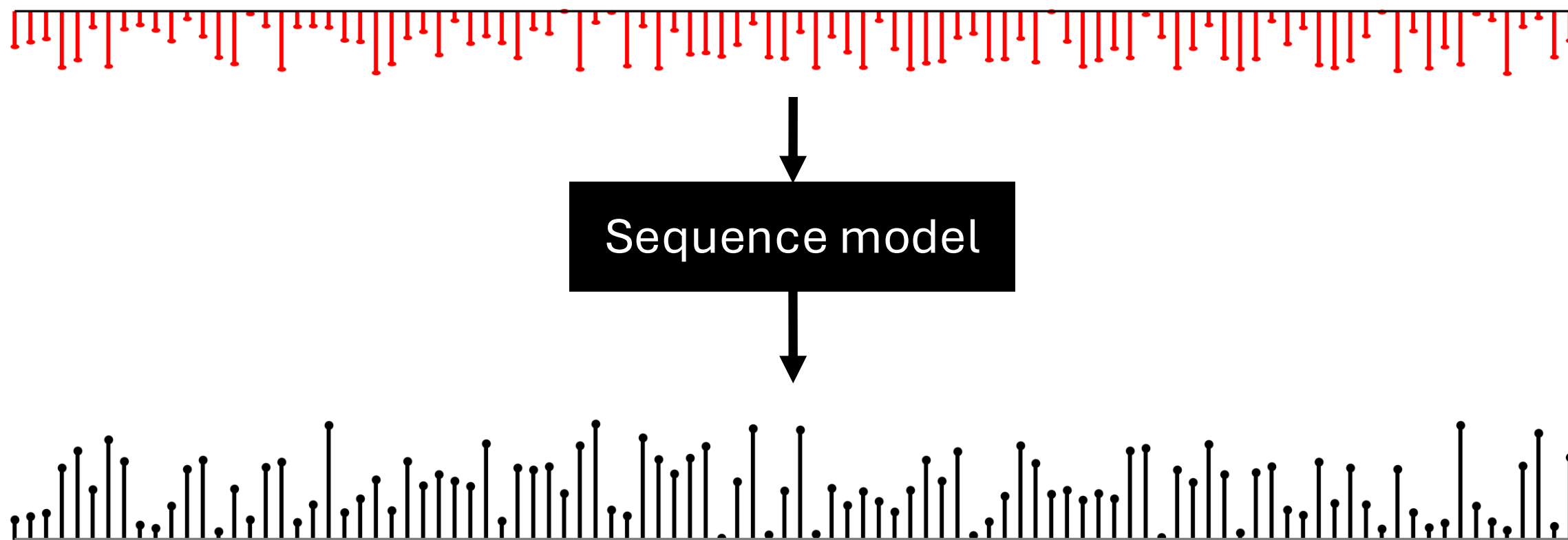




Unifying ChIP-exo DNA-Binding and RNA-Seq Coverage Predictions with a Multi-Species Fungal Language Model

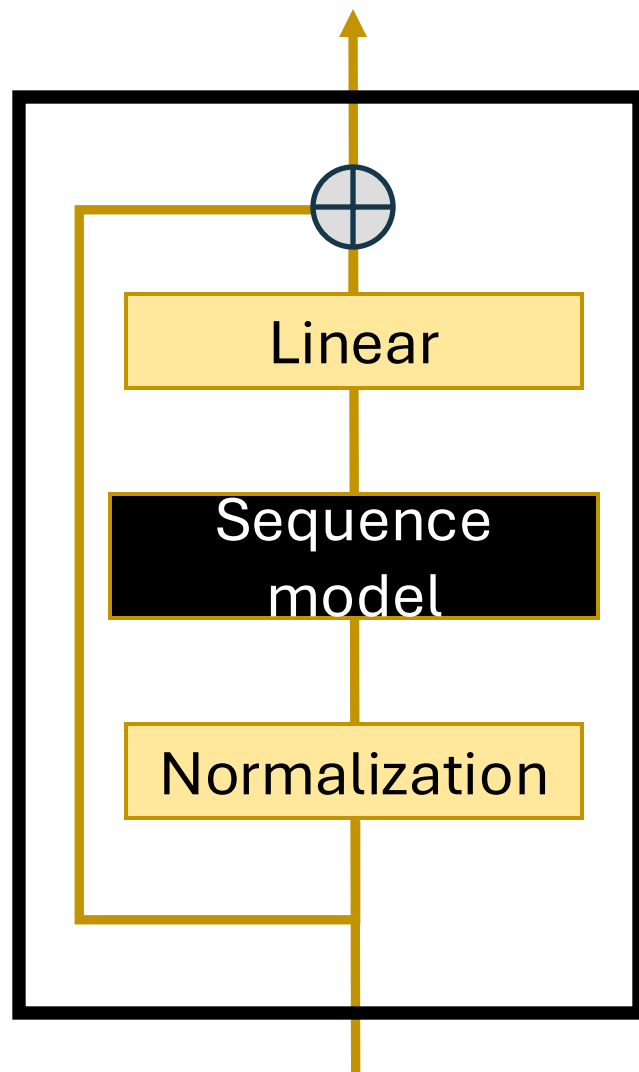
Kuan-Hao Chao

2025.01.22



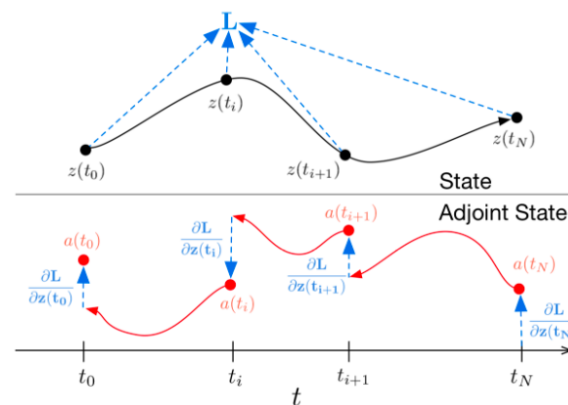
Sequence models map a sequence to a sequence

(batch, length, dim)

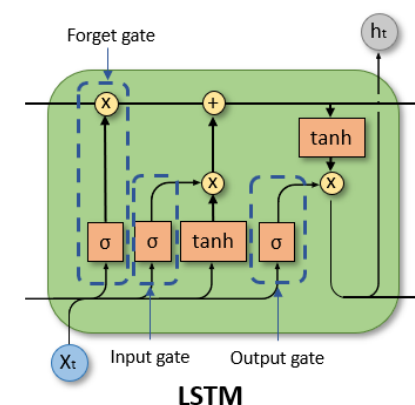


(batch, length, dim)

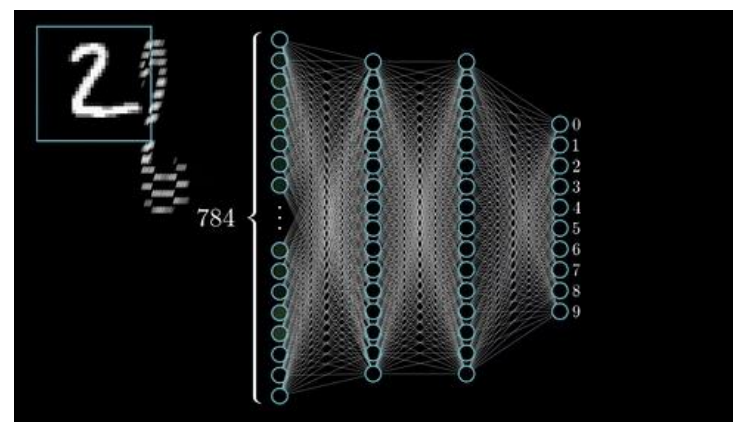
Neural ODEs



RNN



CNNs



Transformers

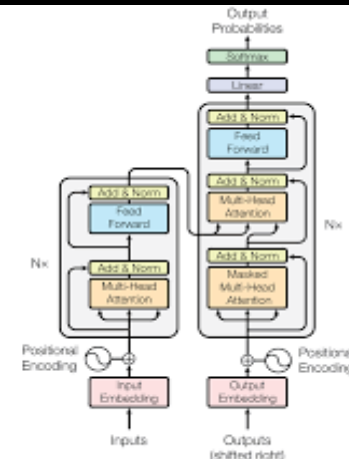


Figure 1: The Transformer - model architecture.

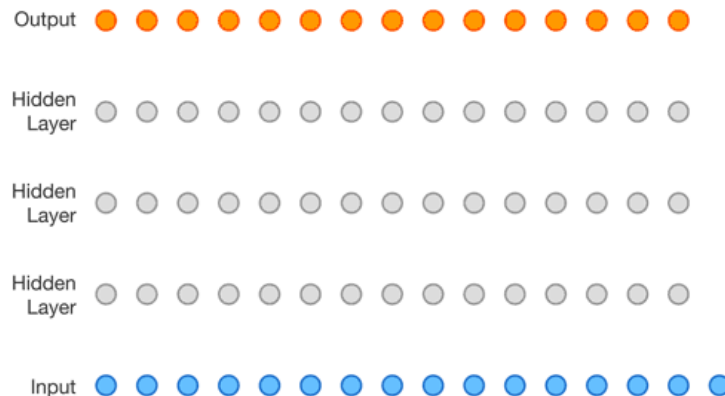
Gemini
LLaMA

Input Prompt:

Recite the first law of robotics

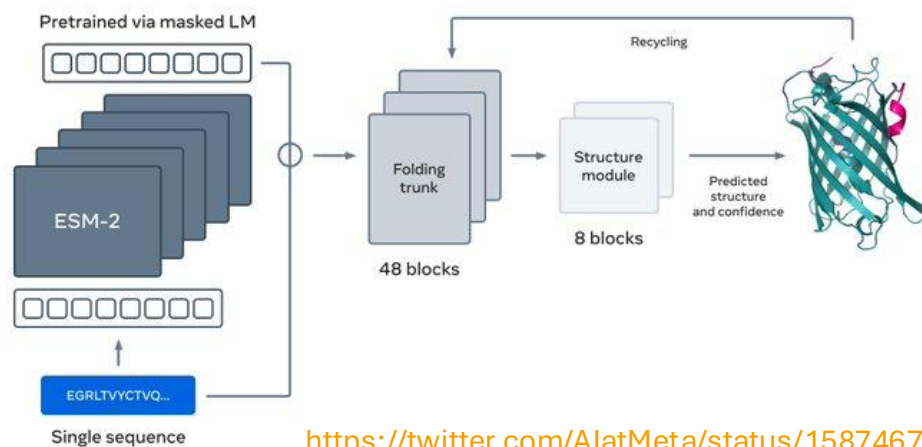


Output:



<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>

<https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/>



ESMFold



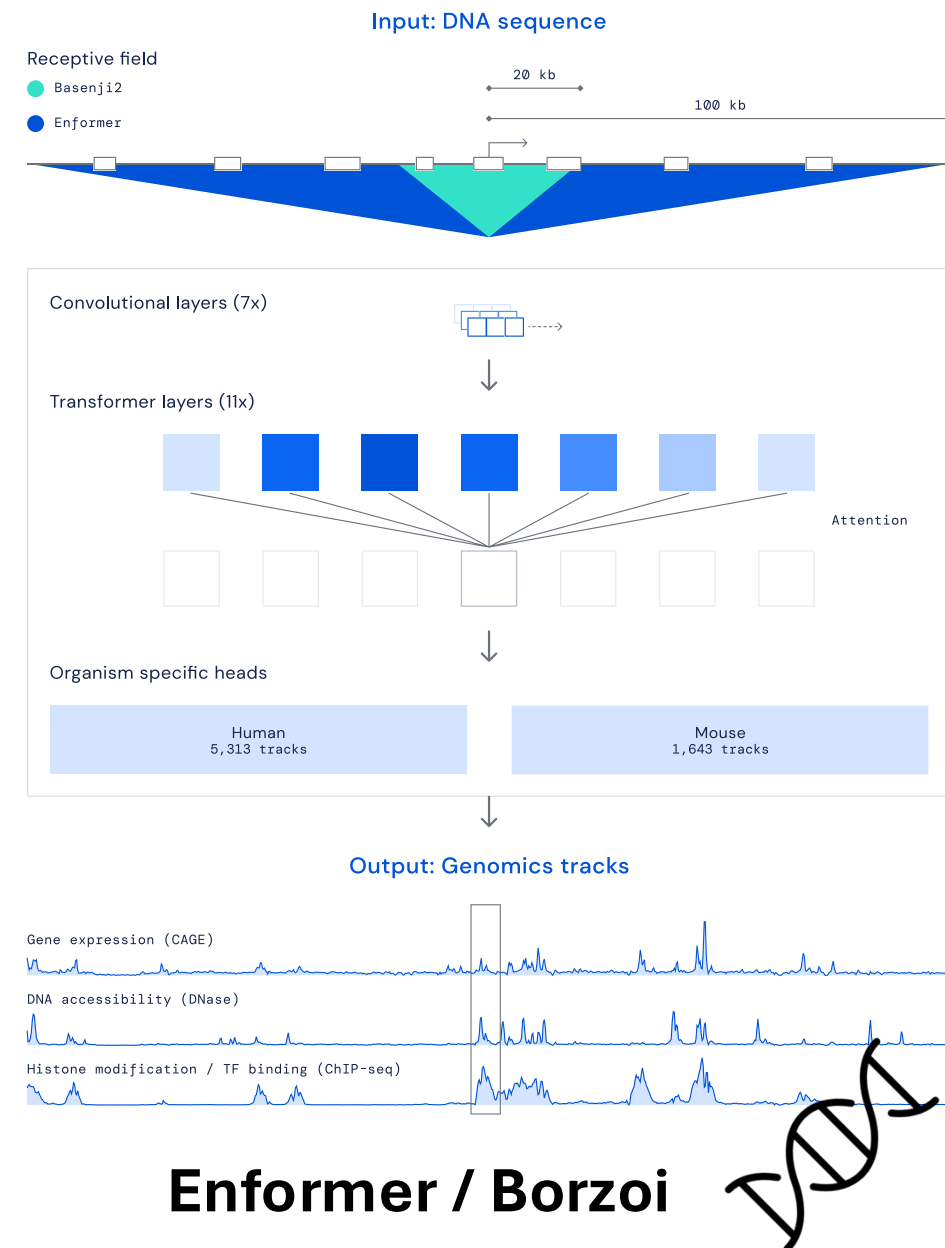
<https://twitter.com/AlatMeta/status/1587467600413351937/photo/1>

Introduction

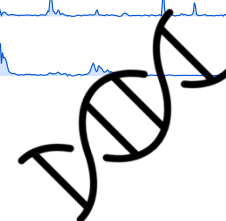
Pre-training LM

Fine-tuning LM

Applications & Conclusions

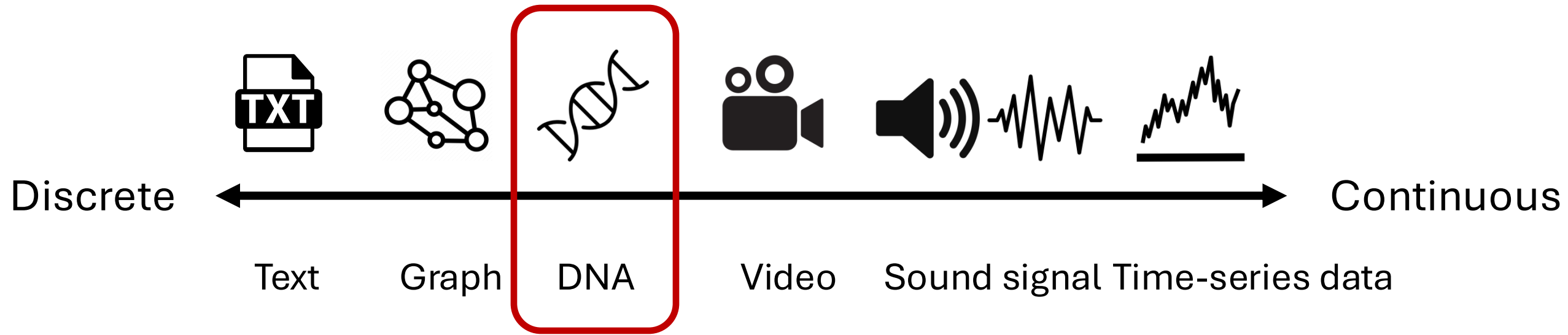


Enformer / Borzoi



<https://deepmind.google/discover/blog/predicting-gene-expression-with-a/>

Spectrum of Sequential Data



Deep learning-based DNA sequence model

nature methods

Explore content

Troyanskaya Lab Princeton

DeepSEA

2015

Predicting effects of noncoding DNA sequence on gene expression with deep learning-based sequence models

Jian Zhou & Olga G Troyanskaya

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR

Institution: MILTON S EISENHOWER LIBRARY

Basset: learning the regulatory grammar of accessible genome with deep convolutional neural networks

David R. Kelley¹, Jasper Snoek² and John L. Rinn¹

Calico

Basset

2016

Cell

Volume 176, Issue 3, 24 January 2019, Pages 535-548.e24

Article

Predicting Splicing from Primary RNA-seq with Deep Learning

Kishore Jaganathan^{1,4}, Sofia Kyriazopoulou Panagiotopoulou^{1,4}, Siavash Fazel Darbandi², David Knowles³, Yang Li^{1,3}, Jack Wenwu Cui^{1,3}, Grace B. Schwartz², Eric D. Chow¹, Efsthios Efthymiou^{1,3}, Serafim Batzoglou¹, Stephan J. Sanders², Kyle Kai-How Farh^{1,2,3,4}

Calico

SpliceAI

2019

Agarwal and Kelley Genome Biology (2022) 23:245
https://doi.org/10.1186/s13059-022-02811-x

RESEARCH

The genetic and biochemical determinants of mRNA degradation rates in mammals

Vikram Agarwal^{1,2*} and David R. Kelley^{1*}

Calico

Saluki

2022

nature biotechnology

Explore content

About the journal

Publish with

FUToronto

DeepBind

2015

Predicting the sequence specificity of DNA- and RNA-binding proteins with deep learning

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch & Brendan J. Frey

CSH PRESS GENOME RESEARCH

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR

Institution: MILTON S EISENHOWER LIBRARY

Sequential regulatory activity across chromosomes with convolutional neural networks

David R. Kelley¹, Yakir A. Reshef², Maxwell Bileschi³, David Cory Y. McLean³ and Jasper Snoek³

Calico

Basenji

2018

nature methods

Predicting 3D genome folding from Hi-C data with Akita

Geoff Fudenberg^{1,2,3,4}, David R. Kelley^{2,5,6,7} and Katherine S. Pollard^{1,2,3,4}

Calico

Akita

2020

nature methods

Explore content

About the journal

Publish with us

Calico

scBasset

2022

Article | Published: 08 August 2022

scBasset: sequence-based model of chromatin accessibility using convolutional neural networks

Han Yuan & David R. Kelley

Bioinformatics

Article Navigation

JOURNAL ARTICLE

Convolutional neural network for predicting protein binding

Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford

Gifford Lab MIT

DNA-TF binding

2016

nature genetics

Explore content

Troyanskaya Lab Princeton

Deep learning sequence-based model of variant effects on expression and splicing

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. C.

Calico

ExPecto

2018

ARTICLES

Effective gene expression prediction from DNA sequence by integrating chromatin accessibility and gene expression

Žiga Avsec^{1,2,3,4}, Vikram Agarwal^{1,2,4}, Daniel Viswanathan^{1,2,3,4}, Agnieszka Grabska-Barwinska^{1,2,3,4}, Kyle R. Taylor^{1,2,3,4} and David R. Kelley^{1,2,3,4}

Calico

Enformer

2021

nature genetics

Article

Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation

Johannes Linder^{1,2,3,4}, Divyanshi Srivastava^{1,2,3,4}, Han Yuan^{1,2,3,4}, Vikram Agarwal^{1,2,3,4} and David R. Kelley^{1,2,3,4}

Calico

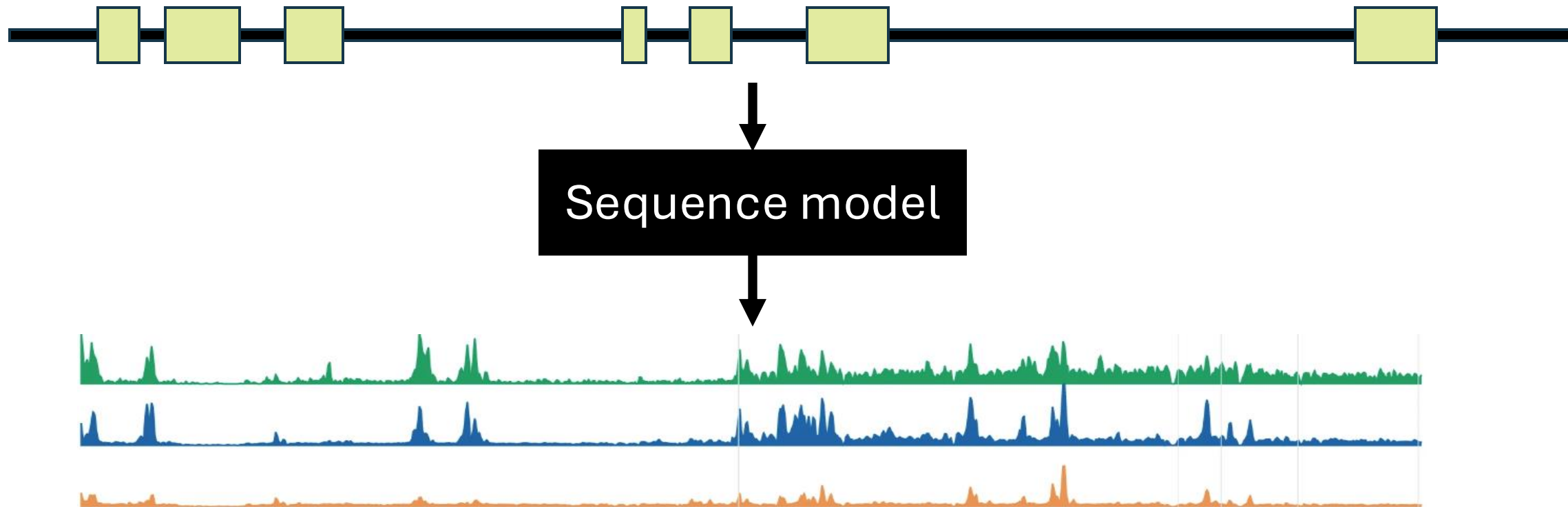
Borzo

2025

Input: DNA sequences

$$(N_{batch_size} * L_{input} * 4)$$

DeepMind + Calico	Calico
Enformer 2021	Borzoi 2025



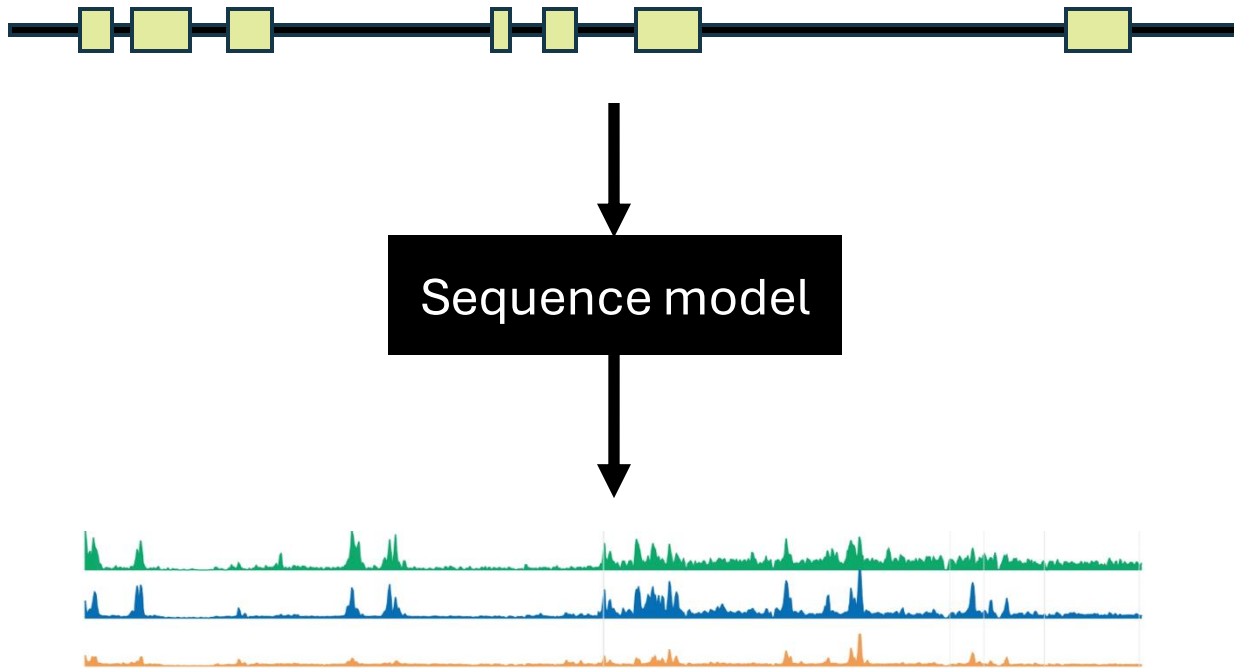
Output: Genomics tracks

$$(N_{batch_size} * L_{output} * T_{track_number})$$

Supervised learning

Input: DNA sequences

Output: Genomics tracks



Stage 1 Self-supervised learning

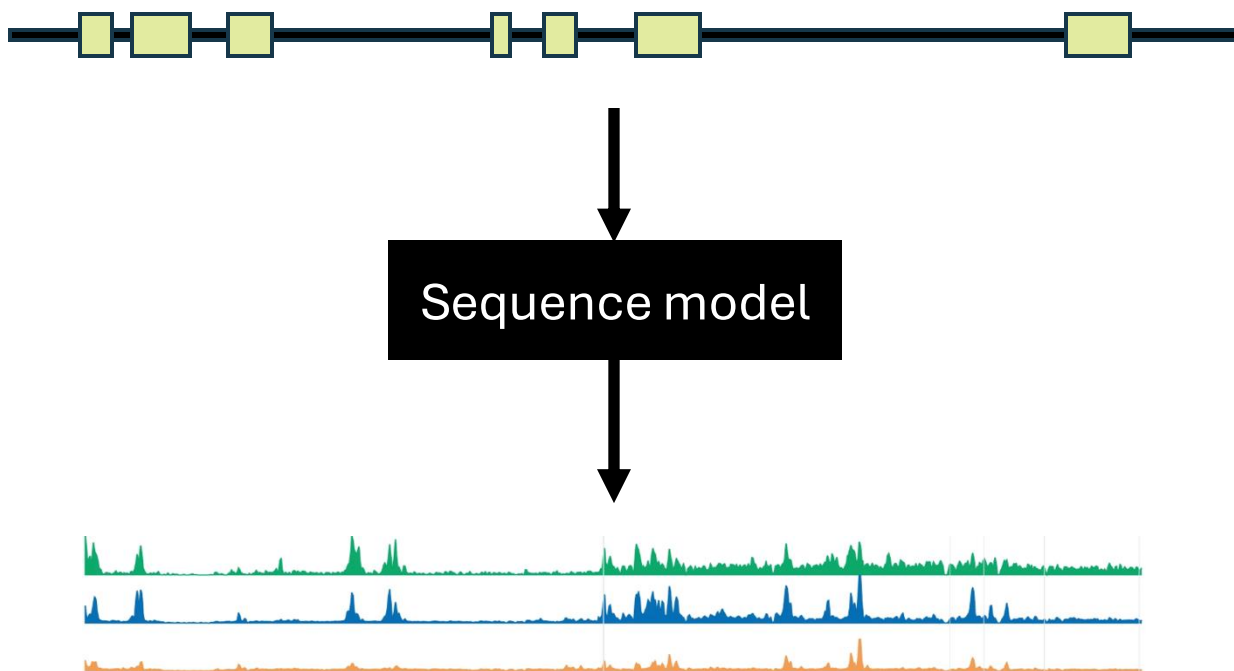


Language model (LM) /
Foundation model

Stage 2 Fine-tuning LM



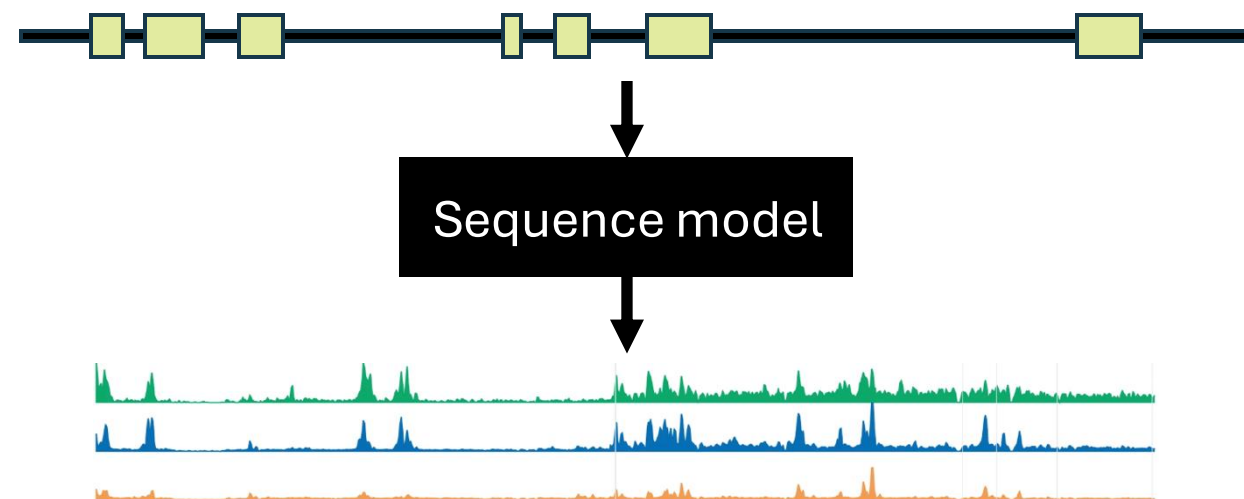
Supervised learning



Stage 1 Self-supervised learning



Stage 2 Fine-tuning LM



Protein Language model

Facebook
First LM attempt
PNAS 2020

Höcker Lab
ProtGPT2
Nat Commun 2022

Meta
ESMFold
Science 2023

Facebook
Transformer protein LM
ICLR 2021

Harvard + Columbia
AminoBERT
Nat Biotechnol 2022

Profluent
PLM
Nat Biotechnol 2024

Saleforce + UIUC
BERTology
ICLR 2021

Saleforce
ProGen
Nat Biotechnol 2023

Peking Uni
ProLLaMA
arxiv, 2024

MIT
Review
Cell Systems 2021

Saleforce + Profluent
ProGen2
Cell System 2023

~ 6000 citations

DNA Language model

DNABERT: pre-trained Bidirectional Encodings of Representations from Transformers for DNA-language in genomic context
Yanrong Ji, Zhihan Zhou, Han Liu
DNABERT
Bioinformatics 2021

