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

Kuan-Hao Chao

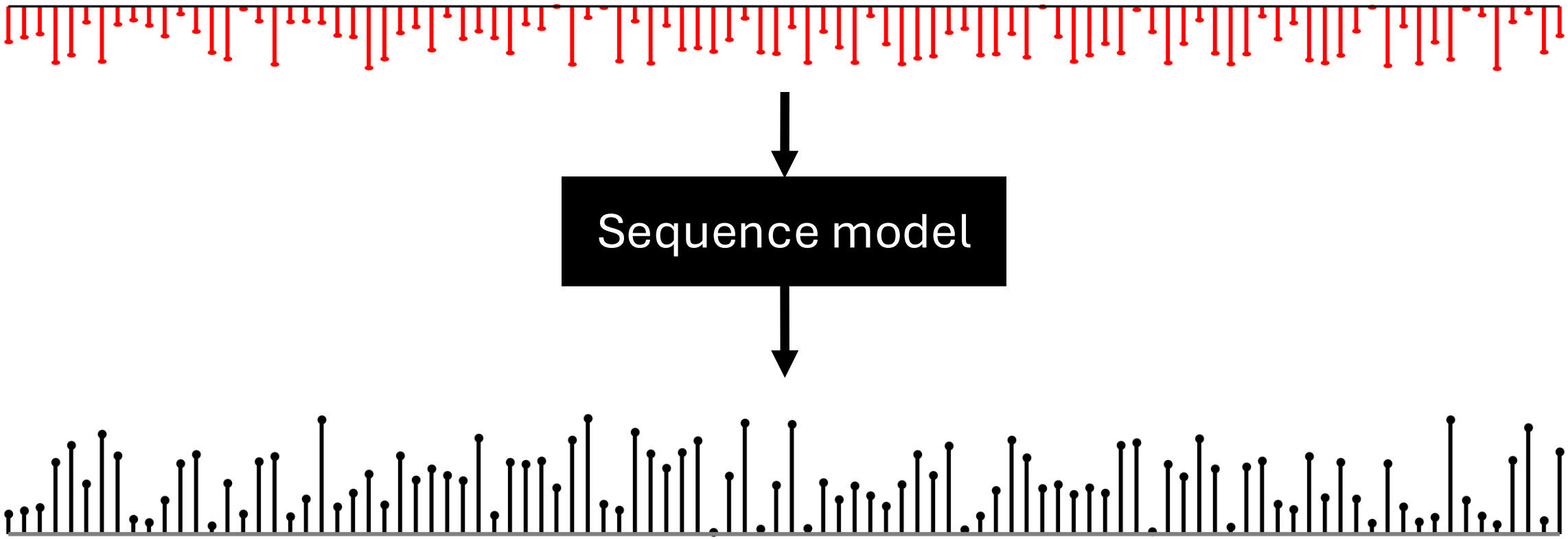
2025.01.22

khchao.com

@KuanHaoChao

Kuanhao-Chao

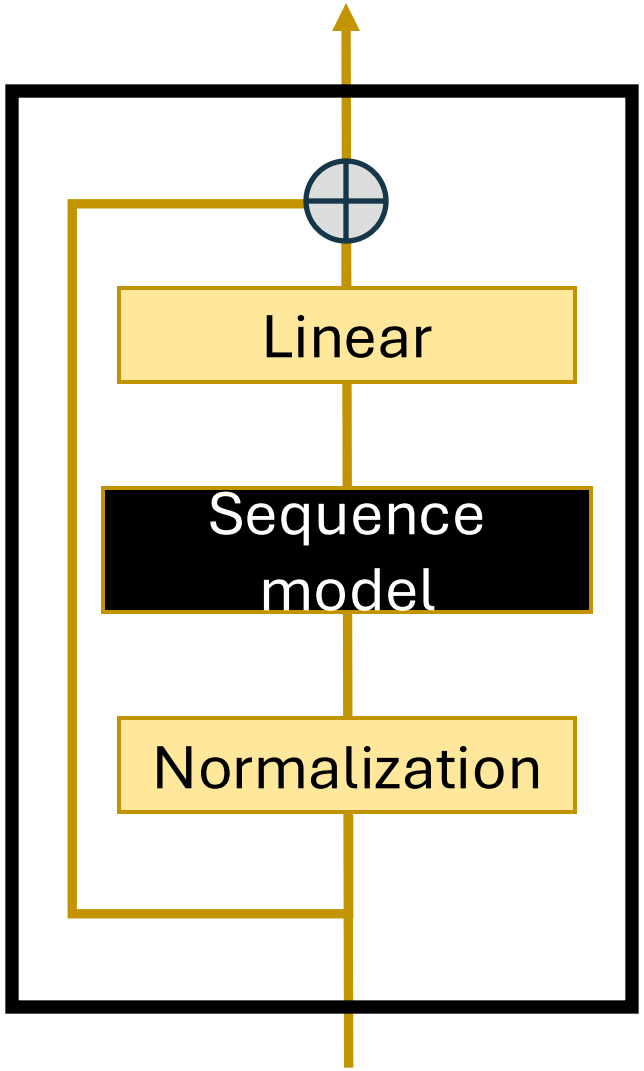
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s



Sequence models map a sequence to a sequence

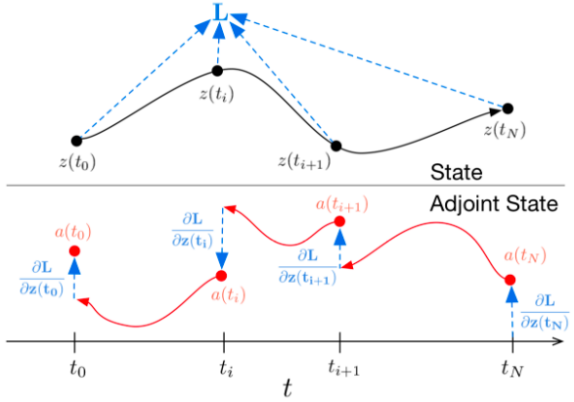
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

(batch, length, dim)

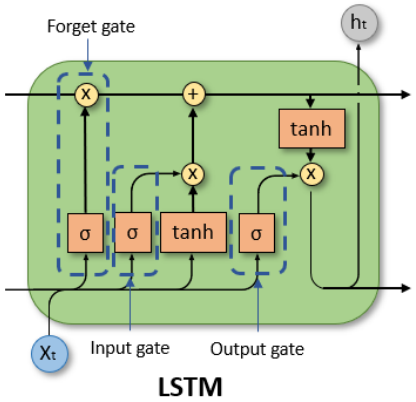


(batch, length, dim)

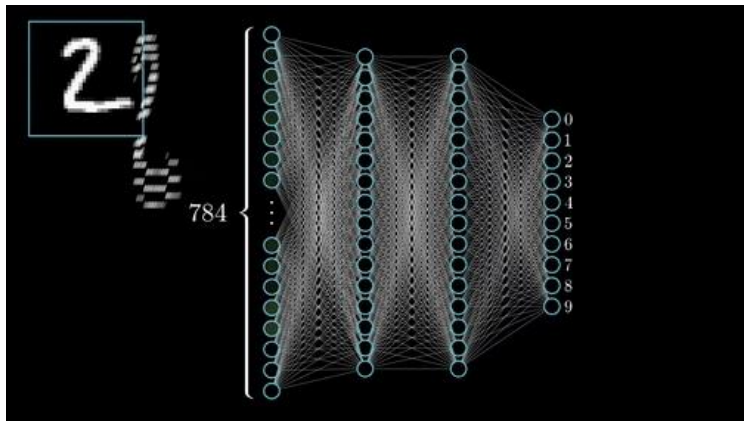
Neural ODEs



RNN



CNNs



Transformers

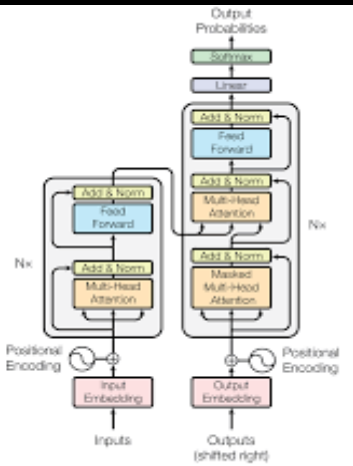
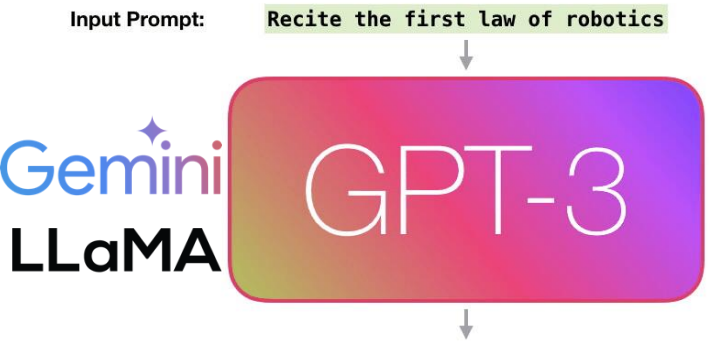
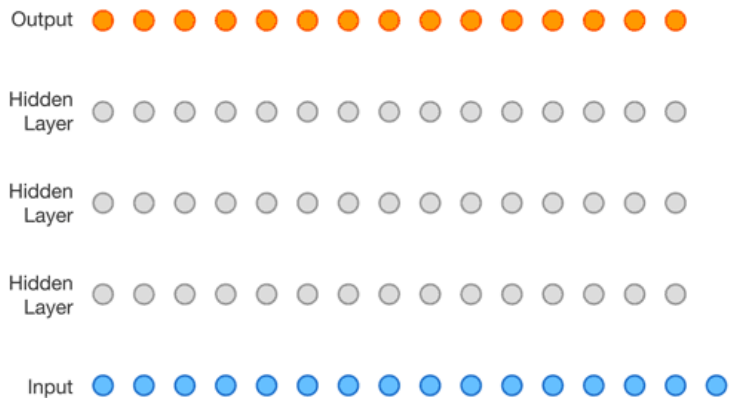


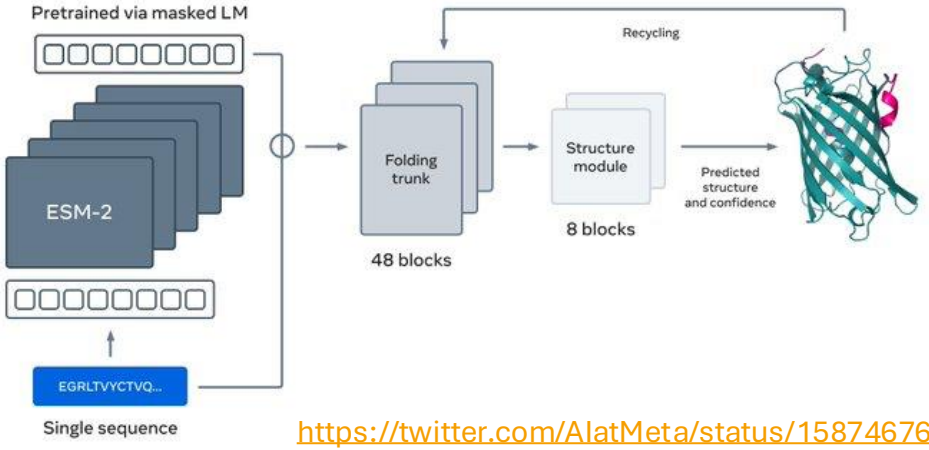
Figure 1: The Transformer - model architecture.





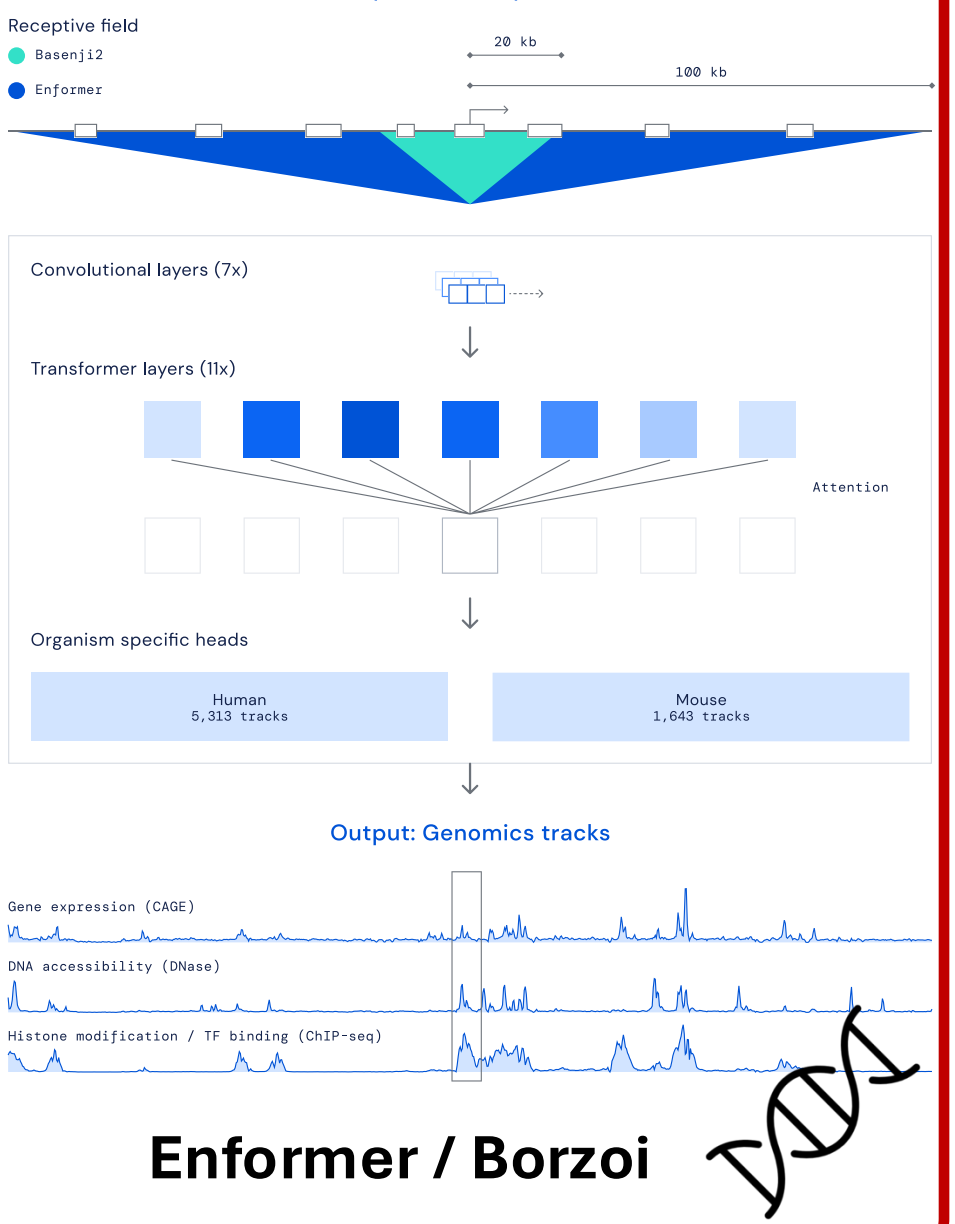
https://jalammar.github.io/how-gpt3-works-visualizations-animations/

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



ESMFold

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

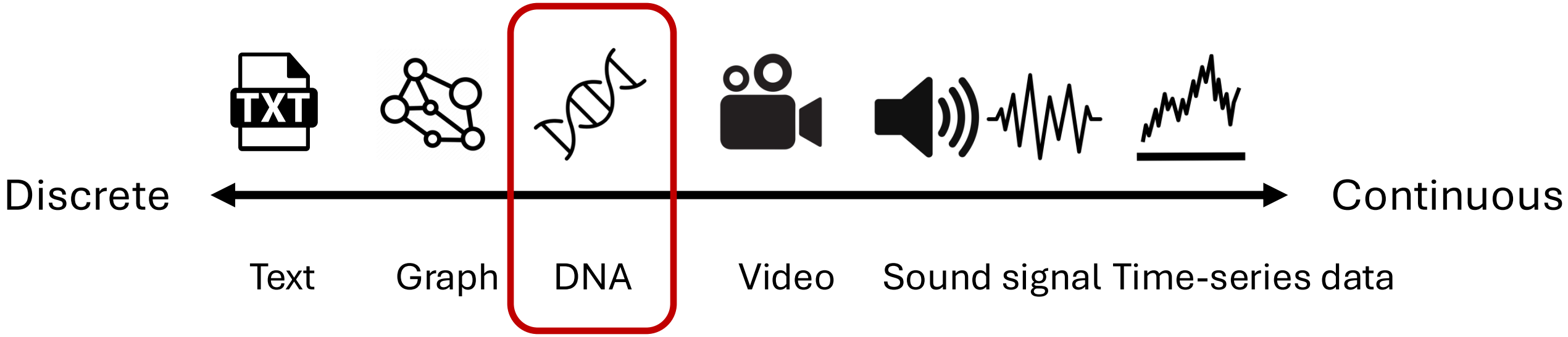


Enformer / Borzoi

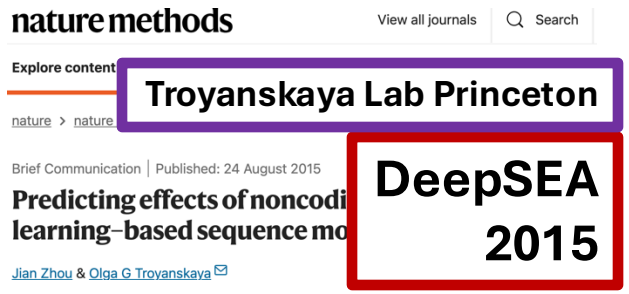
https://deepmind.google/discover/blog/predicting-gene-expression-with-a/

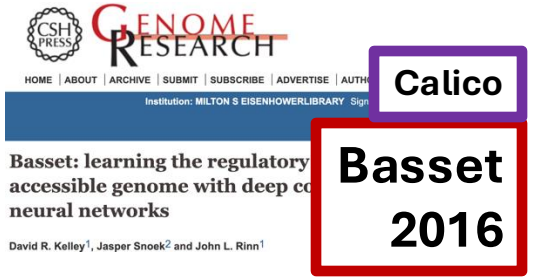
Reference: https://www.youtube.com/watch?v=luCBXCErkCs&t=197s

