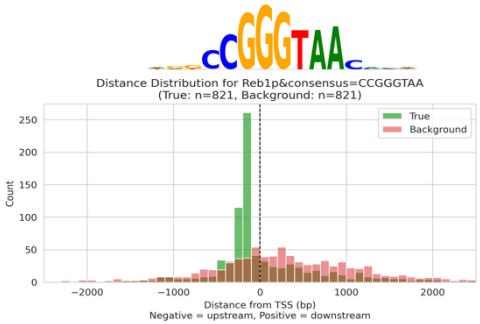
Zero-Shot Regulatory Grammar: motifs are enriched in promoter & genic regions

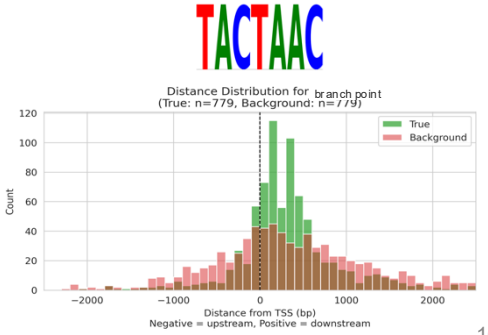








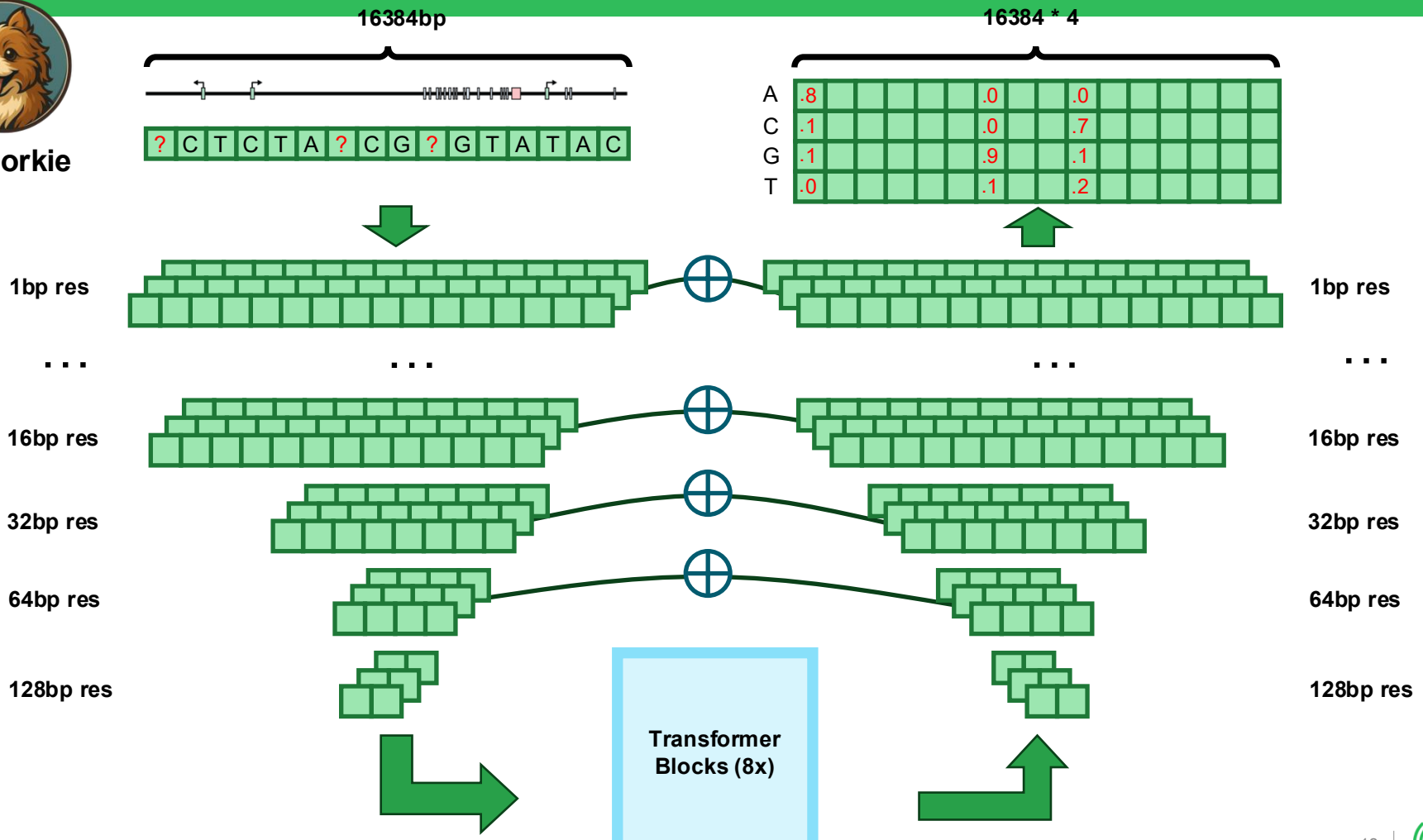




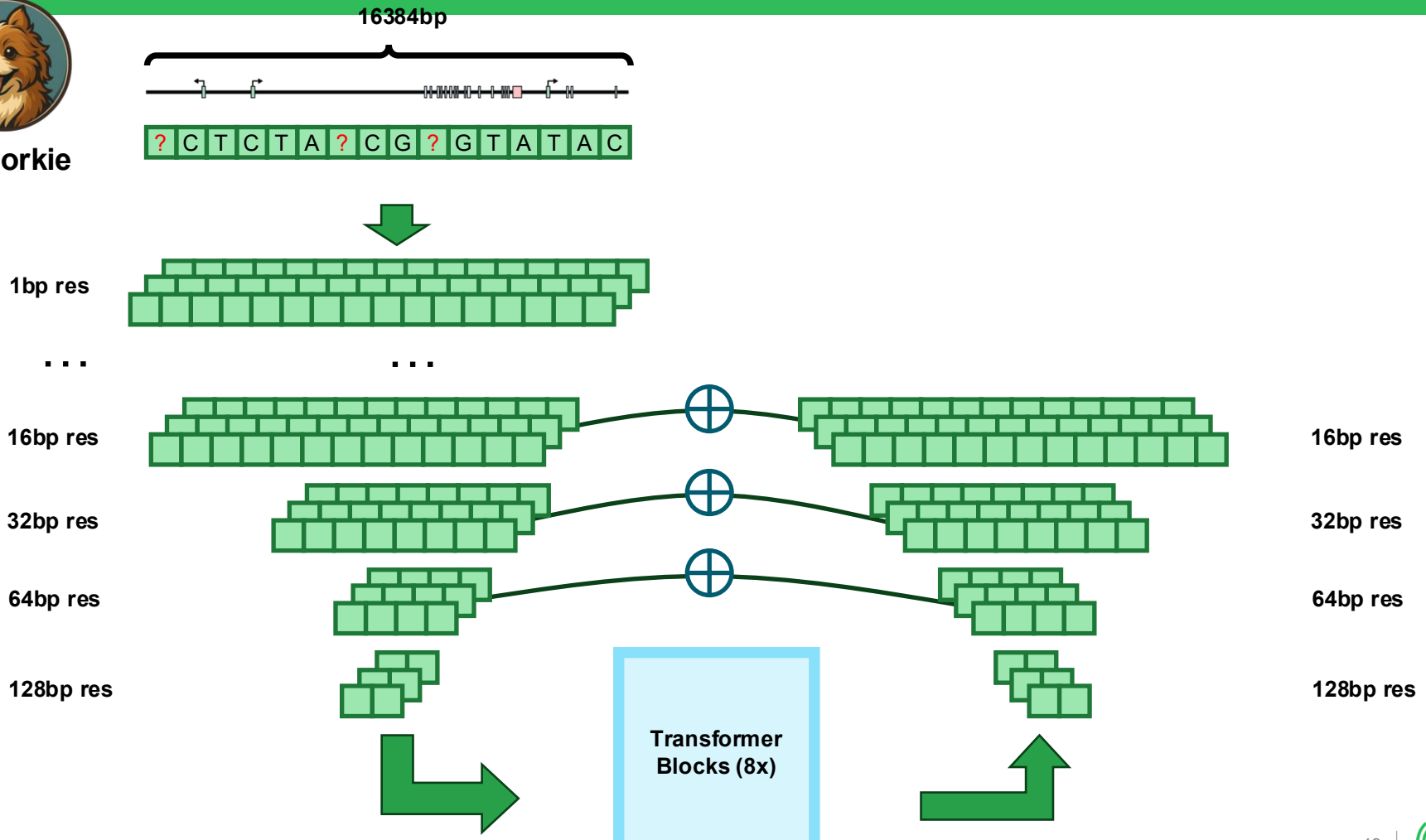
Shorkie

Fine-tuning Shorkie LM with thousands of RNA-Seq

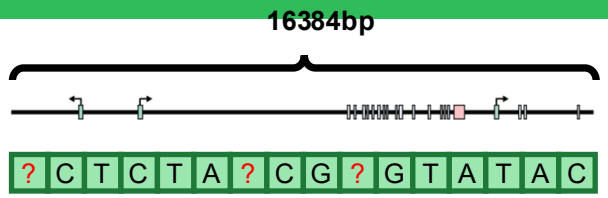
Shorkie

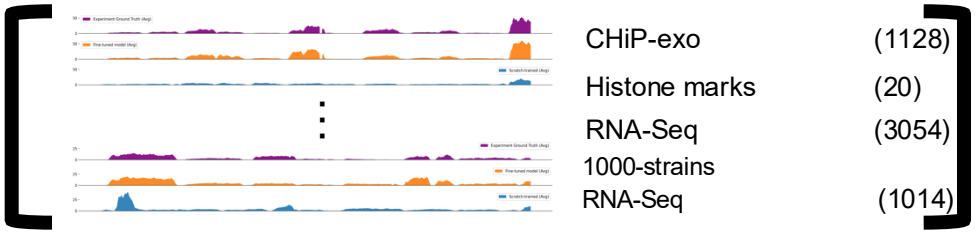

16384bp
16384 * 4
? C T C T A ? C G ? G T A T A C
A .8 .0 .0
C .1 .0 .7