DNABERT-2: Efficient Foundation Models for DNA Language Benchmark For Multi-Species
Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik
DNABERT-2
ICLR 2024

Sequence evolution of DNA language models
Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin
Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman P.
Stephen A. Baccus, Tina Hernandez-Boussard, Christo
doi: <https://doi.org/10.1101/2024.02.27.582234>
Evo
bioRxiv 2024

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY |
DNA language models are powerful predictors of genome-wide variation
Gonzalo Benegas, Sanjit Singh Batra, and Yun S. Song
GPN
PNAS 2023

Article | [Open access](#) | Published: 23 July 2024
DNA language models capture genomic context in the human genome
Melissa Sanabria, Jonas Hirsch, Pierre
GROVER
Nat Mach Intell 2024

[Submitted on 5 Mar 2024]
Caduceus: Bi-Directional Range DNA Sequence Modeling
Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri
Caduceus
ICML 2024

The Nucleotide Transformer Foundation Models for Genomic Language
Hugo Dallan, Adam Henry, Bernardo P. Marie Lope
Nucleotide Transformer
bioRxiv 2023

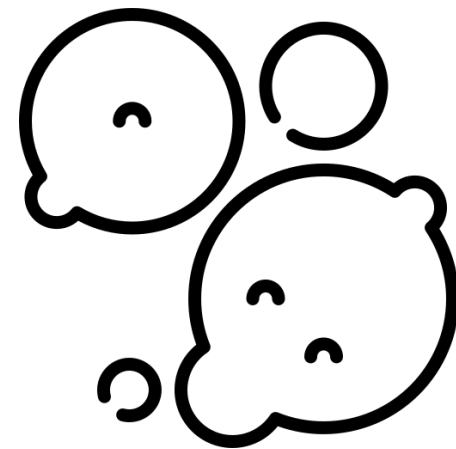
Article | [Open access](#) | Published: 03 April 2024
Genomic language model predicts regulatory elements
Yunha Hwang, Andre L. Cornman, Elizabeth
R. Girguis
Genomic LM
Nat Commun 2024

Cross-species nucleotide dependency analysis
Jingjing Zhai, Aaron Gokaslan, Yair
Michelle C. Stitzer, M. Cinta Rom
doi: <https://doi.org/10.1101/2024.06>
PlantCaduceus
bioRxiv 2024

arXiv:2306.15794 (cs)
[Submitted on 27 Jun 2023 (v1), last revised 14 Nov 2023 (this version, v2)]
HyenaDNA: Long-Range Context Modeling at Single Nucleotide Resolution
Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas
Aman Patel, Clayton Rabideau, Stefano Massaroli, Yosh
Baccus, Chris Ré
HyenaDNA
NeurIPS 2023

Research | [Open access](#) | Published: 02 April 2024
Species-aware DNA language models capture regulatory elements across species
Alexander Karol
Julien Gagneur
Species-aware DNA LM
Genom Biol 2024

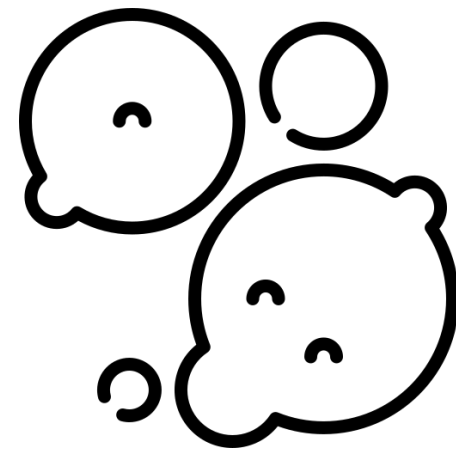
Nucleotide dependency analysis of DNA language models
Nucleotide dependency
bioRxiv 2024



Study Goals

- Building a SOTA **yeast** gene expression model.
- **Part I:** Exploring DNA **Language Model (LM)**.
- **Part II:** Fine-tuning DNA LM to predict gene expression (RNA-Seq tracks).
 - Is *self-supervised learning with fine-tuning* better than *training from scratch*?
- Understanding what models learn at each stage: (LM, fine-tuning, and scratch-training)
- **Part III:** Applications:
 - Predicting the influence of context and distal regulatory elements on gene expression.
 - Assessing the variant effects on eQTLs and negatively selected eQTLs

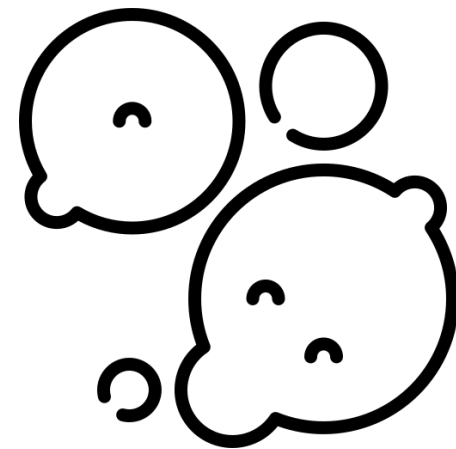
Why **yeast** ?



- Computation is expensive for human (human: 3B nt; yeast: 12 M nt)

Human + Mouse genomes = **5.7 B nt**

	1 sequence (All tracks)	All sequences(All tracks)	Practical runtime
Enformer	$(\# \text{ 128-resolution bin}) \times (\# \text{ tracks})$ $896 \times 6,956$ 6,232,576 float32 (4 bytes) $\cong$ 25 MB	$5.7B \div 196,608 \cong 28992$ $28992 * 2.5MB \cong$ 724.8 GB	1 model: 64 TPU v3 cores (16GB GPU memory each) (~ 3 days)
Borzoï	$(\# \text{ 32-resolution bin}) \times (\# \text{ tracks})$ $16,384 \times 10,219$ 167,428,096 float32 (4 bytes) $\cong$ 670 MB	$5.7B \div 524,288 \cong 10872$ $10872 * 670 MB \cong$ 7.28 TB	1 model: 2 A100 GPUs (40 GB GPU memory each) (~25 days)

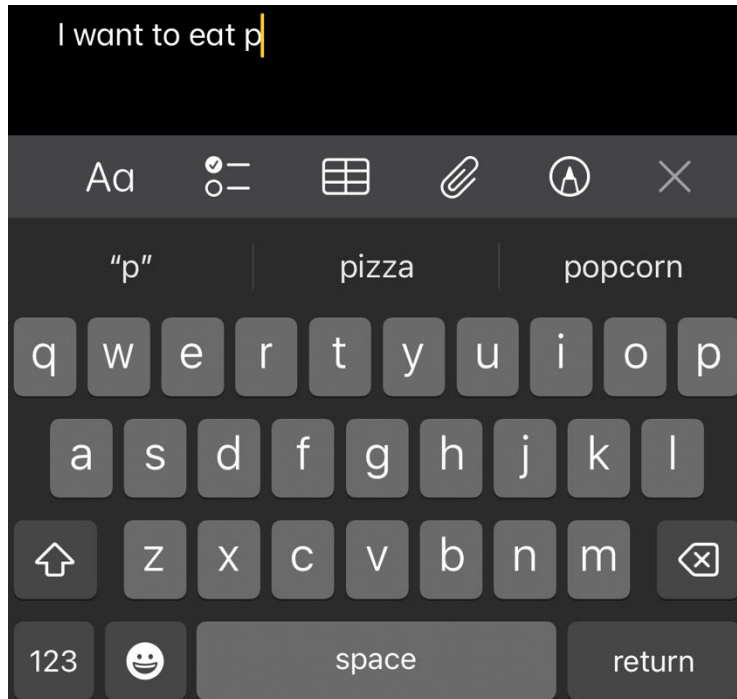


Why **yeast** ?

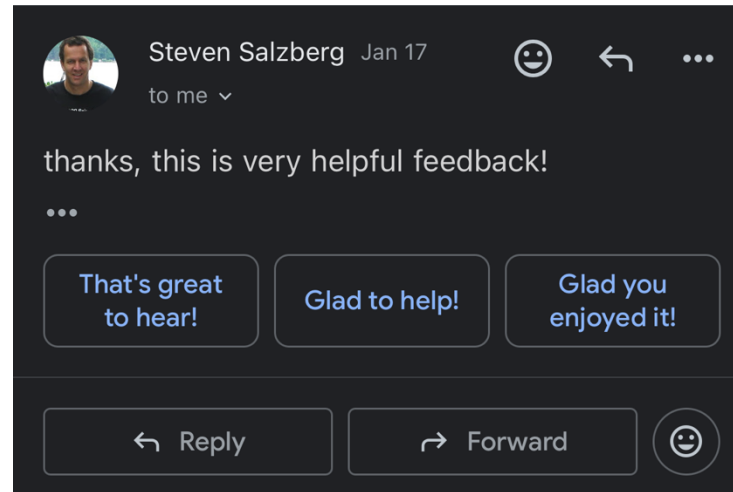
- Computation is expensive for human (human: 3B nt; yeast: 12 M nt)
- In human, we can't do large scale TF perturbation study
- Yeast is a great model organism to generate data & training models
 - Simple Eukaryotic Model: cost-effectiveness and scalability
 - Rapid Growth and Easy Culturing and Quick Lift Cycle
 - Genetic Manipulability
 - Well-Characterized Genome
 - Conserved Regulatory Mechanisms
 - Great species to study aging! (bud scar / cell size)

Language Model (LM) in our daily life

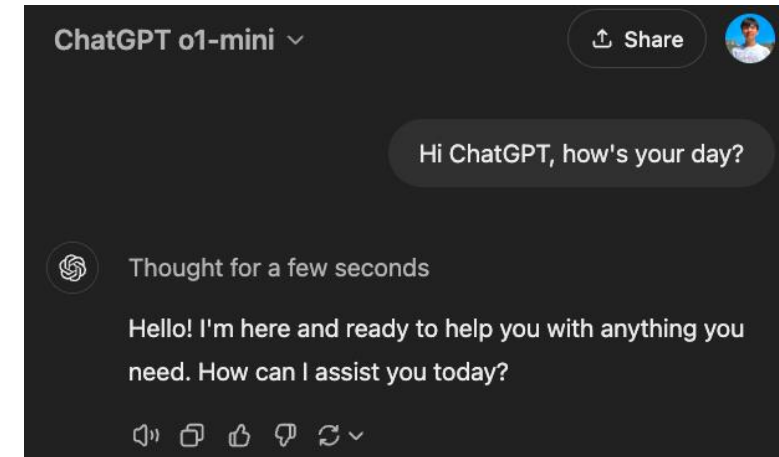
Next word prediction



Email suggested reply



ChatGPT



What is a **language model (LM)**?

$$\mathbf{P}(\underbrace{X_t}_{\text{Next word}} \mid \underbrace{X_1, X_2, \dots, X_{t-1}}_{\text{Context or prefix}})$$

Next word

Context or prefix

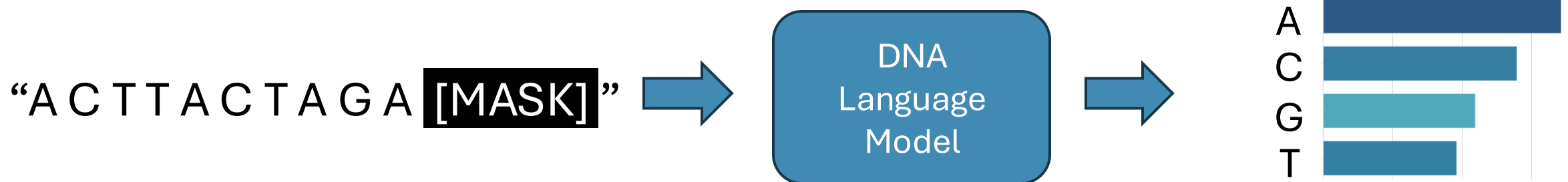
What is a **language model (LM)**?

- Directly we train models on “**marginals**”
- We are implicitly learning the **full/joint distribution** of language.

$$P(\underbrace{X_t}_{\text{Next word}} \mid \underbrace{X_1, X_2, \dots, X_{t-1}}_{\text{Context or prefix}})$$

The chain rule:

$$P(X_1, \dots, X_t) = P(X_1) \prod_{i=1}^t P(X_i \mid X_1, X_2, \dots, X_{i-1})$$



Language Modeling $\triangleq$ learning prob distribution over language sequence.

Part I



Fungal Language Model

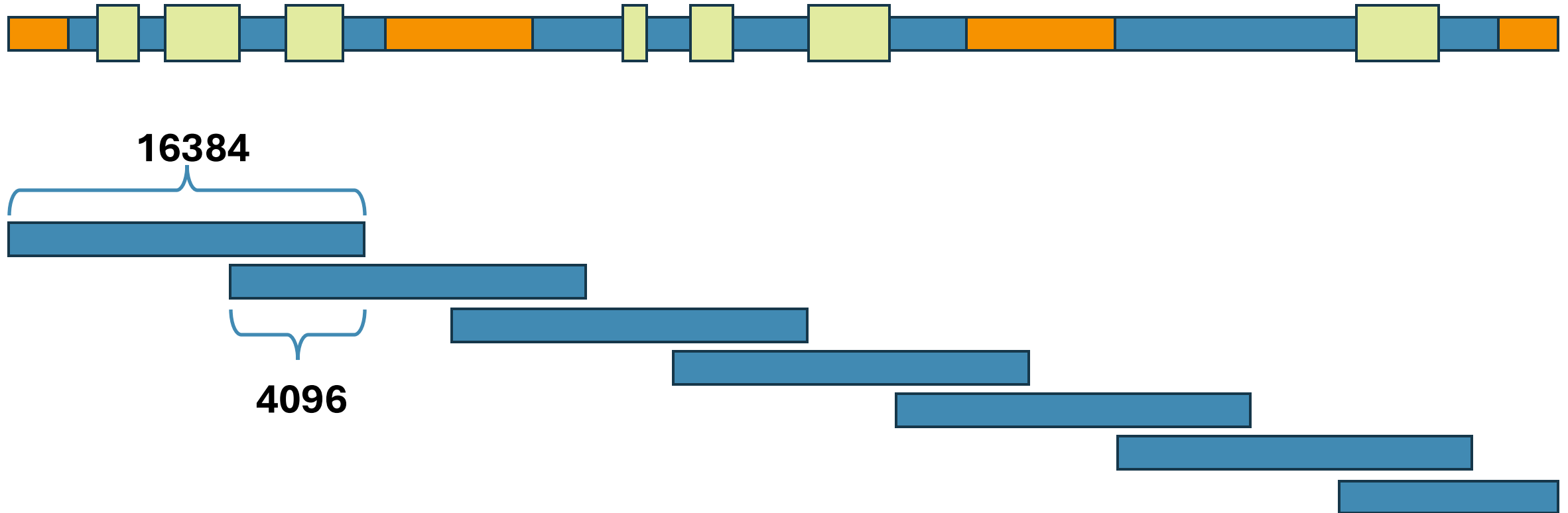
Summary

We need to understand the genome & carefully preprocess genome to correctly train a Fungi language model.

Data preprocessing

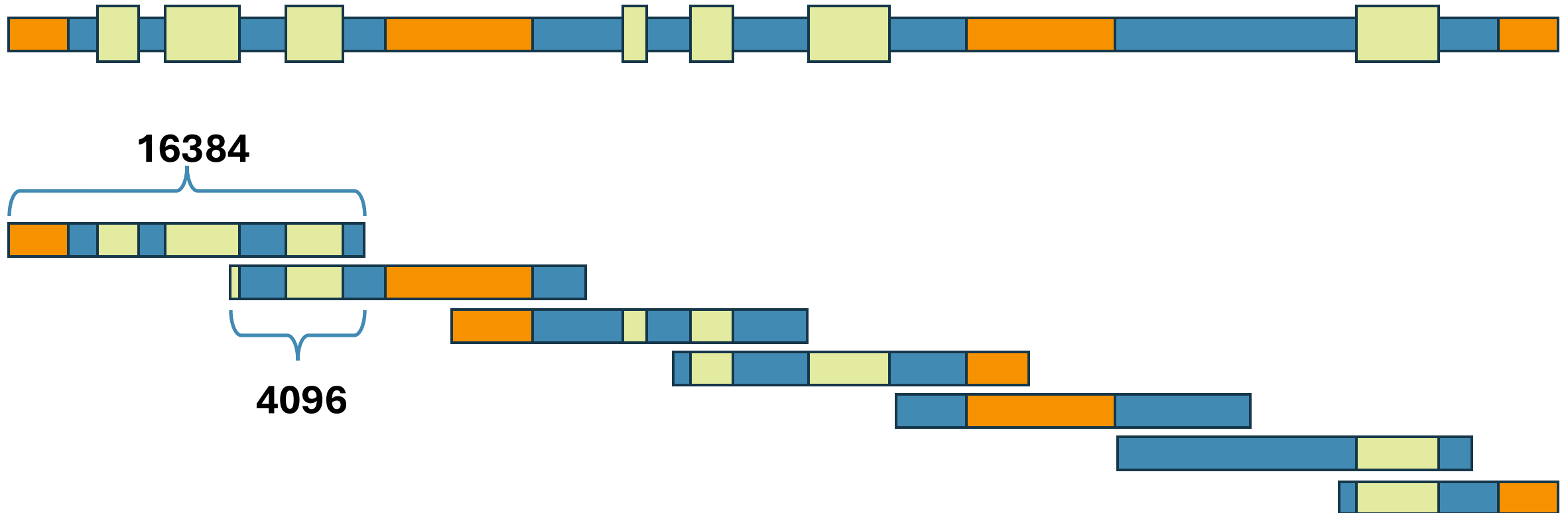


Data preprocessing



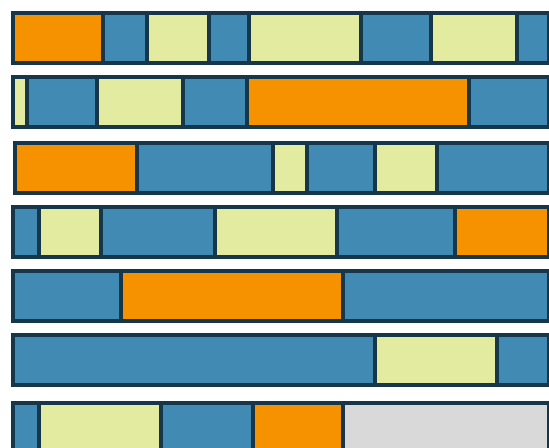
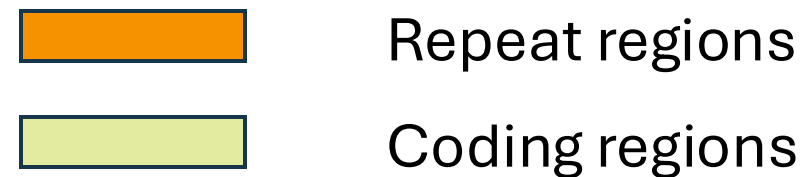
~ 8 genes per window

Data preprocessing

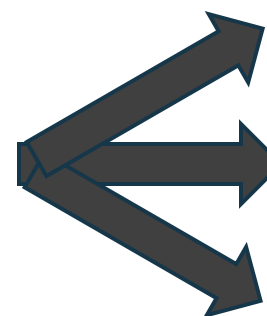
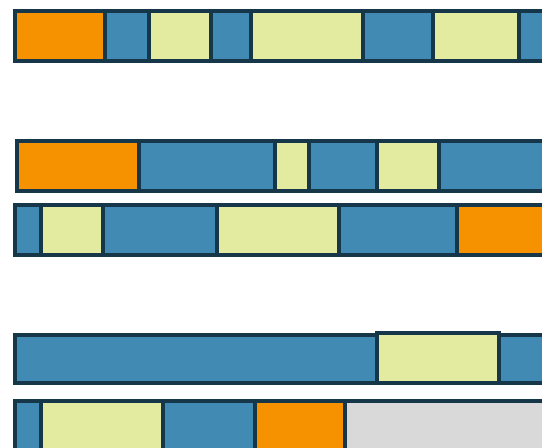


~ 8 genes per window

Data preprocessing



7% repeat
threshold



Training

Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)

Why building a Fungal Language Model (LM)?

- Yeast genome is small. 12Mbps. 
- Thousands of fungal genomes with high quality. No supervised measurements
- Language model pre-training on all available genomes followed by transfer learning to the smaller yeast genome.

Selected Genomes for LM

Fungi diverged from other life around 1.5 billion years ago

**Same species,
Different strains**



Order level

Kingdom level

Dataset 1

Dataset 2

Dataset 3

Dataset 4

Q1: Diversity of strains?

Q2: Diversity of species?

Q3: Even more diverse?

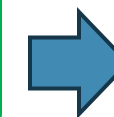
R64
reference yeast



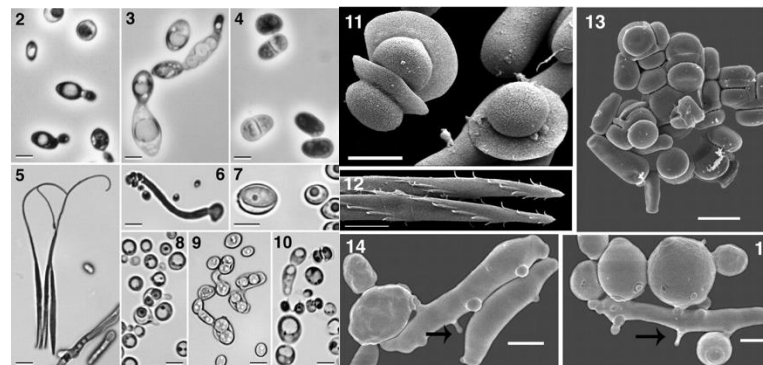
80 strains
of yeasts



165
Saccharomycetales



1361
Fungus genomes



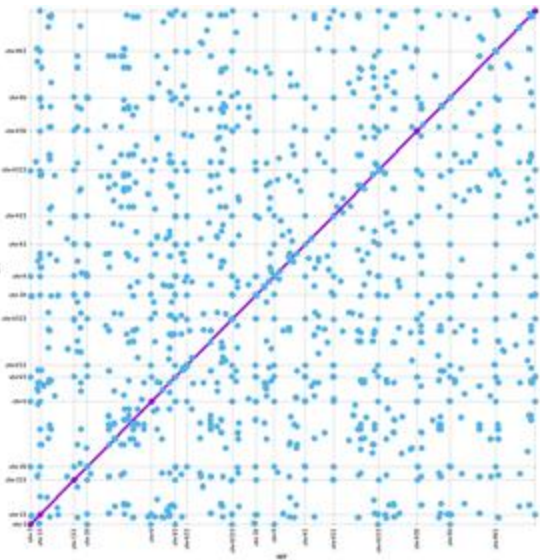
Suh, S. O., Blackwell, M., Kurtzman, C. P., & Lachance, M. A. (2006). Phylogenetics of Saccharomycetales, the ascomycete yeasts. *Mycologia*, 98(6), 1006-1017.