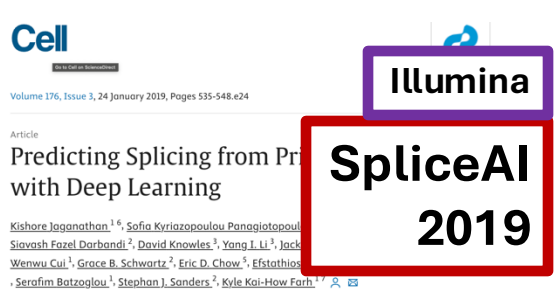
Spectrum of Sequential Data

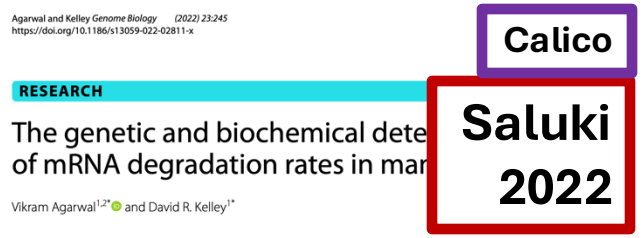


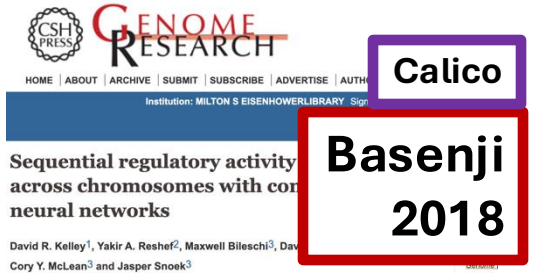
Deep learning-based DNA sequence model

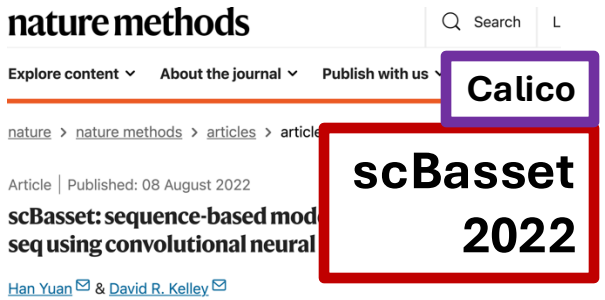


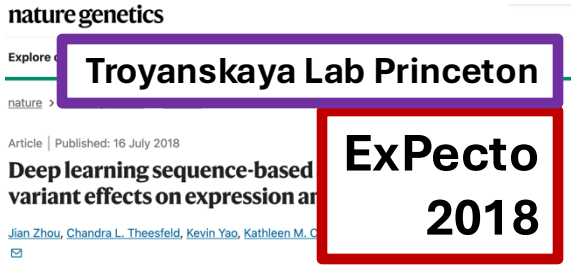


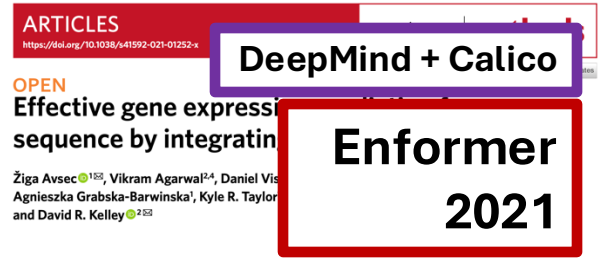


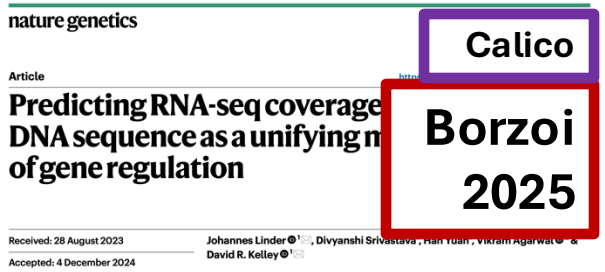








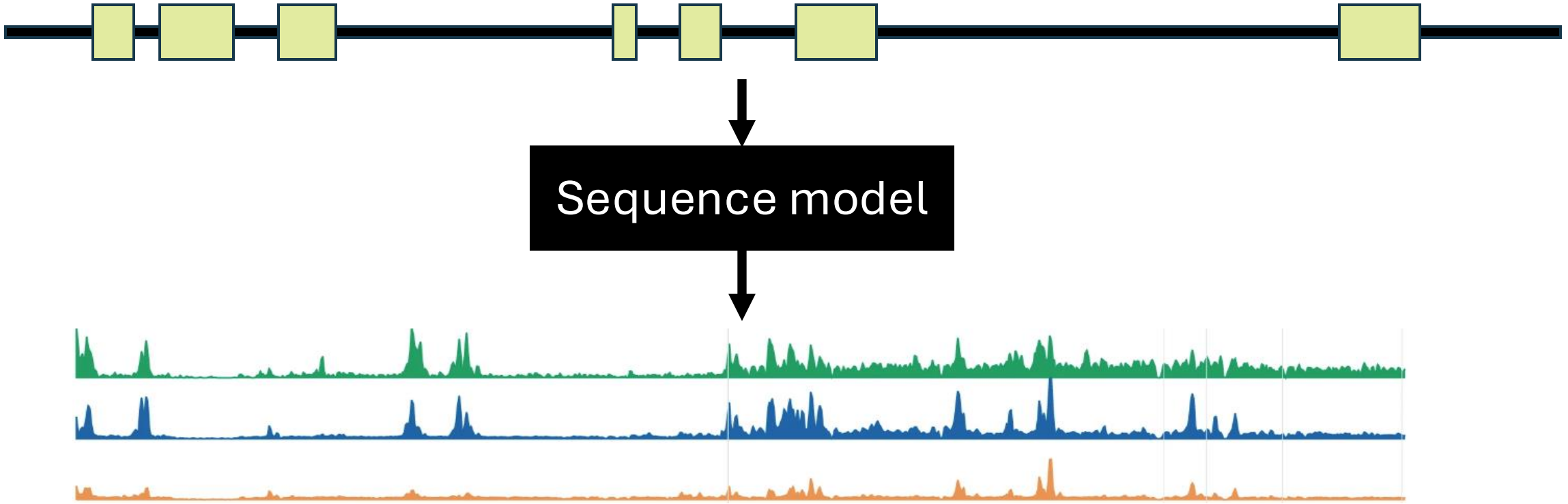




DeepMind + Calico	Calico
Enformer 2021	Borzoi 2025

Input: DNA sequences

$(N_{batch\_size} * L_{input} * 4)$



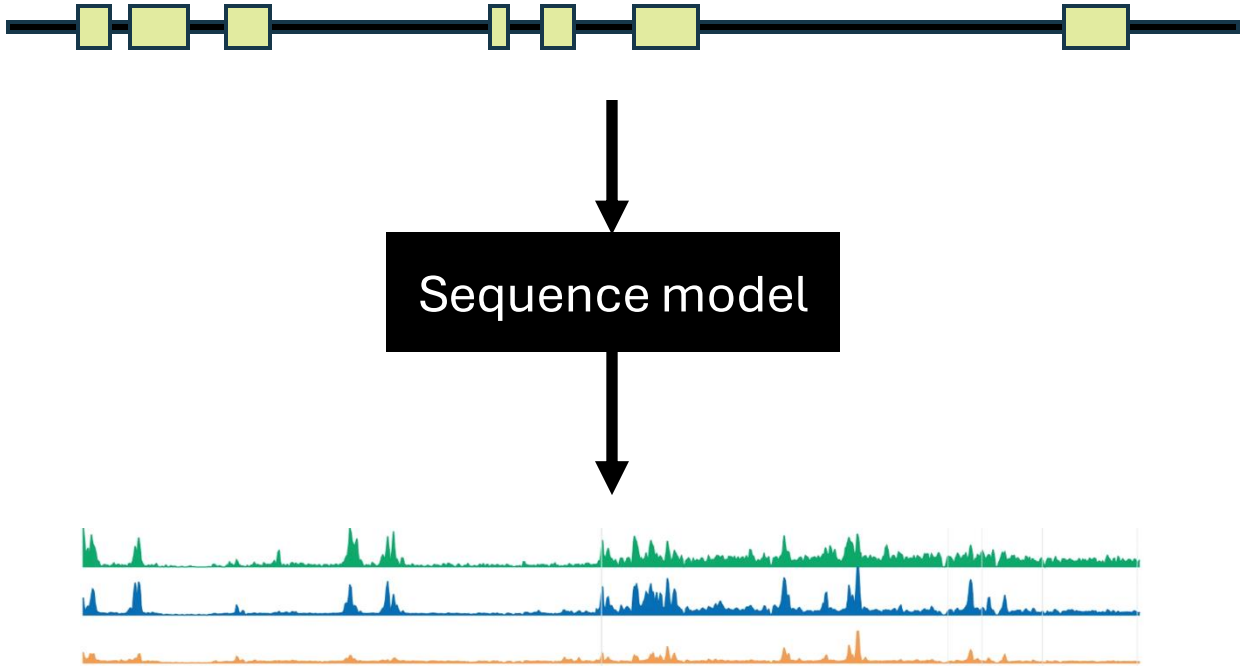
Output: Genomics tracks

$(N_{batch\_size} * L_{output} * T_{track\_number})$

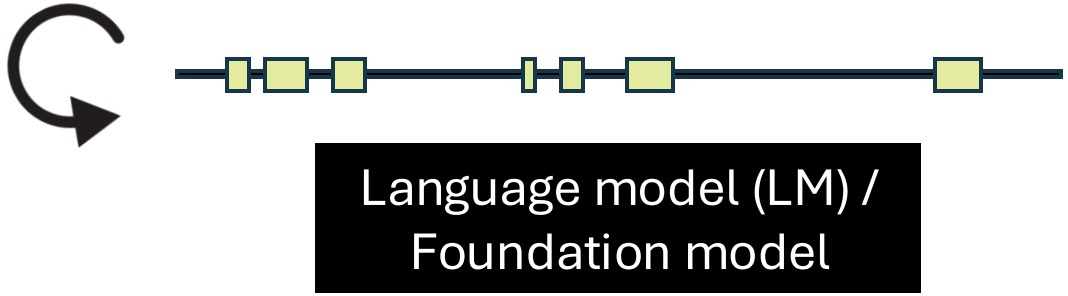
Supervised learning

Input: DNA sequences

Output: Genomics tracks



Stage 1 Self-supervised learning

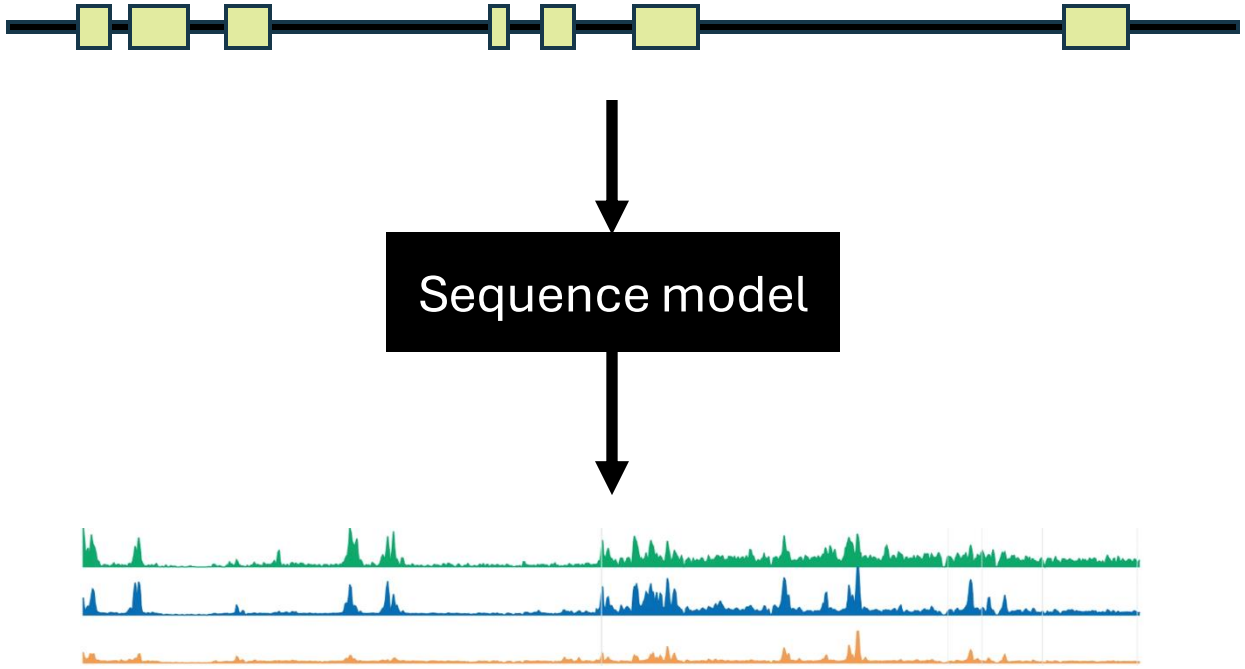


Stage 2 Fine-tuning LM

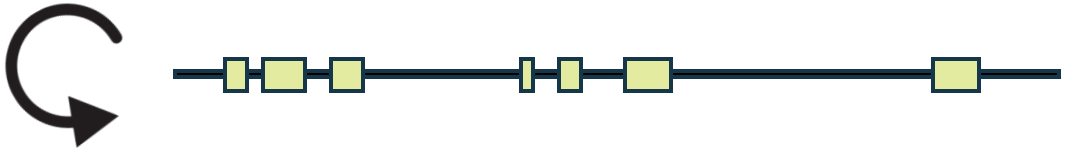
Enformer

Borzoi

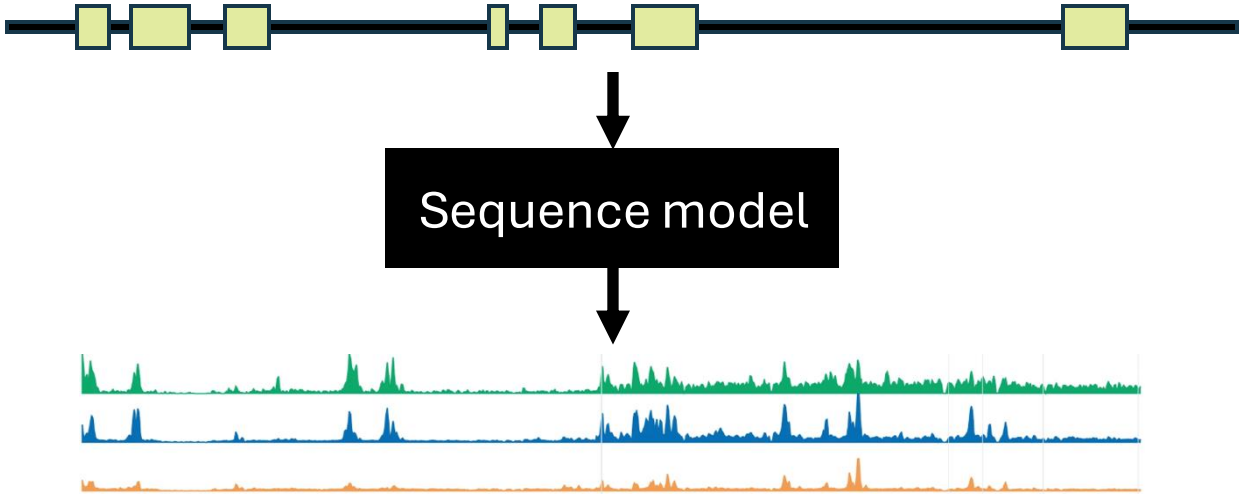
Supervised learning



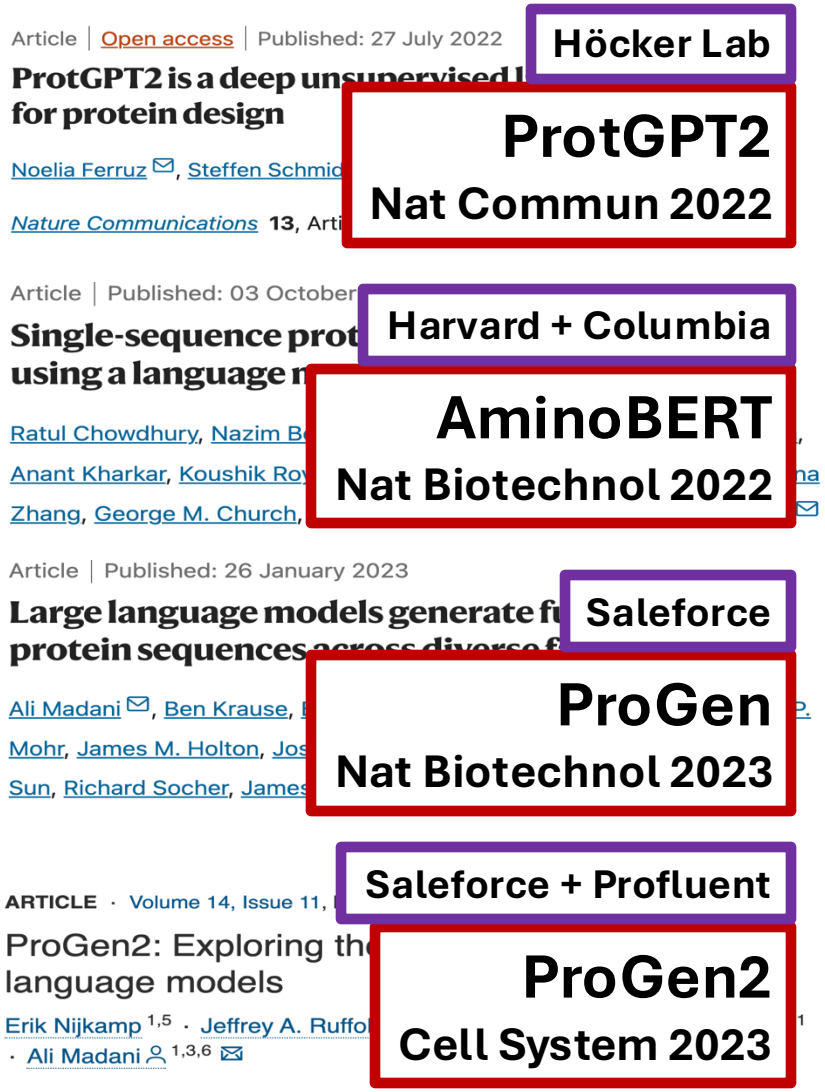
Stage 1 Self-supervised learning

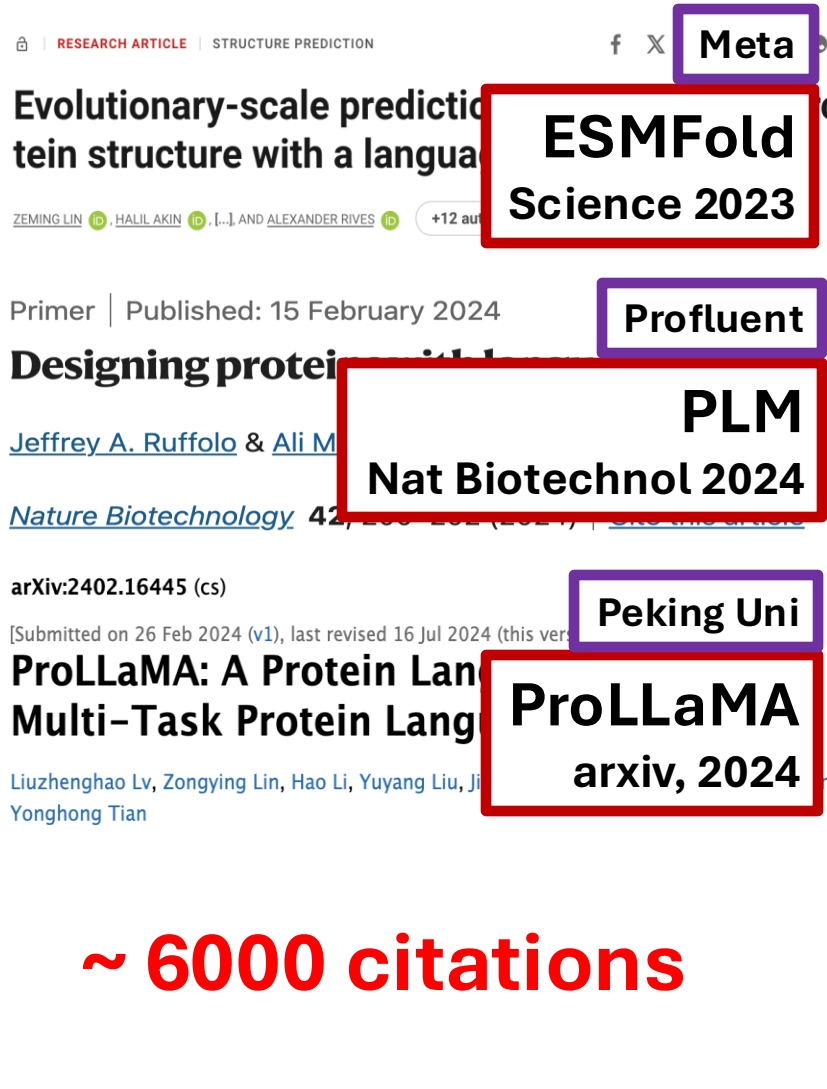


Stage 2 Fine-tuning LM



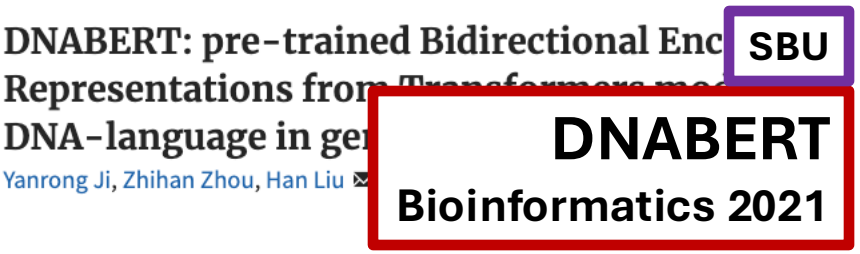
Protein Language model





~ 6000 citations

DNA Language model



SBU

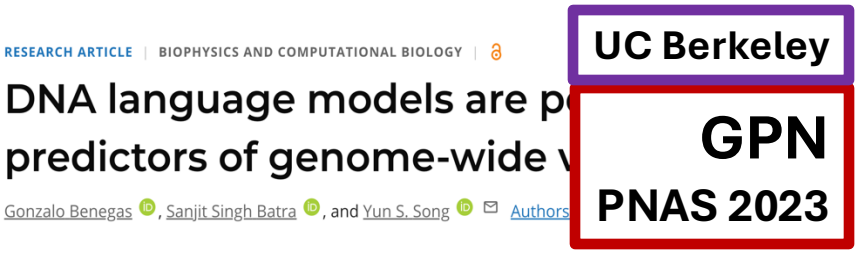
DNABERT
Bioinformatics 2021

SBU + NU

DNABERT-2
ICLR 2024

Stanford + Arc Inst + TogetherAI

Evo
bioRxiv 2024

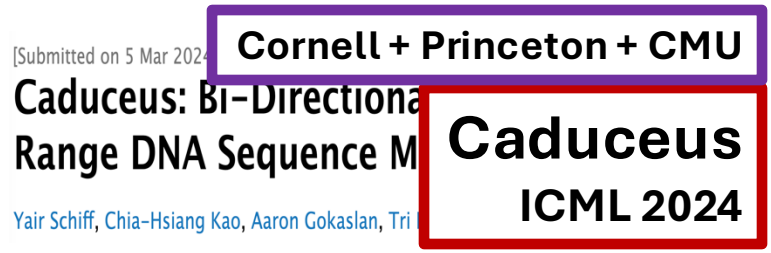


UC Berkeley

GPN
PNAS 2023

TUD

GROVER
Nat Mach Intell 2024



Cornell + Princeton + CMU

Caduceus
ICML 2024

InstaDeep + Nvidia + TUM

Nucleotide Transformer
bioRxiv 2023

Harvard + MIT

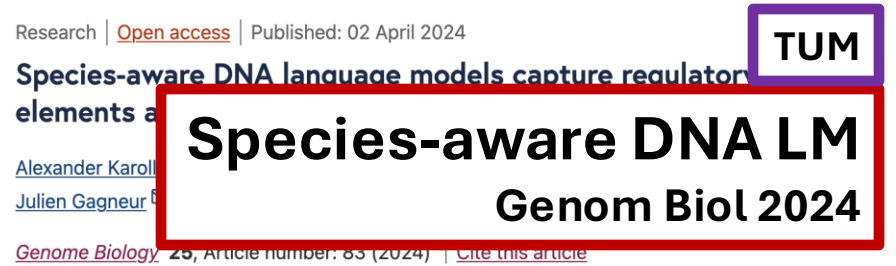
Genomic LM
Nat Commun 2024

Cornell + USDA-ARS + Simons

PlantCaduceus
bioRxiv 2024

Stanford

HyenaDNA
NeurIPS 2023



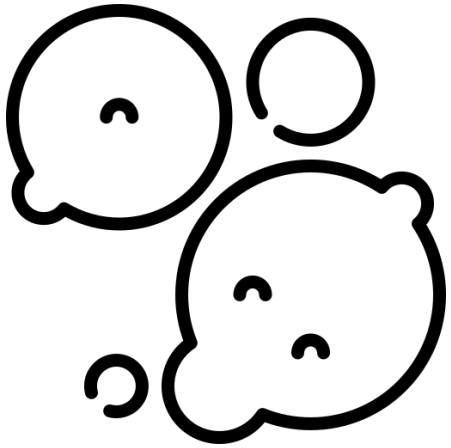
TUM

Species-aware DNA LM
Genom Biol 2024

TUM

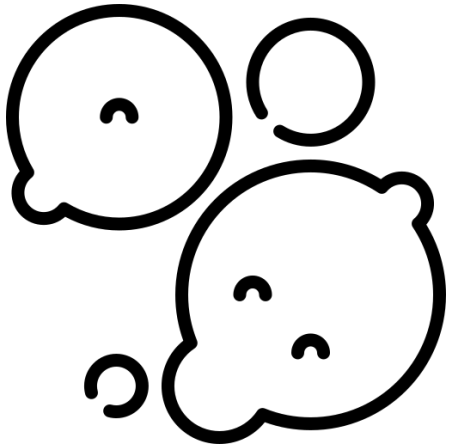
Nucleotide dependency
bioRxiv 2024

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** ?

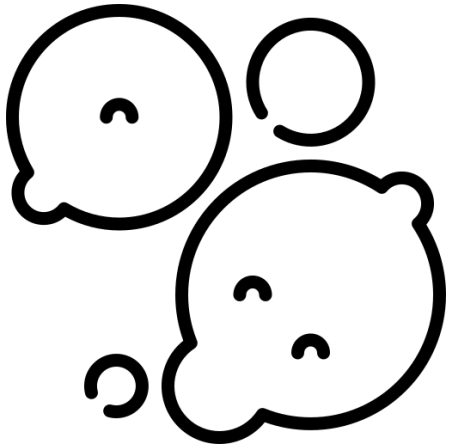


- Computation is expensive for human (human: 3B nt; yeast: 12 M nt)

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

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

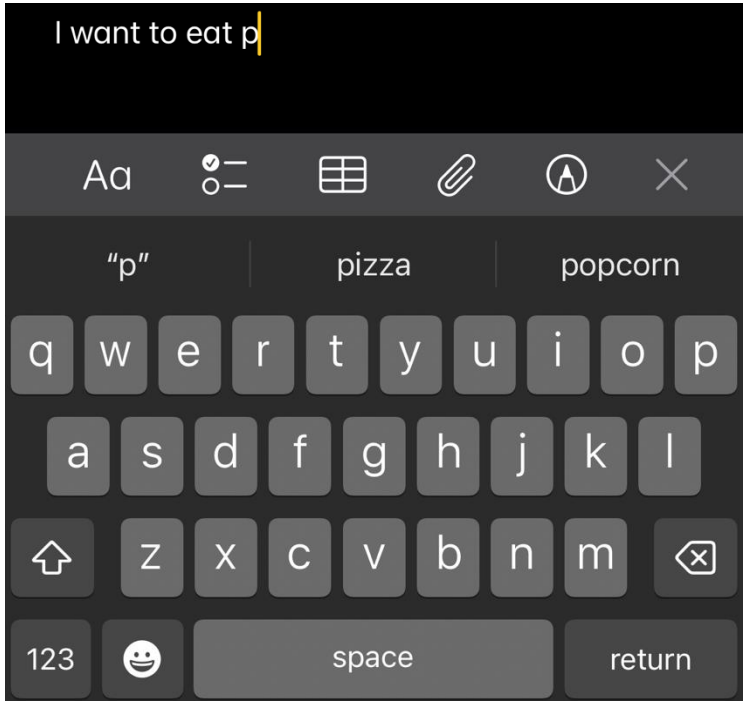
Why **yeast** ?