G .1 .9 .1
T .0 .1 .2
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

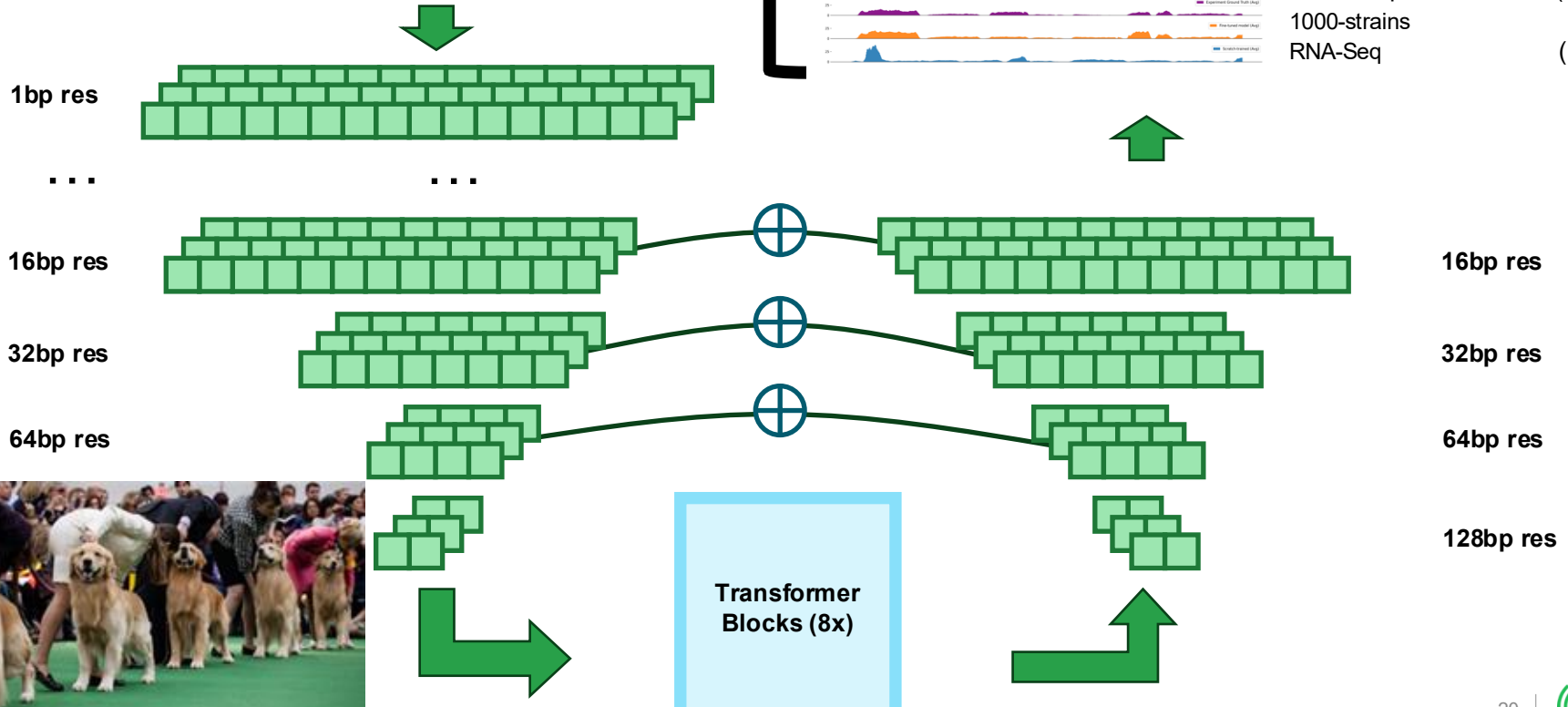
Shorkie


16384bp
? C T C T A ? C G ? G T A T A C
1bp res
. . .
. . .
16bp res
32bp res
64bp res
128bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Shorkie

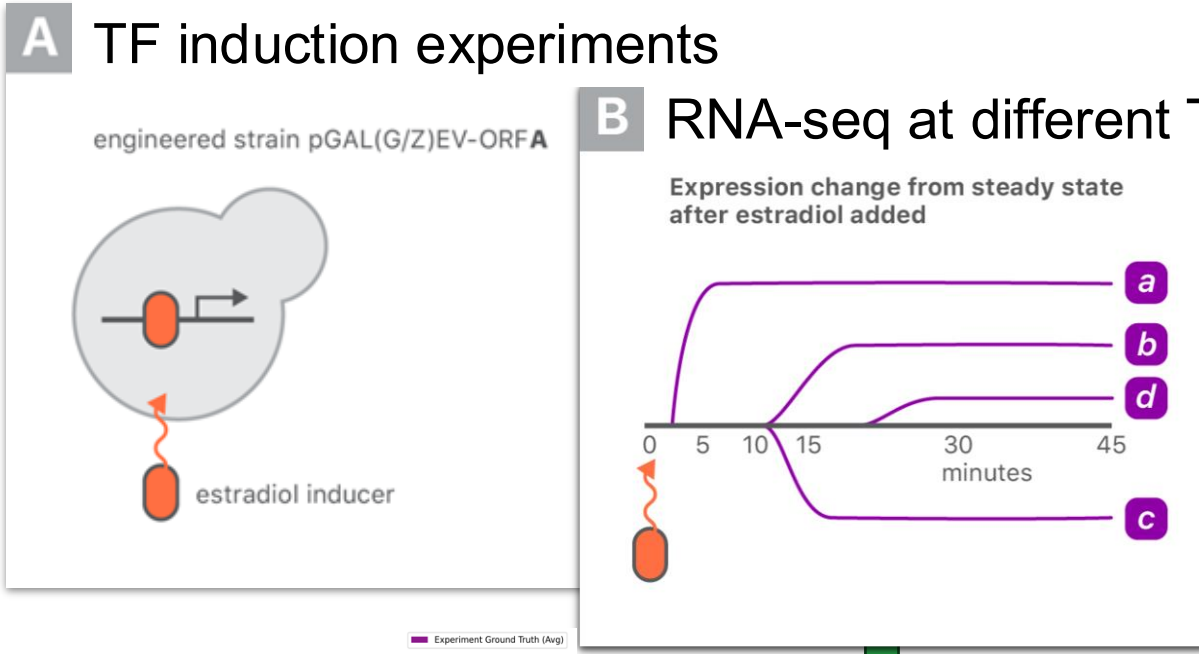

16384bp
? C T C T A ? C G ? G T A T A C


Coverage Tracks (896 * 2488)
CHiP-exo (1128)
Histone marks (20)
RNA-Seq (3054)
1000-strains RNA-Seq (1014)