Genome distance evaluation

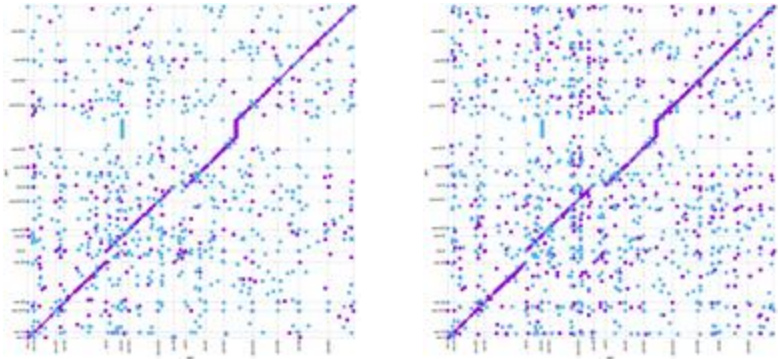
R64 Reference Yeast

Mummerplot

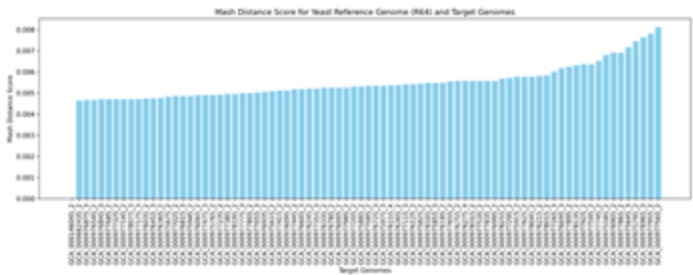


80 strains of yeasts

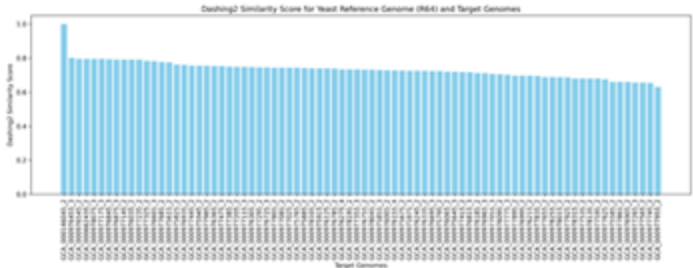
Mummerplot



Mash distance



Dashing2 similarity

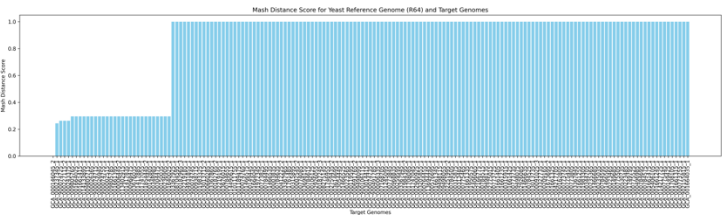


✓ 165 Saccharomycetales

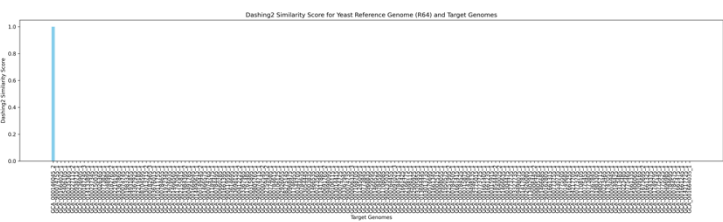
Mummerplot

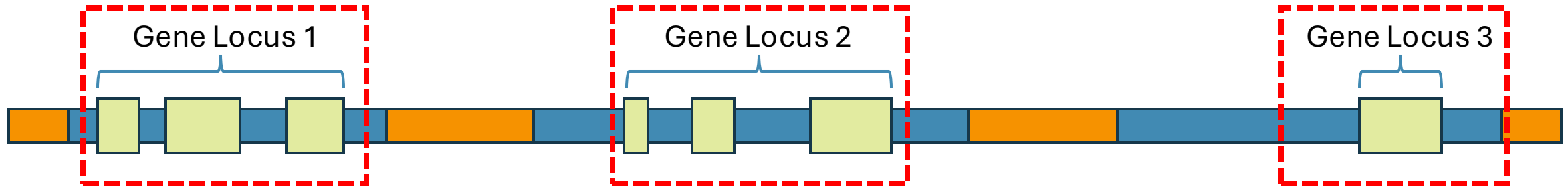


Mash distance



Dashing2 similarity





Q2: How many genes per window?

Q3: What is the quality of the annotation?

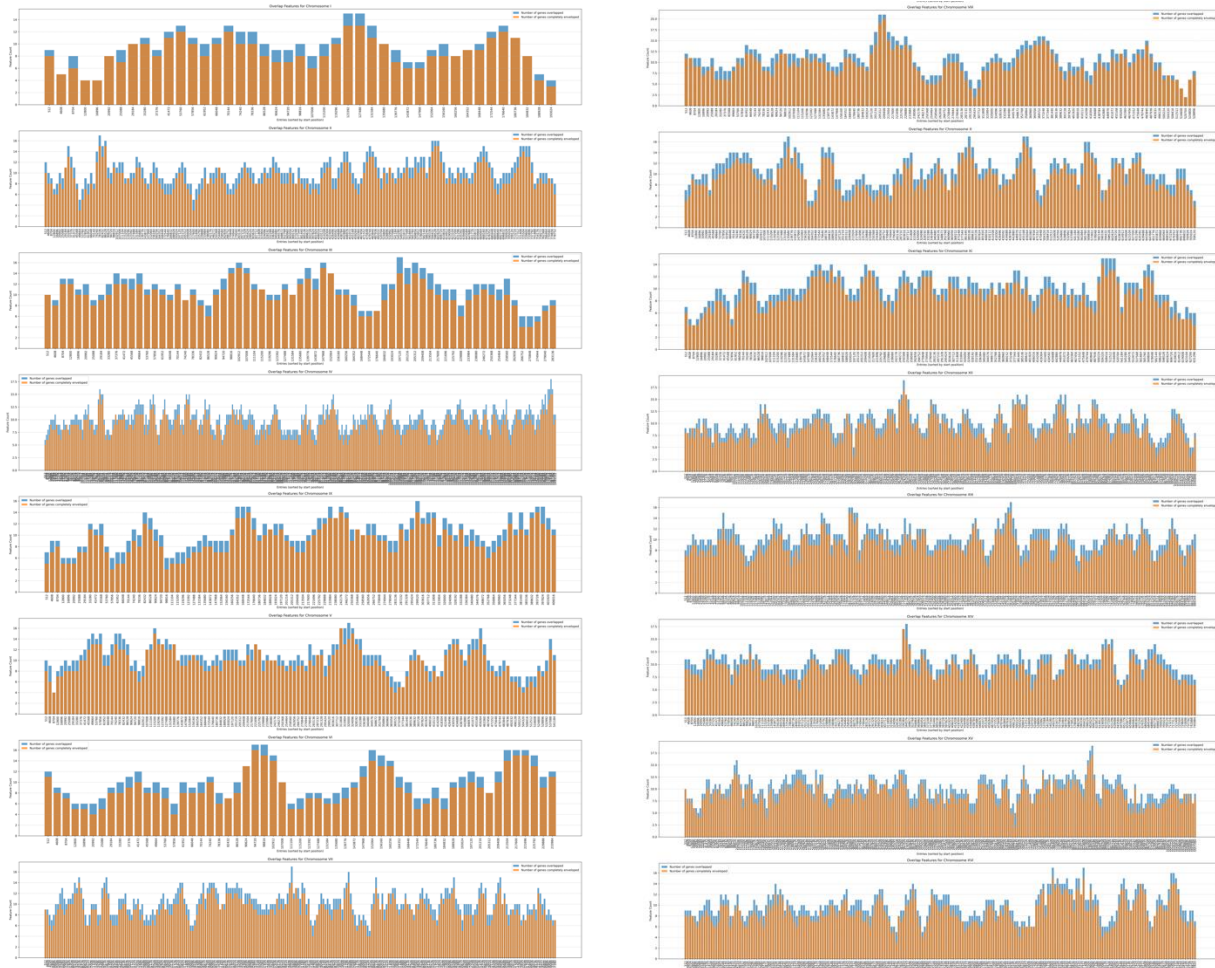
Q4: What is the protein-coding region ratio?



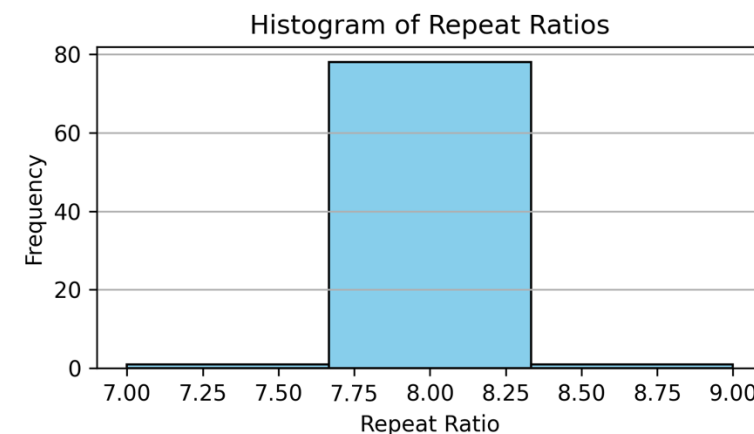
Genome evaluation – # genes per 16K window

R64 Reference Yeast

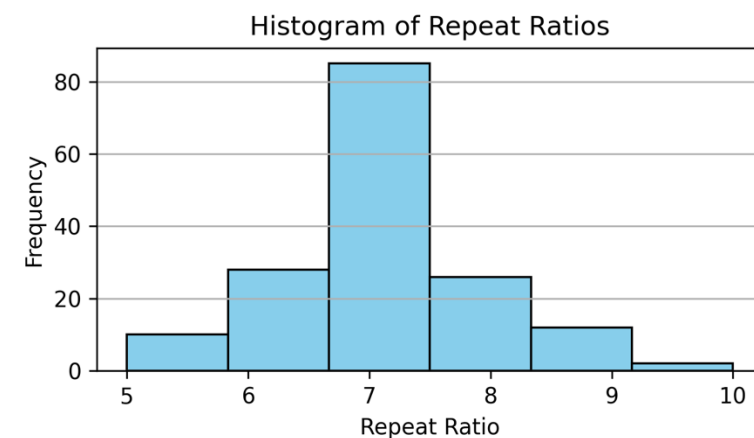
Median: 9.0; Mean: 8.98



80 strains of yeasts



165 Sachramonycetales



Genome annotation completeness evaluation

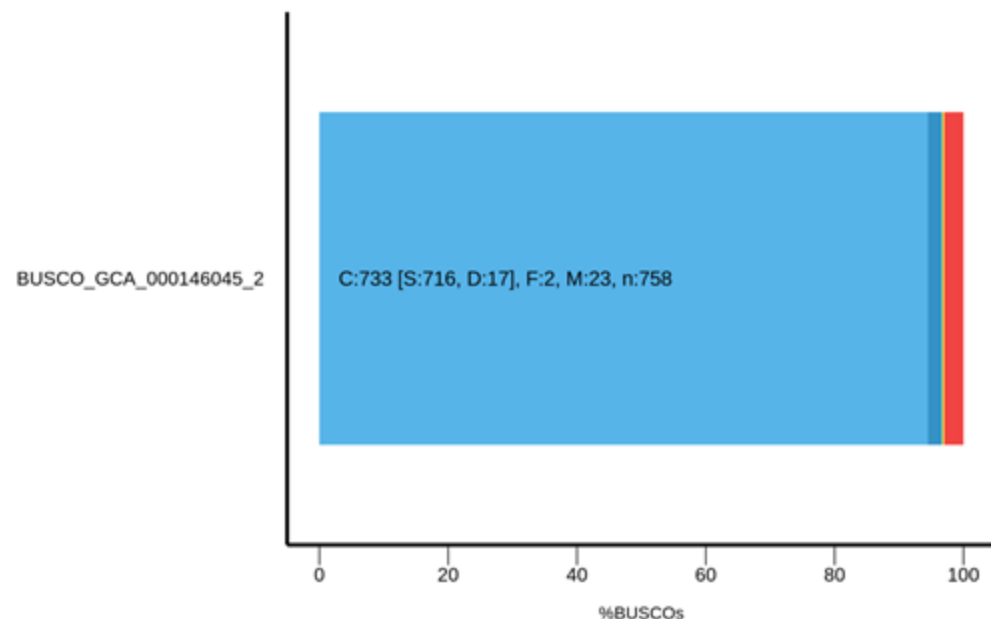
R64 Reference Yeast

80 strains of yeasts

165 Sachramonycetales

BUSCO Assessment Results

Complete (C) and single-copy (S) Complete (C) and duplicated (D)
Fragmented (F) Missing (M)



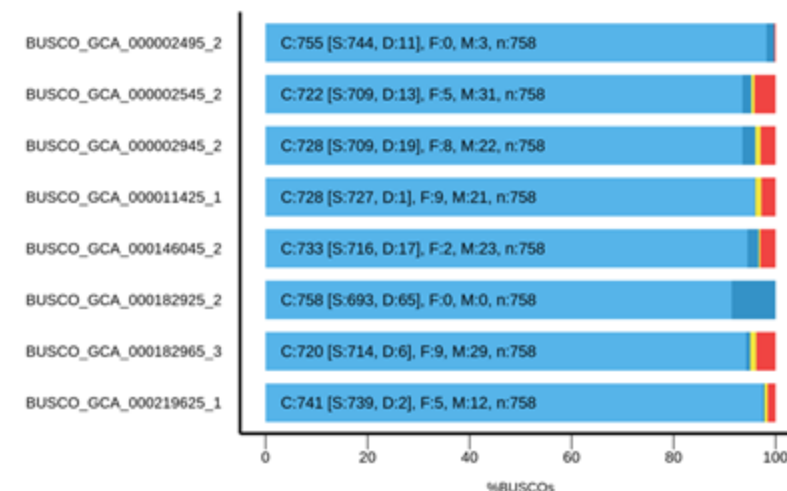
BUSCO Assessment Results

Complete (C) and single-copy (S) Complete (C) and duplicated (D)
Fragmented (F) Missing (M)



BUSCO Assessment Results

Complete (C) and single-copy (S) Complete (C) and duplicated (D)
Fragmented (F) Missing (M)

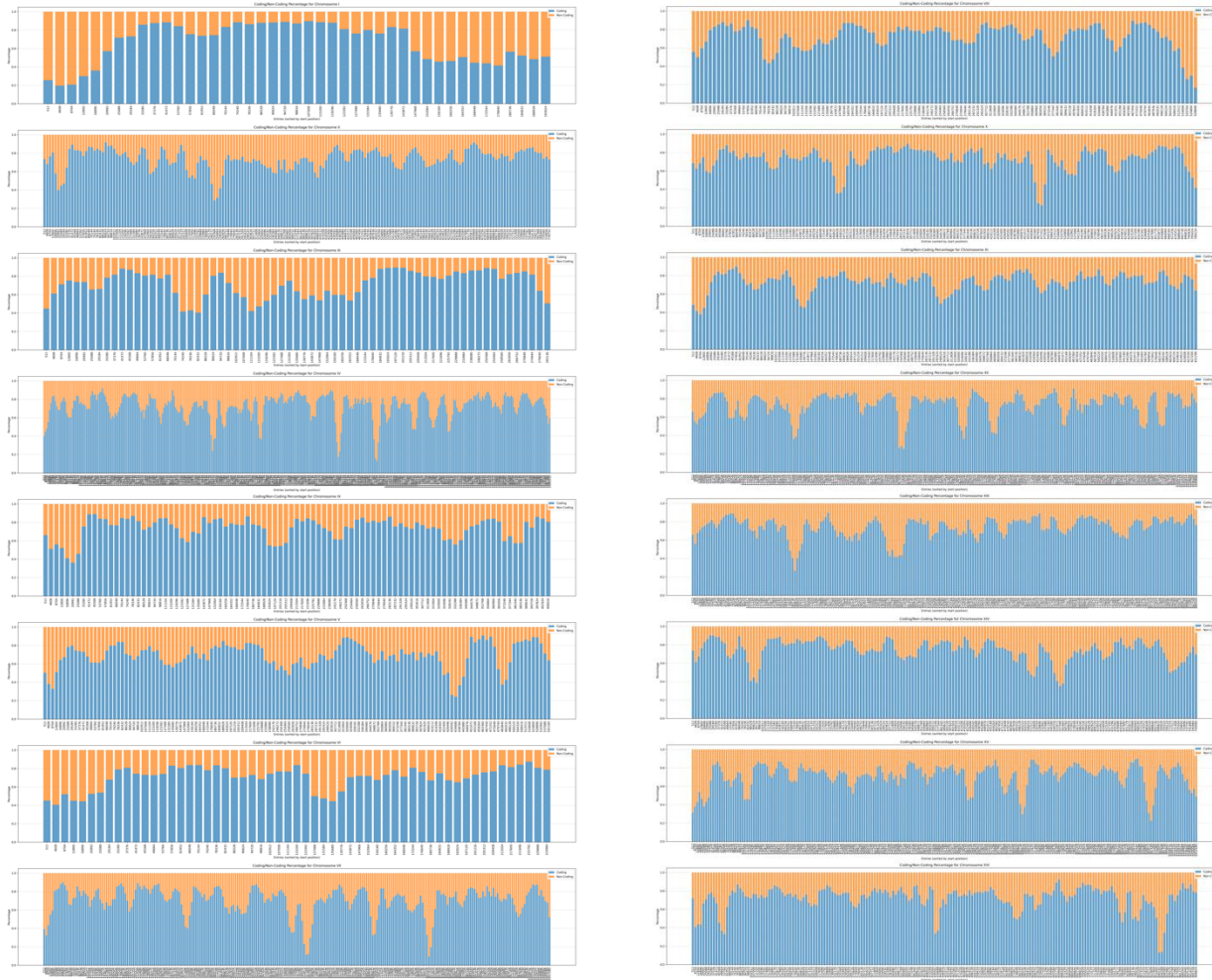


Conclusion:
BUSCO ~95% completeness

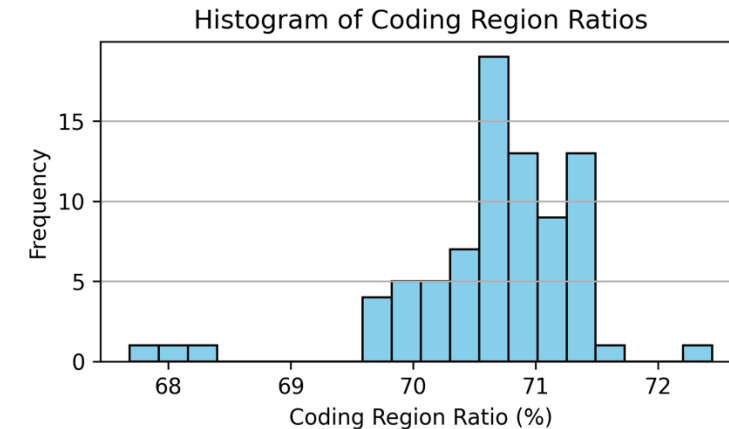
Genome evaluation – coding / noncoding regions

R64 Reference Yeast

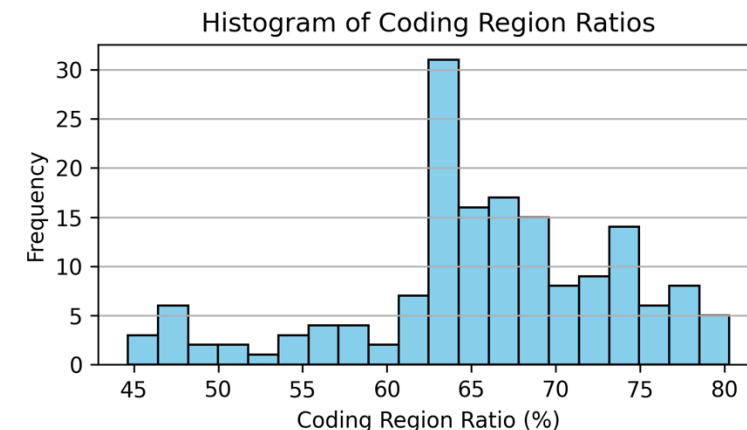
72.46% coding regions

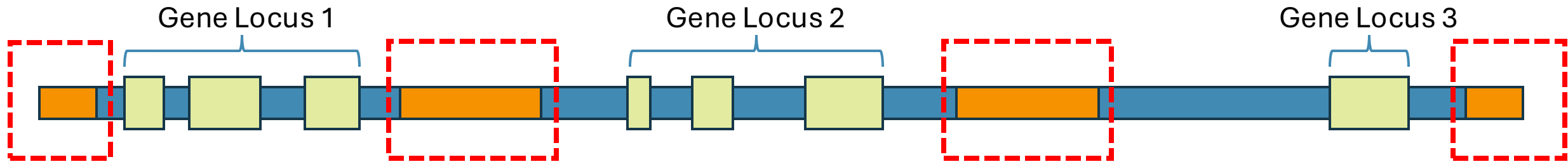


80 strains of yeasts

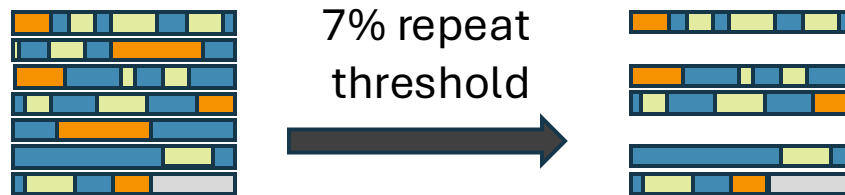


165 Sachramonycetales



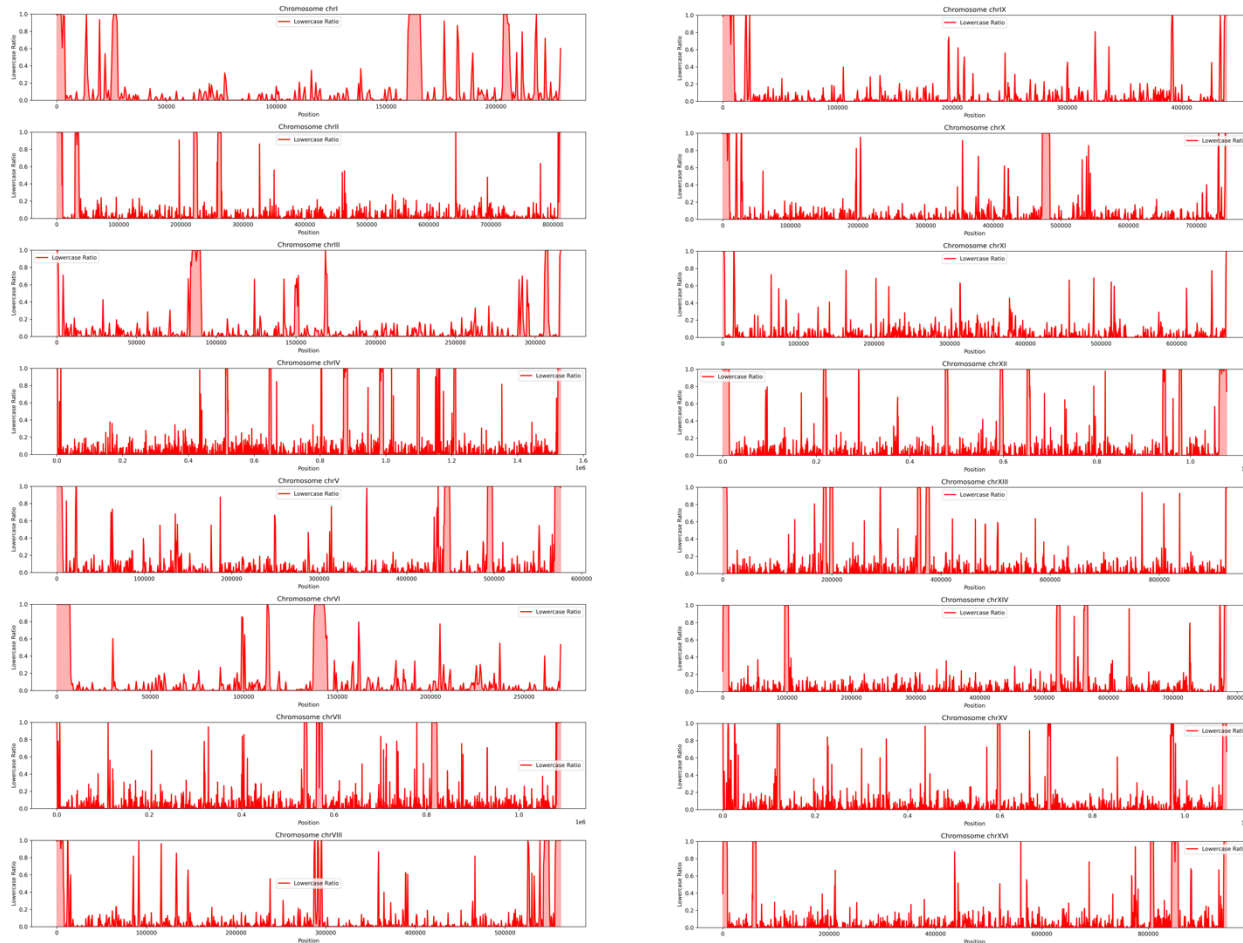


Q5: How repetitive are the genomes?

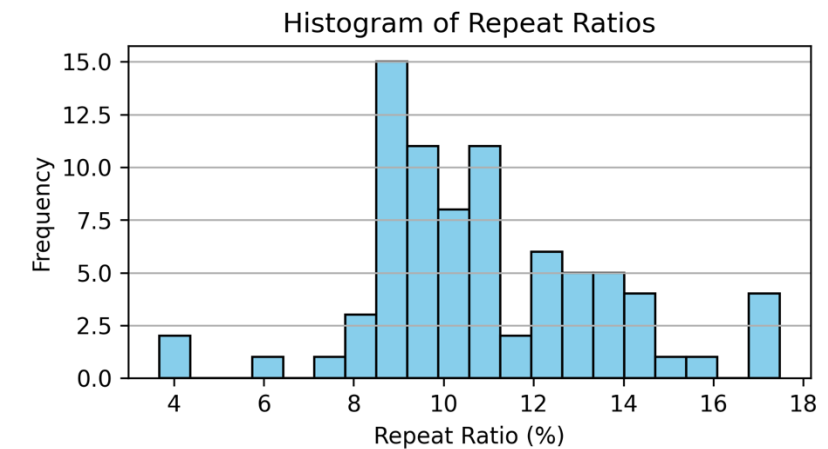


Genome evaluation – repeat regions

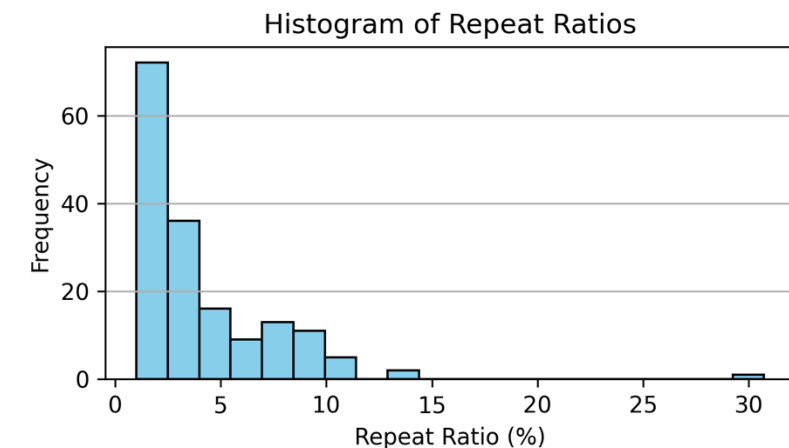
R64 Reference Yeast 7.39% *repeat regions*

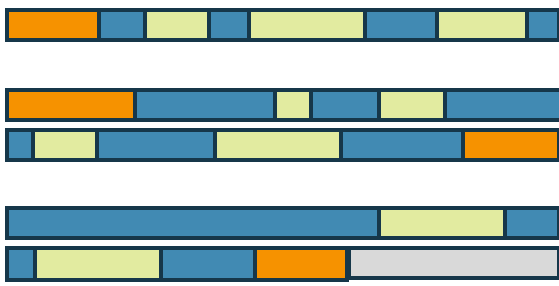


80 strains of yeasts



165 Sachramonvcetales





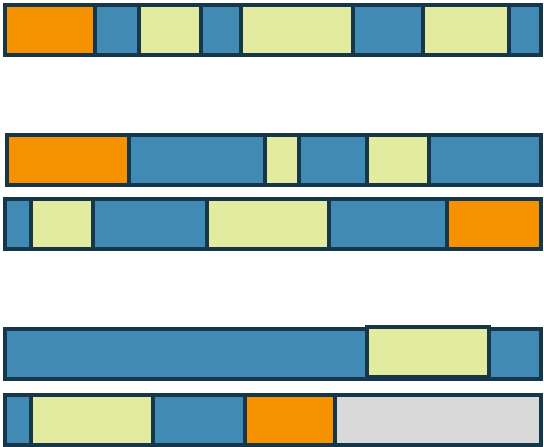
Training

Validation (chrXI, chrXIII, chrXV)

Testing (chrXII, chrXIV, chrXVI)

Q6: How many homologous sequences are there between training and testing?