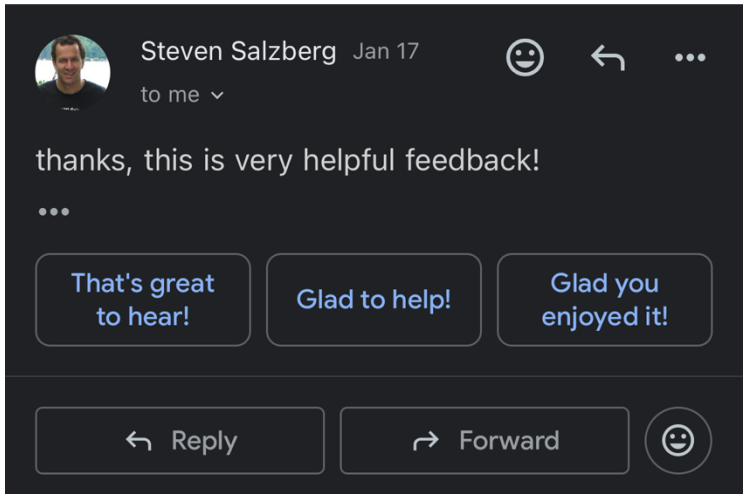
- Computation is expensive for human (human: 3B nt; yeast: 12 M nt)
- In human, we can't do large scale TF perturbation study
- Yeast is a great model organism to generate data & training models
  - Simple Eukaryotic Model: cost-effectiveness and scalability
  - Rapid Growth and Easy Culturing and Quick 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

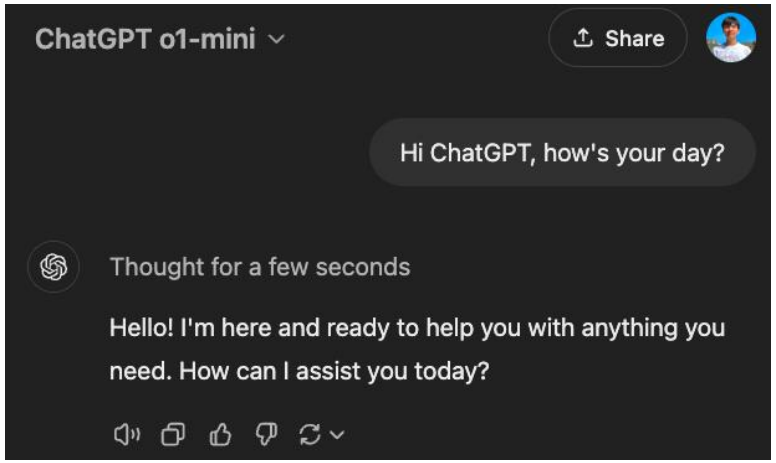
Next word prediction



Email suggested reply



ChatGPT



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

$$\mathbf{P}(\ \underbrace{X_t}\ \mid\ \underbrace{X_1, X_2, \ldots, X_{t-1}}\ )$$

Next word **Context or prefix**

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

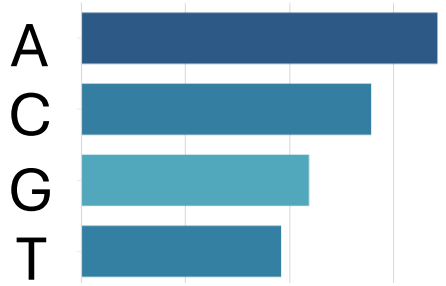
- Directly we train models on "**marginals**"
- We are implicitly learning the **full/joint distribution** of language.

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

The chain rule:

$$\mathbf{P}(X_1, \ldots, X_t) = \mathrm{P}(X_1) \prod_{i=1}^{t} \mathrm{P}(X_i \mid X_1, X_2 \ldots, X_i)$$

"A C T T A C T A G A [MASK]" → DNA Language Model →



Language Modeling ≙ learning prob distribution over language sequence.

Part I

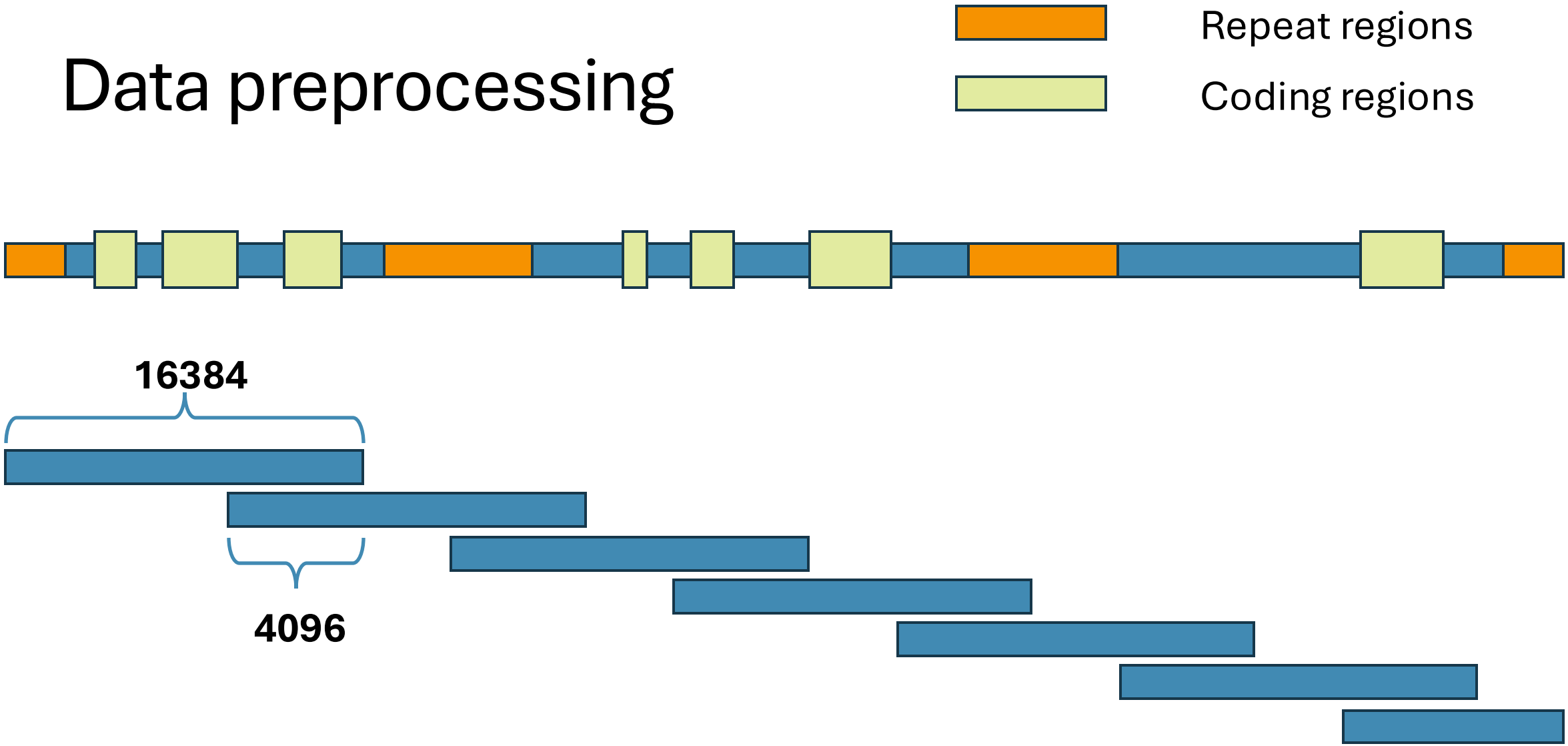
Fungal Language Model

Summary

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

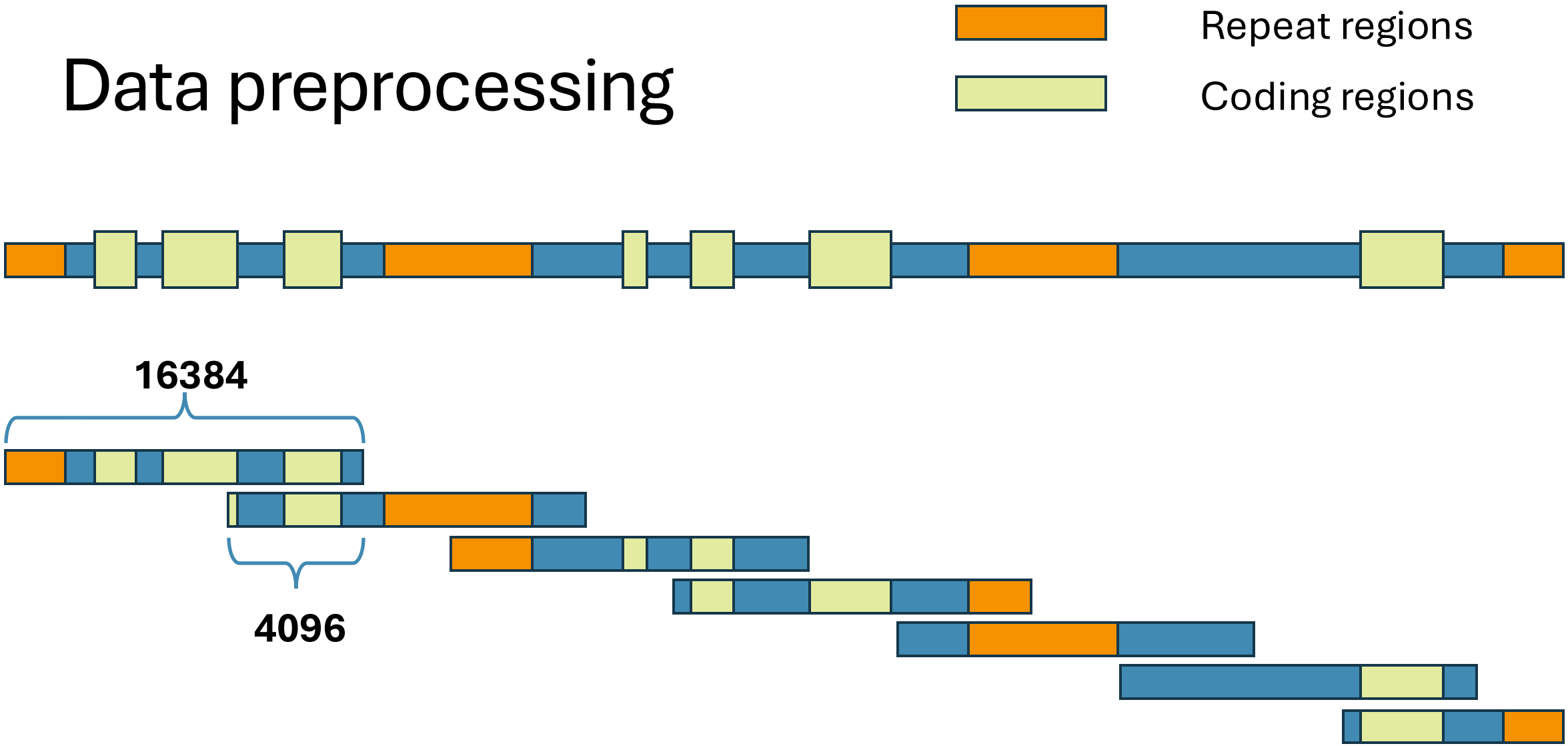
Data preprocessing

Data preprocessing



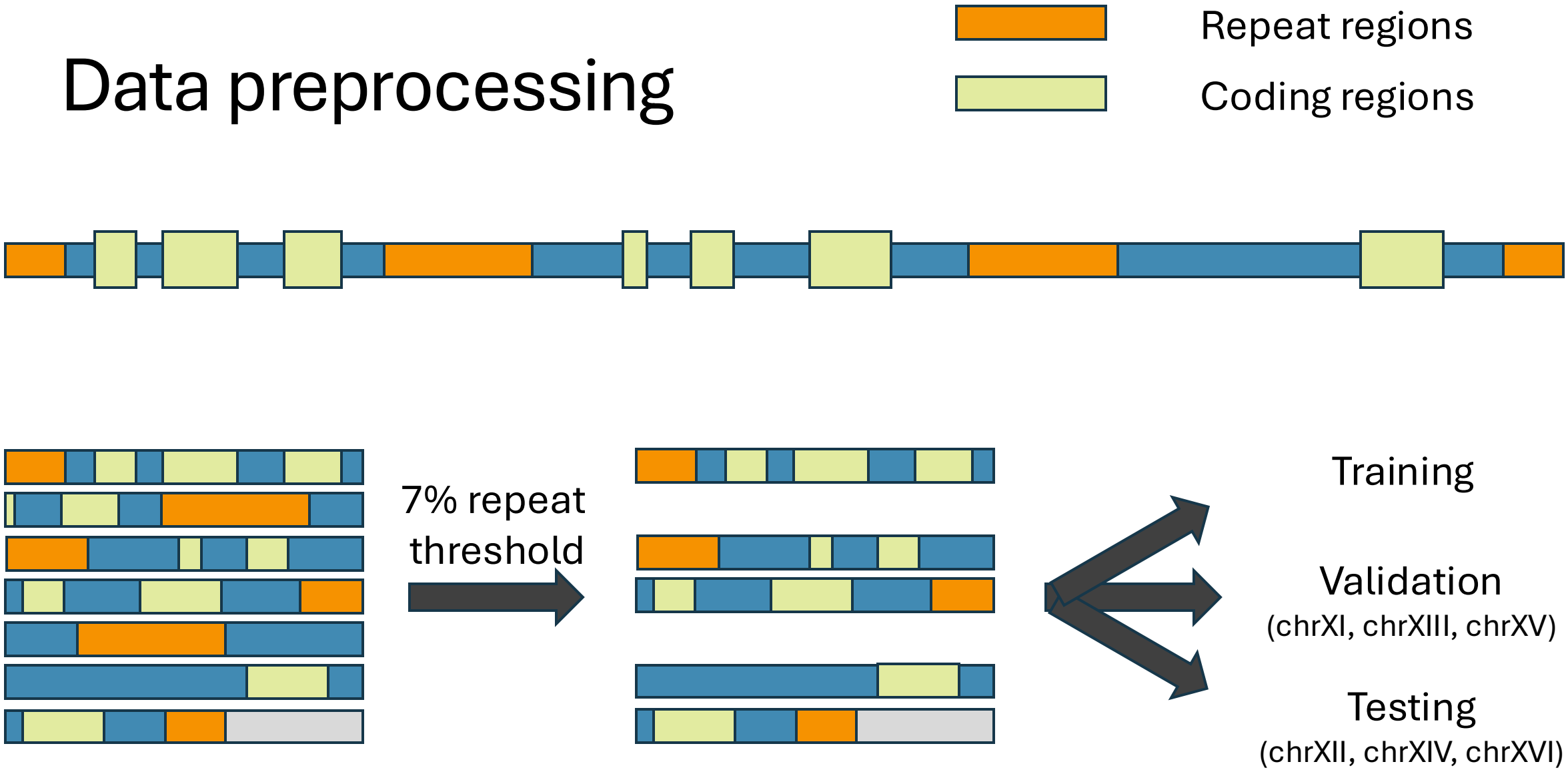
~ 8 genes per window

Data preprocessing



~ 8 genes per window

Data preprocessing



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.

Q1: To what evolutionary distance should we include in our LM?

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?

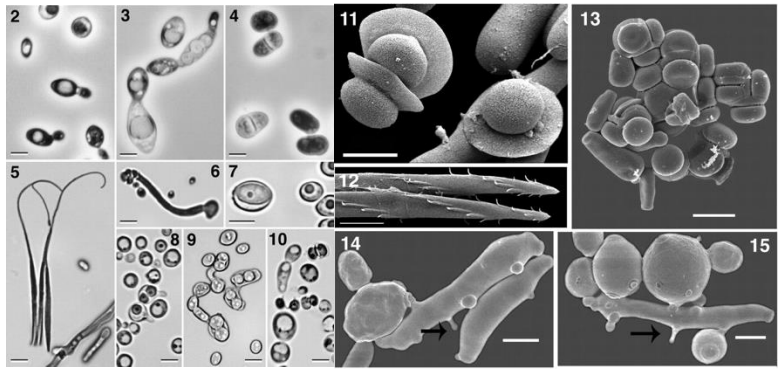
Q3: Even more diverse?

R64 reference yeast

80 strains of yeasts

165 Saccharomycetales

1361 Fungus genomes

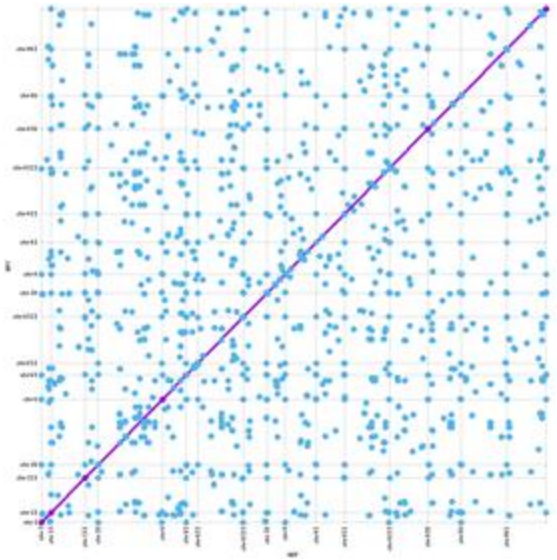


Suh, S. O., Blackwell, M., Kurtzman, C. P., & Lachance, M. A. (2006). Phylogenetics of Saccharomycetales, the ascomycete yeasts. *Mycologia*, *98*(6), 1006-1017.

Genome distance evaluation

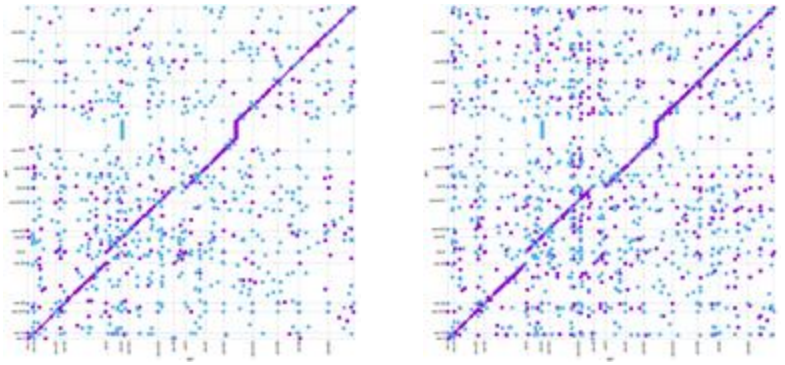
R64 Reference Yeast

Mummerplot

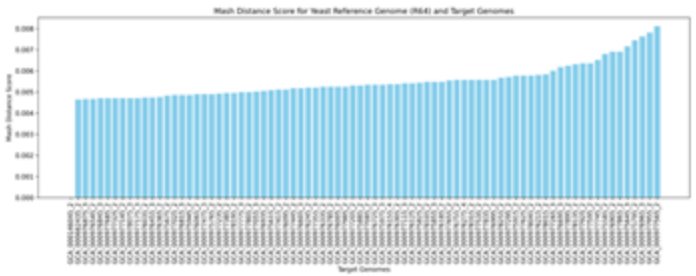


80 strains of yeasts

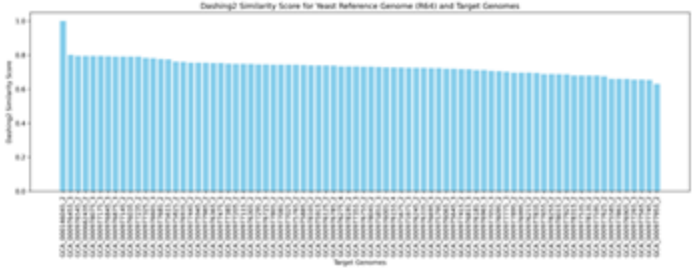
Mummerplot



Mash distance



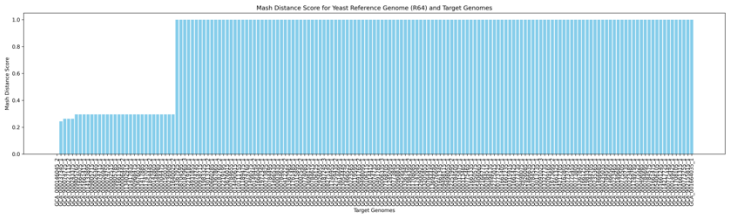
Dashing2 similarity



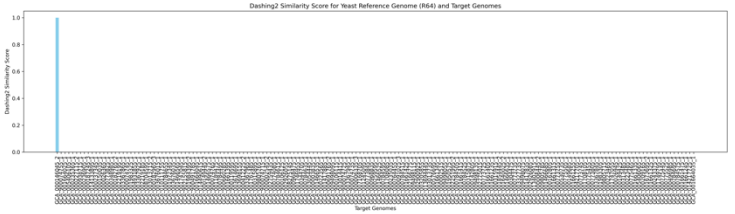
✅165 Saccharomycetales

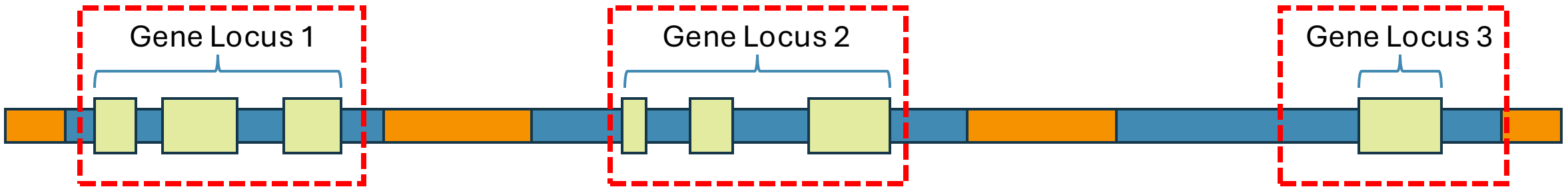
Mummerplot

Mash distance



Dashing2 similarity





Q2: How many genes per window?

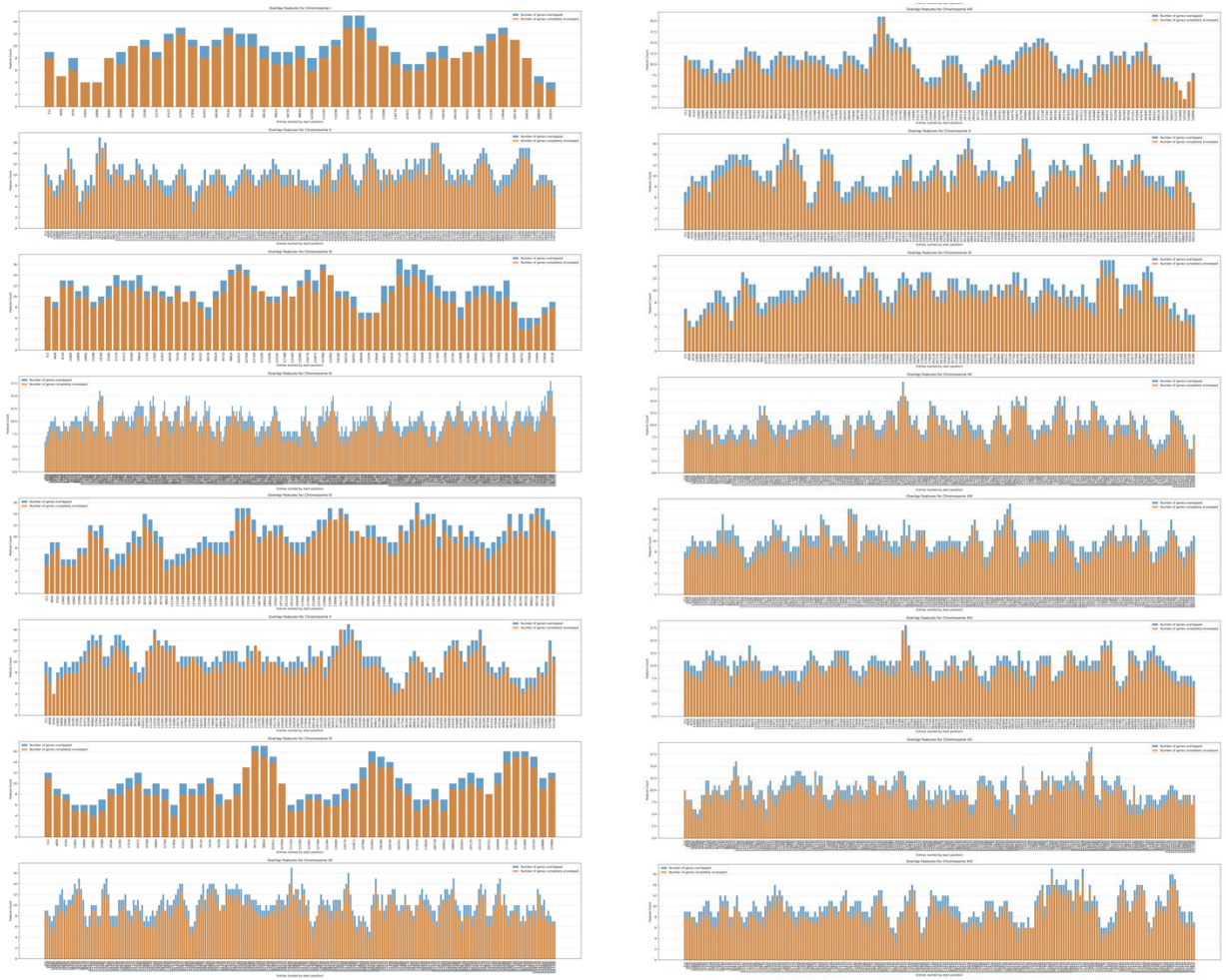
Q3: What is the quality of the annotation?