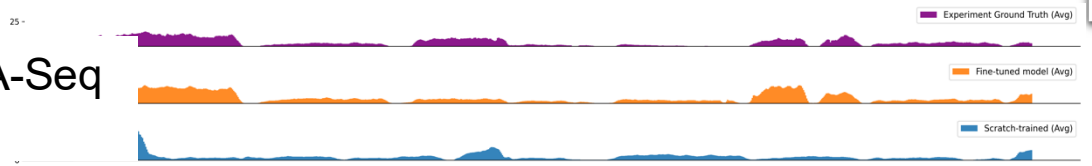

1bp res
16bp res
32bp res
64bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)

Calico In-house Large-Scale TF-Induction RNA-seq Enables Dynamic Modeling

- Genome-scale transcription factor perturbation (**1340** experiments; **3054** RNA-Seq readouts)
- Aggregating dynamics across many time-courses



3054 RNA-Seq readouts

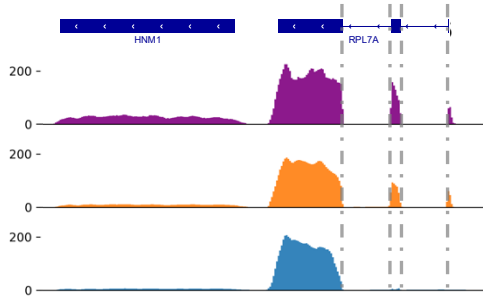


Pretraining Substantially Improves RNA-seq Prediction

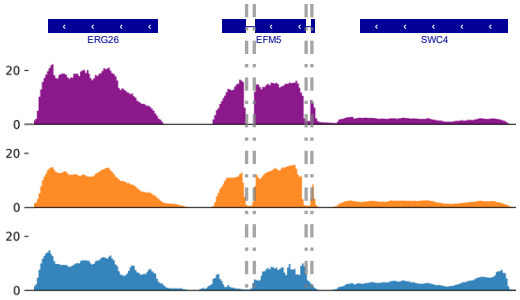
RNA-Seq tracks

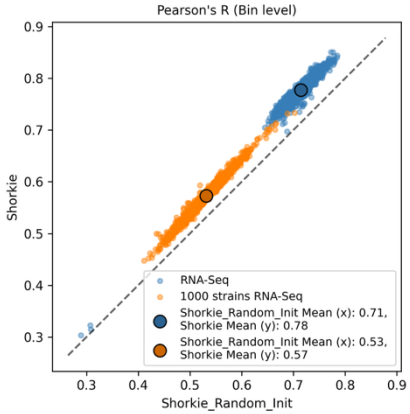
Experiment Ground Truth (Avg) | Fine-tuned model (Avg) | Scratch-trained (Avg)

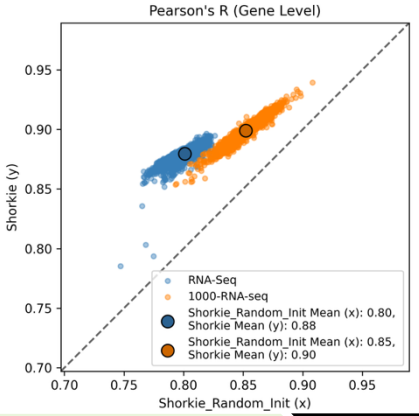
chrVII:362,180-366,023 (RNA-Seq tracks)

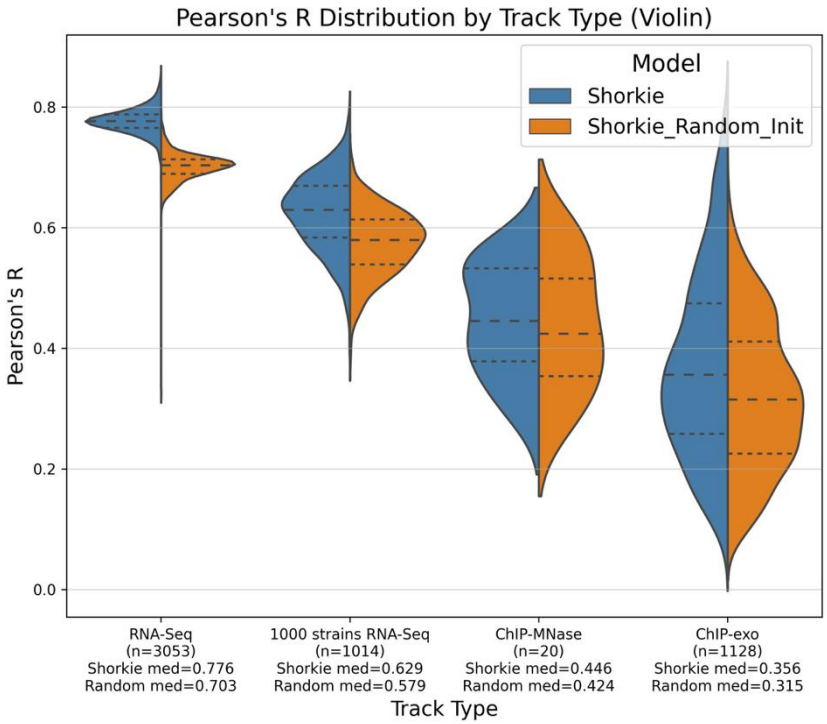


chrVII:495,374-499,965 (RNA-Seq tracks)