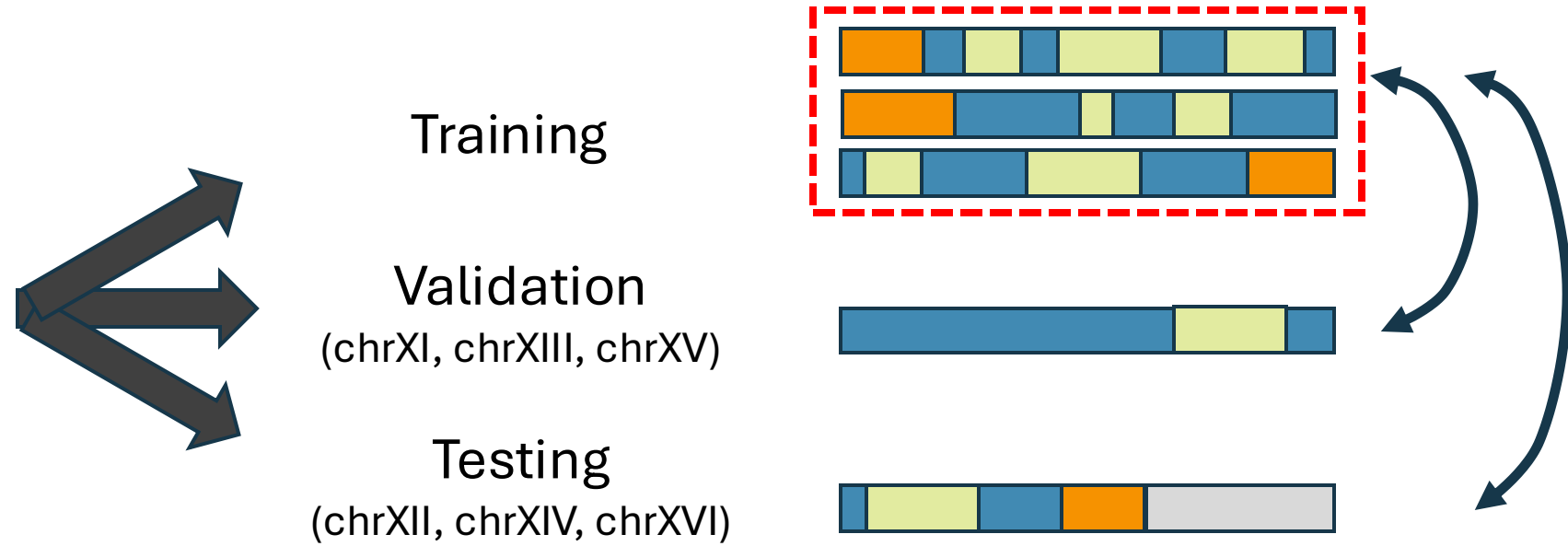




Training

Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)



Detect homologous sequence using **Minimap2** (5% sequence divergence)

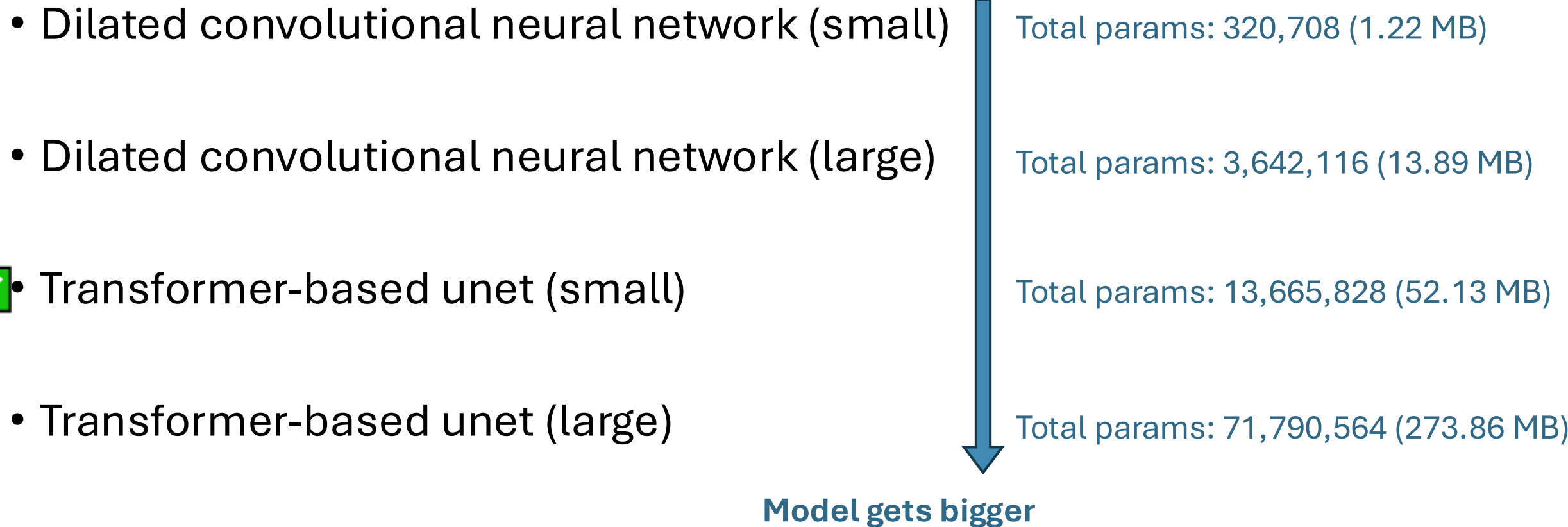
Final sequence for training / testing / validation

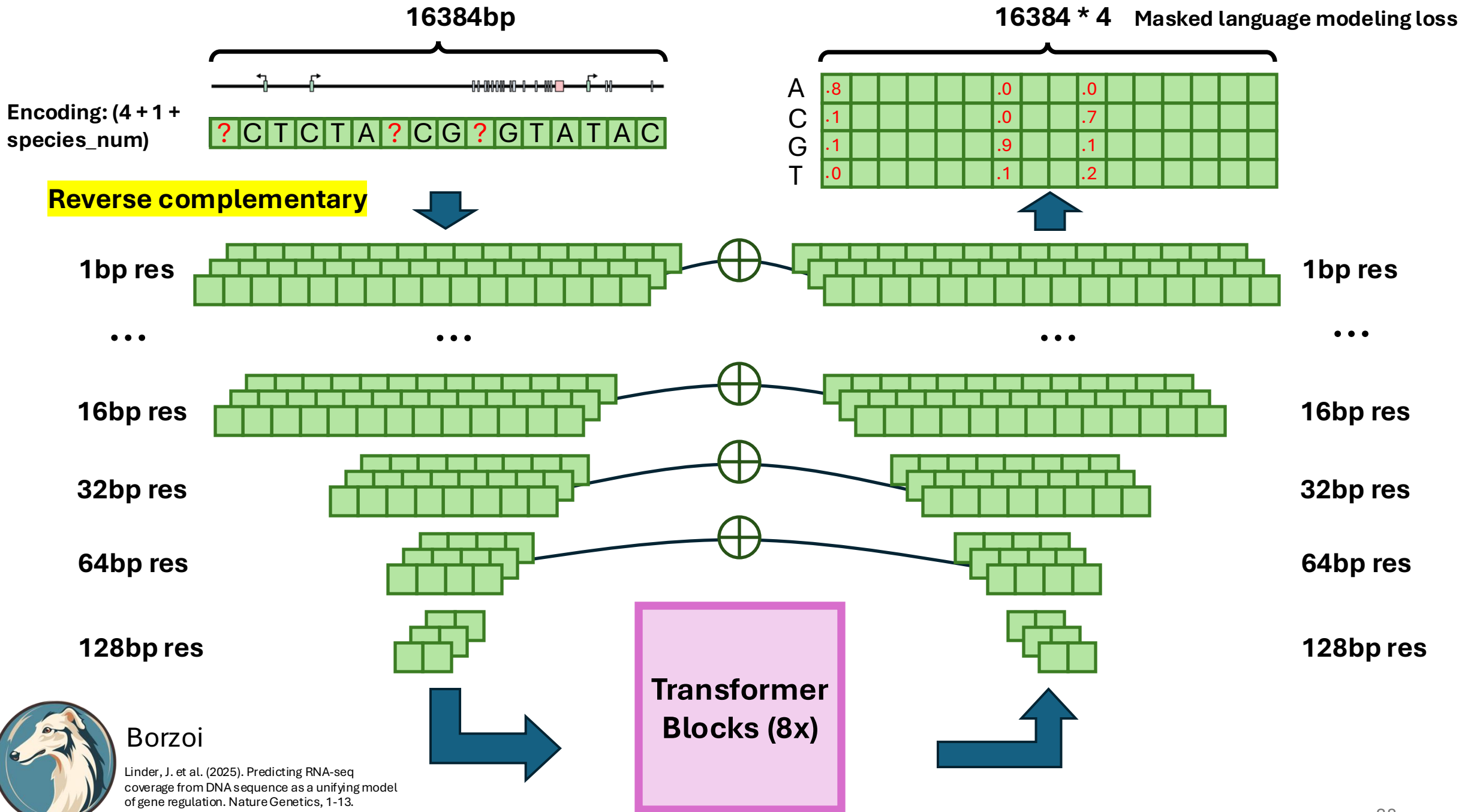
	r64	80 strains	165 Saccharomycetales
Before cleaning	Train : 1440 Test : 608 Validation : 576	Train : 108960 Test : 608 Validation : 576	Train : 404608 Test : 608 Validation : 576
	-377 (-14.4%)	-6783 (-6.16%)	-19195 (-4.73%)
After cleaning	Train : 1201 Test : 528 Validation : 518	Train : 102315 Test : 528 Validation : 518	Train : 385551 Test : 528 Validation : 518

Fungal Language Model Architecture



Different model architecture we've tried

- Dilated convolutional neural network (small) Total params: 320,708 (1.22 MB)
 - Dilated convolutional neural network (large) Total params: 3,642,116 (13.89 MB)
 - ✓ • Transformer-based unet (small) Total params: 13,665,828 (52.13 MB)
 - Transformer-based unet (large) Total params: 71,790,564 (273.86 MB)
- Model gets bigger
- 



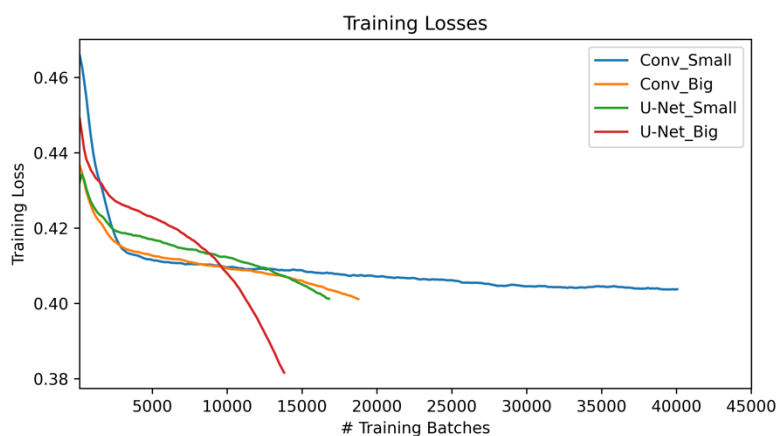
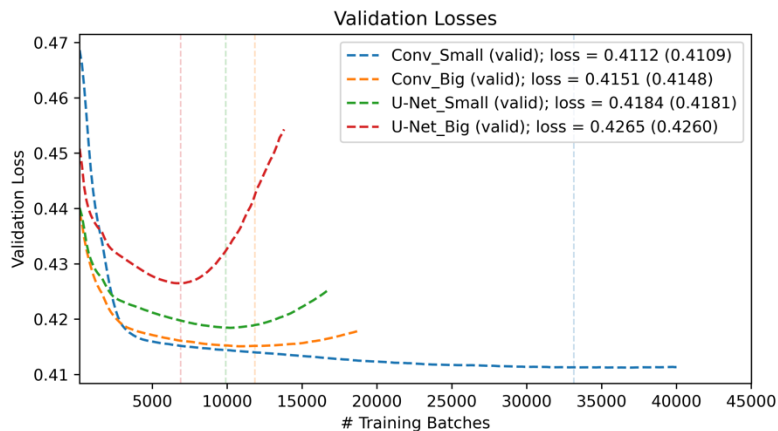
Fungal Language Model

Self-supervised training Results

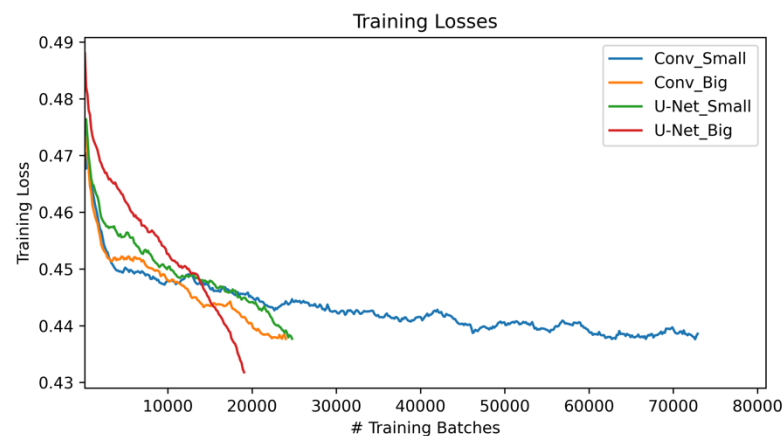
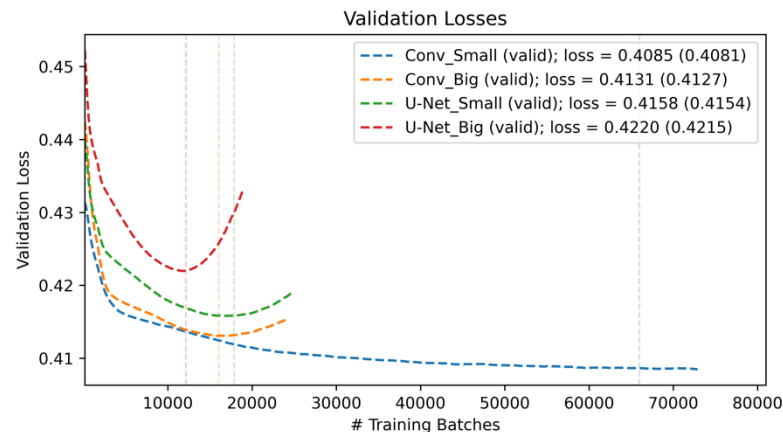


Model comparison

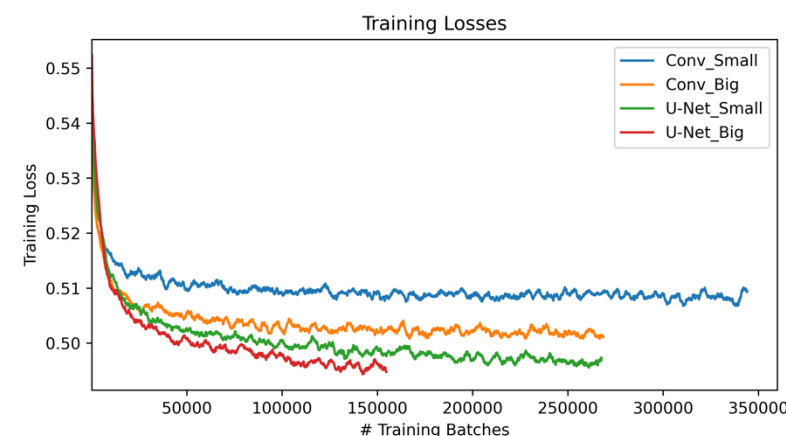
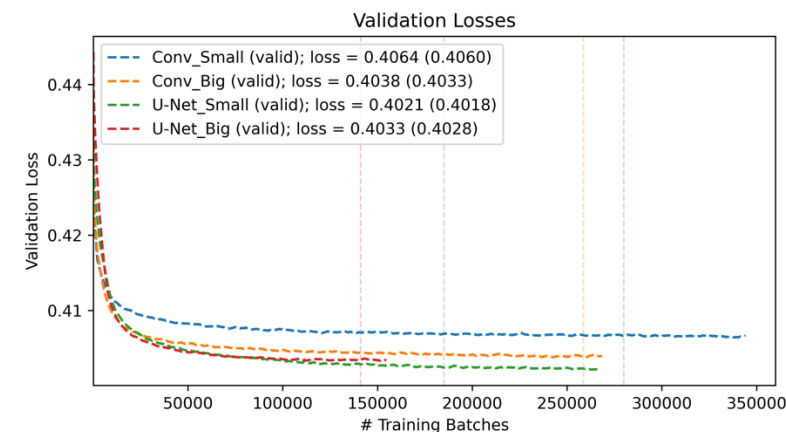
r64



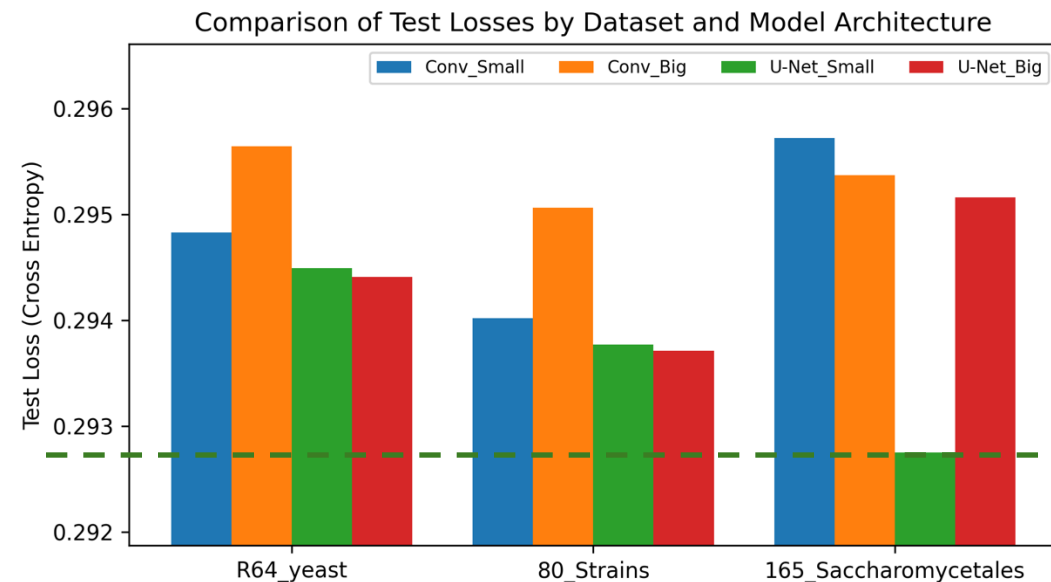
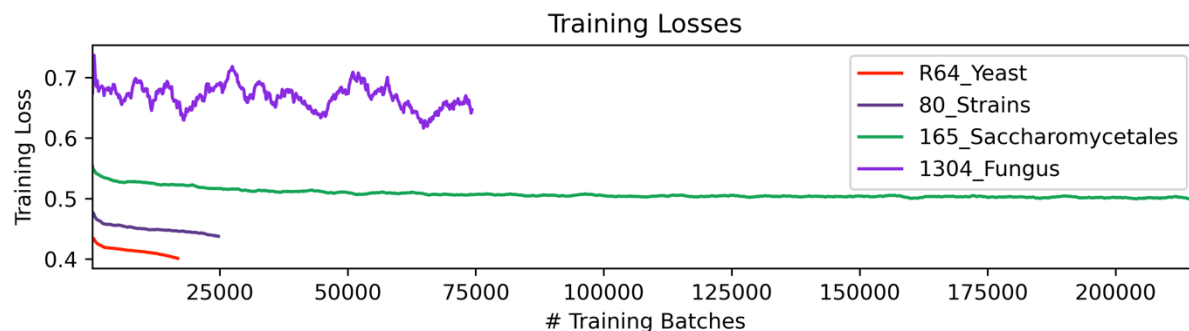
80 strains



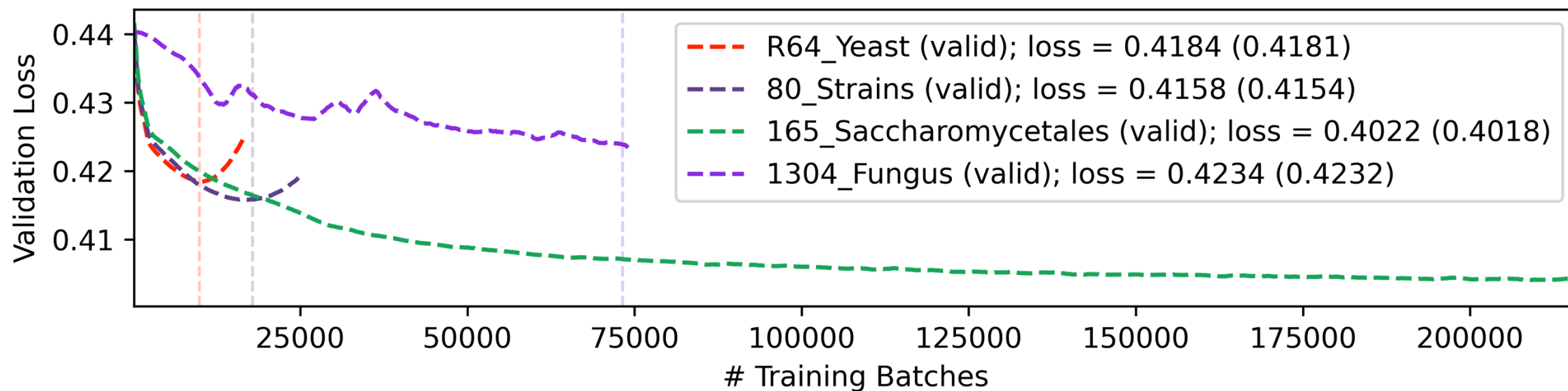
165 Saccharomycetales



Dataset comparison



Validation Losses



Fungal Language Model

Interpretability

1. Motif inference 2. Attention map visualization

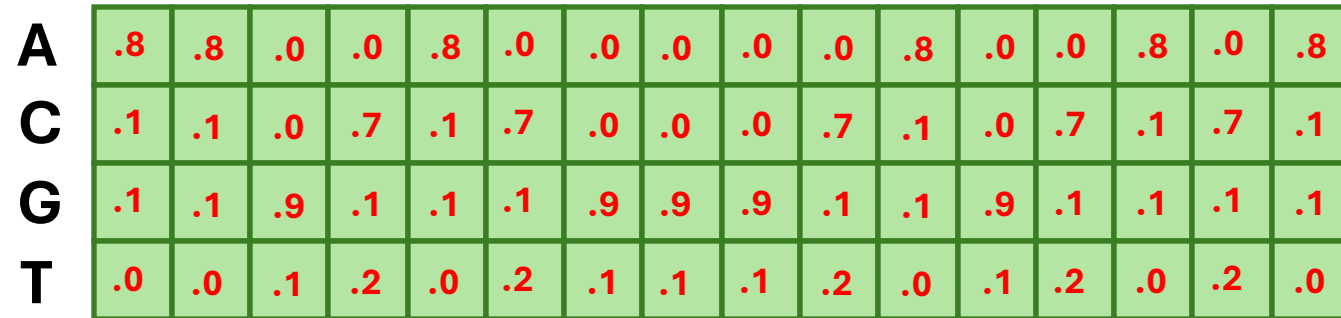


Fungal LM learns sequence conservation

Step 1: Construct PWM from test set

Input sequences in test set
(chrXII, chrXIV, chrXVI)

Predicting **15 %** masked
regions for each iteration



A	.8	.8	.0	.0	.8	.0	.0	.0	.0	.0	.8	.0	.0	.8	.0	.8
C	.1	.1	.0	.7	.1	.7	.0	.0	.0	.7	.1	.0	.7	.1	.7	.1
G	.1	.1	.9	.1	.1	.1	.9	.9	.9	.1	.1	.9	.1	.1	.1	.1
T	.0	.0	.1	.2	.0	.2	.1	.1	.1	.2	.0	.1	.2	.0	.2	.0

Step 2: PWM Normalization

1. Pseudocount Addition

$$p_{i,j} = p_{i,j} + \epsilon$$

$p_{i,j}$ is the read count for
nucleotide i at position j

2. Row Normalization

$$p_{j,i}^{norm} = \frac{p_{j,i}}{\sum_{k=1}^4 p_{j,k}}$$

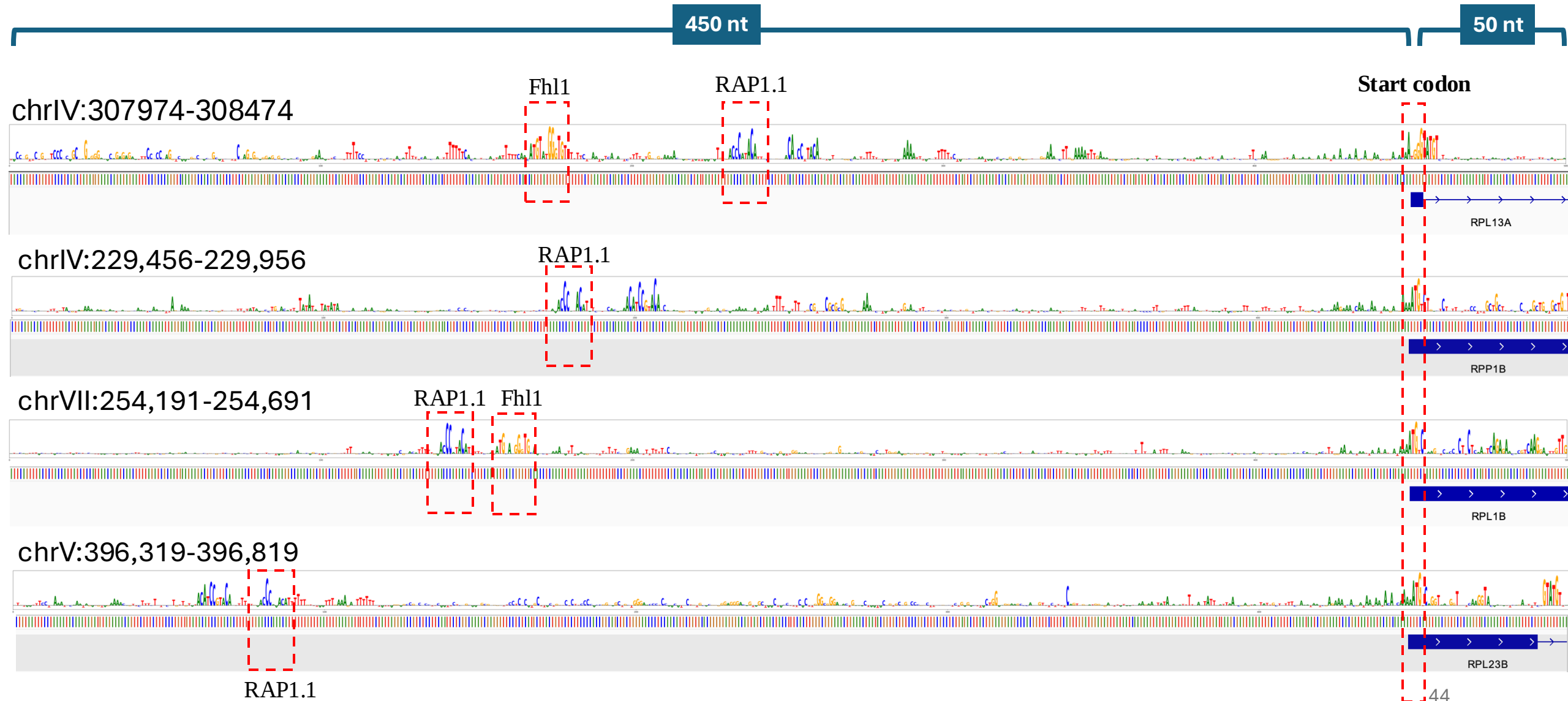
3. Entropy Calculation

$$H_j = - \sum_{i=1}^4 p_{j,i}^{norm} \cdot \log_2(p_{j,i}^{norm})$$

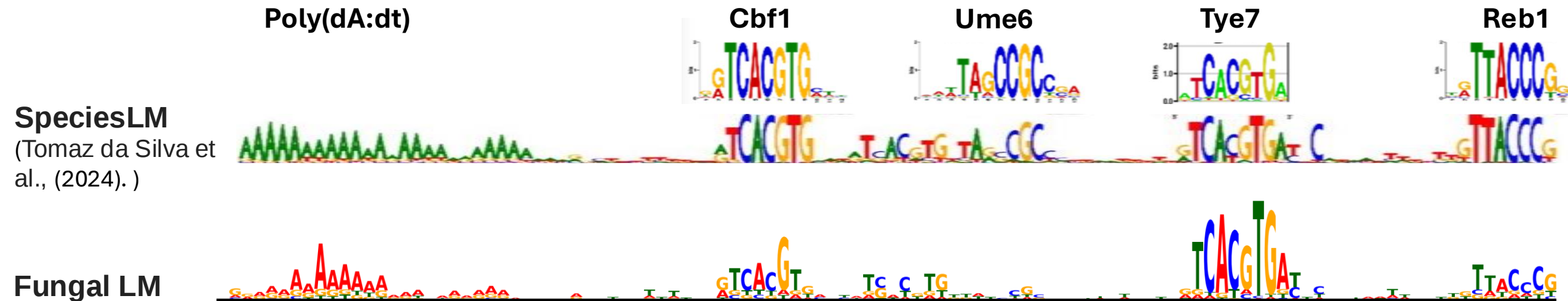
4. Conservation Calculation

$$C_j = 2 - H_j$$

Risbosomal Protein Upstream Promoter regions



Upstream Promoter region of SMT3 gene



SpeciesLM: Trained with 5' and 3' regions only.