Q4: What is the protein-coding region ratio?

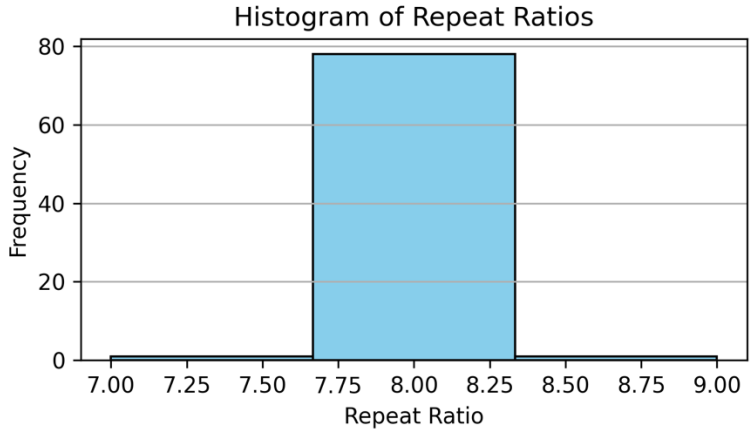
Genome evaluation – # genes per 16K window

R64 Reference Yeast

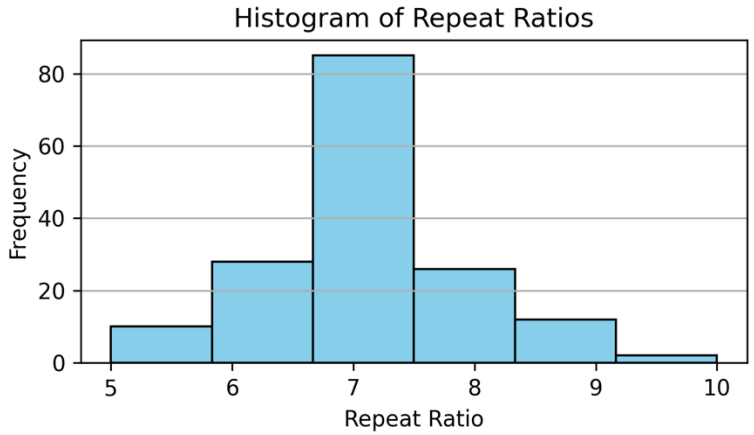
Median: 9.0; Mean: 8.98



80 strains of yeasts

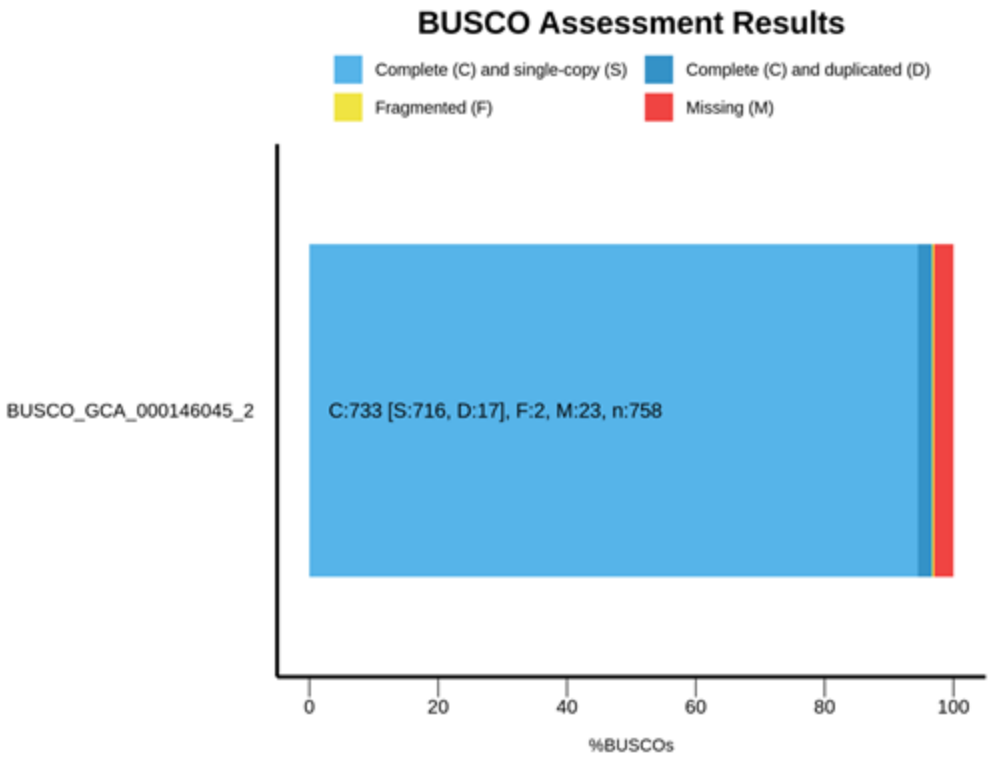


165 Sachramonycetales

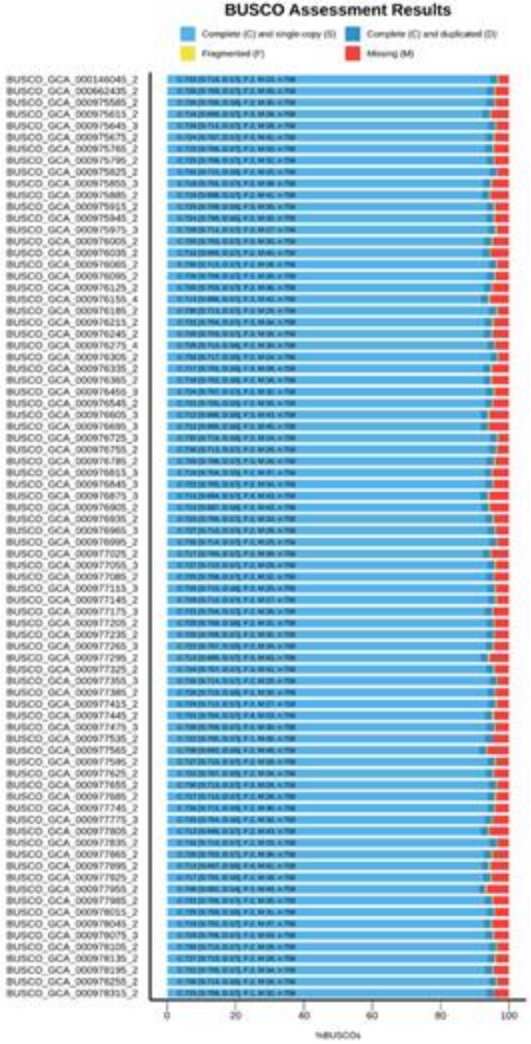


Genome annotation completeness evaluation

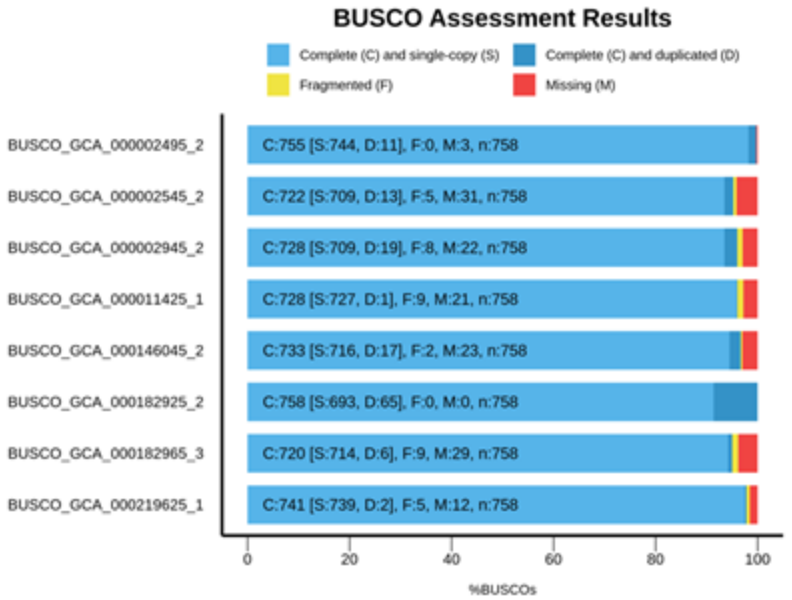
R64 Reference Yeast



80 strains of yeasts



165 Sachramonycetales

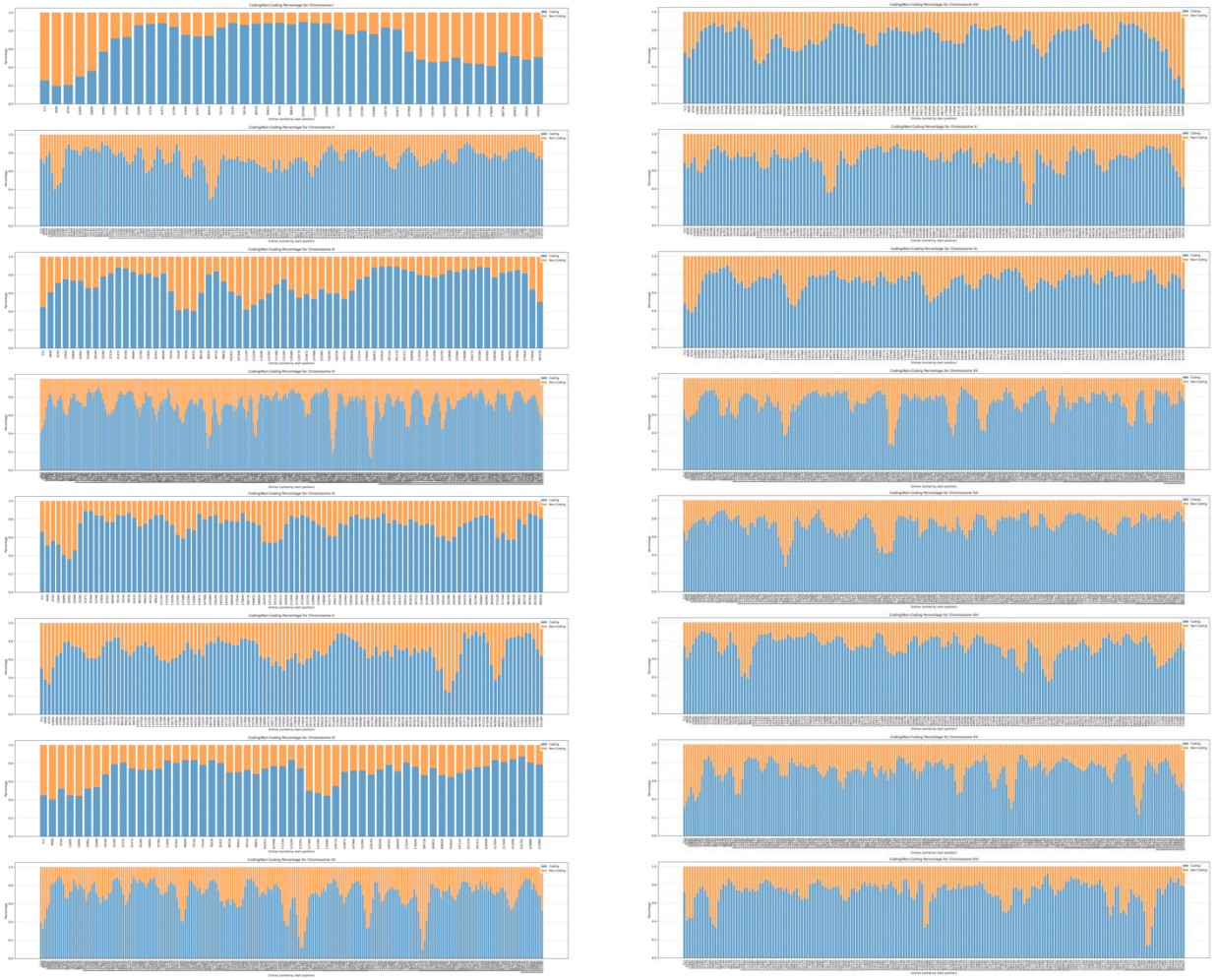


Conclusion:
BUSCO ~95% completeness

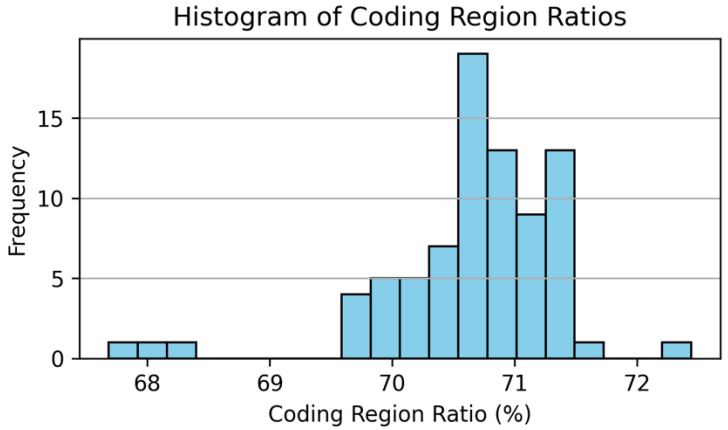
Genome evaluation – coding / noncoding regions

R64 Reference Yeast

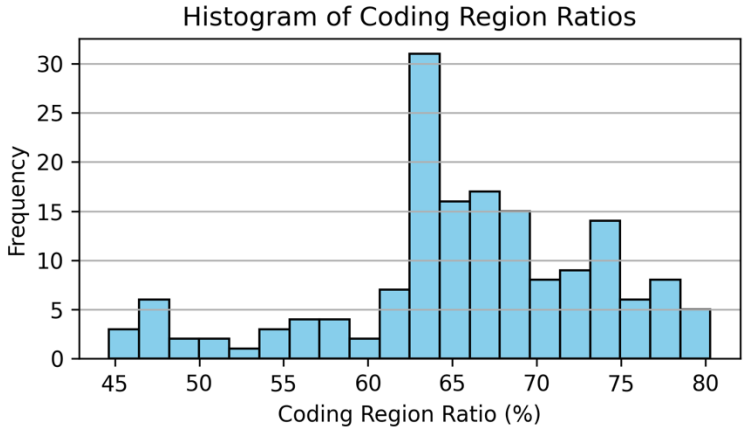
72.46% coding regions

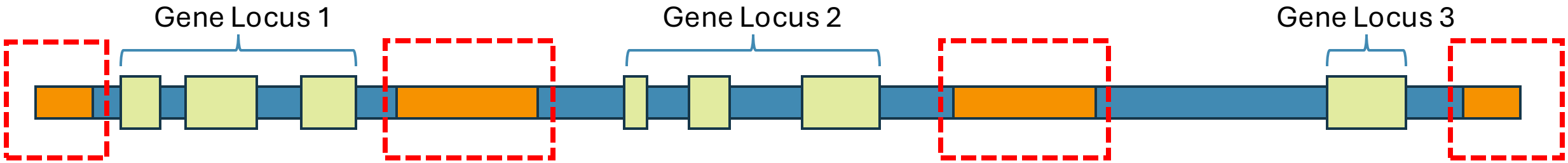


80 strains of yeasts

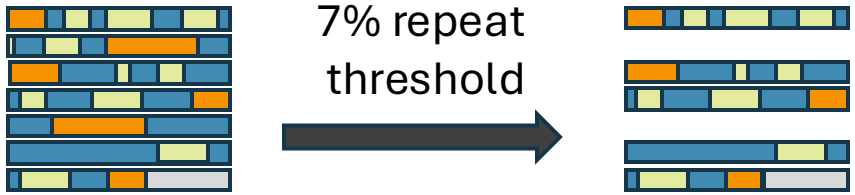


165 Sachramonycetales





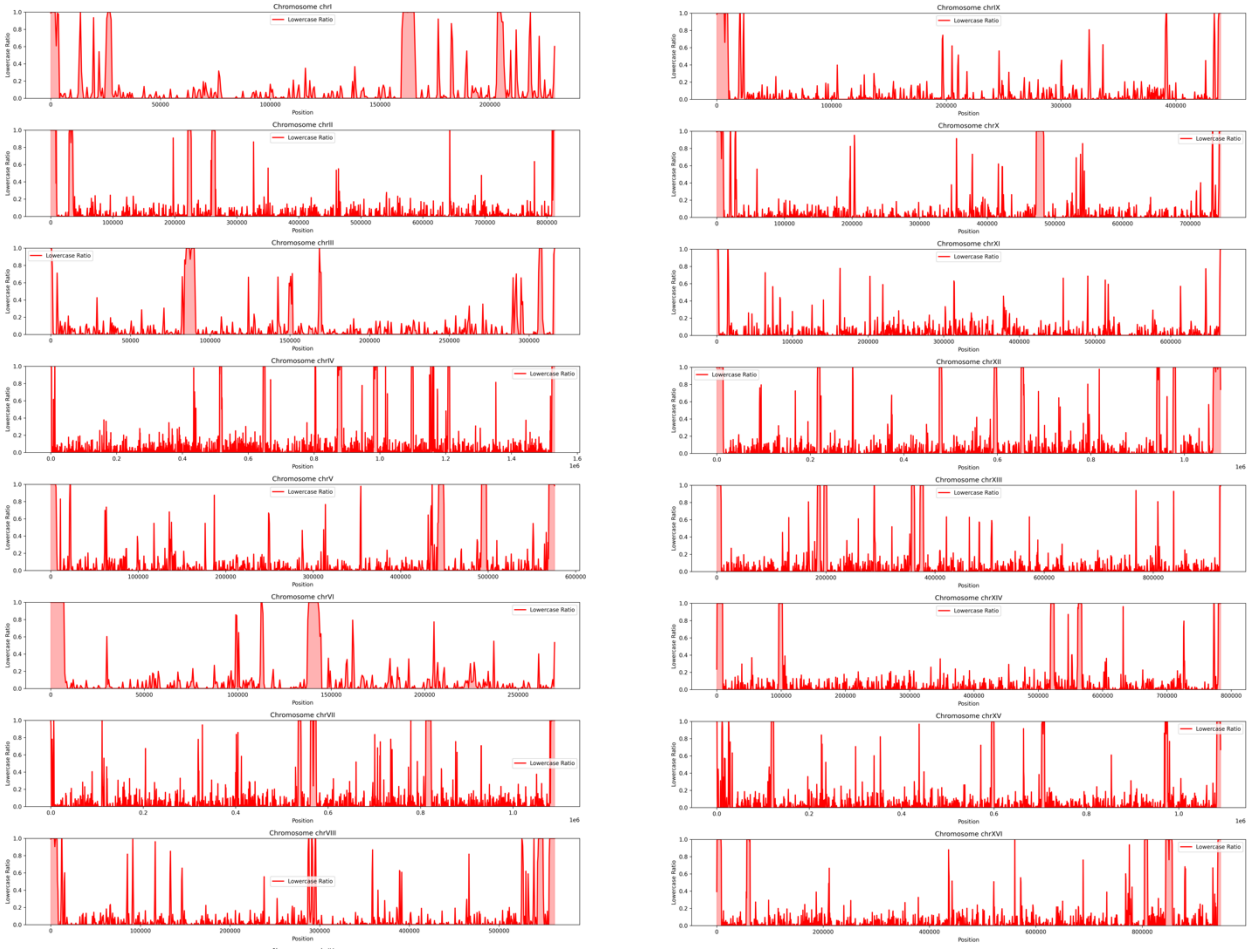
Q5: How repetitive are the genomes?