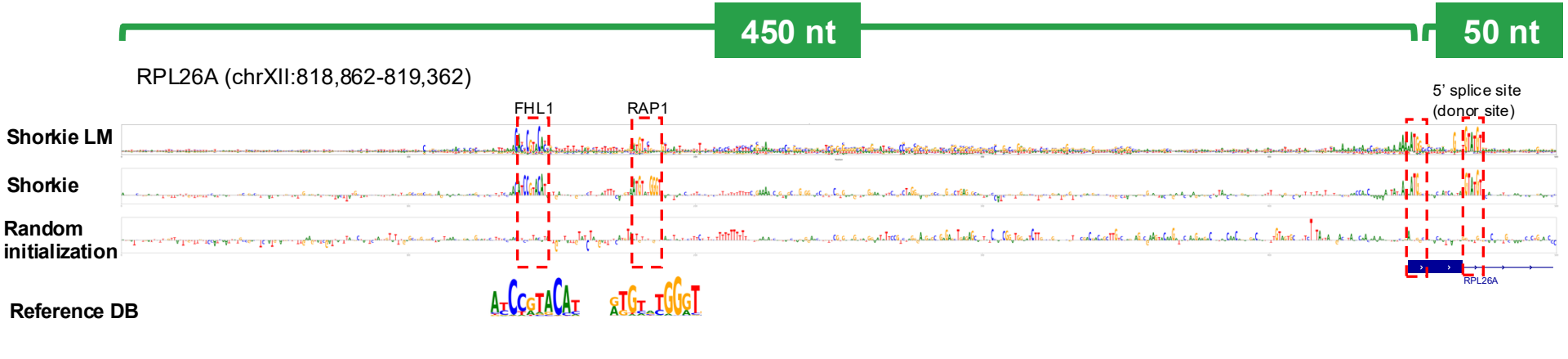


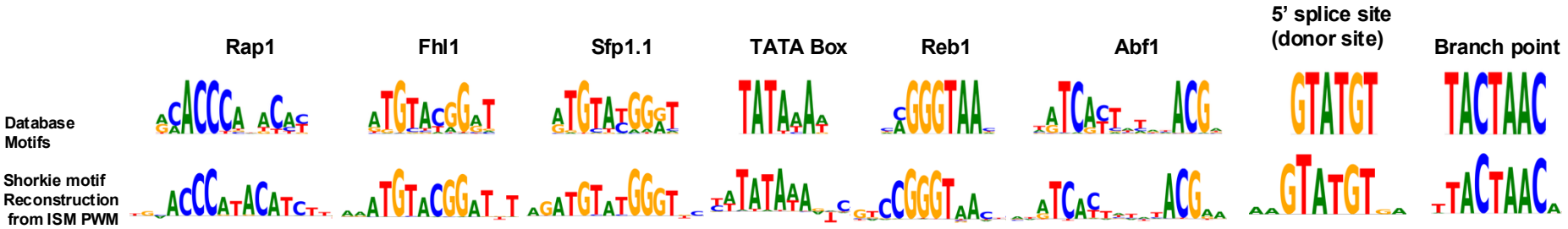






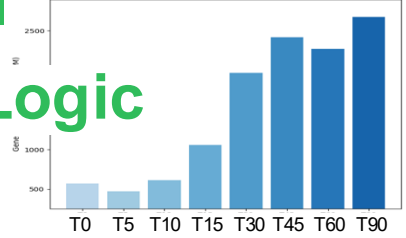
Pretraining Sharpens Interpretable Regulatory Motifs in Ribosomal genes & RRB genes

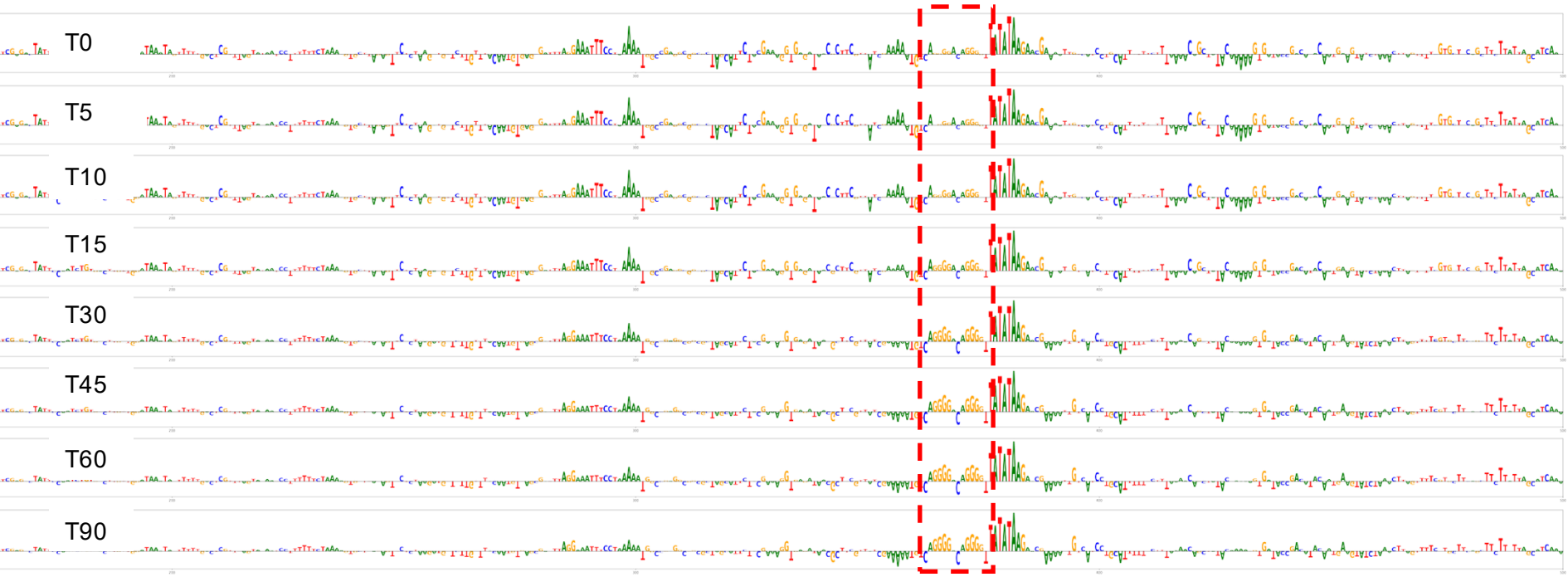




Shorkie Captures Time-Dependent Regulatory Logic

chrII:515214-515714 (Promoter region of ATG42)





YeTFaSCo DB motif

MSN2

TATA Box

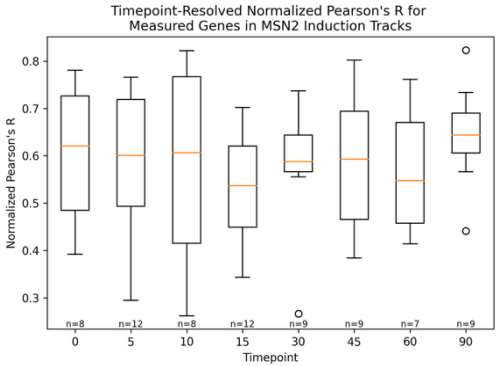
Shorkie predicts average temporal motif changes

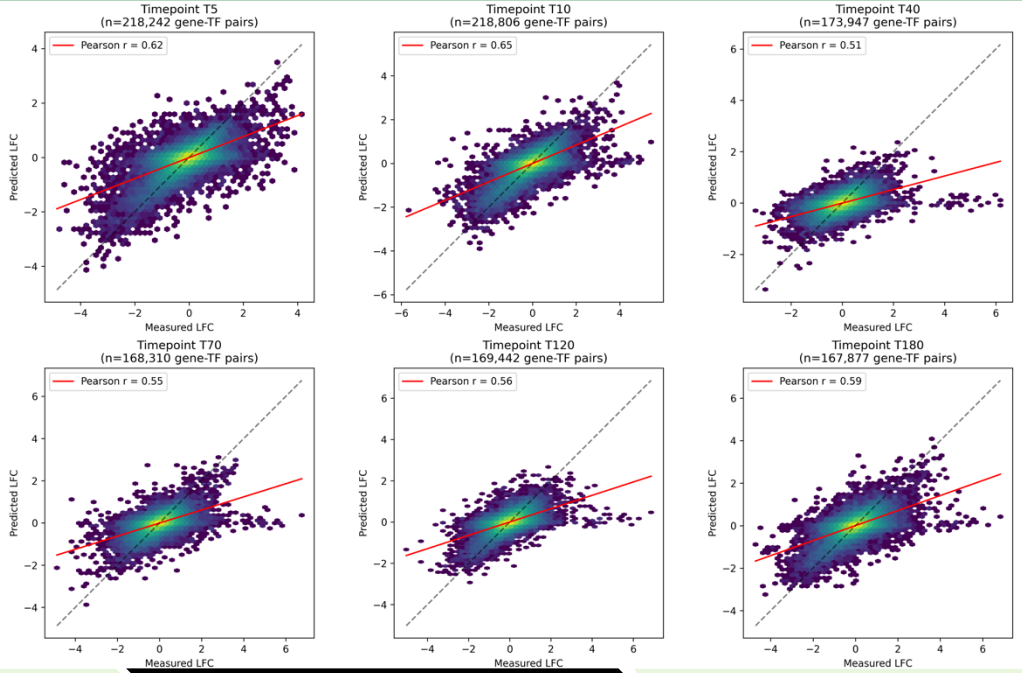
MSN2 Induction

YeTFaSCo DB motif 

TF-Modisco-detected MSN2 binding site

$T_5 - T_0$	$T_{10} - T_0$	$T_{15} - T_0$	$T_{30} - T_0$	$T_{45} - T_0$	$T_{60} - T_0$	$T_{90} - T_0$
Not clustered						