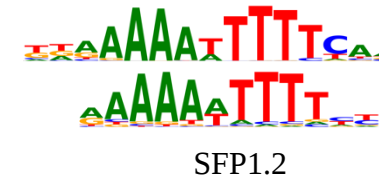
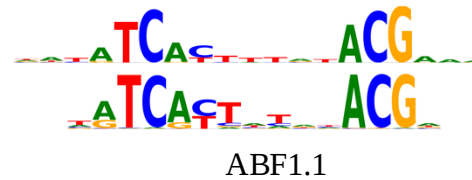
Fungal LM: Trained with 165 full Saccharomycetales genomes (**harder approach**)

Tomaz da Silva et al., (2024). Nucleotide dependency analysis of DNA language models reveals genomic functional elements. bioRxiv

Constructing motifs in *S. cerevisiae* genome

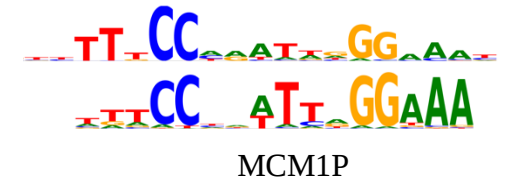
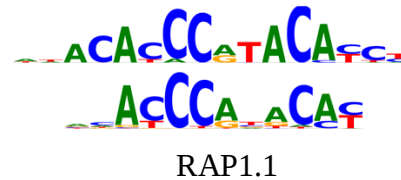
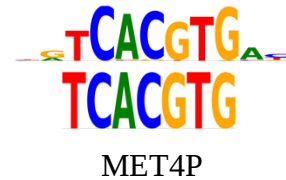
Yeast LM Motifs

DB Motifs



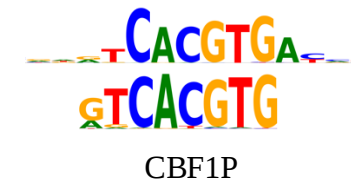
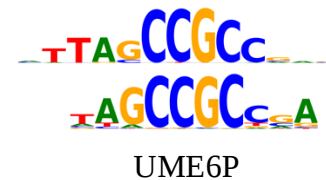
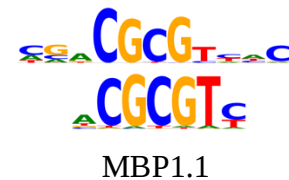
Yeast LM Motifs

DB Motifs



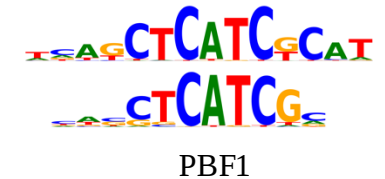
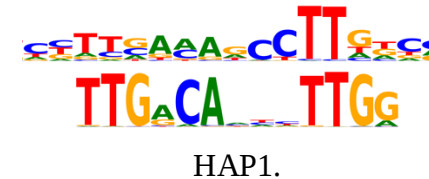
Yeast LM Motifs

DB Motifs



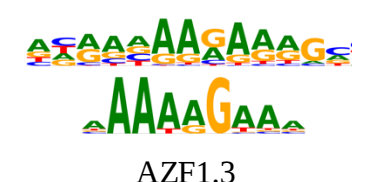
Yeast LM Motifs

DB Motifs



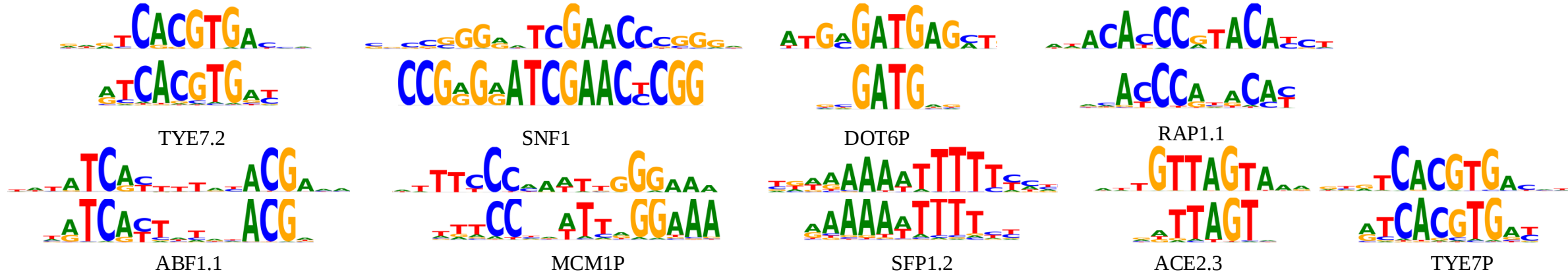
Yeast LM Motifs

DB Motifs

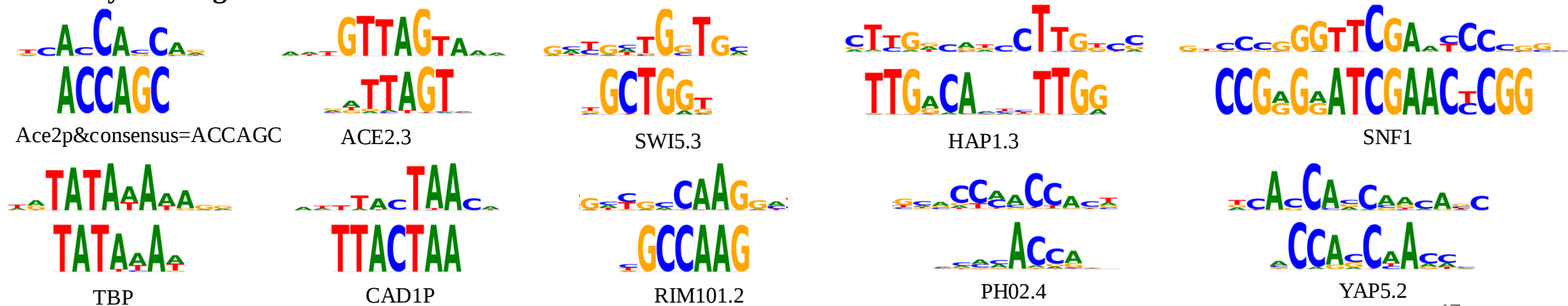


Constructing motifs in unseen genomes

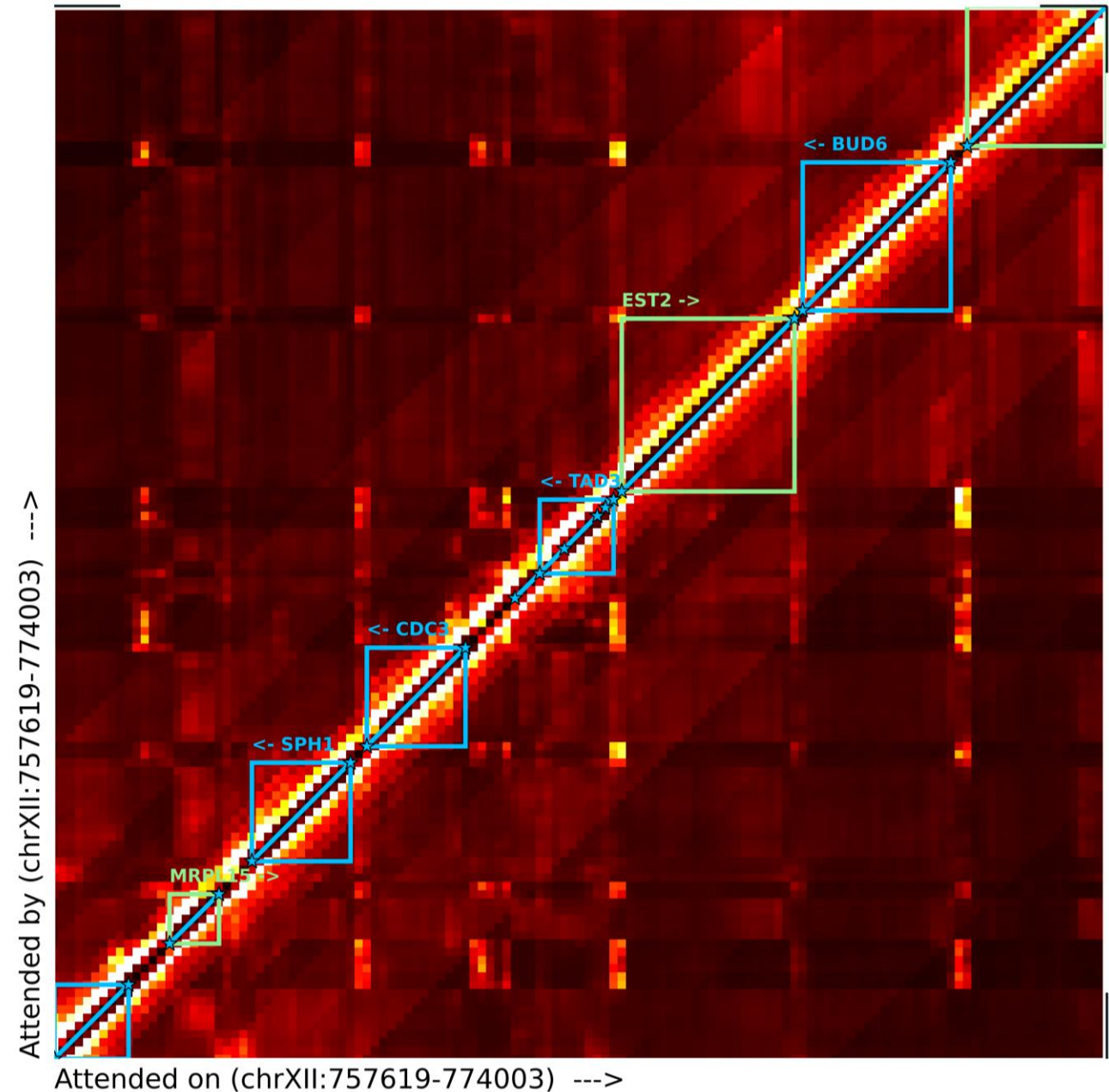
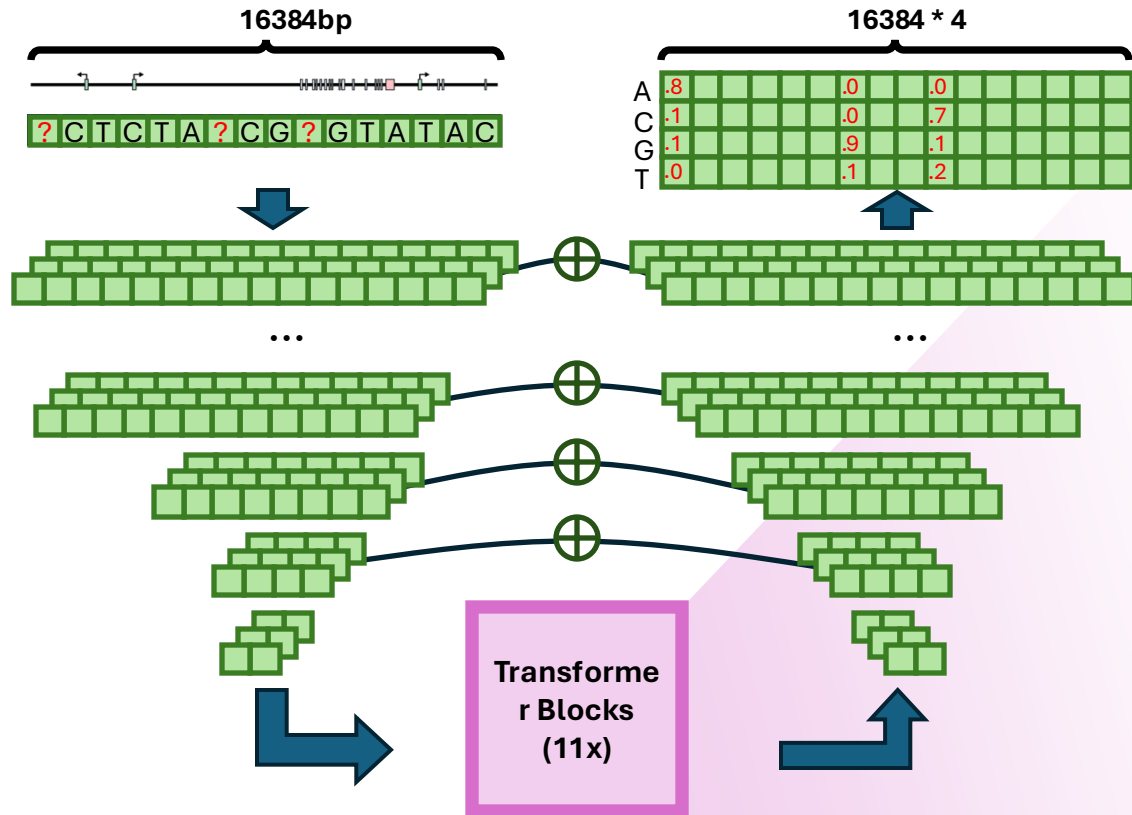
8 *S. cerevisiae* Strains



8 Schizosaccharomycetales genomes



Self-attention maps



Fungal LM: Summary

1. The **Saccharomycetales order** is a good evolutionary distance, offering good species diversity.
2. Thoroughly investigate genomes (protein-coding / repetitive / # gene per window)
3. Homologous sequence removal between train-test/validation is crucial
4. Transformer-based U-Net architecture is the best
5. Model interpretability
 1. LM can capture cis-regulatory motifs
 2. Attention maps highlights potential regulatory elements

Part II

Fine-tuning Fungal Language Model

ChIP-exo, histone marks, RNA-Seq prediction

Q: Does fine-tuning a pretrained LM outperform training a new model from scratch under the exact model architecture?

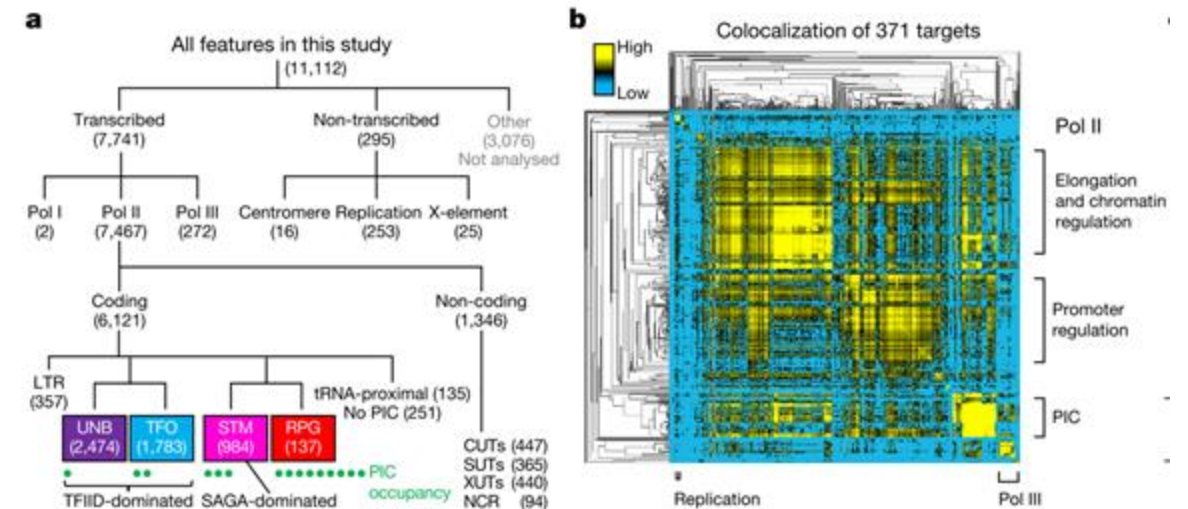
Genomic tracks intro & Data preprocessing



ChiP-exo + Histone Marks + RNA-Seq

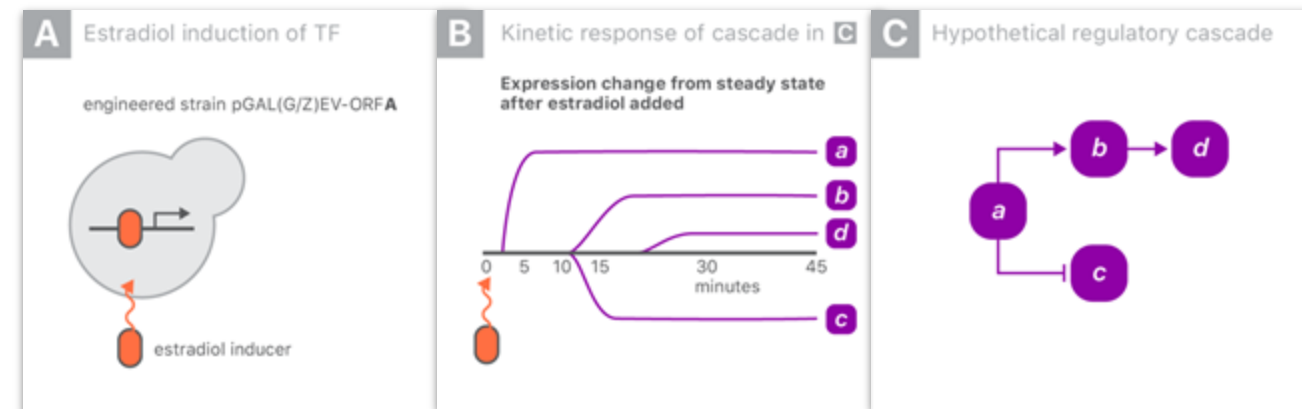
- ChIP-exo provides high res view of protein-DNA binding across the yeast genome.
- Dataset includes 1128 ChIP-exo experiments
- Histone Mods MNase-ChIP-seq

Rossi, M. J., Kuntala, P. K., Lai, W. K., Yamada, N., Badjatia, N., Mittal, C., ... & Pugh, B. F. (2021). A high-resolution protein architecture of the budding yeast genome. *Nature*, 592(7853), 309-314.



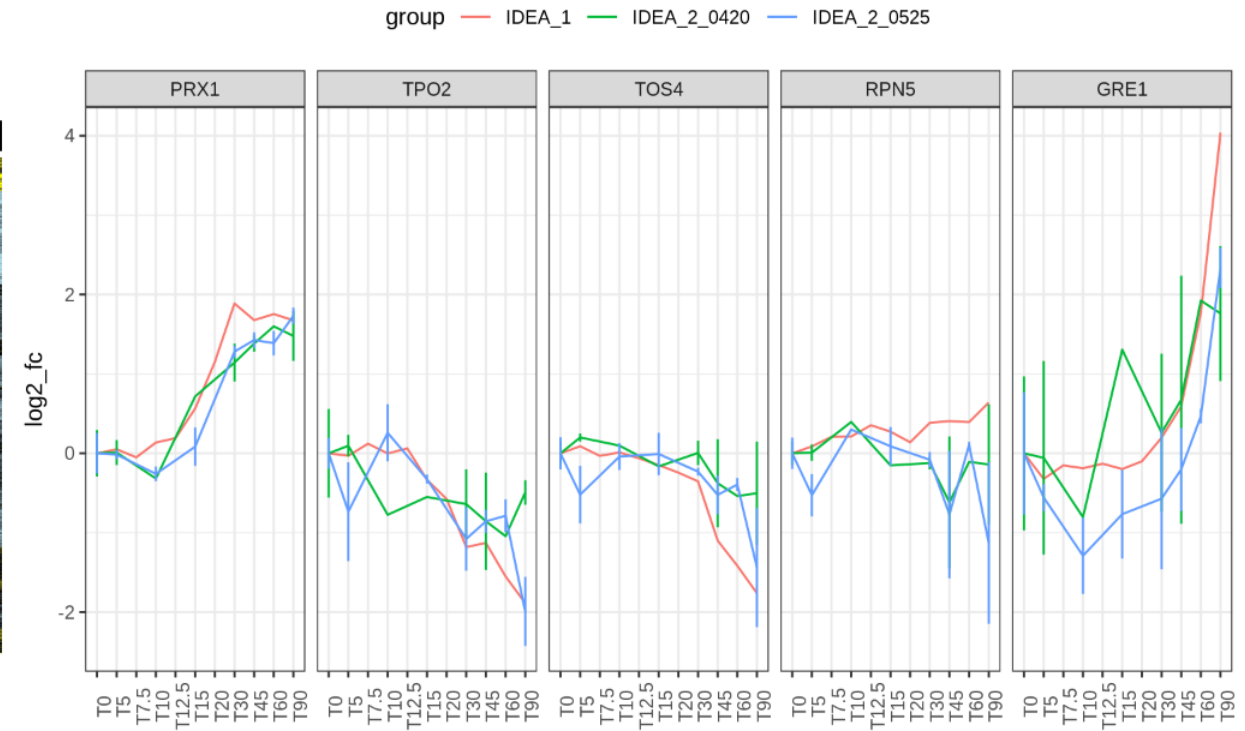
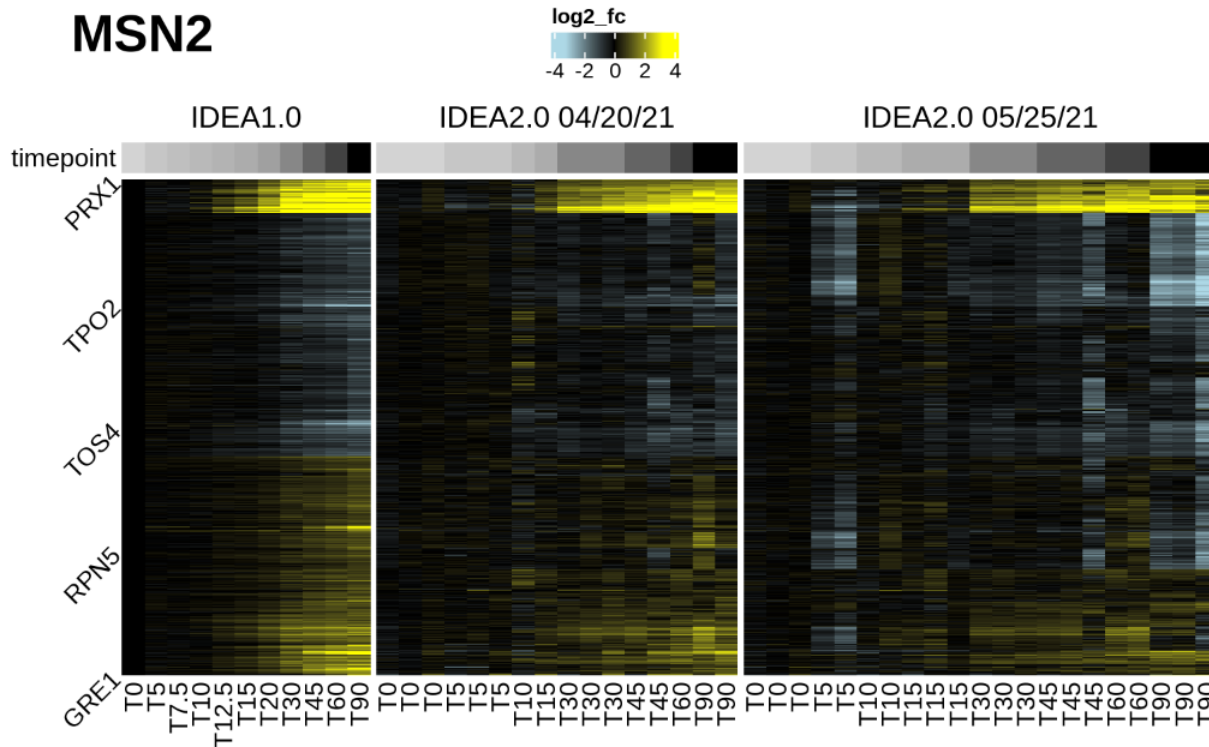
- Genome-scale **perturbation dynamics** propagate signals across regulatory networks (1340 experiments)
- Aggregating dynamics across many **time-courses** enables disambiguation of cause > effect relationships

Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., ... & McIsaac, R. S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Molecular systems biology*, 16(3), e9174.



RNA-Seq

- IDEA (the Induction Dynamics gene Expression Atlas)



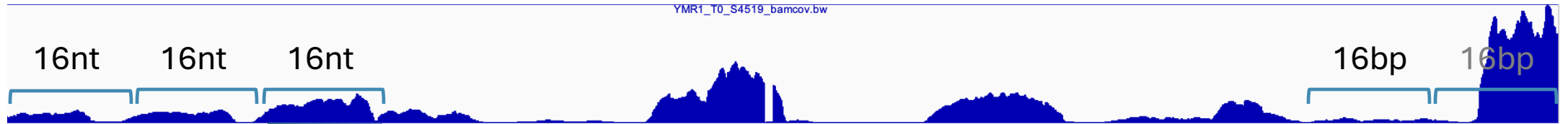
Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., ... & McIsaac, R. S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Molecular systems biology*, 16(3), e9174.

$$\text{Borzoï: squashed scale } y_{j,t}^{(\text{squashed})} = \begin{cases} y_{j,t}^{(3/4)} & \text{if } y_{j,t}^{(3/4)} \leq 384, \\ \text{otherwise } 384 + \sqrt{y_{j,t}^{(3/4)} - 384} \end{cases}$$

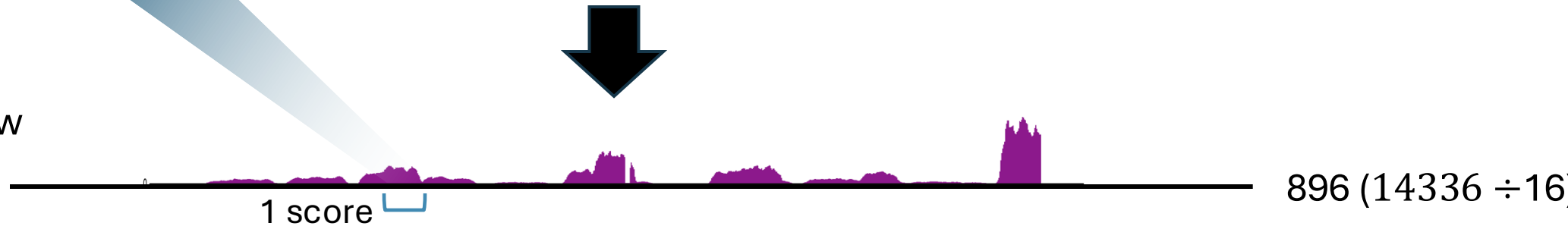
Track transformation

chrXVI:670,897-685,233

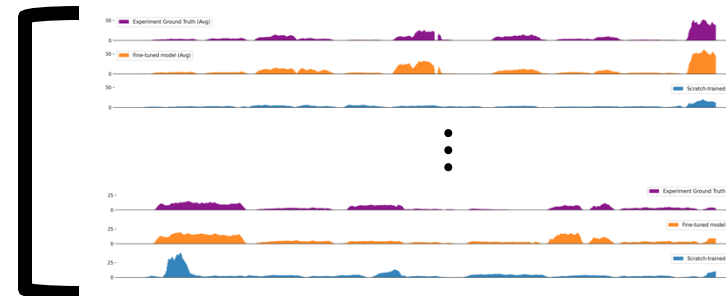
14336 nt (16384 - 1024 * 2)



Summation over
each 16bp window



Coverage Tracks



ChIP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

No normalization across tracks

Loss function: Poisson Loss.