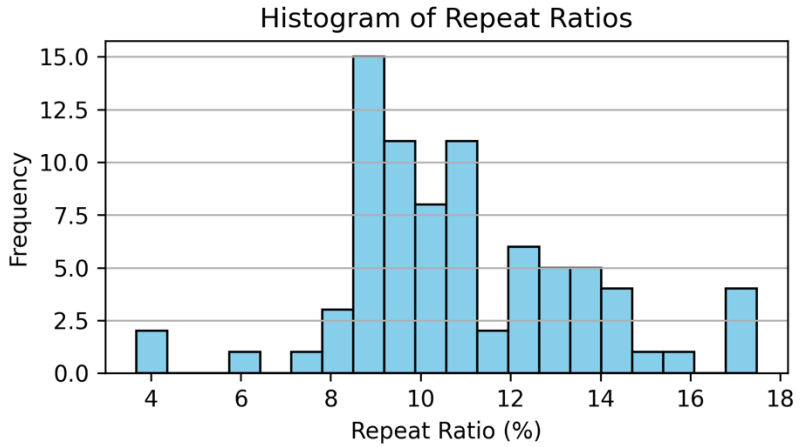
Genome evaluation – repeat regions

R64 Reference Yeast

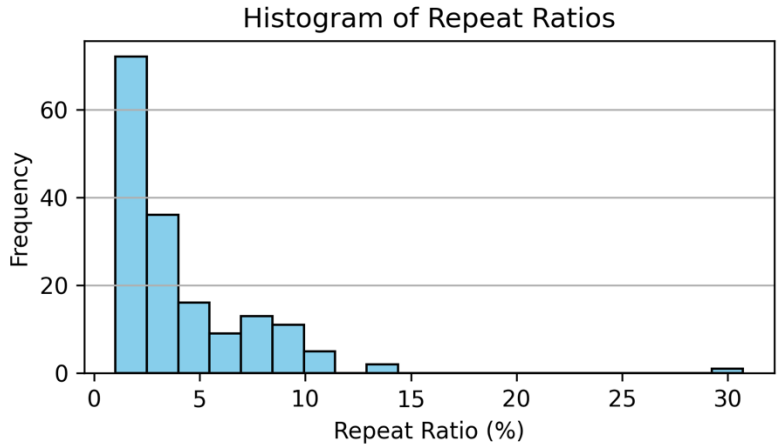
7.39% repeat regions

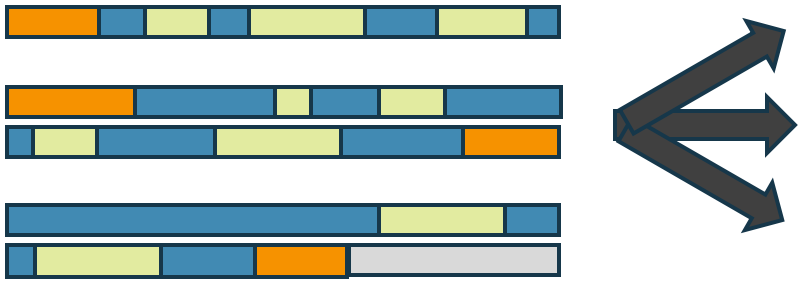


80 strains of yeasts



165 Sachramonvcetales



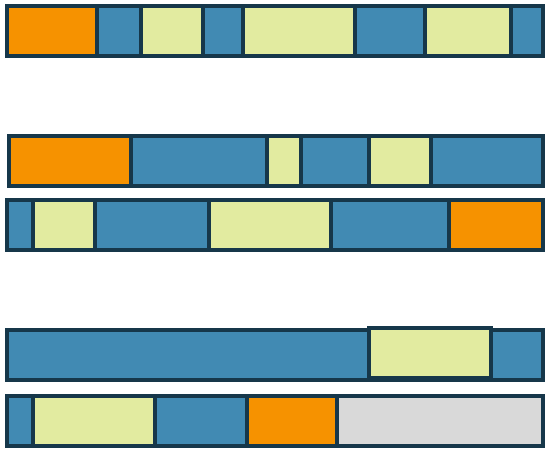


Training

Validation (chrXI, chrXIII, chrXV)

Testing (chrXII, chrXIV, chrXVI)

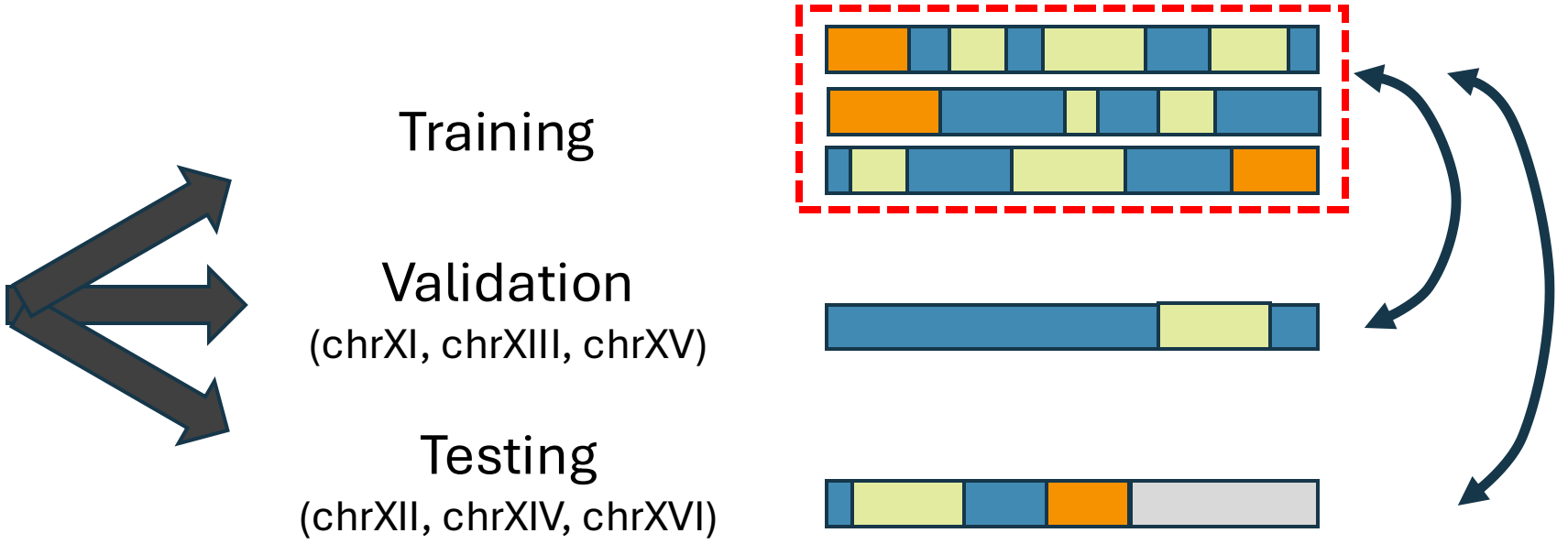
Q6: How many homologous sequences are there between training and testing?



Training

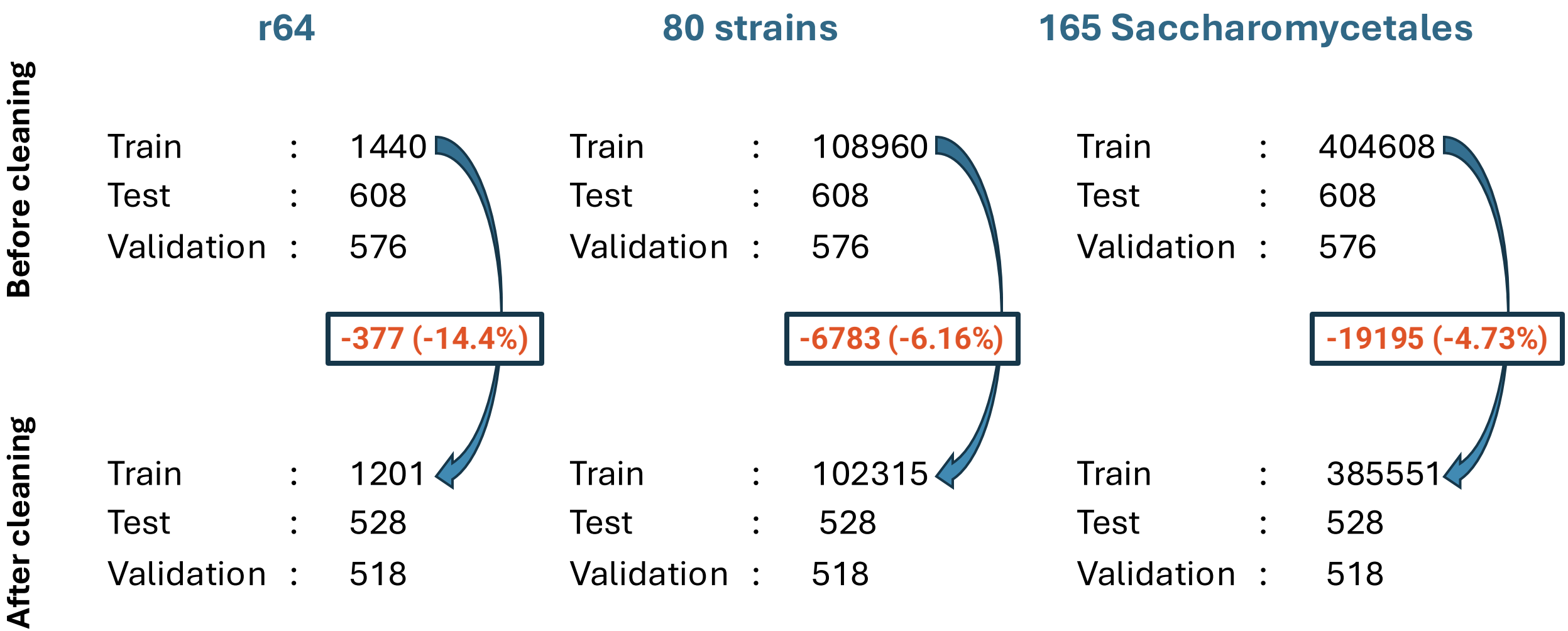
Validation
(chrXI, chrXIII, chrXV)

Testing
(chrXII, chrXIV, chrXVI)



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

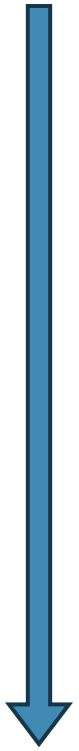
Final sequence for training / testing / validation



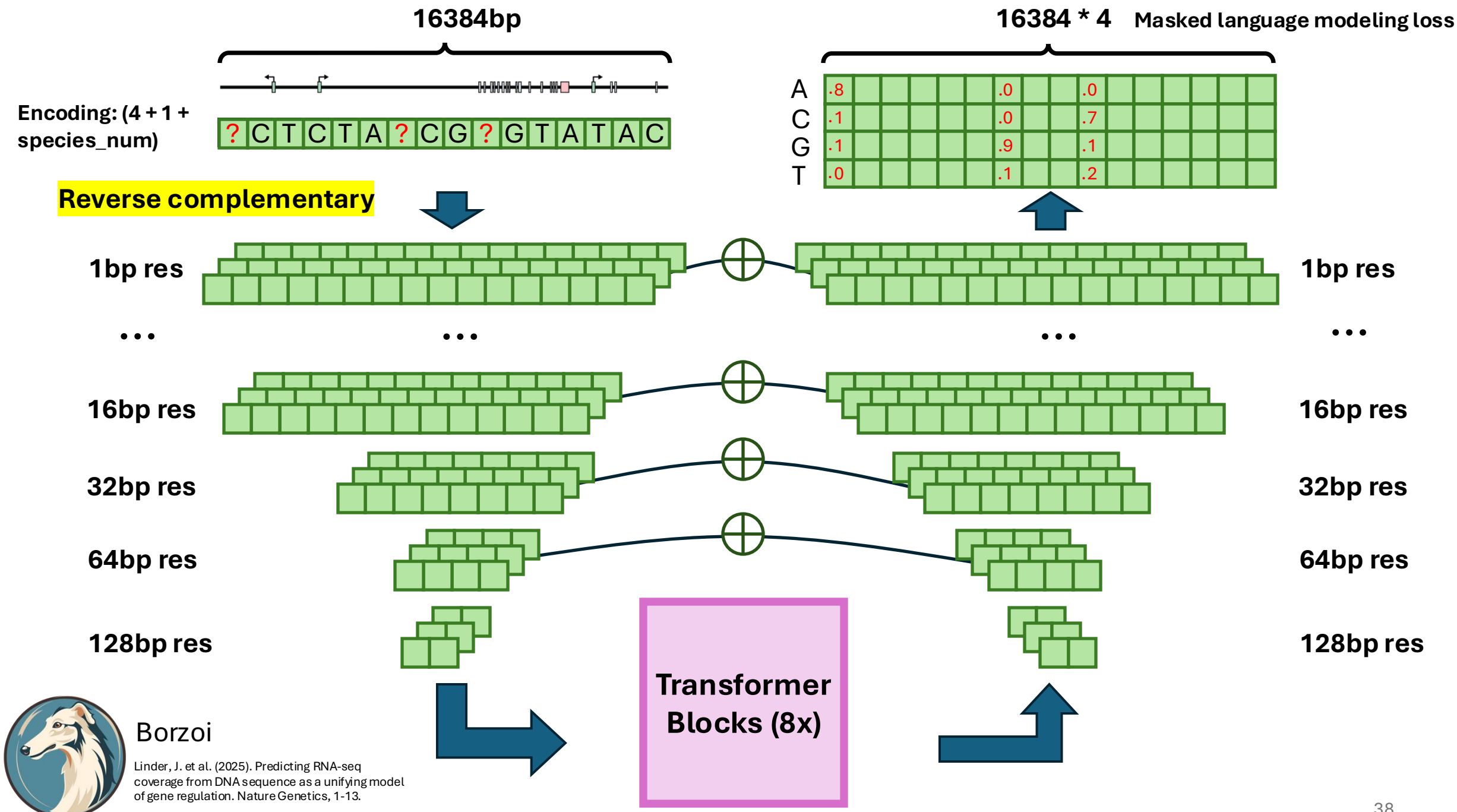
Fungal Language Model Architecture

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

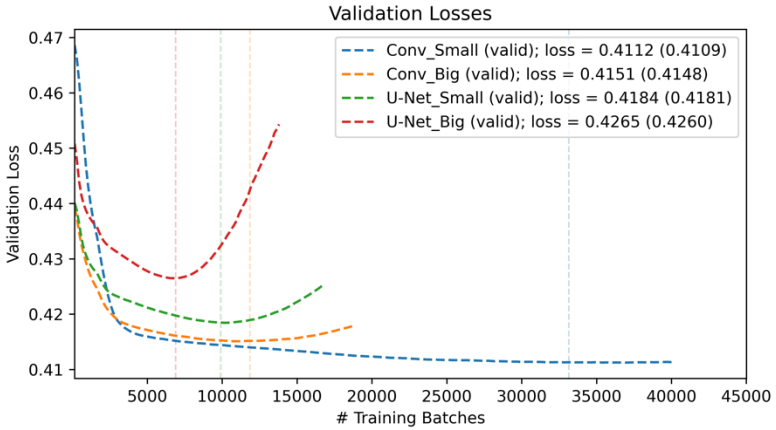

16384bp
16384 * 4
Masked language modeling loss
Encoding: (4 + 1 + species_num)
? C T C T A ? C G ? G T A T A C
A .8 .0 .0
C .1 .0 .7
G .1 .9 .1
T .0 .1 .2
Reverse complementary
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)
Borzoi
Linder, J. et al. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics, 1-13.

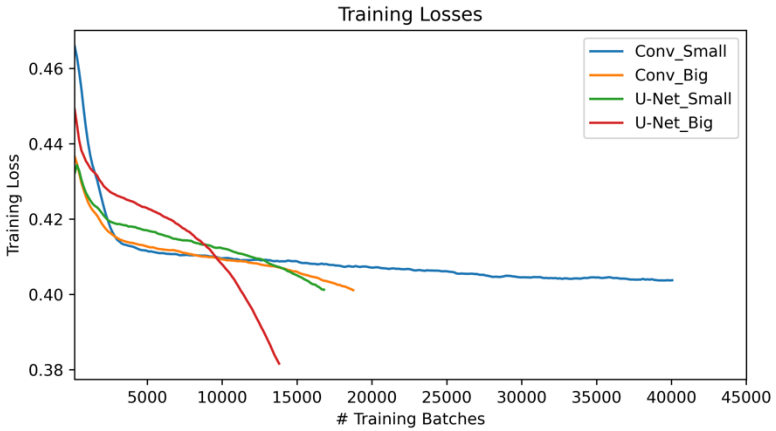
Fungal Language Model

Self-supervised training Results

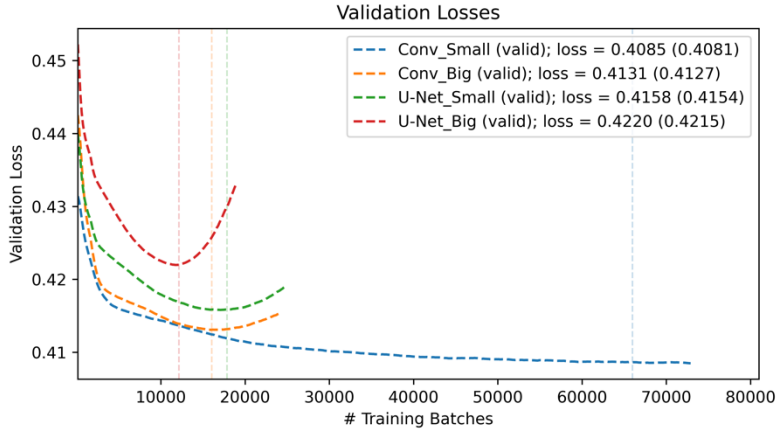
Model comparison

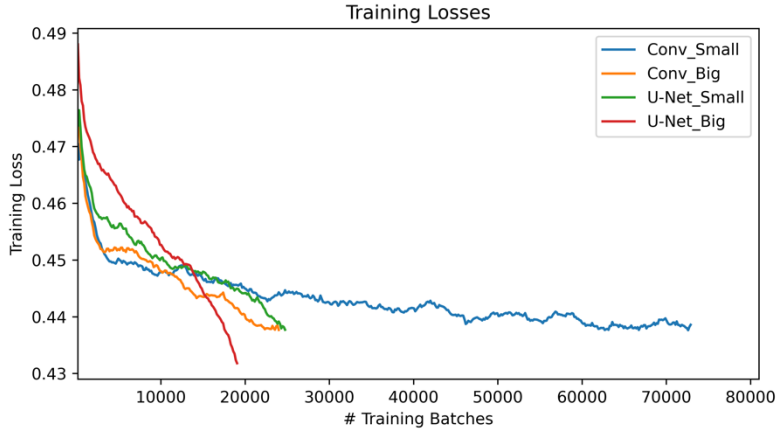
r64



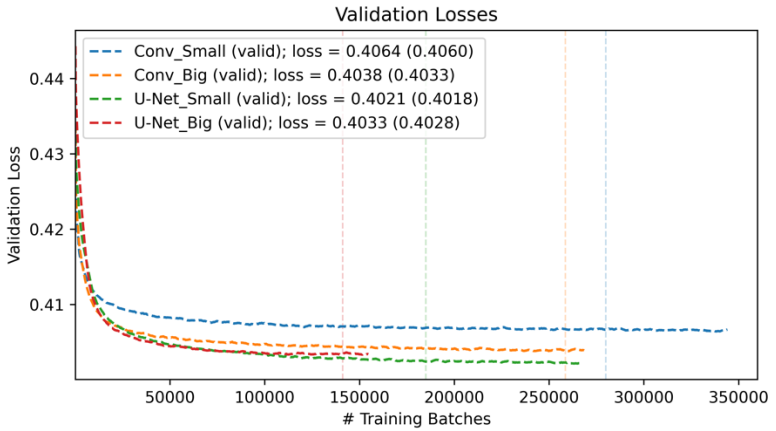


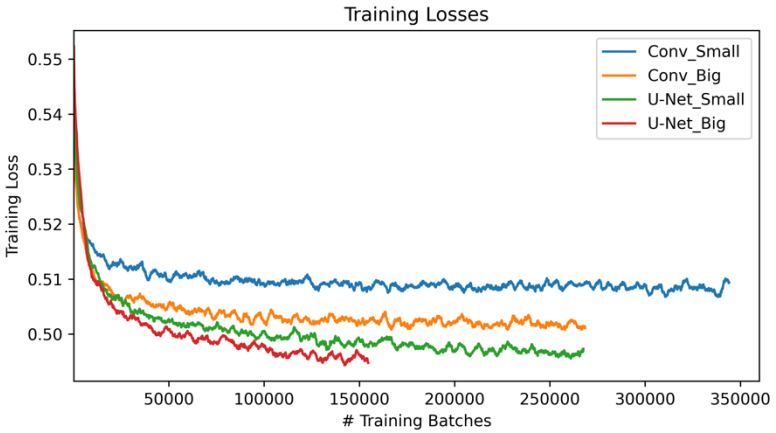
80 strains



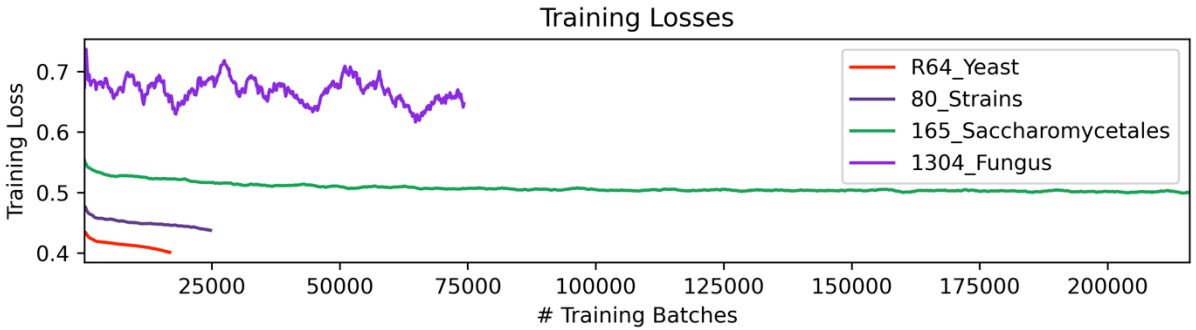


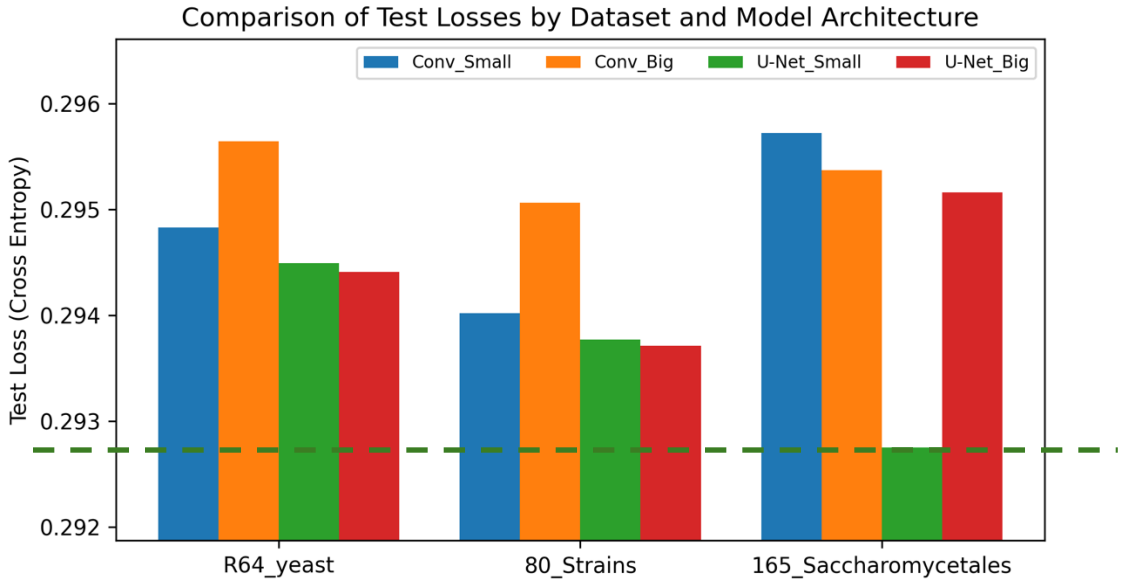
165 Saccharomycetales

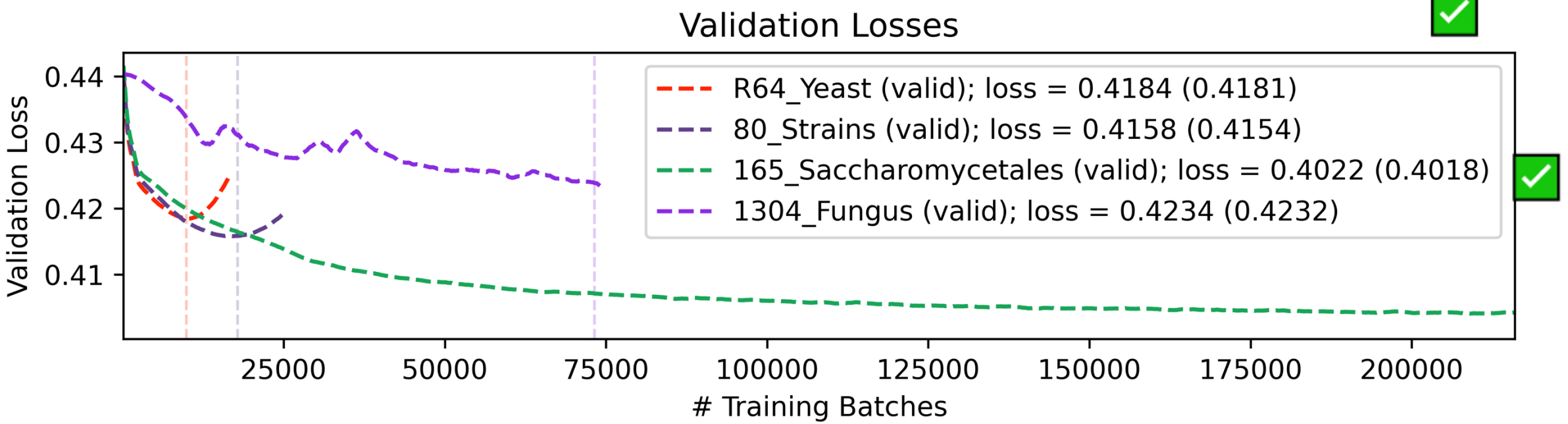




Dataset comparison







Fungal Language Model

Interpretability

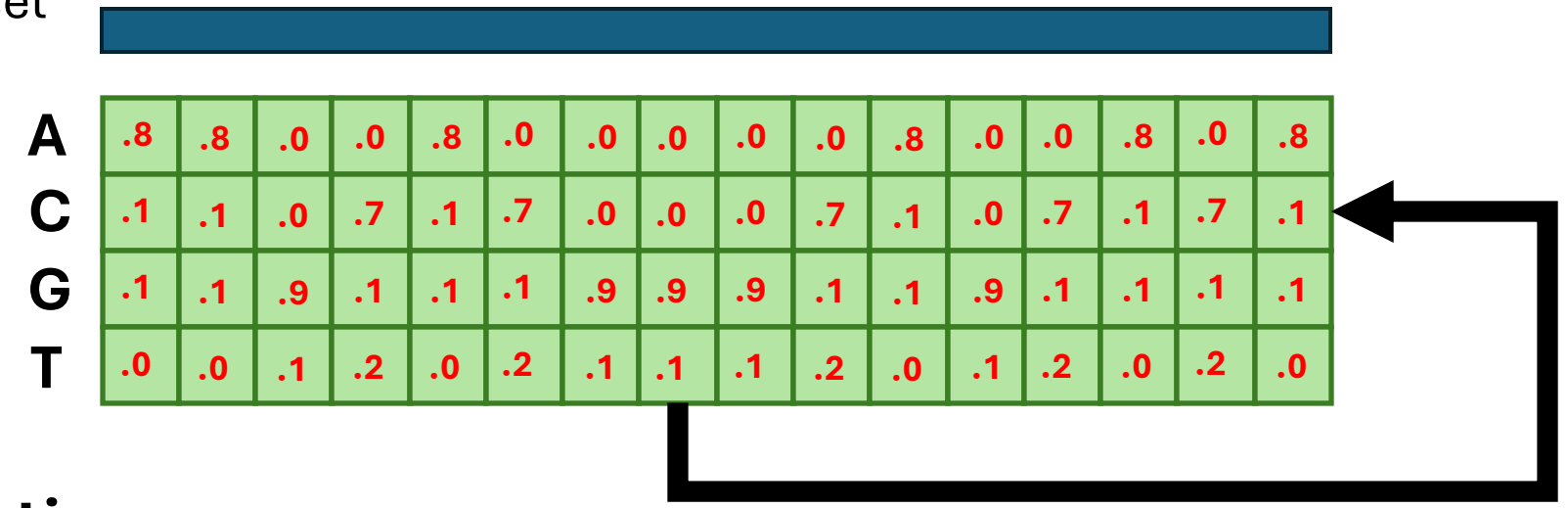
1. Motif inference 2. Attention map visualization