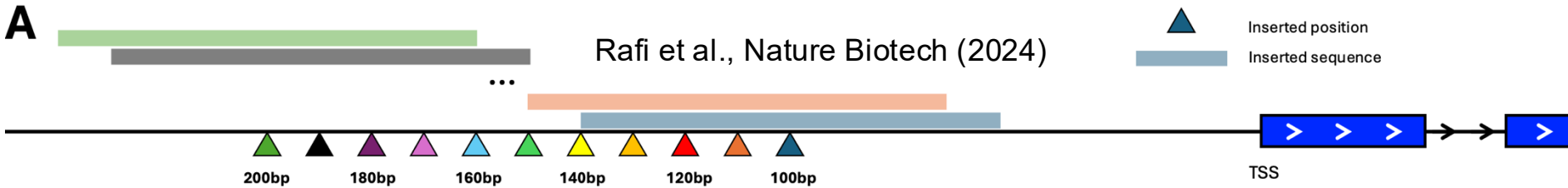




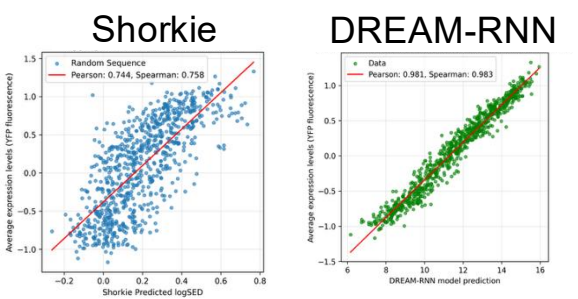
Shorkie Benchmark

Zero-shot prediction on MPRA and eQTL

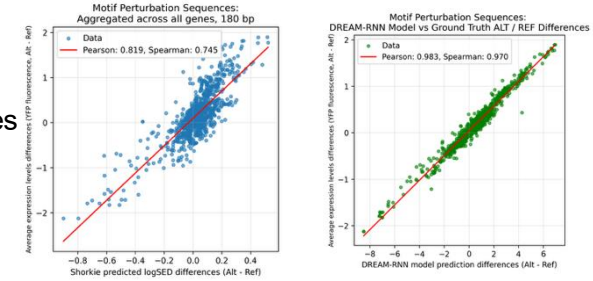
Shorkie Performs Well on Zero-Shot MPRA Prediction



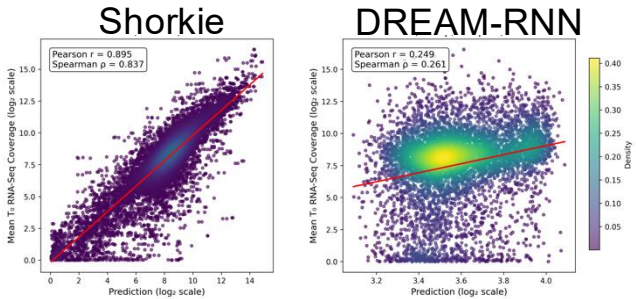
Rafi et al. Random Sequences (single-sequence)



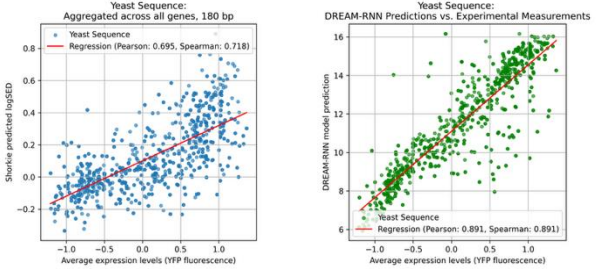
Rafi et al. Motif Perturbation Sequences (dual-sequence)



RNA-Seq Coverage Prediction



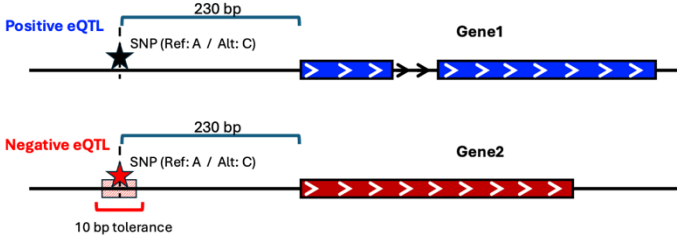
Rafi et al. Yeast Sequences (single-sequence)



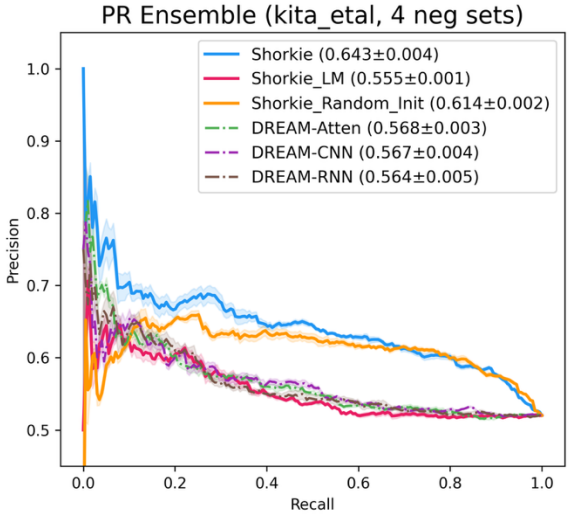
Shorkie Improves Prediction of eQTL Regulatory Effects

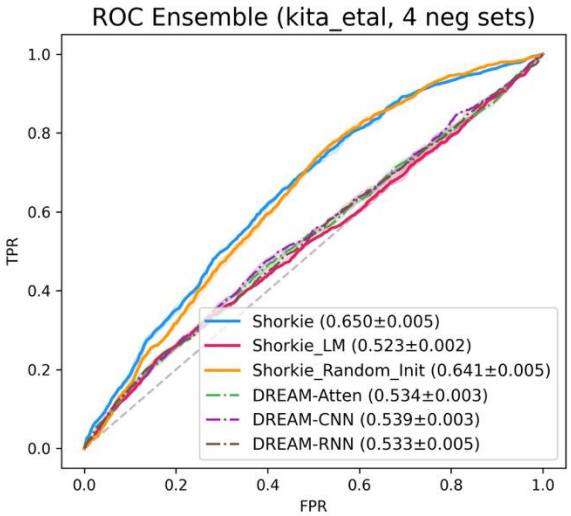
Caudal et al., Nature genetics (2024)

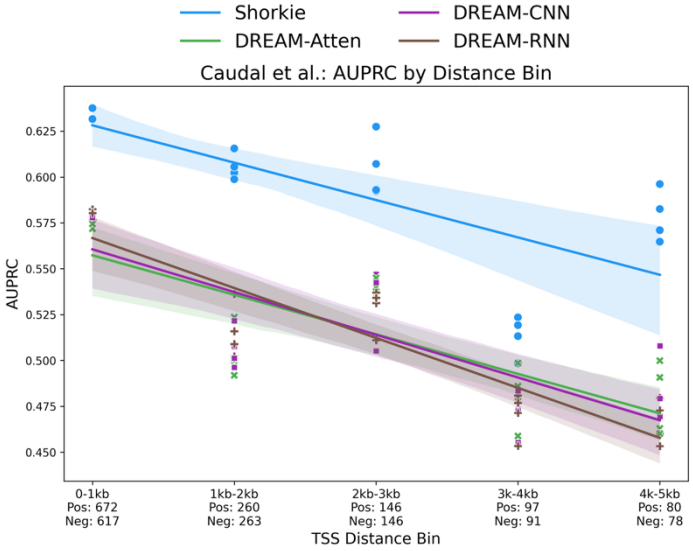
Renganaath et al., eLife 2020



Kita et al., Nature genetics (2024)