The depth of the track somehow reflects the
quality of the data.

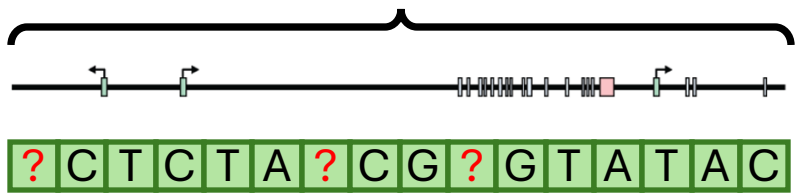
Scratch-trained model vs. Fine-tuned Fungal LM: *Initialization & Training*





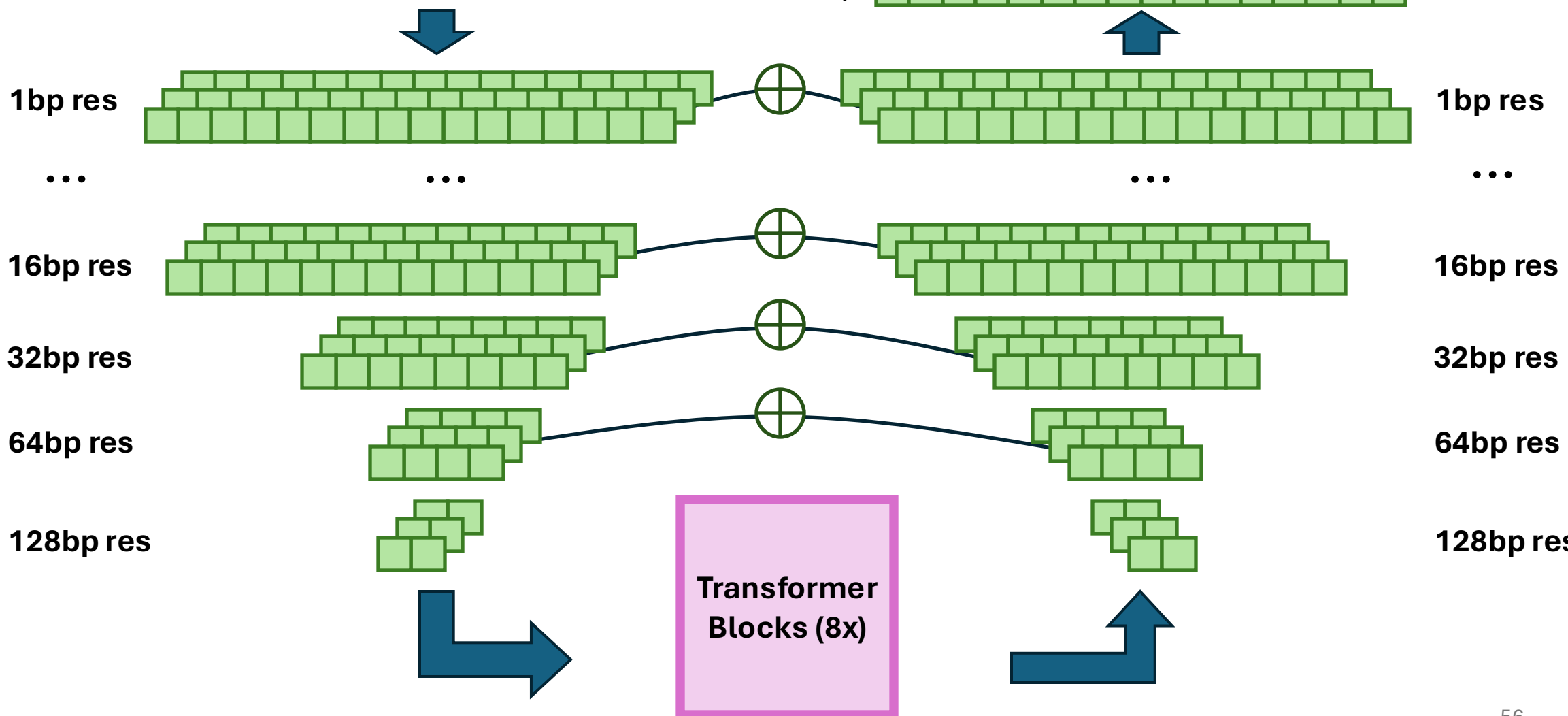
Borzo

16384bp



16384 * 4

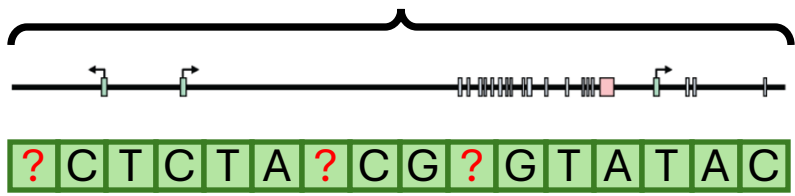
A	.8					.0			.0						
C	.1					.0			.7						
G	.1					.9			.1						
T	.0					.1			.2						



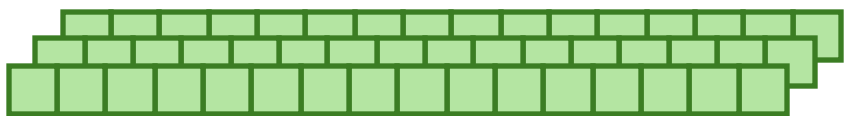


Borzoi

16384bp



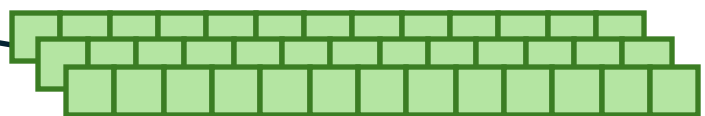
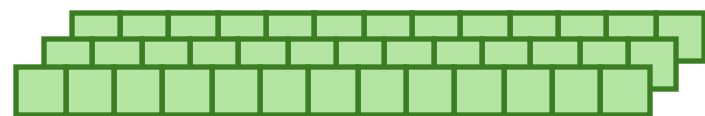
1bp res



...

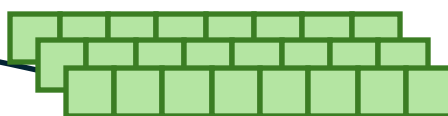
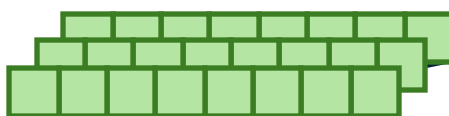
...

16bp res



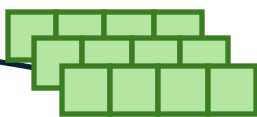
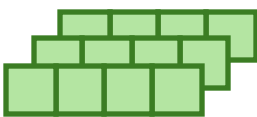
16bp res

32bp res



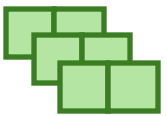
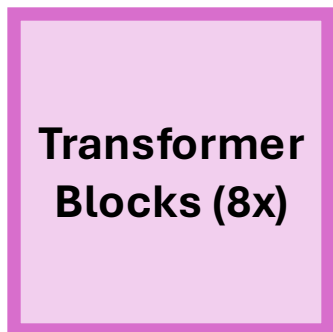
32bp res

64bp res

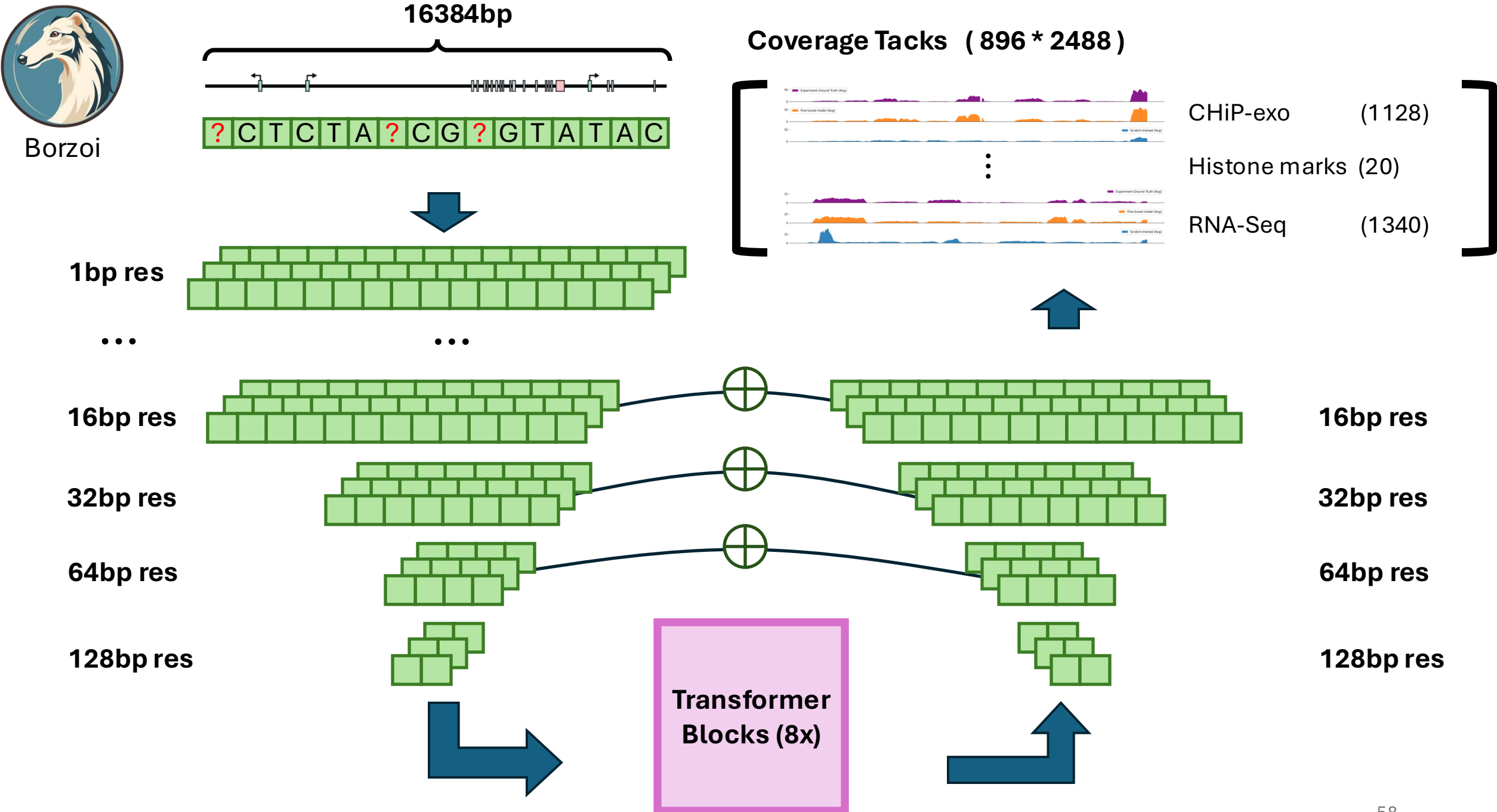


64bp res

128bp res



128bp res



Borzoi

16384bp

Coverage Tacks (896 * 2488)

1bp res

16bp res

32bp res

64bp res

128bp res

16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

CHiP-exo (1128)

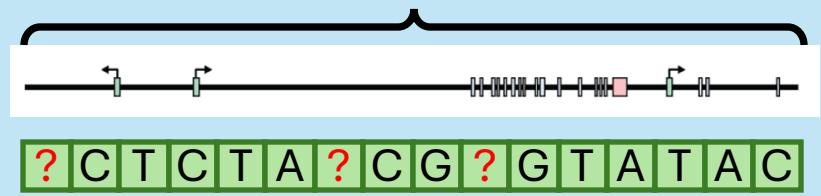
Histone marks (20)

RNA-Seq (1340)

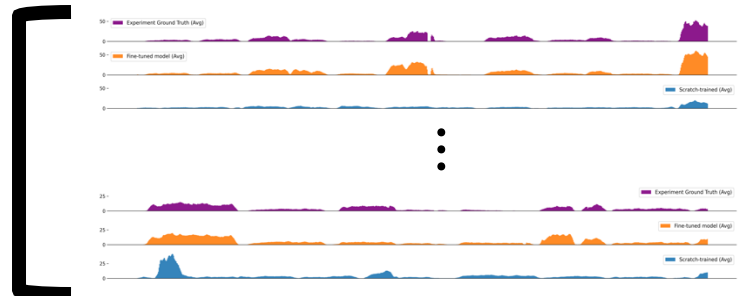


Borzo

16384bp



Coverage Tracks (896 * 2488)

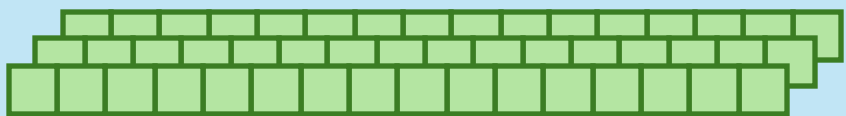


CHiP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

1bp res



...

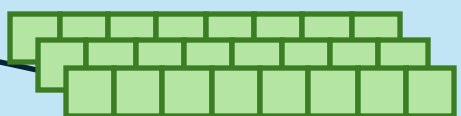
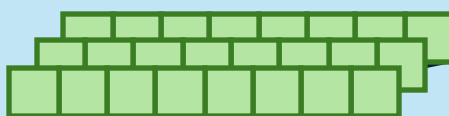
...

16bp res



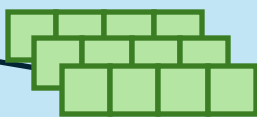
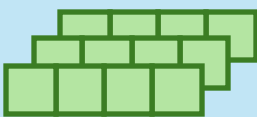
16bp res

32bp res



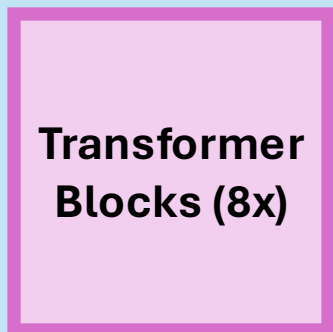
32bp res

64bp res



64bp res

128bp res



128bp res

- 1. Initialized with LM weights (Fine-tuned)
- 2. Random initialization (Scratch-trained)

Model training: 8-fold cross validation

- Divide genome into 8 folds.
- Train 8 models with distinct validation and test folds.



Fold0: 743 seq, 1406020 nt (0.1244)

chrXIV: 0-628758
chrX: 0-436307
chrXI: 440246-666816
chrIII: 0-114385

Fold1: 736 seq, 1433427 nt (0.1268)

chrXI: 0-440129
chrV: 0-151987
chrV: 152104-576874
chrXIII: 0-268031
chrVI: 0-148510

Fold2: 806 seq, 1521492 nt (0.1346)

chrII: 238323-813184
chrVII: 0-496920
chrIV: 0-449711

Fold3: 755 seq, 1408276 nt (0.1246)

chrXVI: 0-555957
chrIV: 449821-990877
chrVI: 48627-270161
chrVIII: 0-105586
chrIX: 355745-439888

Fold4: 732 seq, 1444997 nt (0.1278)

chrIV: 990877-1531933
chrXII: 614562-1078177
chrII: 0-238207
chrIII: 114501-316620

Fold5: 742 seq, 1284157 nt (0.1136)

chrVII: 497038-1090940
chrX: 436425-745751
chrI: 0-151465
chrI: 151582-230218
chrXII: 0-150828

Fold6: 785 seq, 1446481 nt (0.1280)

chrXIII: 268149-924431
chrXII: 150947-614562
chrXV: 0-326584

Fold7: 733 seq, 1360020 nt (0.1203)

chrVIII: 105703-562643
chrXVI: 556073-948066
chrIX: 0-355629
chrXIV: 628875-784333

Scratch-trained

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

Input

$16,384 * 4$

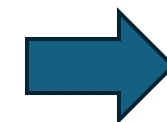
A	1																			
C	0																			
G	0																			
T	0																			

Model

Fine-tuned

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

$16,384 * (4 + 1 + 165)$



Masked encoding

Species encoding
(r64 : **109**)

A																				
C																				
G																				
T																				
...																				

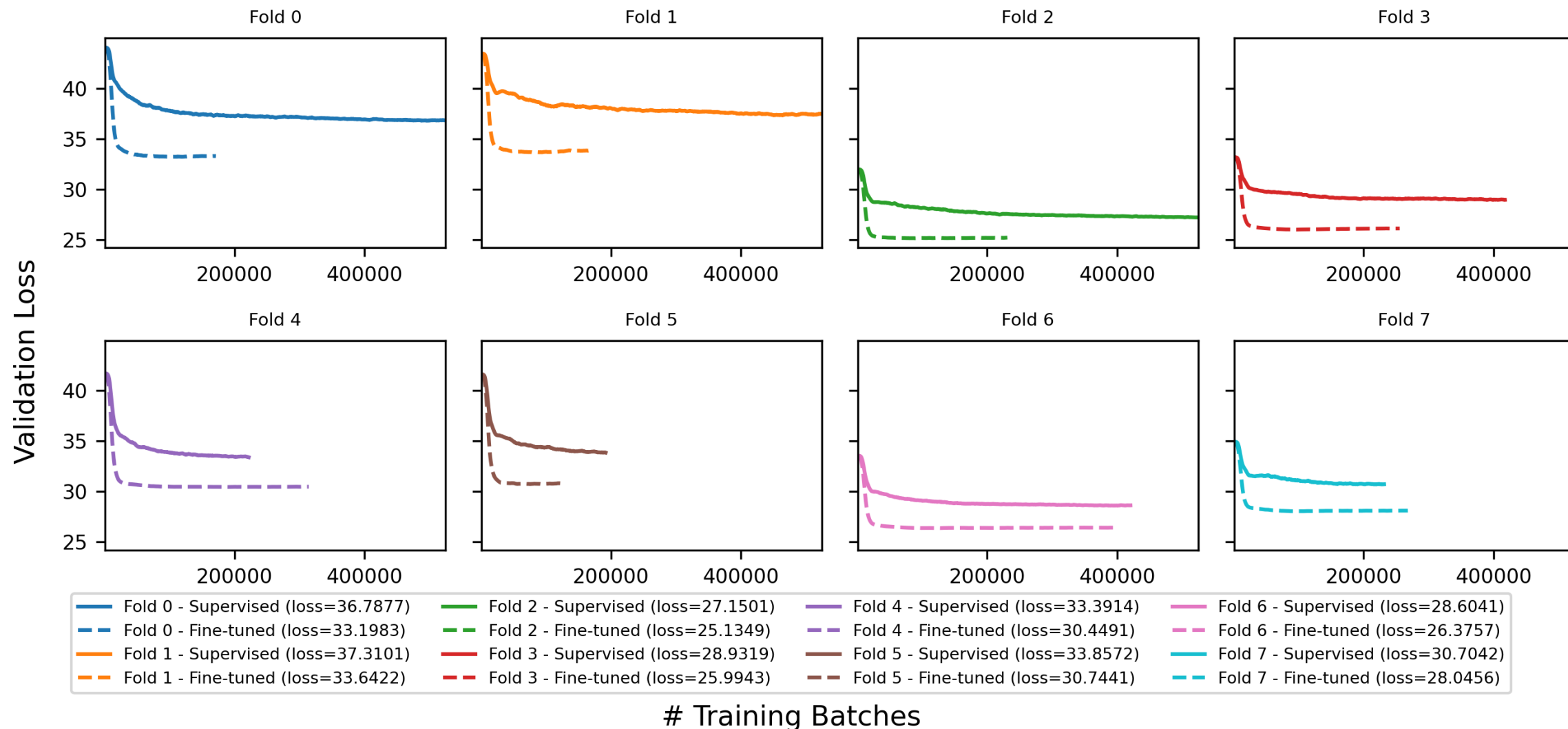
Model

Scratch-trained model vs. Fine-tuned Fungal LM: *Training Results*



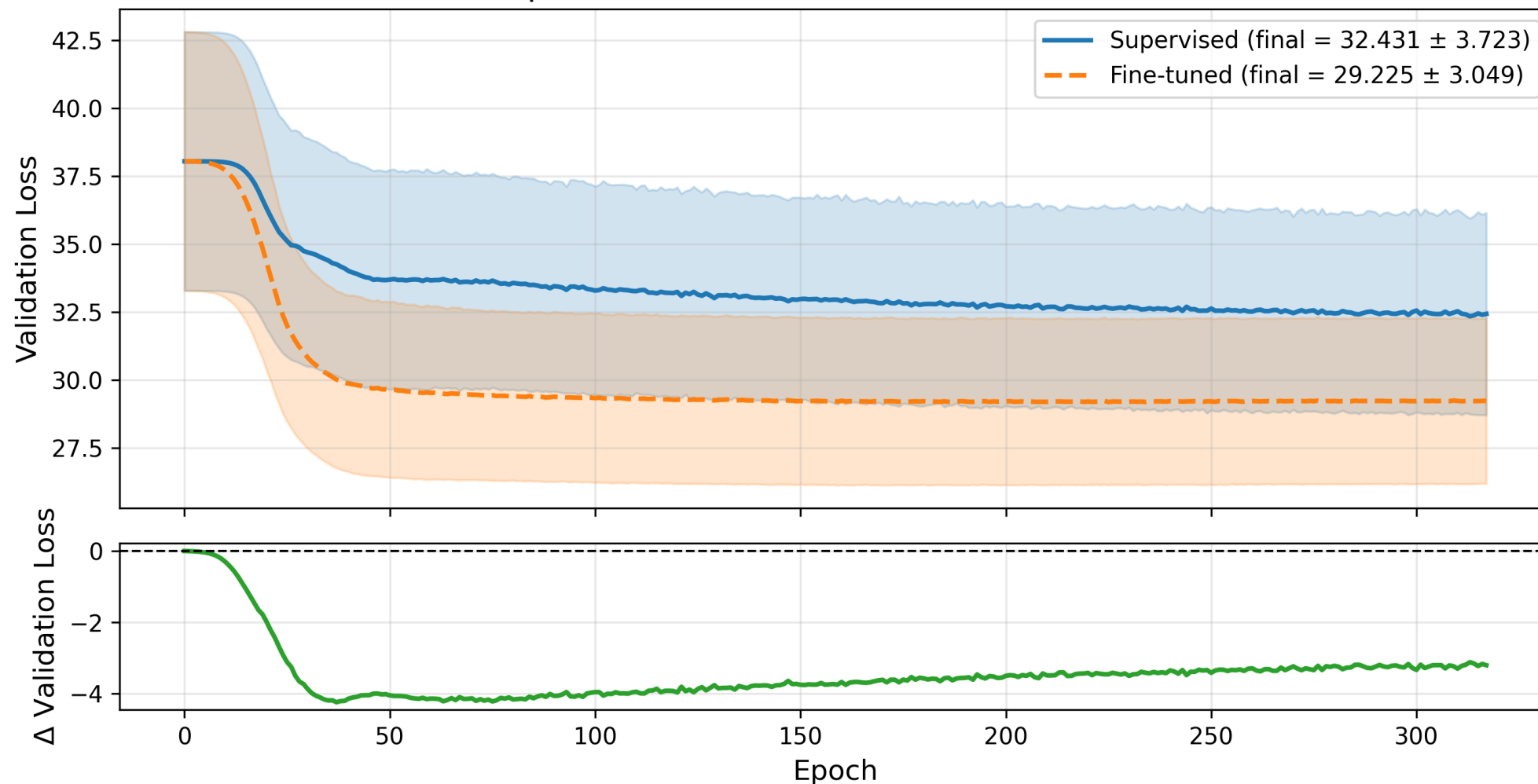
Fine-tuned vs Scratch-Trained

(16 bp resolution)

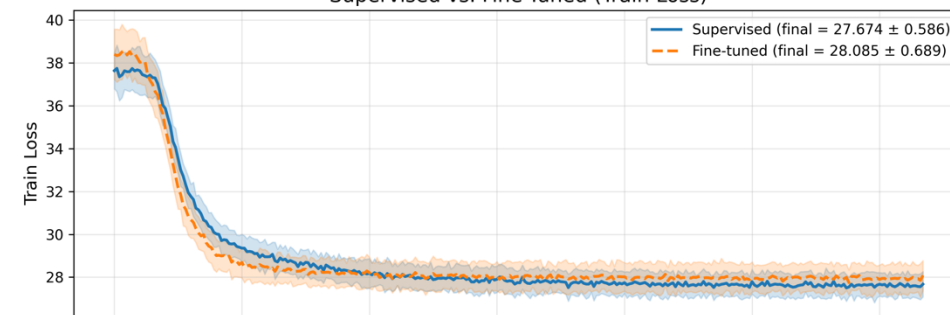


Fine-tuned vs Scratch-Trained

Supervised vs. Fine-Tuned (Validation Loss)



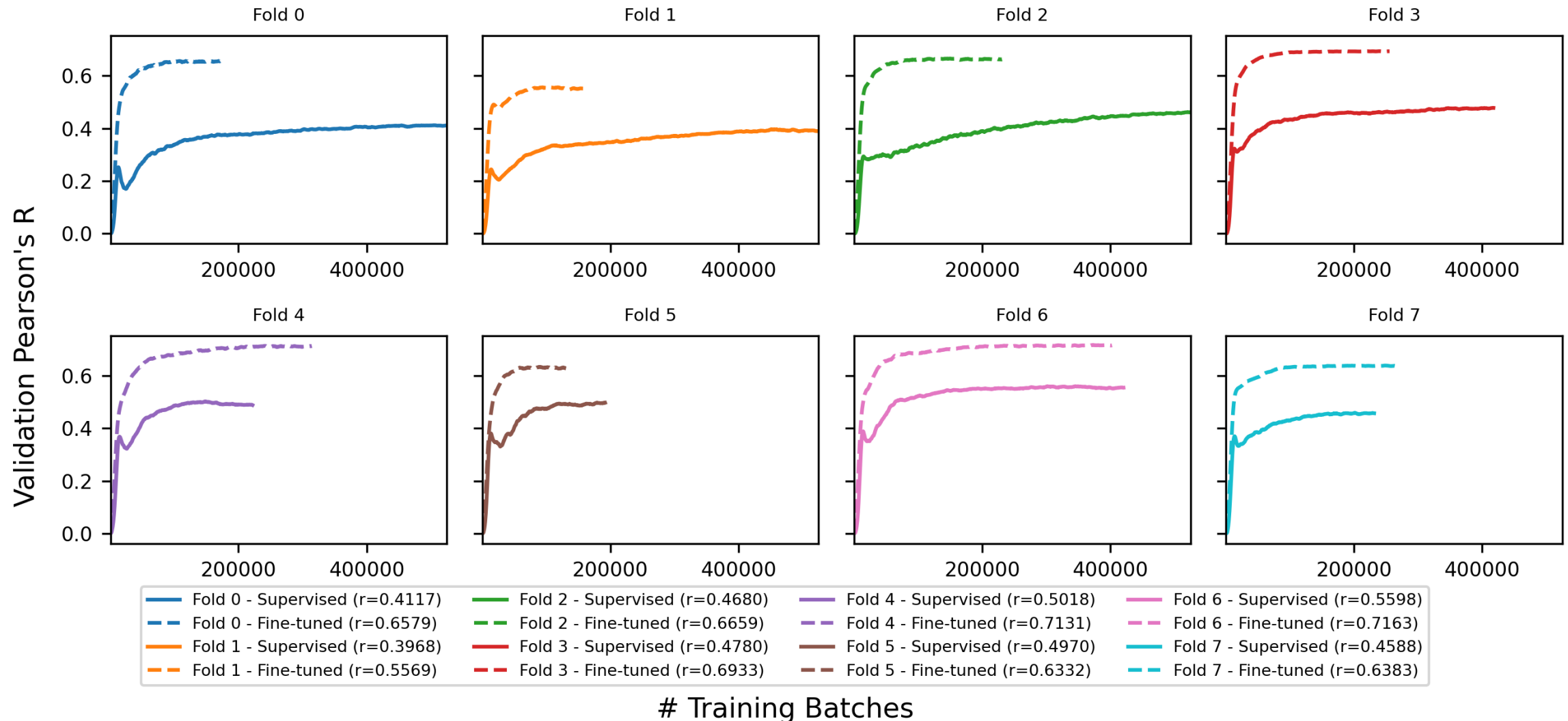
Supervised vs. Fine-Tuned (Train Loss)