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



Step 2: PWM Normalization

1. Pseudocount Addition → **2. Row Normalization** → **3. Entropy Calculation** → **4. Conservation Calculation**

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

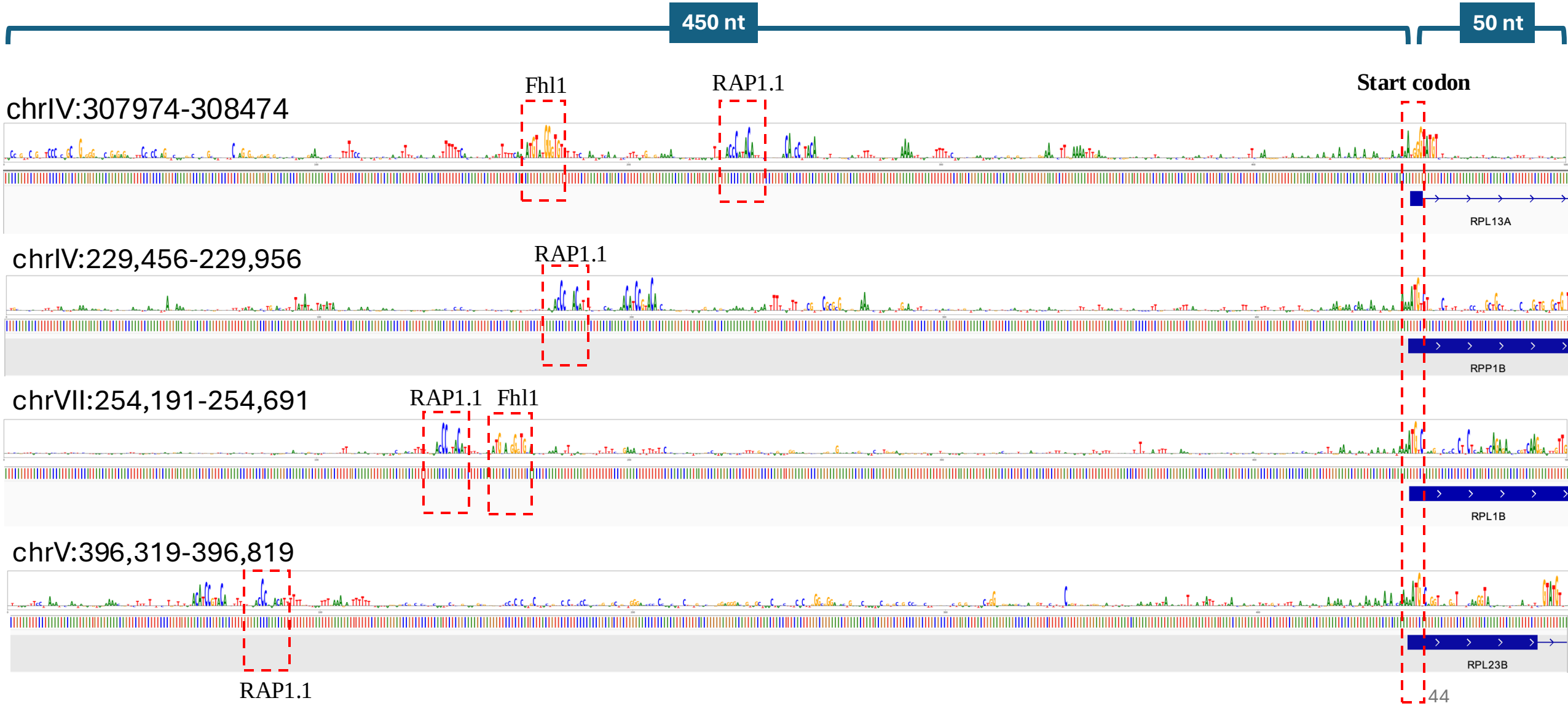
$p_{i,j}$ is the read count for nucleotide i at position j

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

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

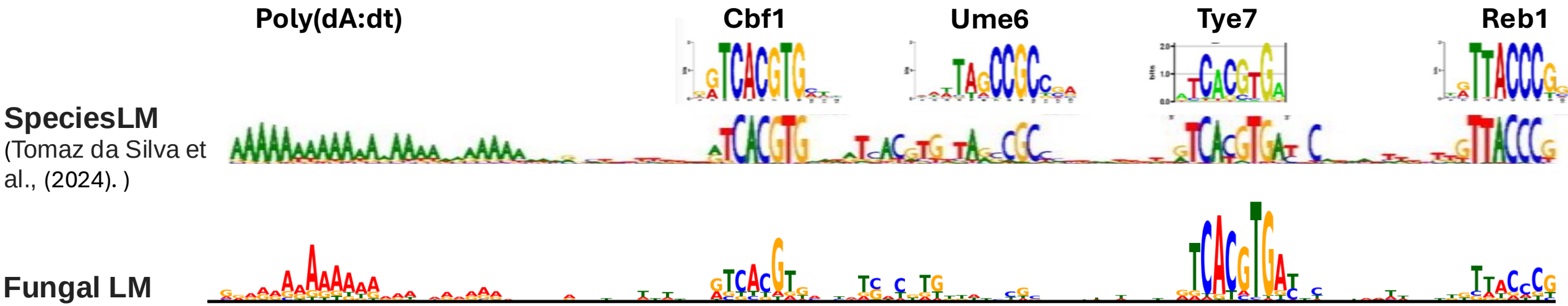
$$C_j = 2 - H_j$$

Risbosomal Protein Upstream Promoter regions



(Ground truth motifs from http://www.yeastepigenome.org/.)

Upstream Promoter region of SMT3 gene

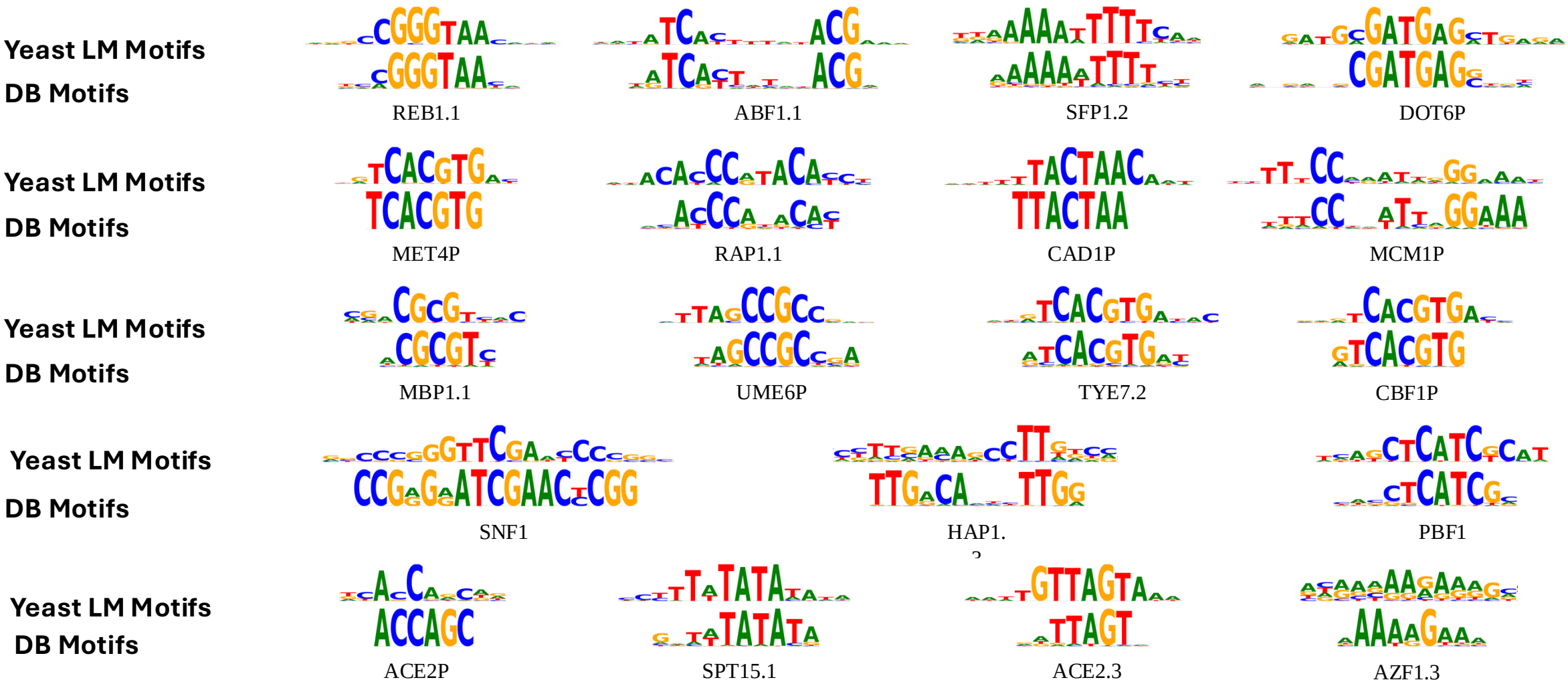


SpeciesLM: Trained with 5’ and 3’ regions only.

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



Constructing motifs in unseen genomes

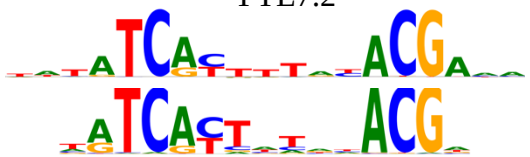
8 *S. cerevisiae* Strains

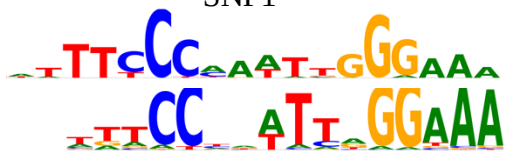

TYE7.2

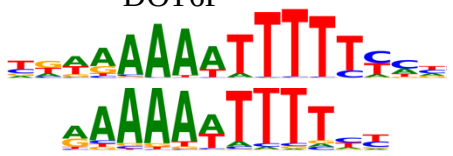

SNF1


DOT6P


RAP1.1


ABF1.1


MCM1P


SFP1.2


ACE2.3

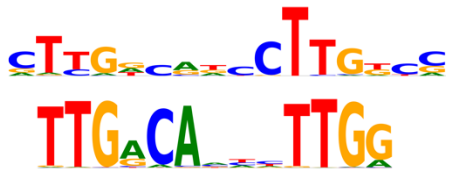

TYE7P

8 Schizosaccharomycetales genoms


Ace2p&consensus=ACCAGC


ACE2.3


SWI5.3

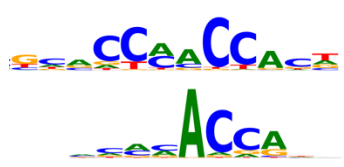

HAP1.3


SNF1


TBP

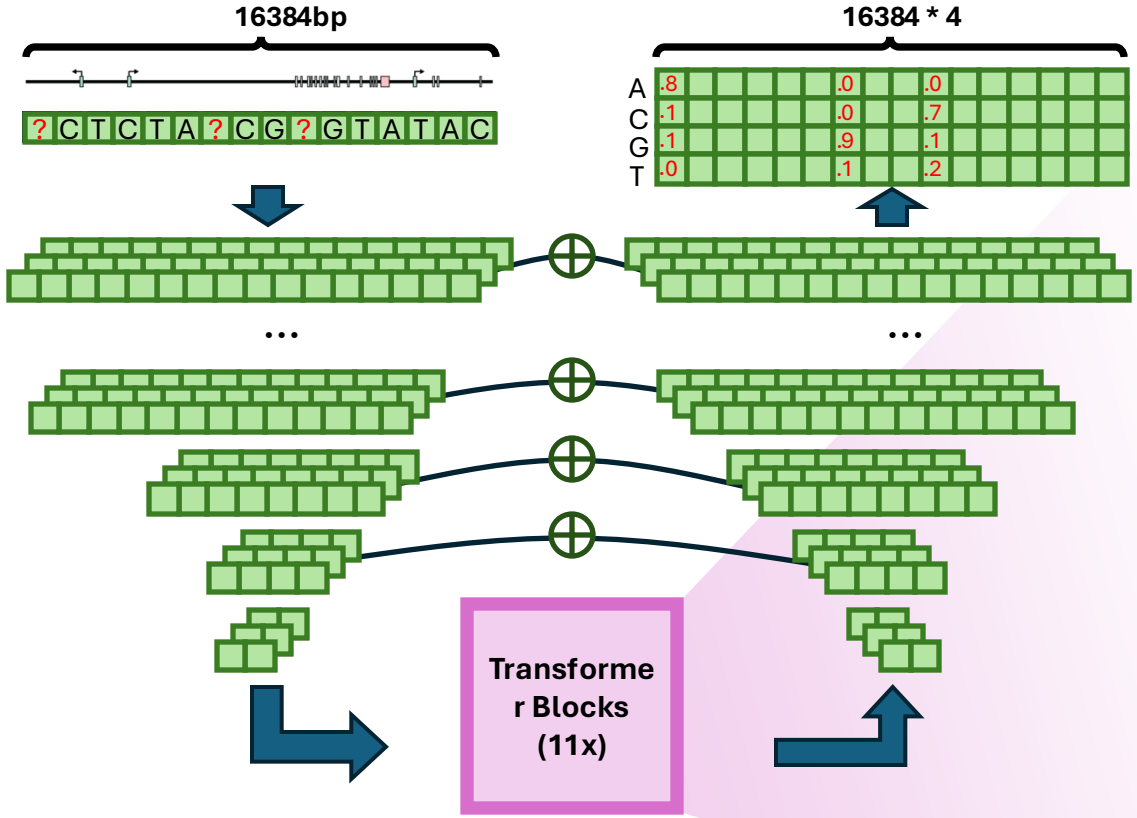

CAD1P

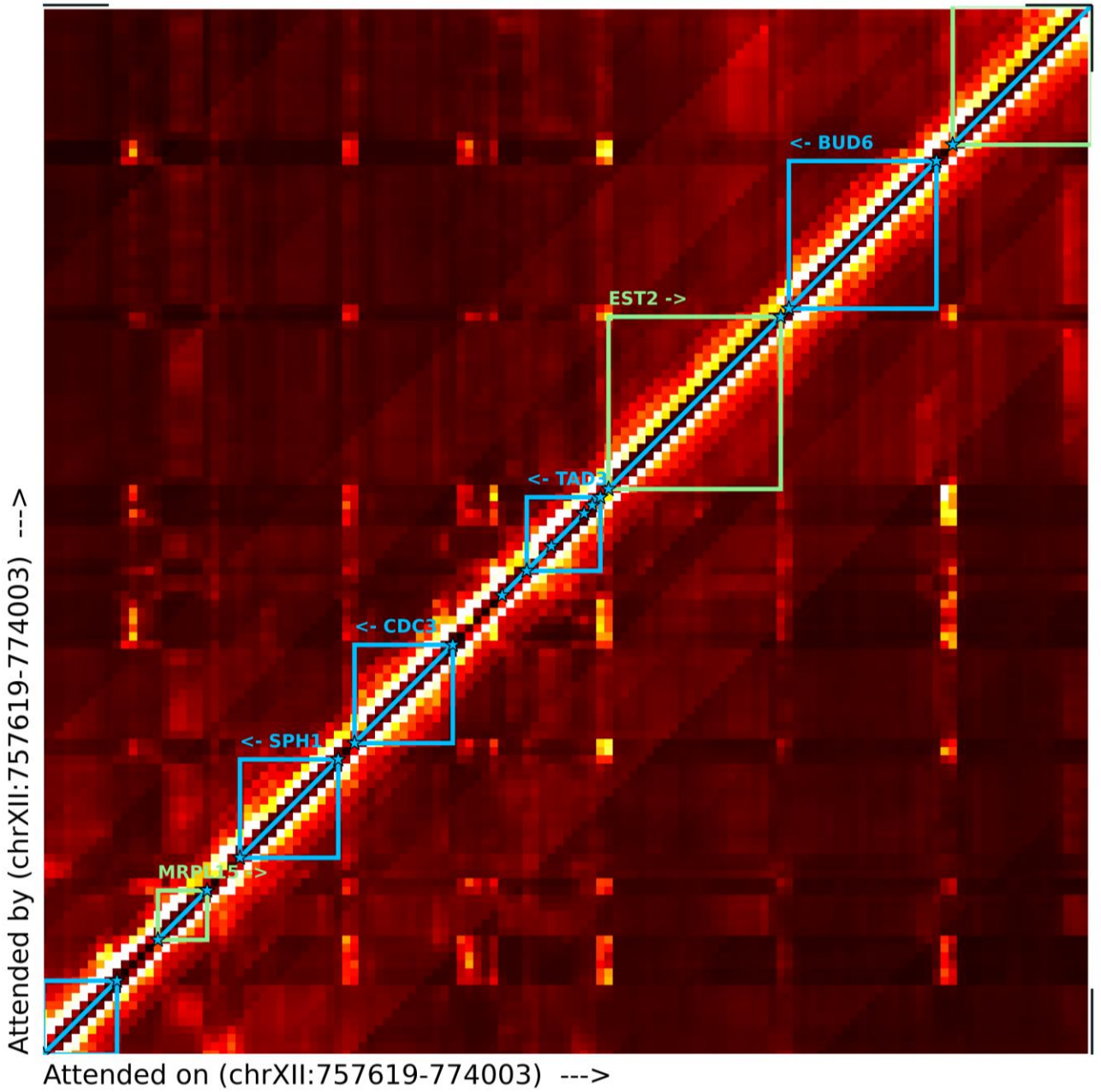

RIM101.2


PH02.4


YAP5.2

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?

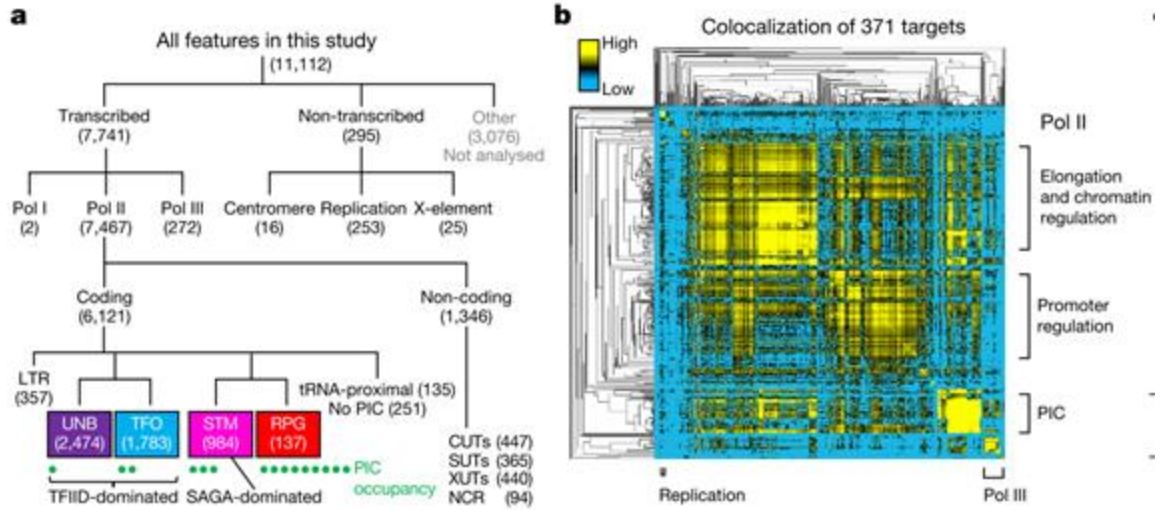
Genomic tracks intro &

Data preprocessing

ChiP-exo + Histone Marks + RNA-Seq

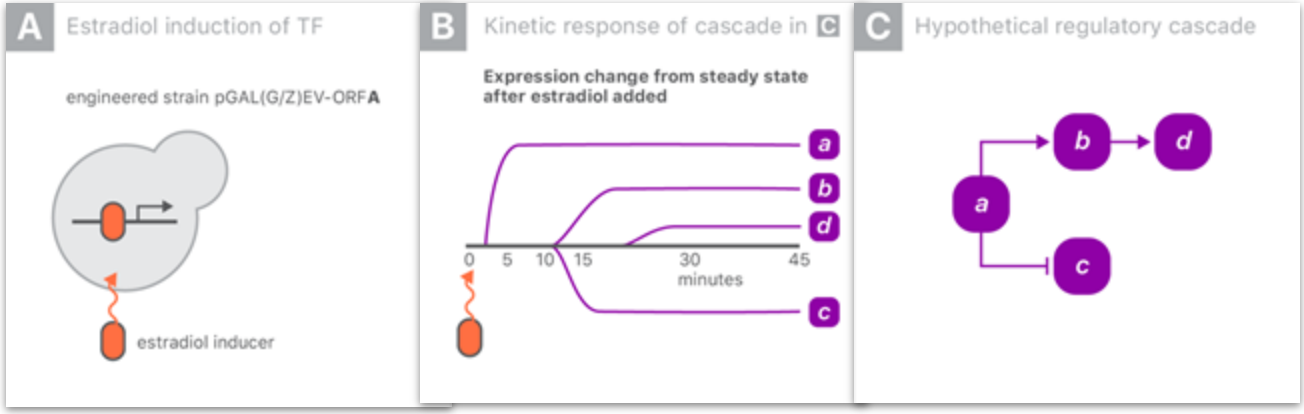
- ChIP-exo provides high res view of protein-DNA binding across the yeast genome.
- Dataset includes 1128 ChIP-exo experiments
- Histone Mods MNase-ChIP-seq

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.