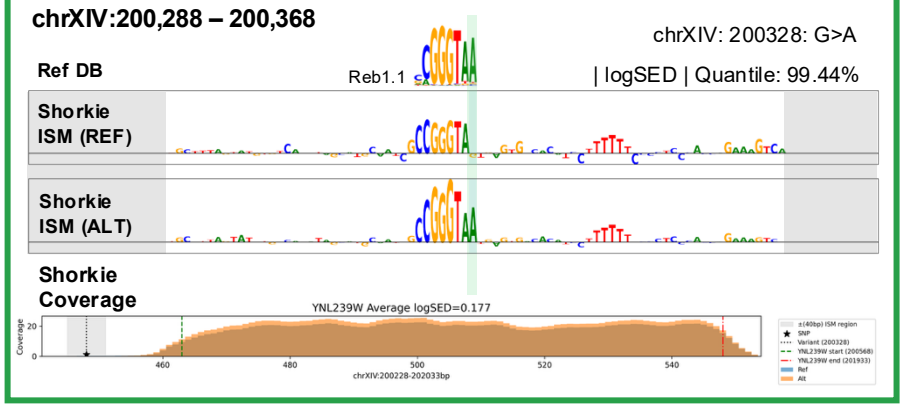




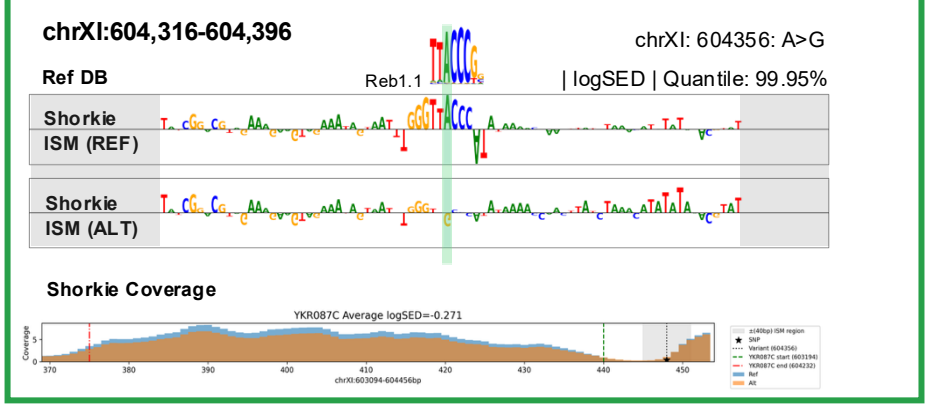


Interpreting Variant Effects via Regulatory Motif Changes

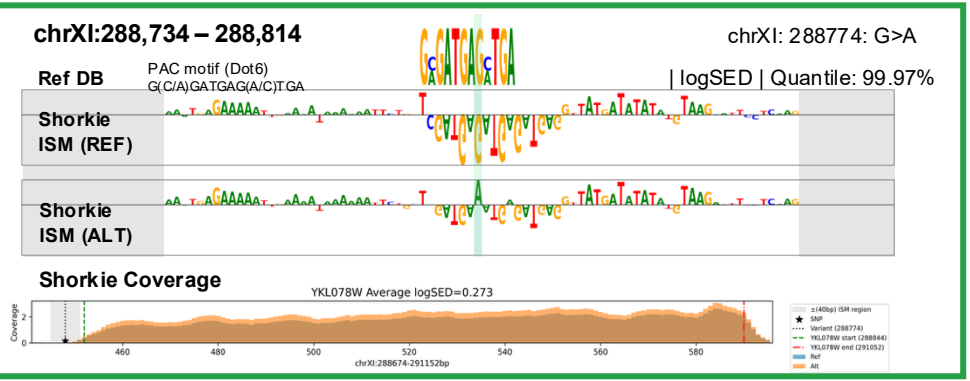
Motif gain → increased expression



Motif loss → decreased expression



Repressor motif loss → increased expression



Conclusion

- **The Takeaway:** For small-genome organisms, phylogenetically informed pretraining combined with transfer learning is a promising framework to decode complex gene regulation and noncoding variants. Shorkie improved dynamic RNA-seq prediction and enabled strong zero-shot variant-effect prediction on MPRA and eQTL benchmarks.
- **Shorkie** → learns native yeast regulatory grammar through phylogenetically informed pretraining and transfer learning.
  - (*Chao et al, bioRxiv, Sep 2025*)
- **Yorzoi** → shows that exogenous sequence expression can be learned directly.
  - (*Schneider et al, bioRxiv, Sep 2025*)
- **ExoShorkie** → shows that the Shorkie framework can be transferred into the exogenous setting.
  - (*Mandl et al, bioRxiv, Jan 2026*)