(16 bp resolution)

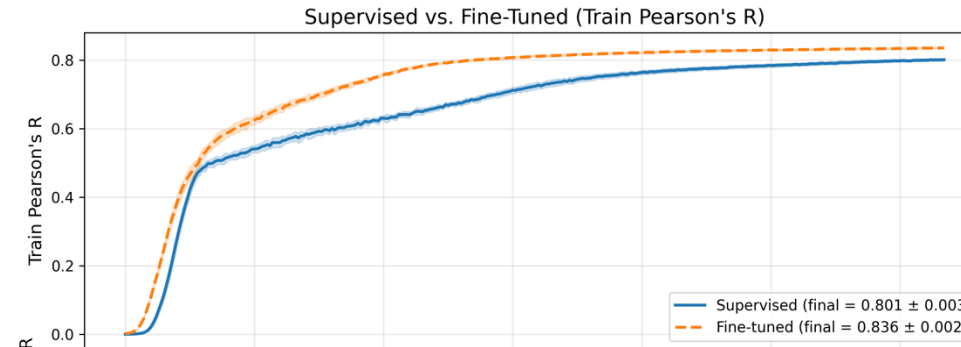
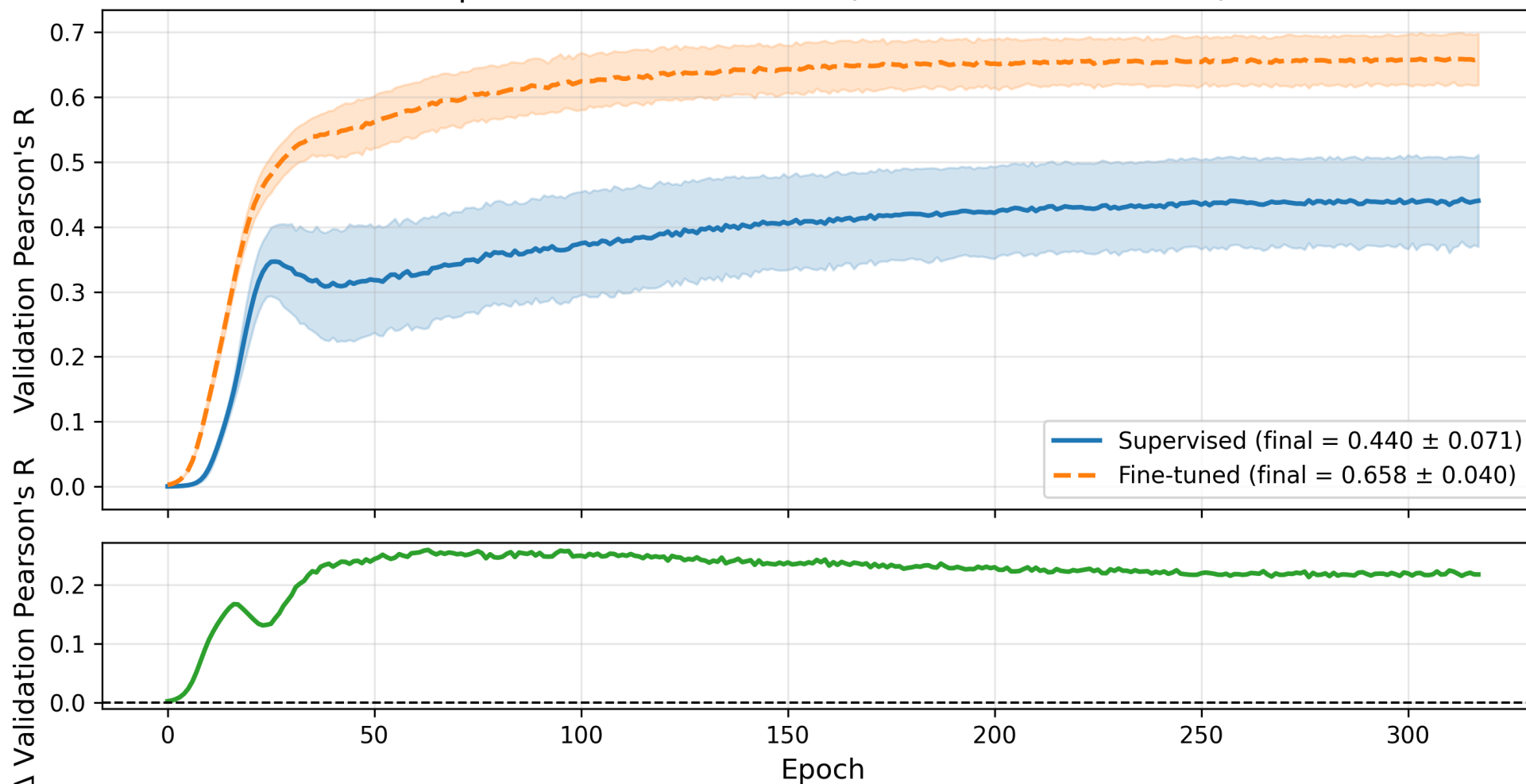
Fine-tuned vs Scratch-Trained

(16 bp resolution)



Fine-tuned vs Scratch-Trained

Supervised vs. Fine-Tuned (Validation Pearson's R)

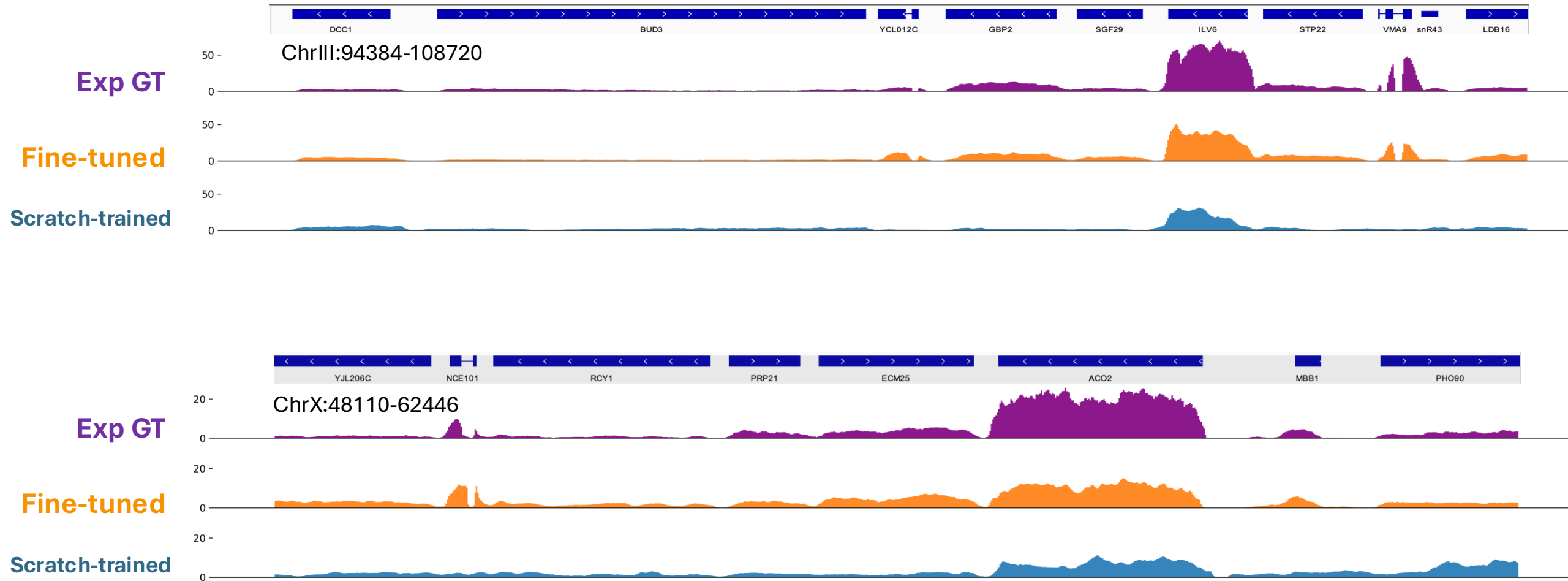


(16 bp resolution)

Scratch-trained model vs. Fine-tuned Fungal LM: *Track-level prediction evaluation*

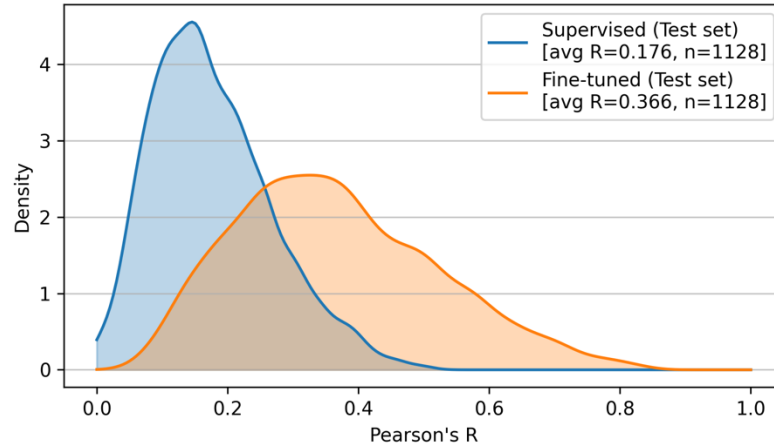


Fine-tuned vs Scratch-Trained (Test set)

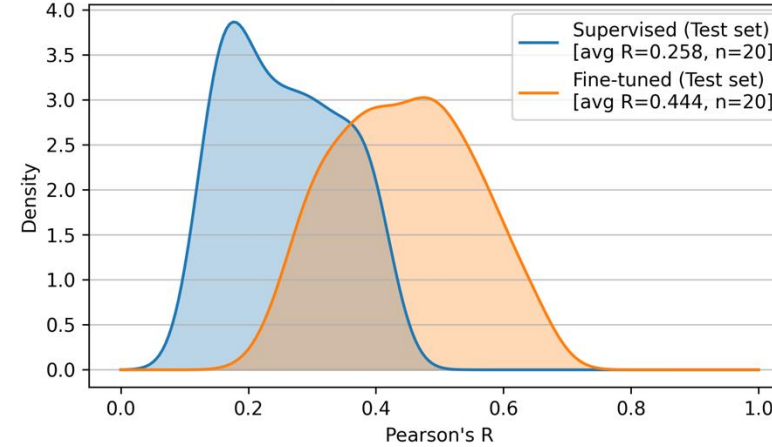


Fine-tune vs Scratch-Trained (Test set)

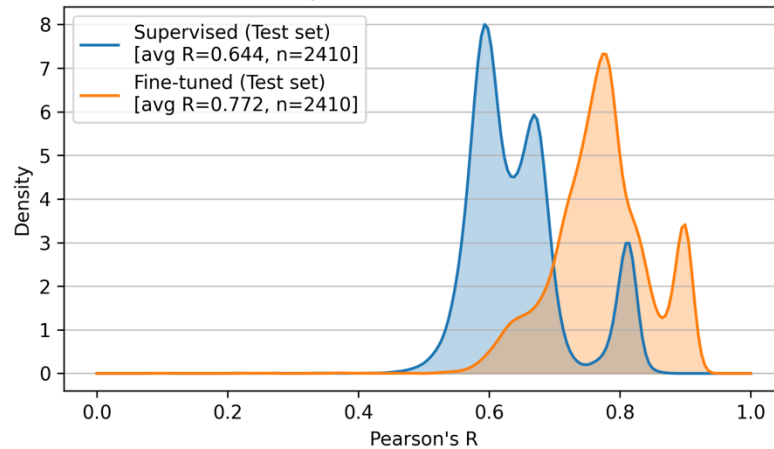
Density of Bin-level Predictions (ChIP-exo)
Supervised vs. Fine-tuned



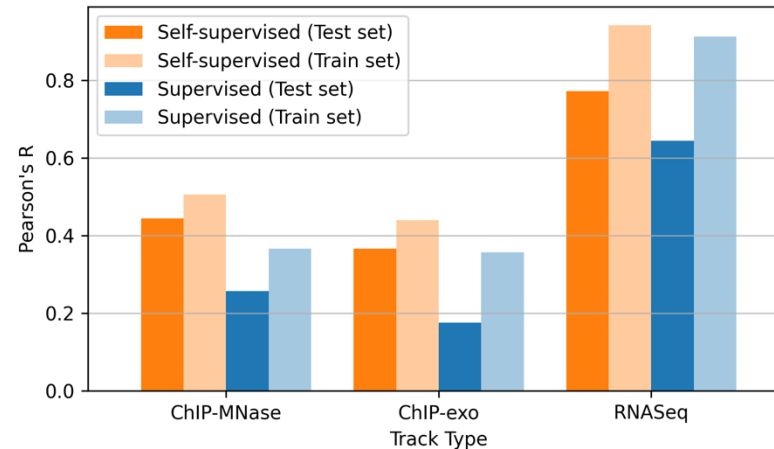
Density of Bin-level Predictions (ChIP-MNase)
Supervised vs. Fine-tuned



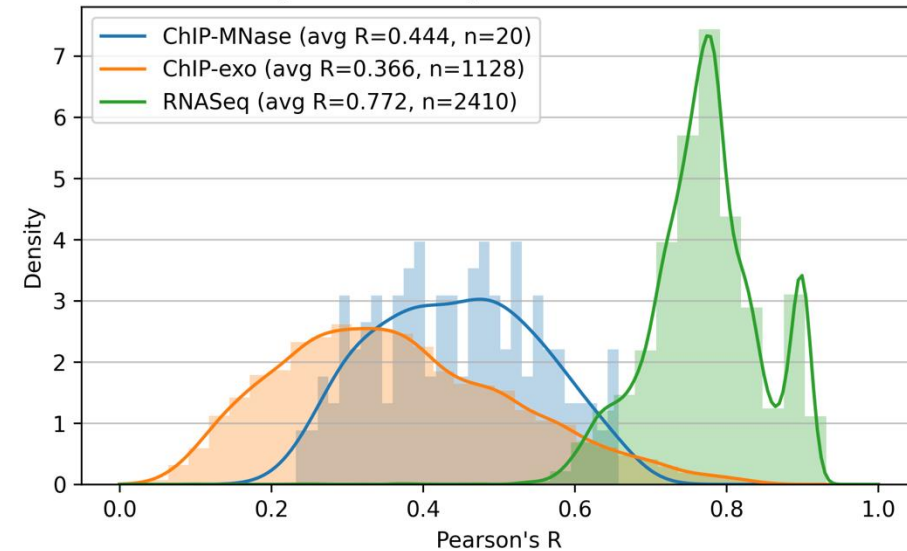
Density of Bin-level Predictions (RNASeq)
Supervised vs. Fine-tuned



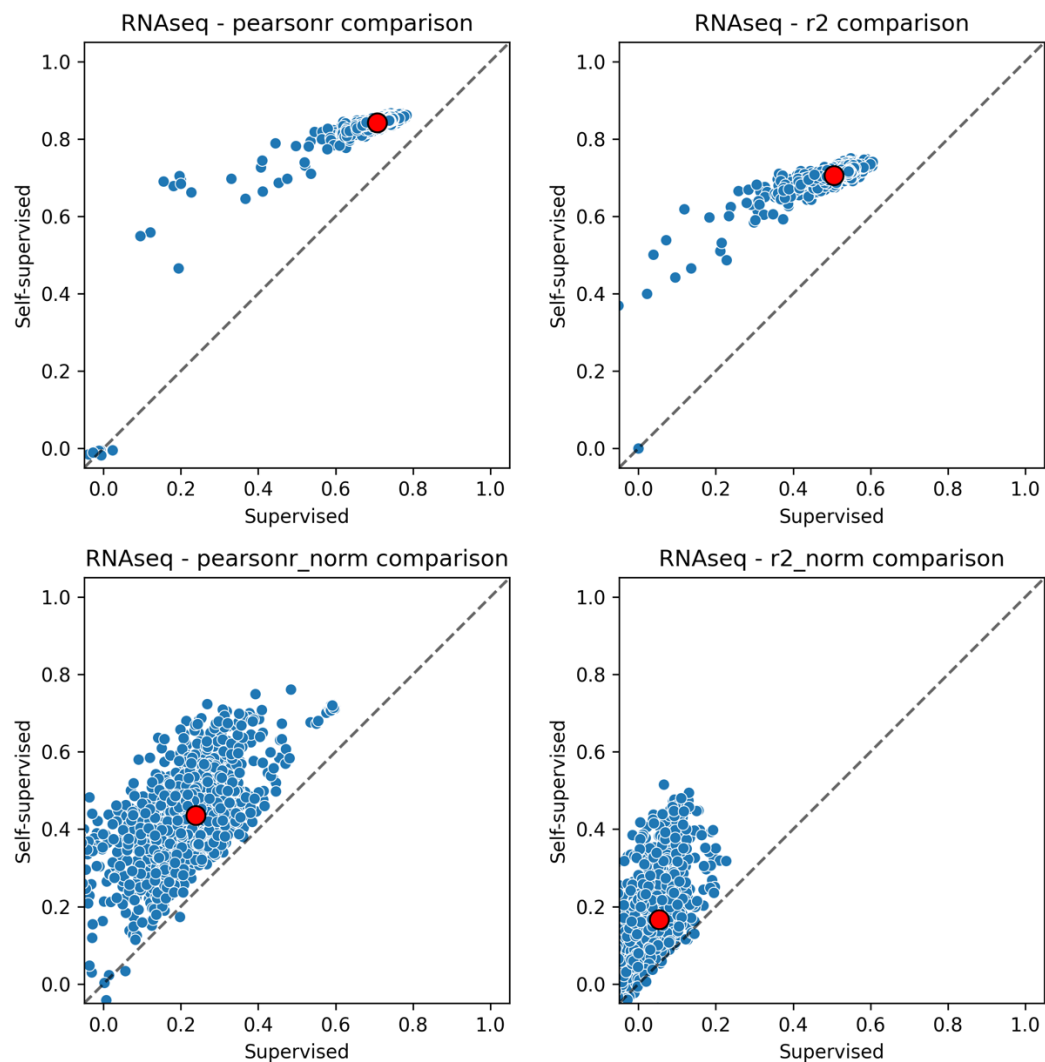
Comparison of Average Pearson's R by Track Type



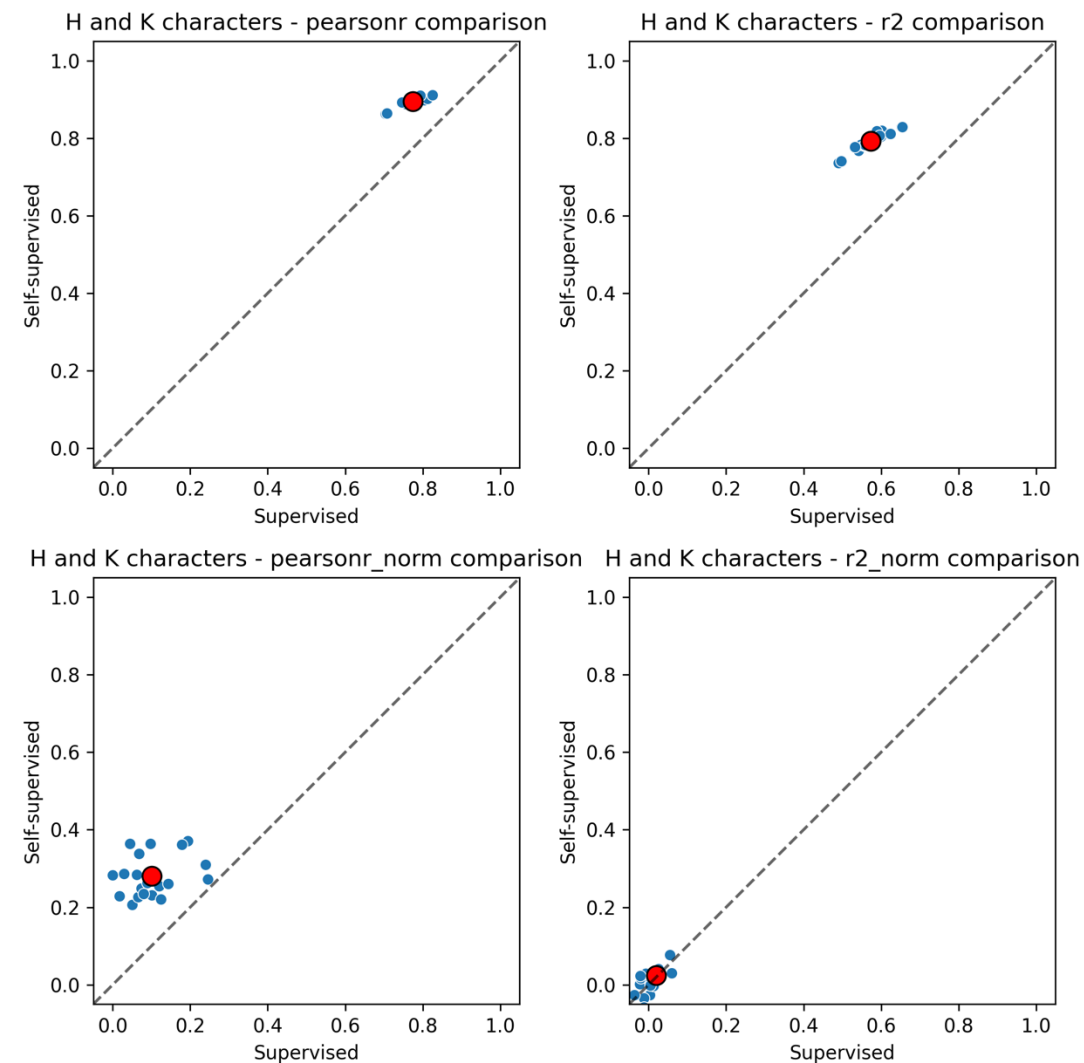
Histogram + Density of Bin-level Predictions



RNA-Seq



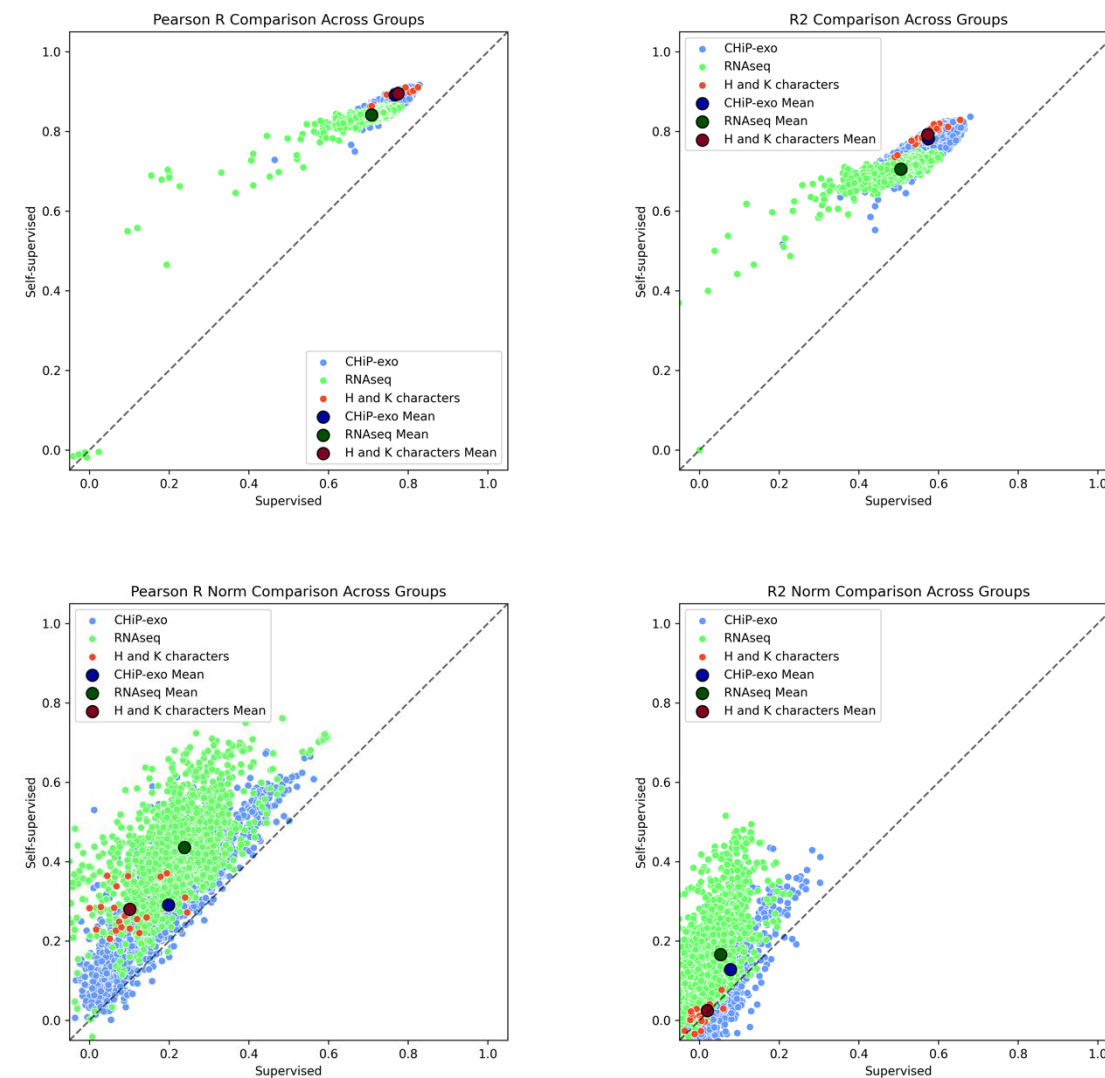
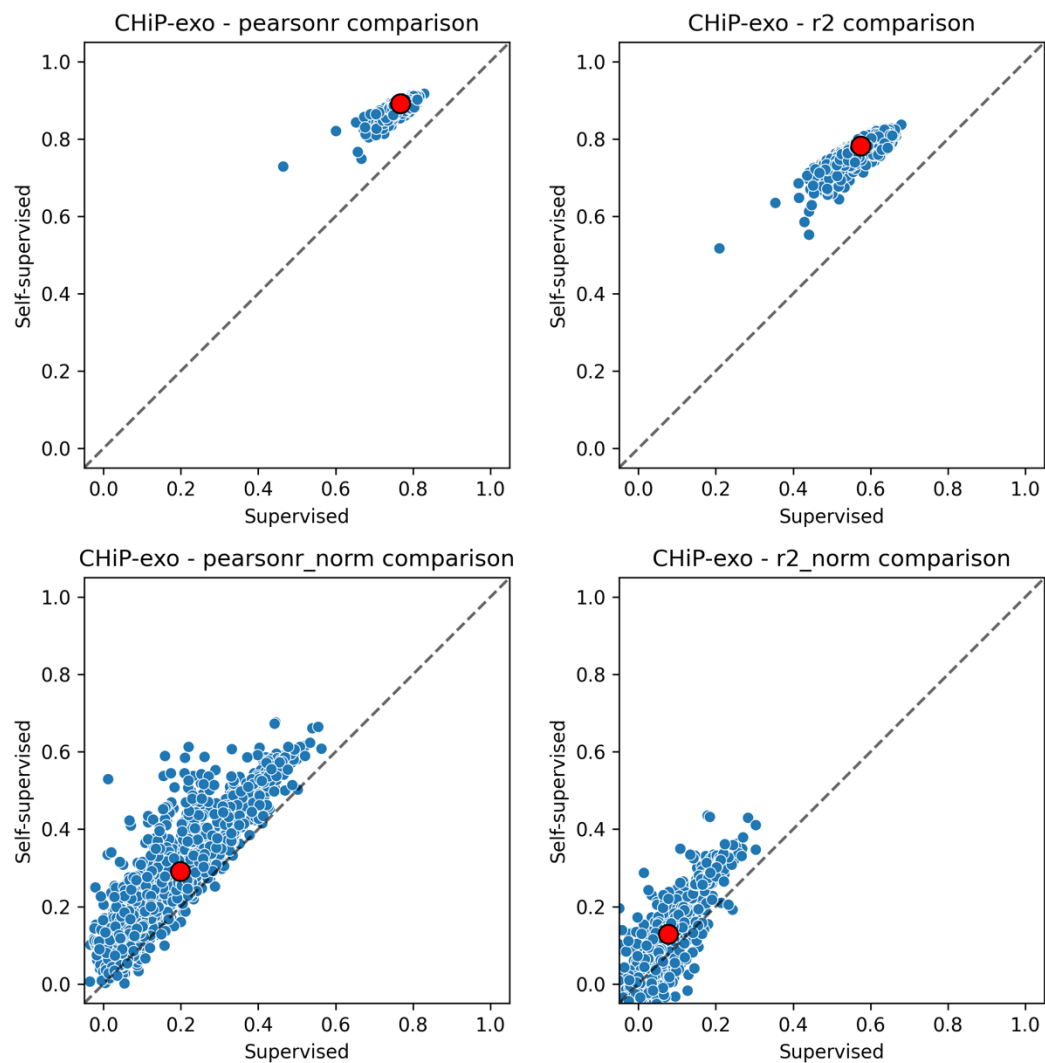
Histone Marks