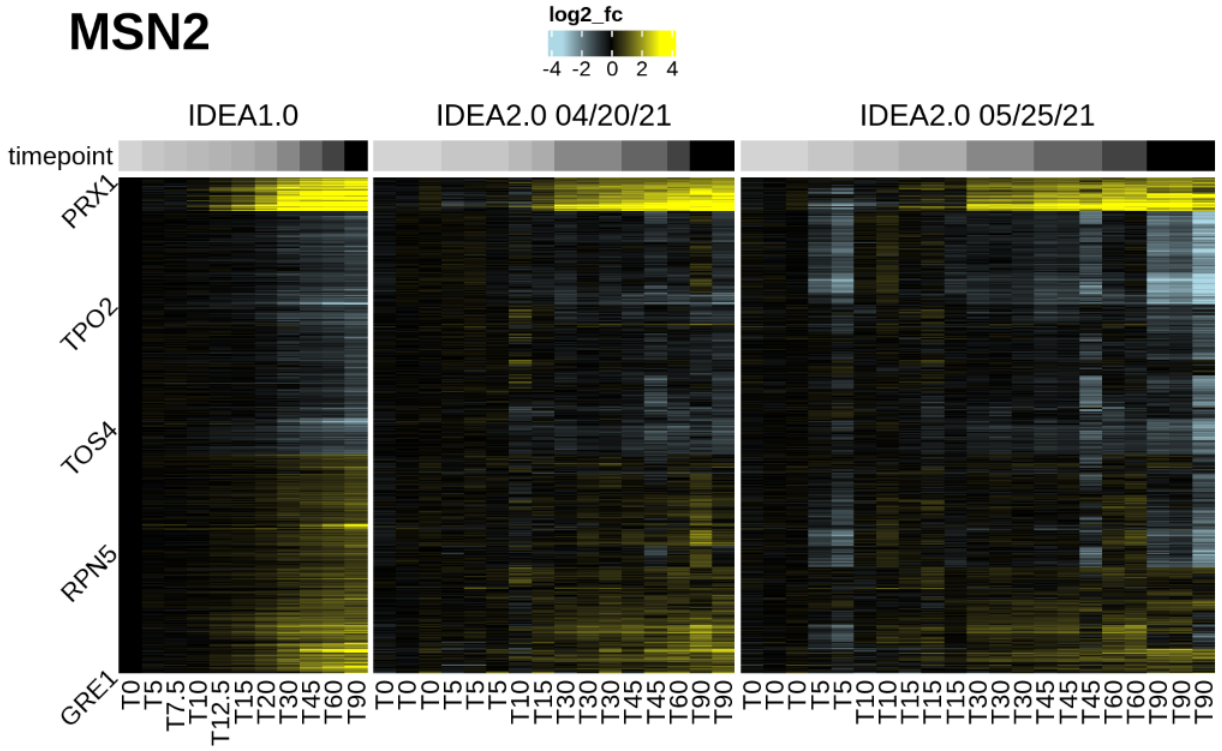
- Genome-scale perturbation dynamics propagate signals across regulatory networks (1340 experiments)
- Aggregating dynamics across many time-courses enables disambiguation of cause > effect relationships

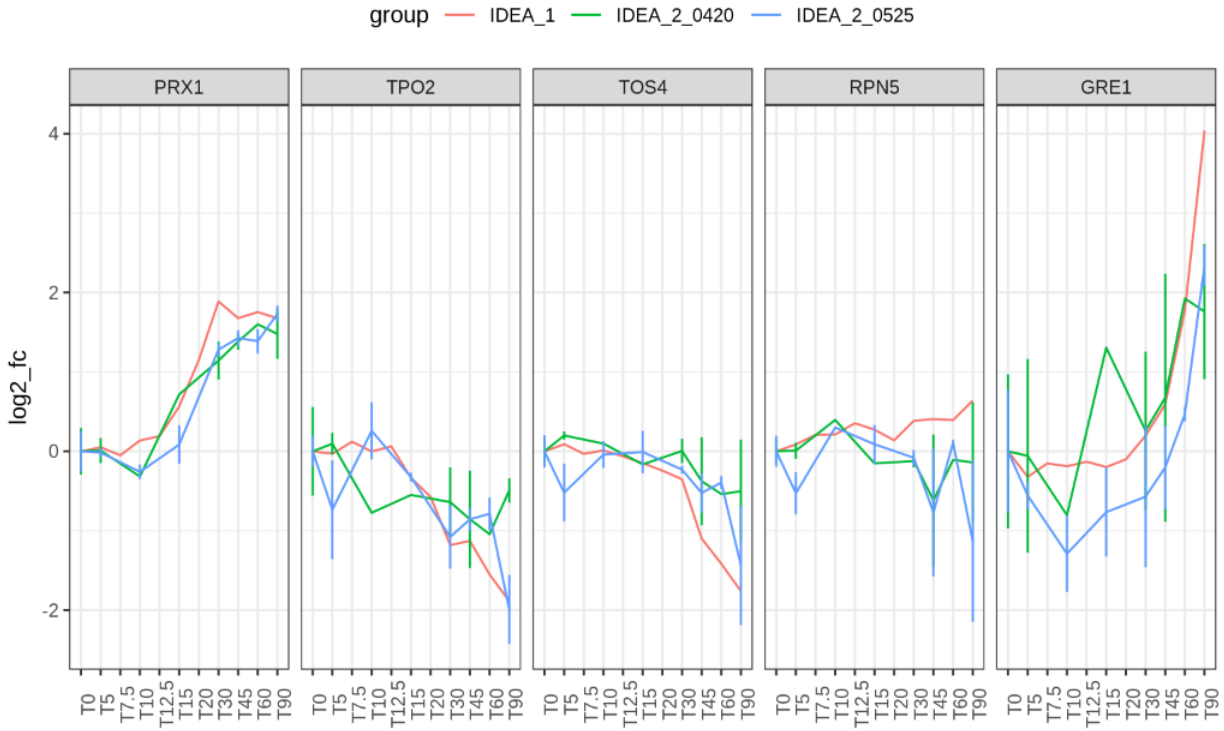
Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., ... & McIsaac, R. S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. Molecular systems biology, 16(3), e9174.



RNA-Seq

- IDEA (the Induction Dynamics gene Expression Atlas)





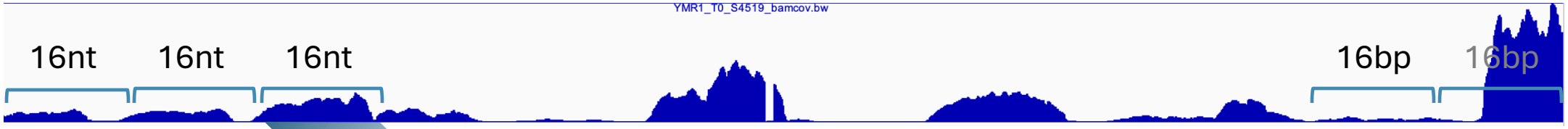
Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., ... & McIsaac, R. S. (2020). Learning causal networks using inducible transcription factors and transcriptome-wide time series. Molecular systems biology, 16(3), e9174.

Borzoi: squashed scale $y_{j,t}^{(\text{squashed})} = \left\{ y_{j,t}^{(3/4)} \text{ if } y_{j,t}^{(3/4)} \leq 384, \text{ otherwise } 384 + \sqrt{y_{j,t}^{(3/4)} - 384} \right\}$

Track transformation

chrXVI:670,897-685,233

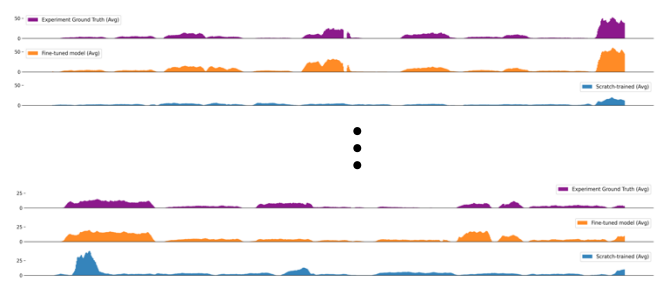
14336 nt $(16384 - 1024 * 2)$



Summation over each 16bp window

896 $(14336 \div 16)$

Coverage Tacks



CHiP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

No normalization across tracks

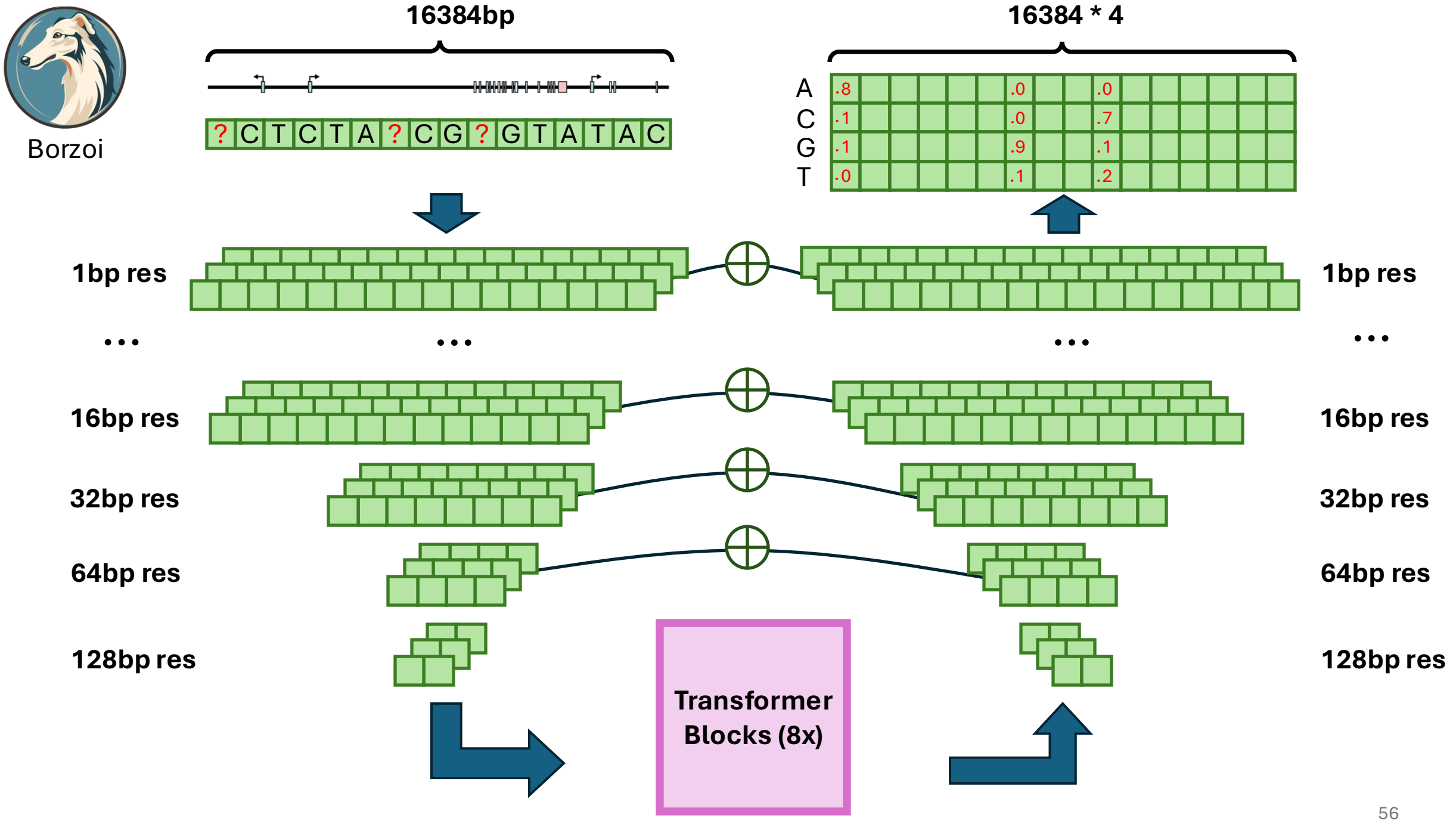
Loss function: Poisson Loss.

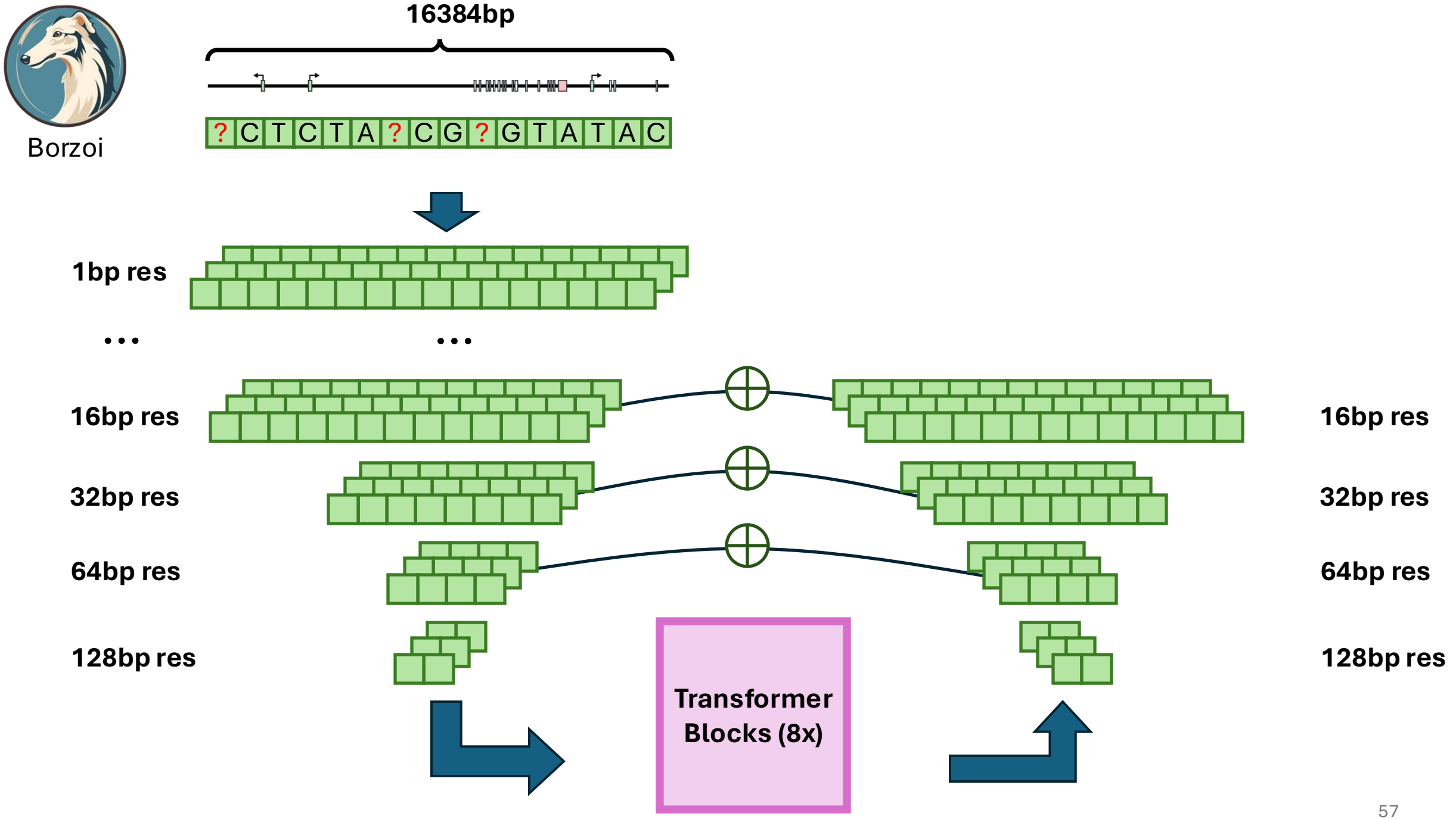
The depth of the track somehow reflects the quality of the data.

Scratch-trained model vs.

Fine-tuned Fungal LM:

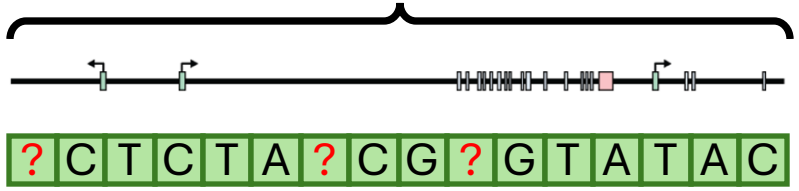
Initialization & Training


Borzoi
16384bp
? C T C T A ? C G ? G T A T A C
16384 * 4
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)
1bp res
16bp res
32bp res
64bp res
128bp res

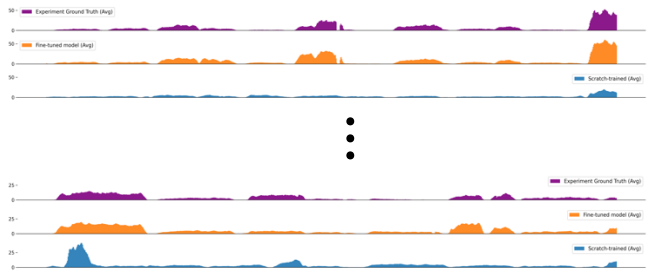

Borzoi
16384bp
? C T C T A ? C G ? G T A T A C
1bp res
16bp res
32bp res
64bp res
128bp res
Transformer Blocks (8x)
16bp res
32bp res
64bp res
128bp res

Borzoi

16384bp



Coverage Tacks (896 * 2488)



CHiP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

1bp res

...

...

16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

16bp res

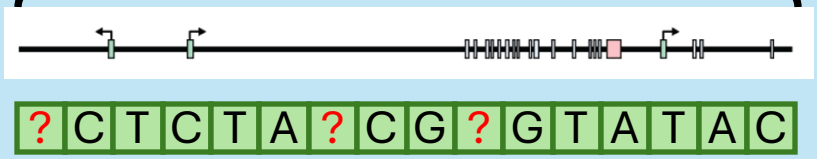
32bp res

64bp res

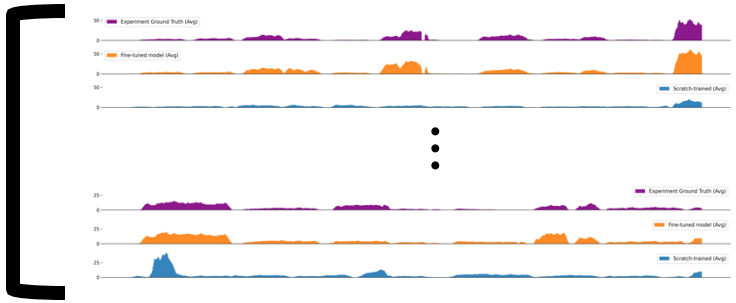
128bp res

Borzoi

16384bp



Coverage Tacks (896 * 2488)



CHiP-exo (1128)

Histone marks (20)

RNA-Seq (1340)

1bp res

...

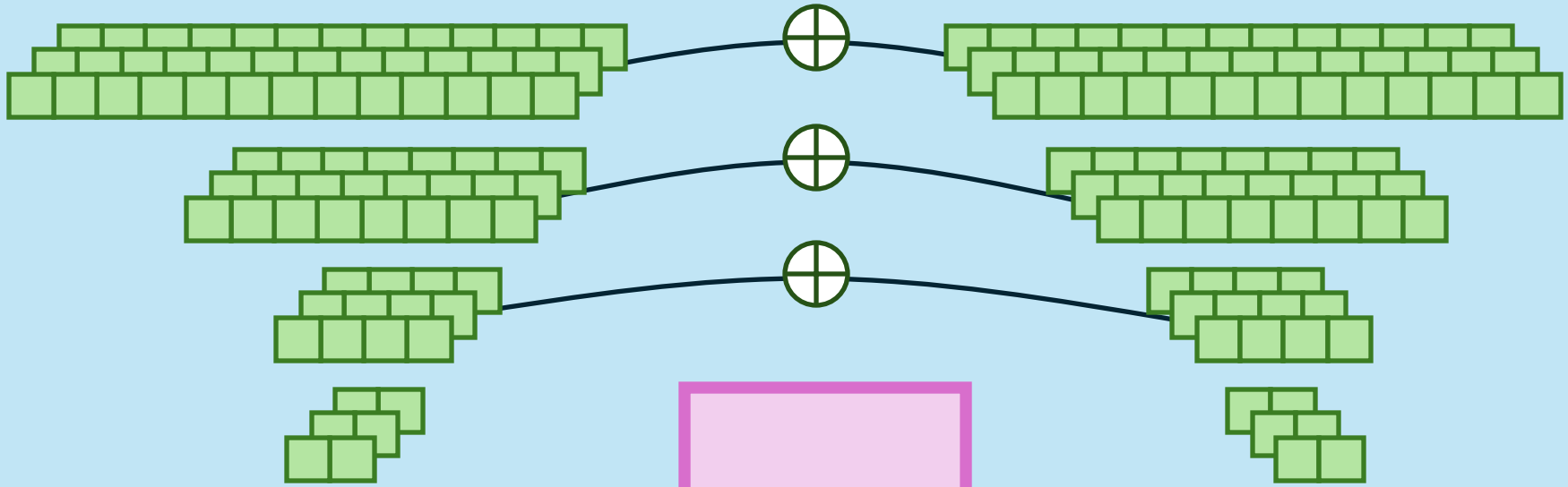
...