ACKNOWLEDGEMENTS

Calico
NIH National Institutes of Health
NSF


David Kelley

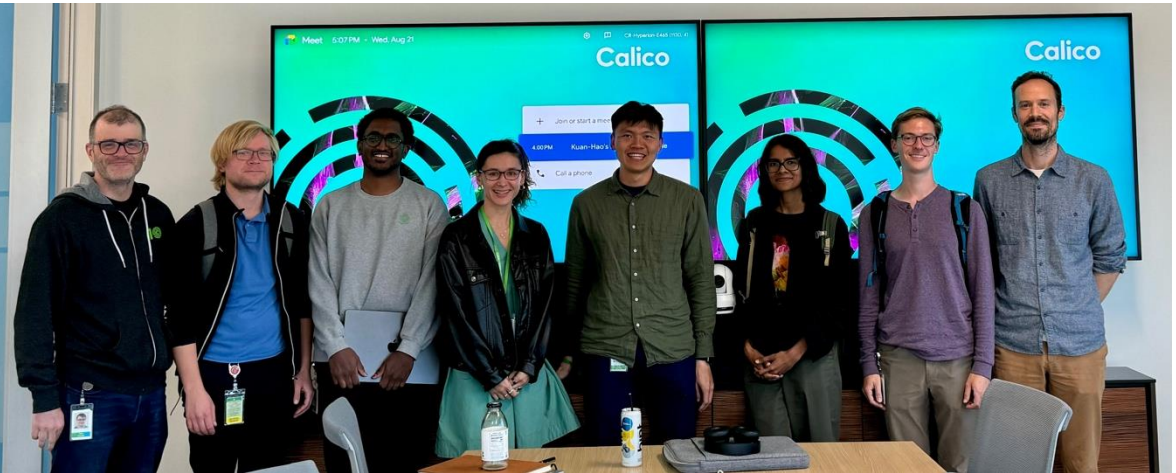

Johannes Linder


Majed Magzoub





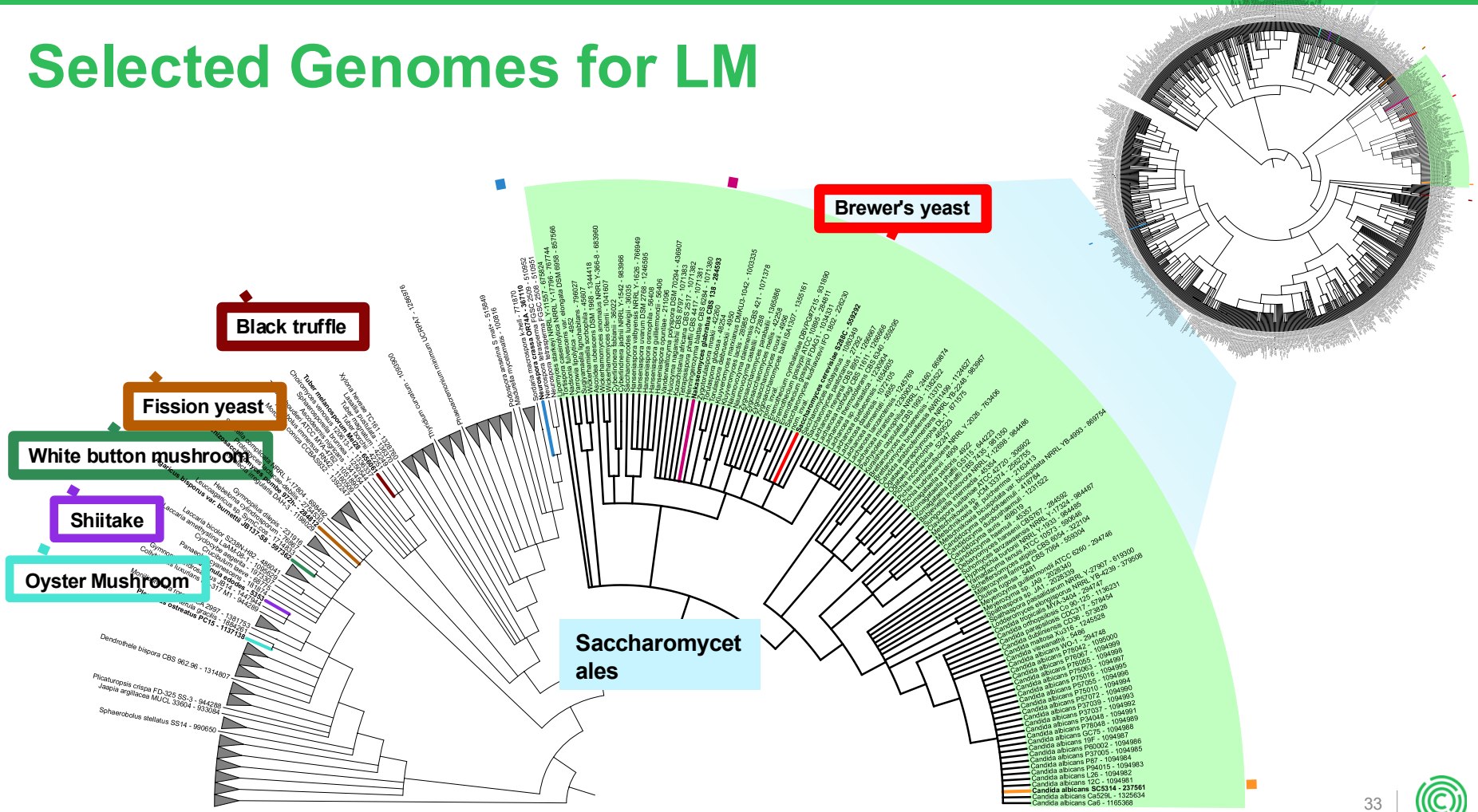


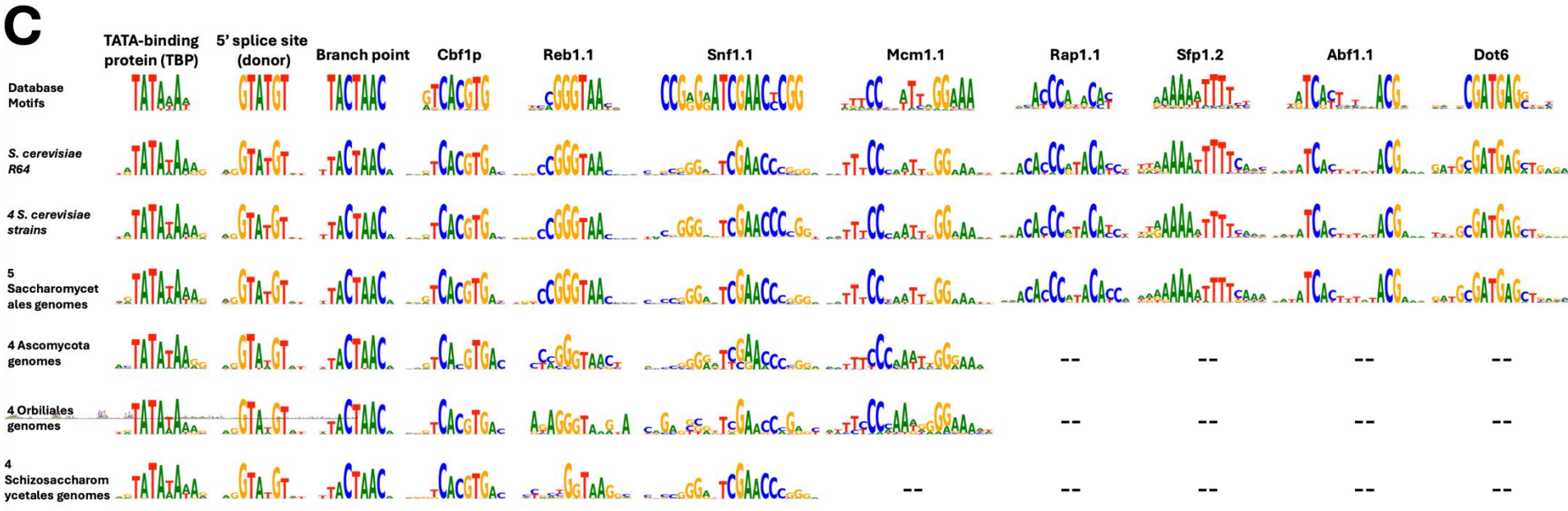

Steven Salzberg


Mihaela Pertea

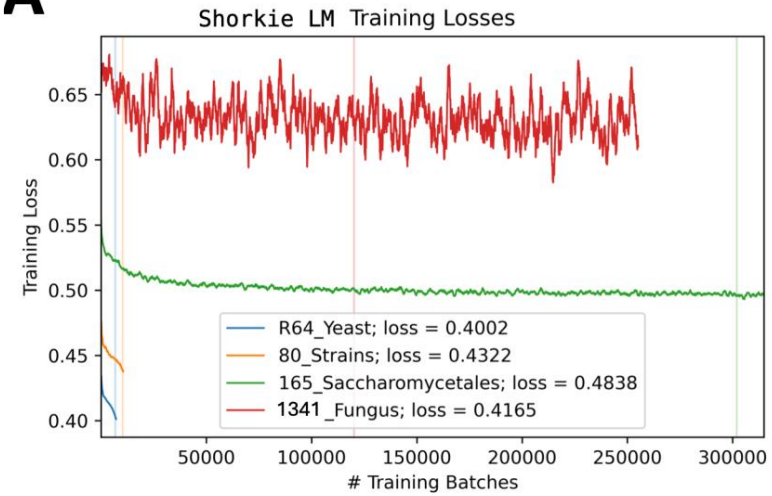
Selected Genomes for LM



Inference of Shorkie LM Across Species with Varying Evolutionary Distances



A

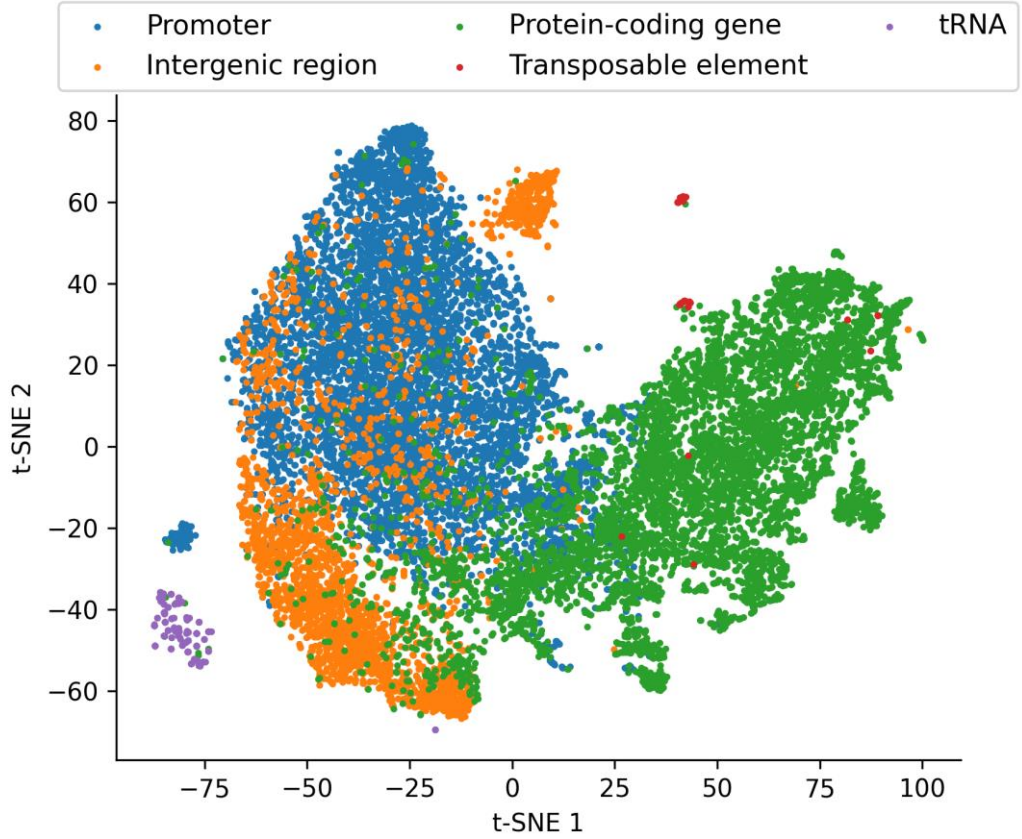


B

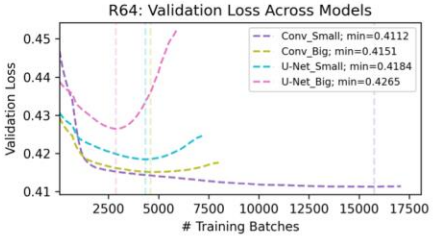
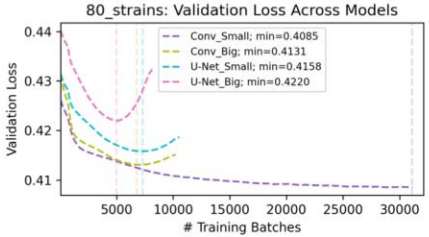
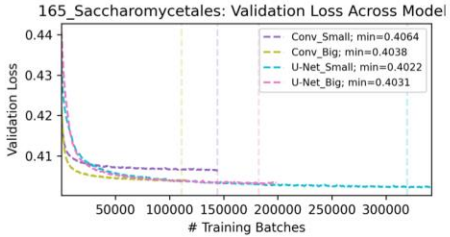
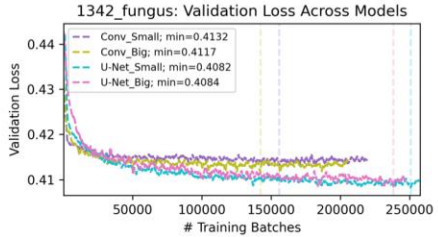
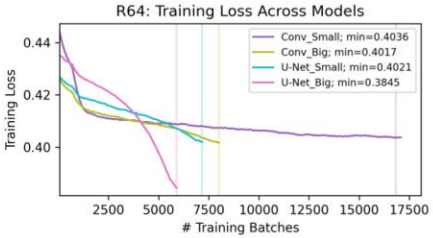
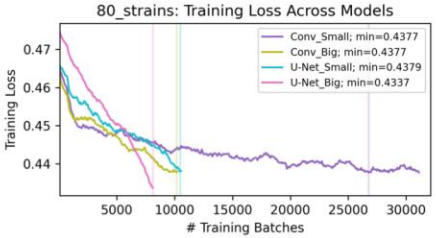
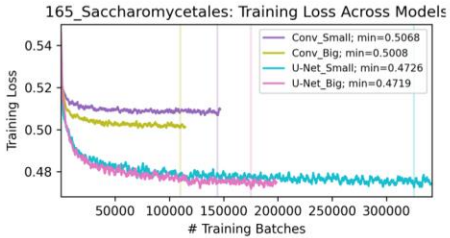
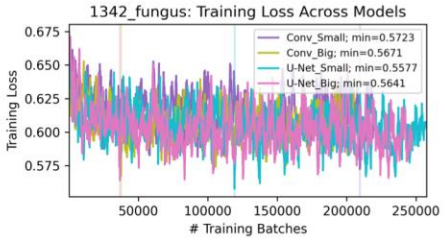
Model architecture comparison

Model	Architecture	Total Params	Key Features
Conv_Small	Conv1D + Residual blocks	320,708	Smaller conv width (64 channels).