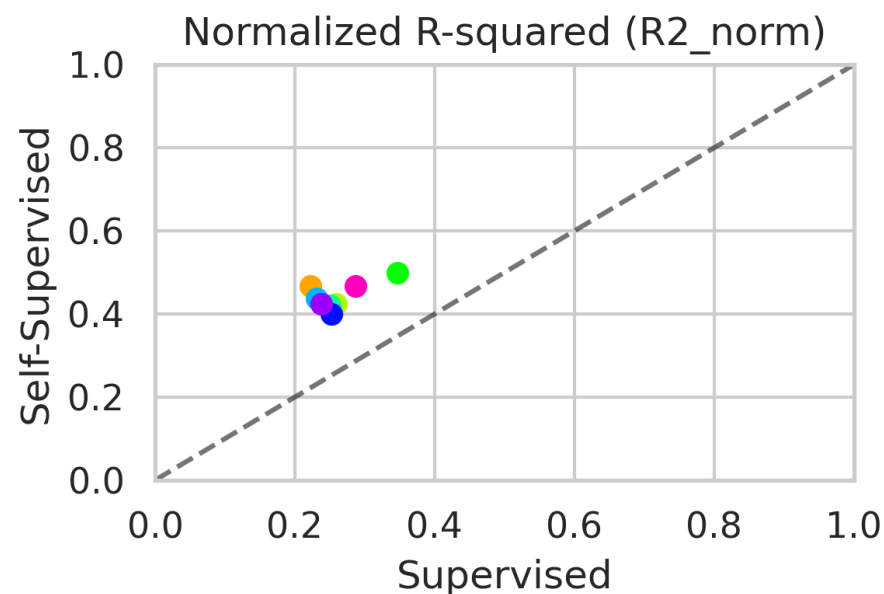
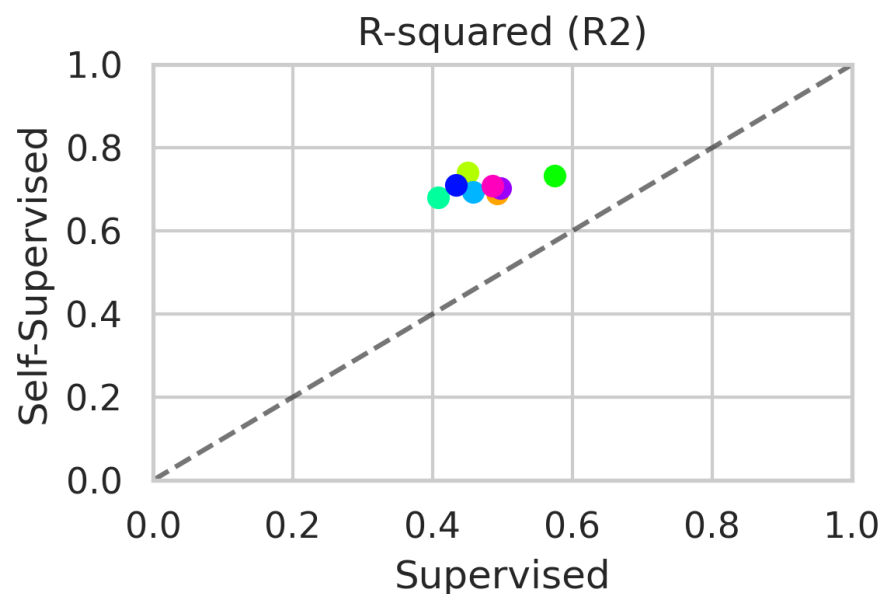
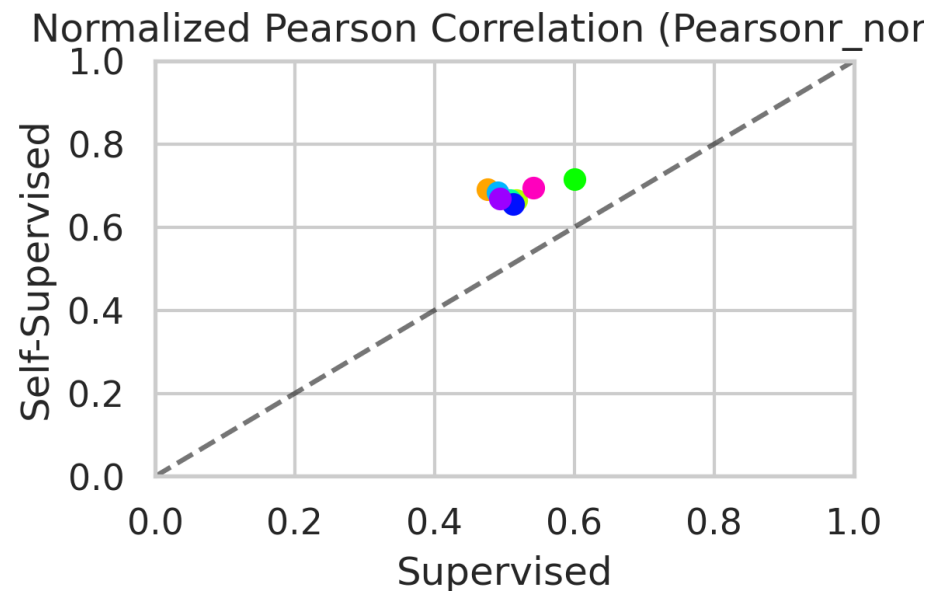
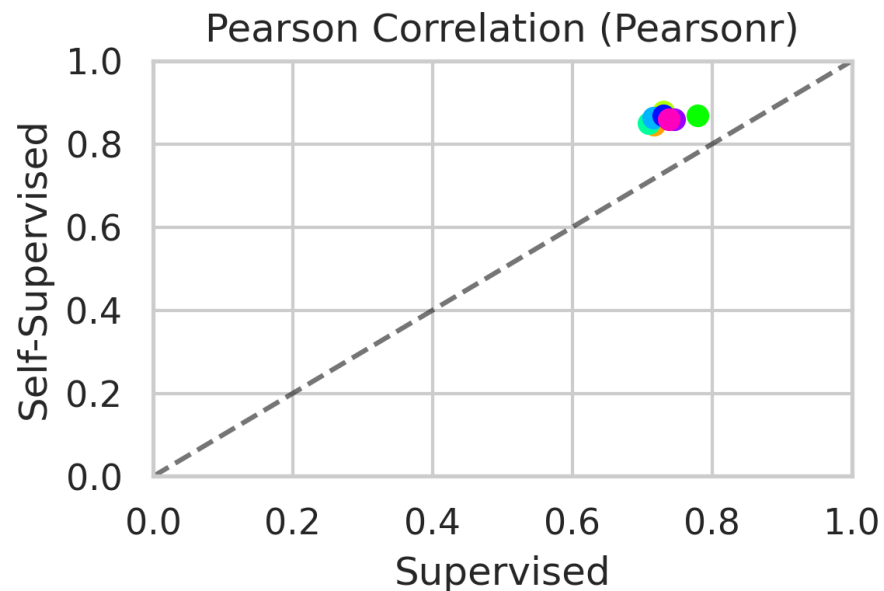
Average results across 8 folds. Each dot is a track.

ChIP-exo

All (RNA-Seq + Histone Marks + ChIP-exo)



Average results across 8 folds. Each dot is a track.



Average results
across tracks.
Each dot is a fold

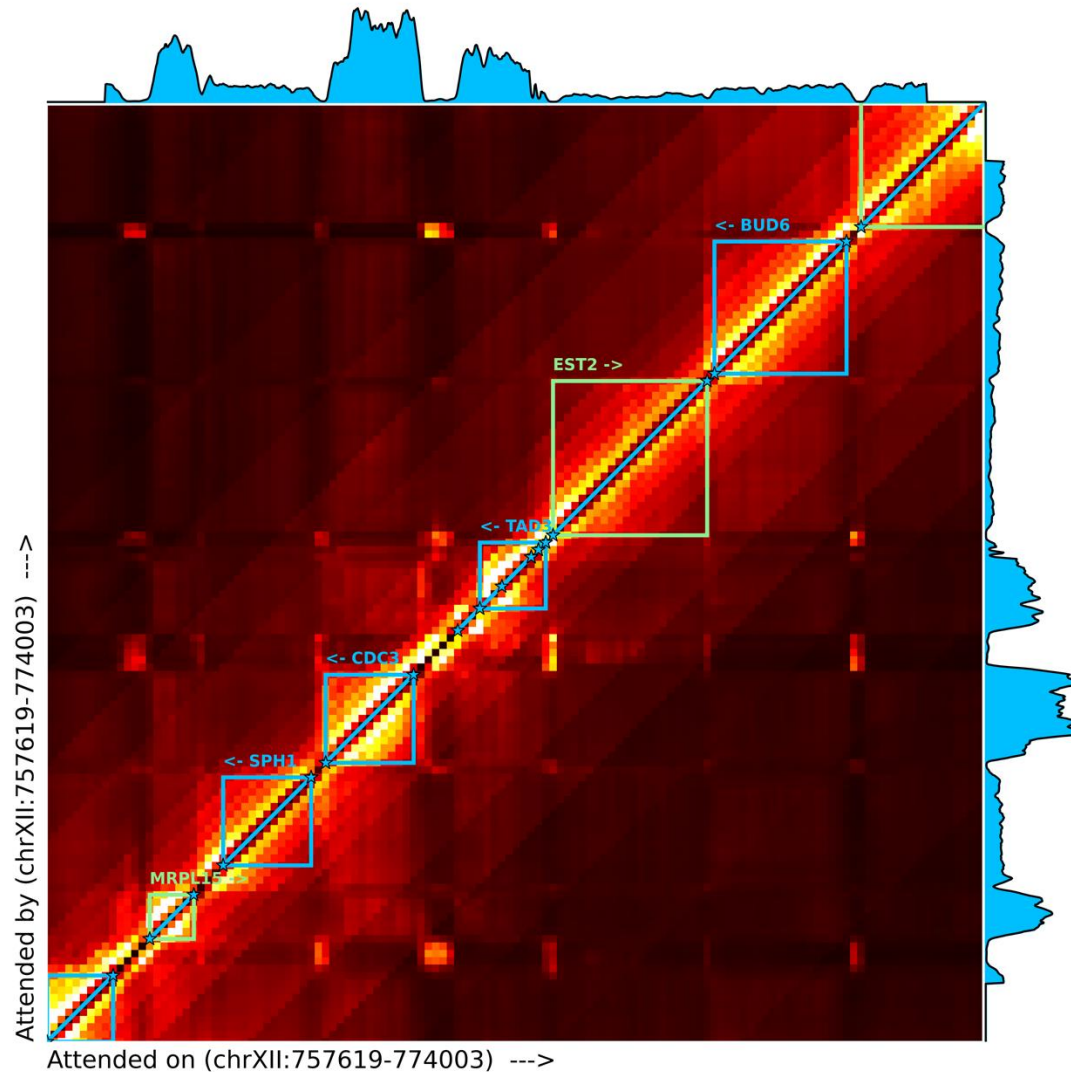


Scratch-trained model vs. Fine-tuned Fungal LM: *Attention maps*

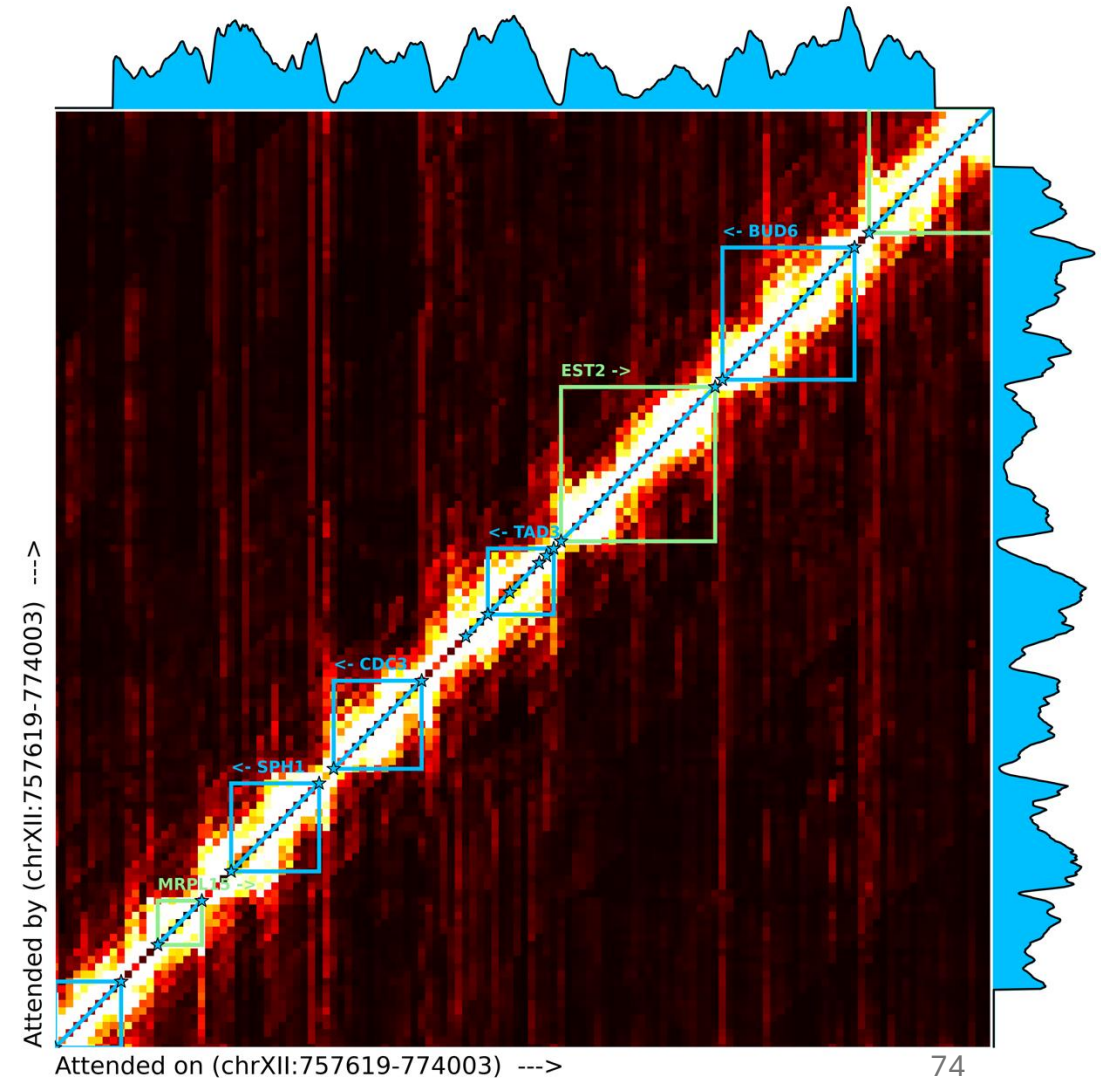


Fine-tuned vs Scratch-Trained

Fine-tuned



Scratch-Trained



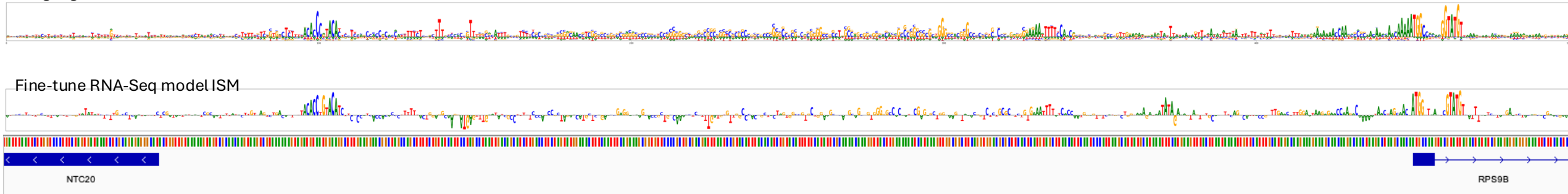
Scratch-trained model vs. Fine-tuned Fungal LM: *Motif usage*



Fine-tuned vs Scratch-Trained

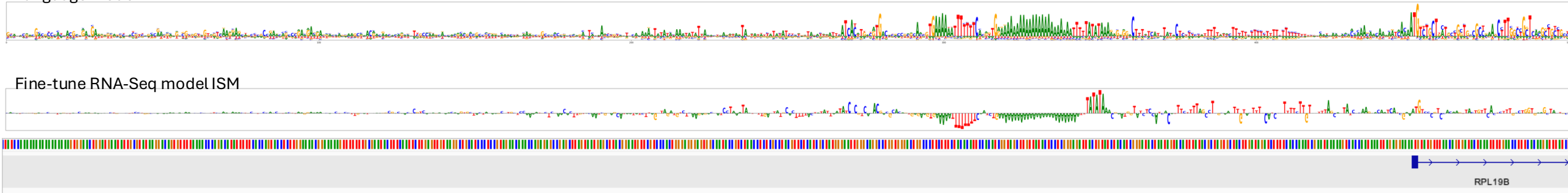
RPS9B

Language model



RPL4A

Language model



Part III

Caudal, É., Loegler, V., Dutreux, F., Vakirlis, N., Teyssonnière, É., Caradec, C., ... & Schacherer, J. (2024). Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nature Genetics*, 1-10.

Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergström, A., ... & Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701), 339-344.

Applications

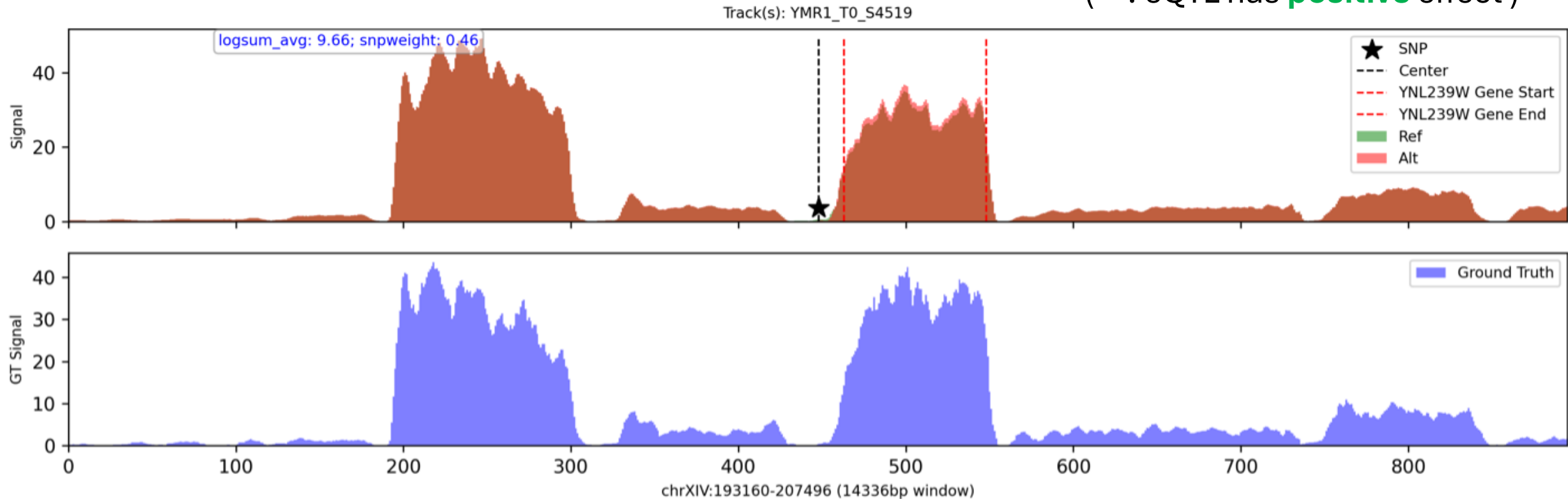
1. Assessing the variant effects on eQTLs and negatively selected eQTLs
2. Predicting the influence of distal regulatory elements (i.e. enhancers) on gene expression.
3. MPRA mutation effect prediction

Predicting eQTLs

SNPWeight from GWAS (+)

Yeast LM prediction (+)

(+: eQTL has **positive** effect)

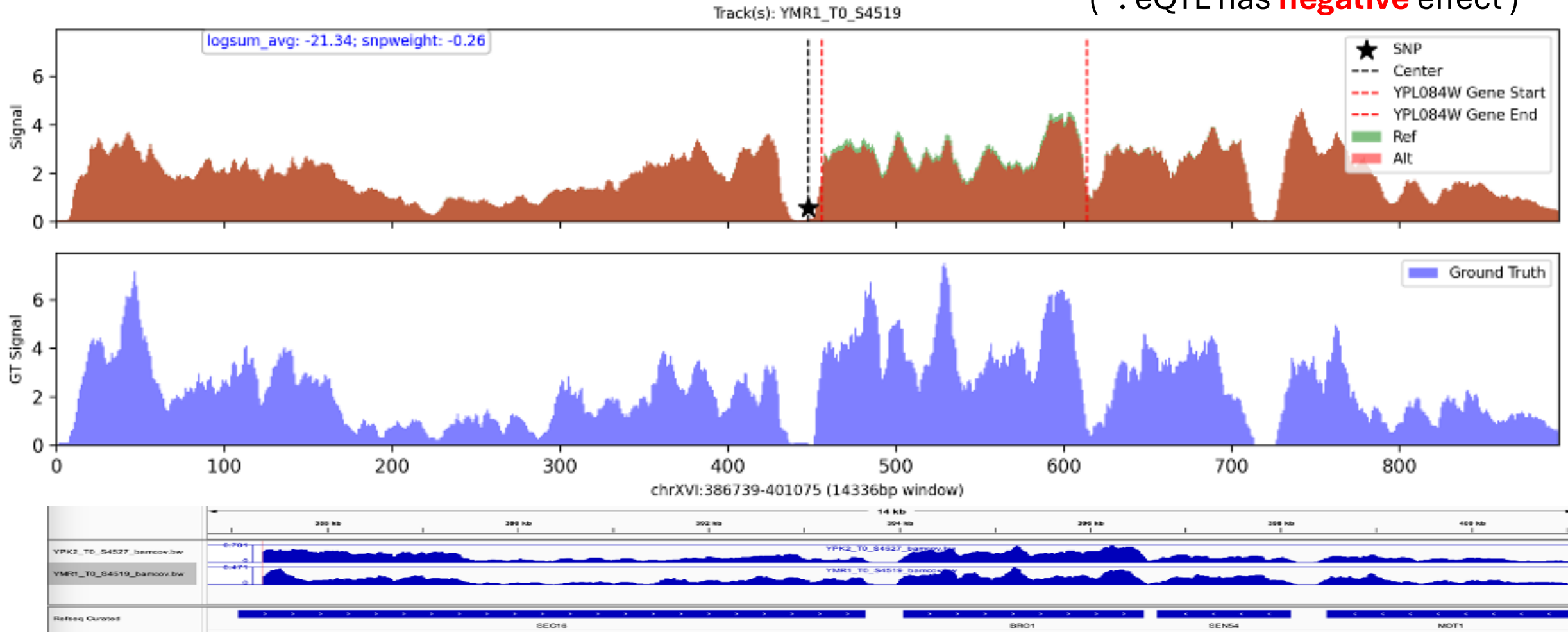


Predicting eQTLs

SNPWeight from GWAS (-)

Yeast LM prediction (-)

(-: eQTL has **negative** effect)



Conclusions

- **Fungal LM:**
 - Learned gene structure
 - Learned conserved regulatory motifs
- **Fine-tuning Fungal LM:**
 - Improved models training from scratch substantially
 - **0.7** Pearson's R in test set
- **Applications**
 - Assessing variant effects on eQTLs
 - Predicting distal regulatory elements influencing gene expression.
 - Predicting mutation effects with MPRA.



ChatGPT prompt:
Generate a figure about deep learning,
genomics, DNA, and language model



khchao.com



@KuanHaoChao



Kuanhao-Chao

Acknowledgement



Johannes Linder



Majed Mohamed Magzoub



David Kelley



Sean Hackett



Steven Salzberg



Mihaela Pertea

Kelley Lab & Calico Computing Team

Great mentors, collaborators and good friends!





JHU Deep Learning in Genomics Study Group

- **Date and Time:** Every other Tuesday , 12:00 pm - 1:00 pm.
 - Next meeting 01/28. Celine's presenting "*A foundation model of transcription across human cell types*"
- **Location:** Room 228 at Malone or on Zoom
- **Slack Channel:** [#deep-learning-reading-group](#)
- Come join us!!

10/22	Kuan-Hao Chao	Predicting RNA-seq coverage from DNA sequence	https://doi.org/10.1038/s41588-024-02053-6
11/05	Mahler Revsine	HyenaDNA: Long-Range Genomic Sequence Model	https://doi.org/10.48550/arXiv.2306.15794
11/19	Cristina Martin Linares	Machine-guided design of cell-type-targeting cis-regulatory elements	https://doi.org/10.1038/s41586-024-08070-z
12/03	Eduarda Vaz	Effective gene expression prediction from sequence	https://doi.org/10.1038/s41592-021-01252-x
1/7	Gus Fridell	Applying interpretable machine learning in computational genomics	https://doi.org/10.1038/s41592-024-02359-7
01/21	Stephen Hwang	Evolutionary-scale prediction of atomic-level protein structure	https://doi.org/10.1126/science.adc2574
02/04	Celine Hoh	A foundation model of transcription across human cell types	https://doi.org/10.1038/s41586-024-08391-z

Presenter Form