Conv_big	Conv1D + Residual blocks	3,642,116	Larger conv width (256 channels).
Unet_small	U-Net + Transformers	13,665,828	Smaller conv width (96 channels) + 8 transformer blocks
Unet_big	U-Net + Transformers	71,790,564	Smaller conv width (384 channels) + 11 transformer blocks

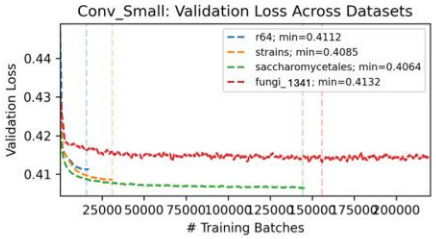
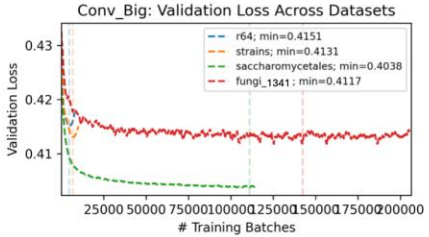
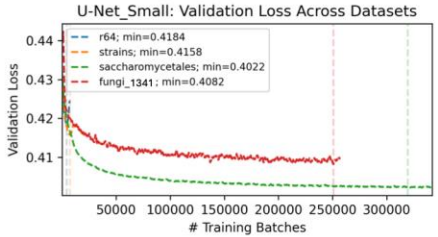
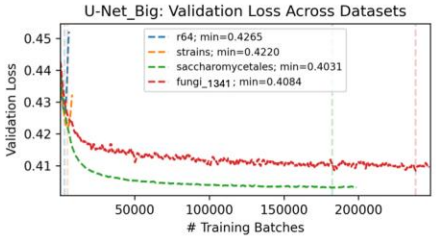
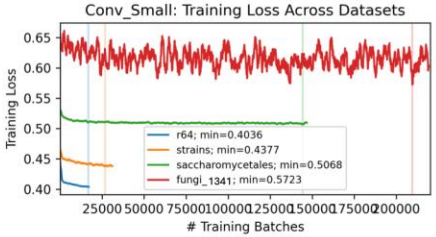
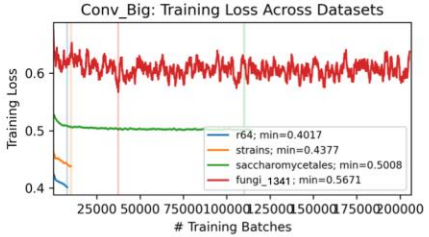
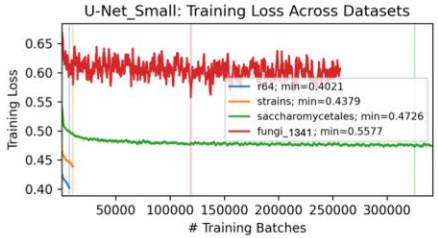
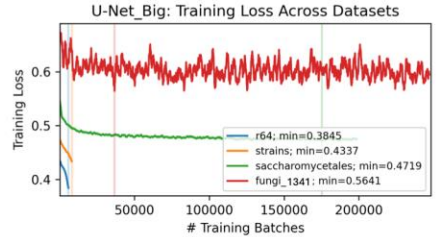
Shorkie LM t-SNE(t-distributed stochastic neighbor embedding)



Shorkie LM, trained on different dataset

	R64	80_Strains	165 Saccharomycetales	1341_Fungus
Validation				
Training				

Shorkie LM, four models we trained

	Conv_Small	Conv_Big	U-Net_Small	U-Net_Big
Validation				
Training				

Shorkie, Shorkie LM, Shorkie_Random_Init Attention Maps