16bp res

32bp res

64bp res

128bp res



16bp res

32bp res

64bp res

128bp res

Transformer Blocks (8x)

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)

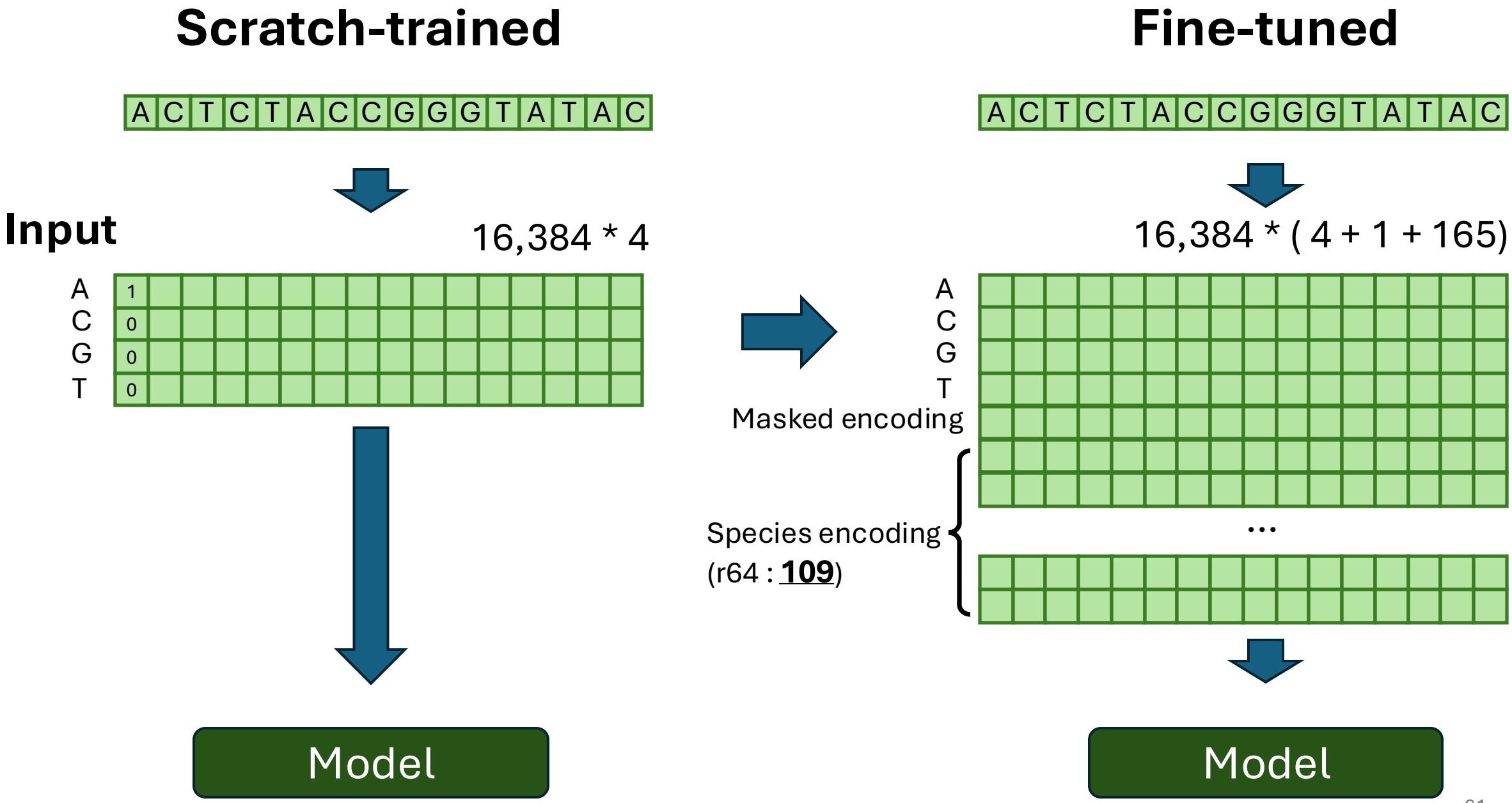
chrVII:	497038-1090940
chrX:	436425-745751
chrI:	0-151465
chrI:	151582-230218
chrXII:	0-150828

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
Fine-tuned
A C T C T A C C G G G T A T A C
A C T C T A C C G G G T A T A C
Input
16,384 * 4
16,384 * (4 + 1 + 165)
A
C
G
T
1
0
0
0
Masked encoding
Species encoding
(r64 : 109)
...
Model
Model

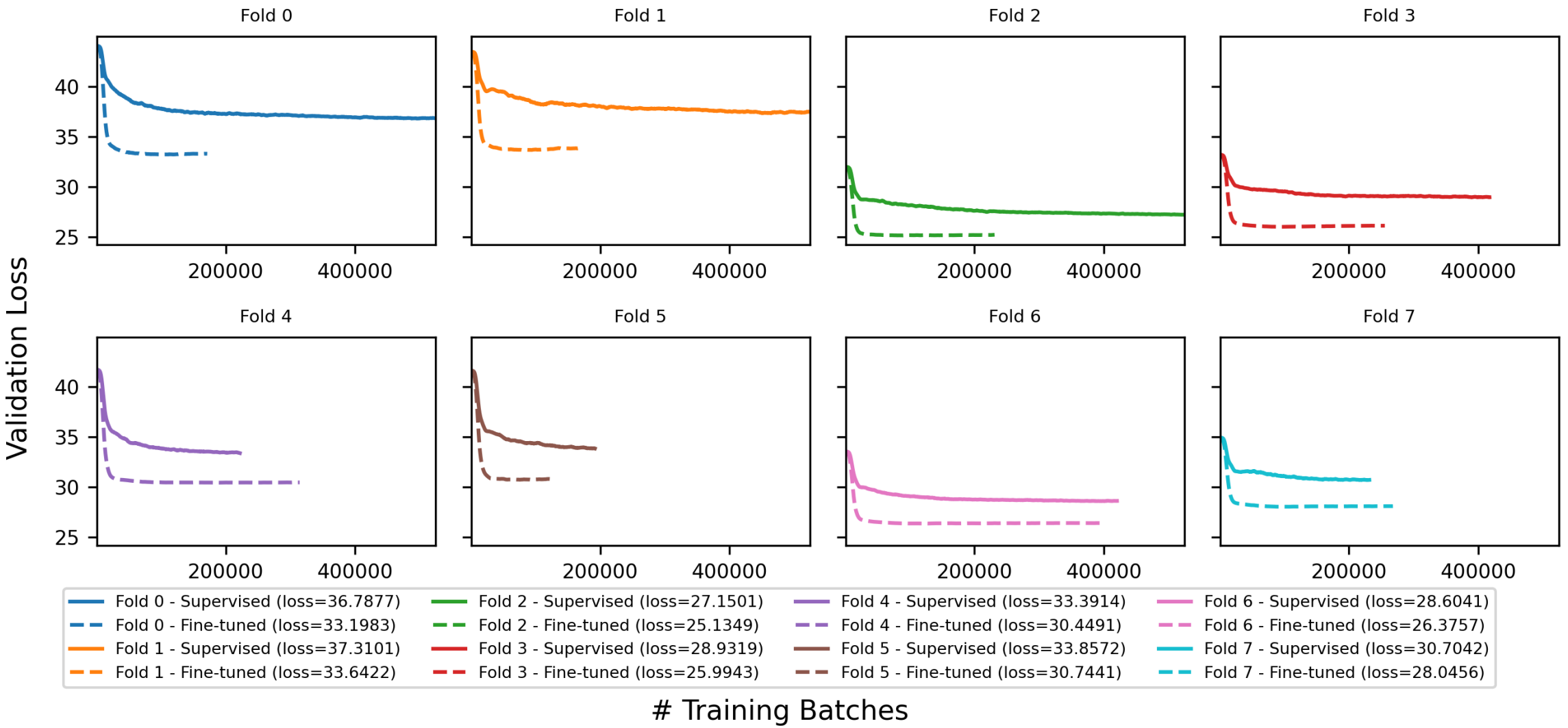
Scratch-trained model vs.

Fine-tuned Fungal LM:

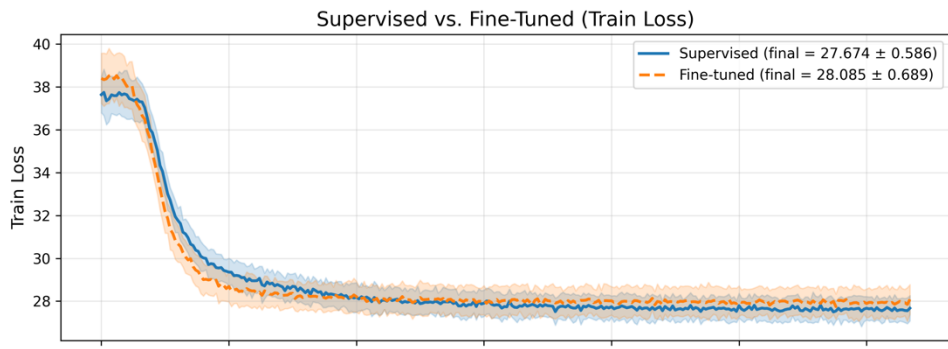
Training Results

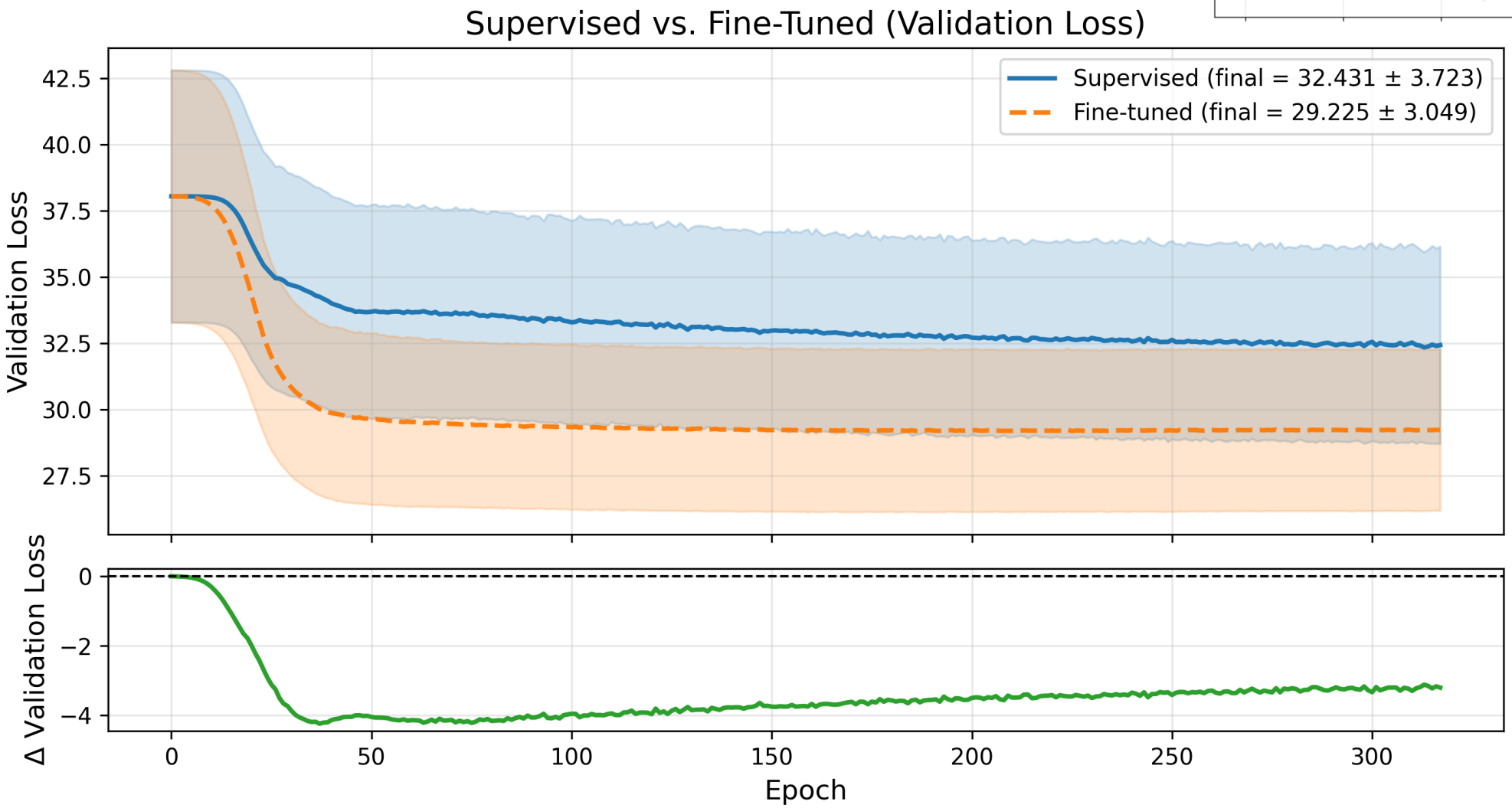
Fine-tuned vs Scratch-Trained

(16 bp resolution)



Fine-tuned vs Scratch-Trained

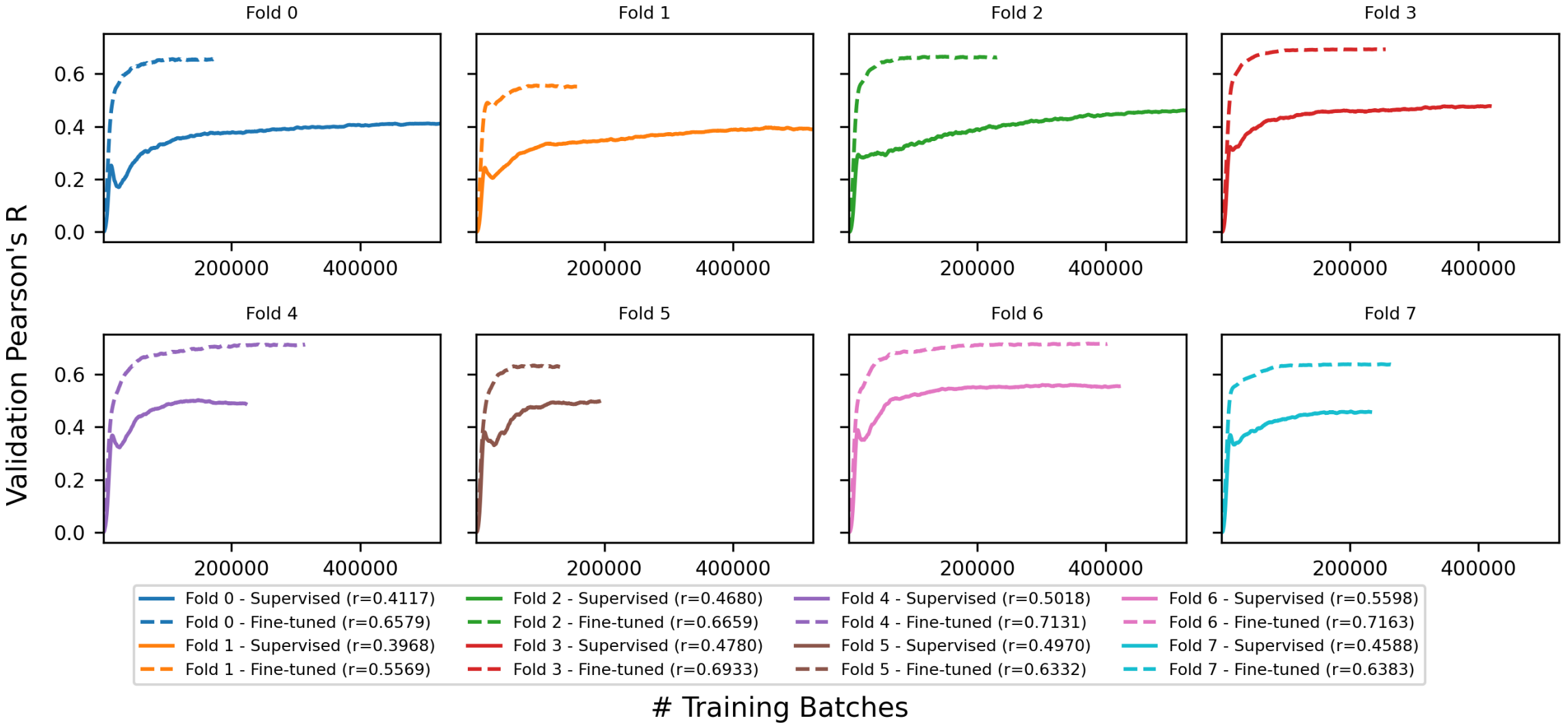




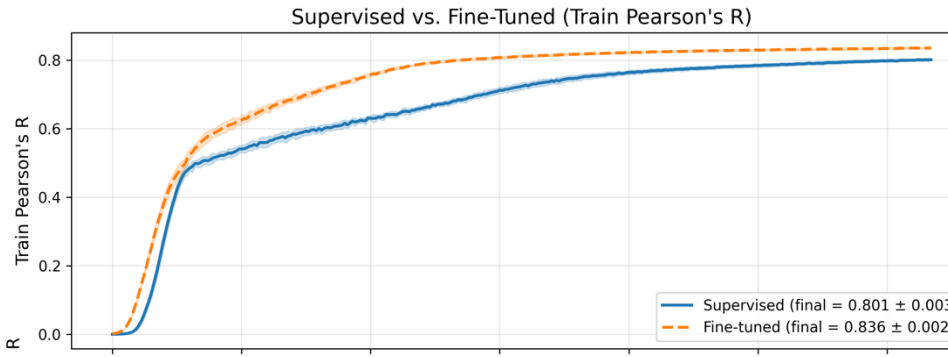
(16 bp resolution)

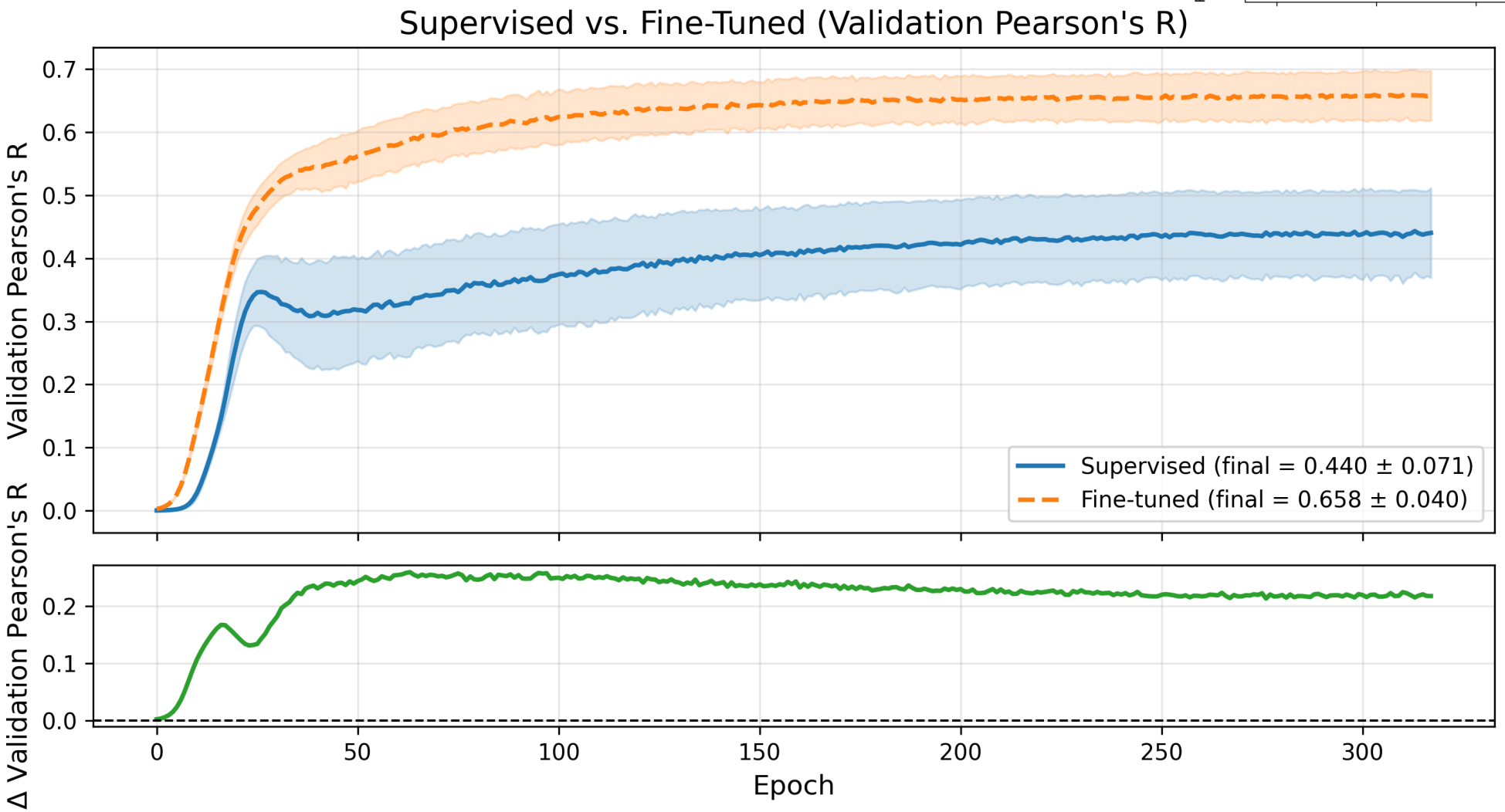
Fine-tuned vs Scratch-Trained

(16 bp resolution)



Fine-tuned vs Scratch-Trained





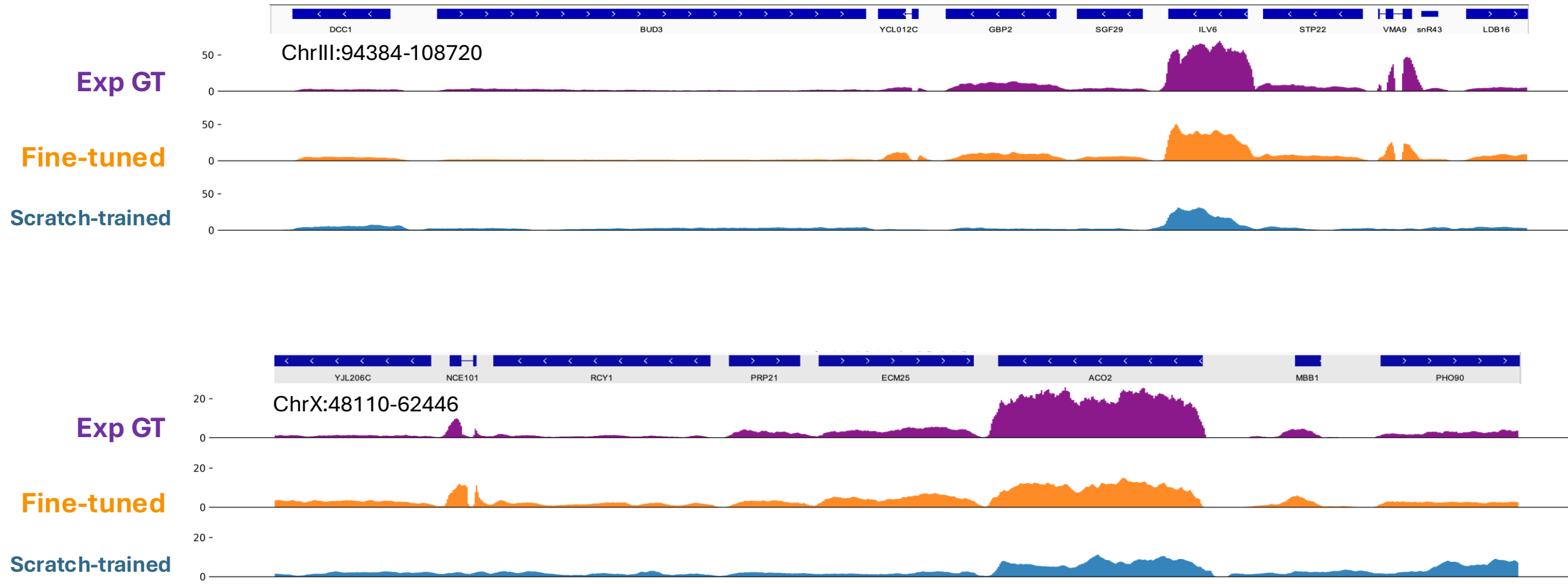
(16 bp resolution)

Scratch-trained model vs.

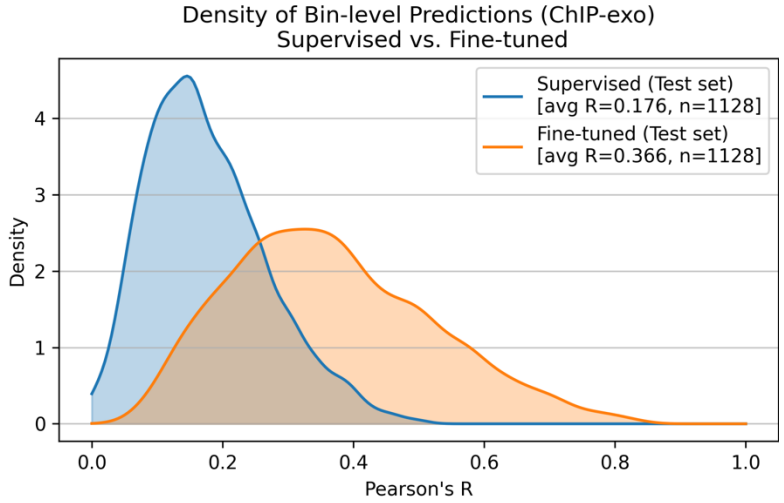
Fine-tuned Fungal LM:

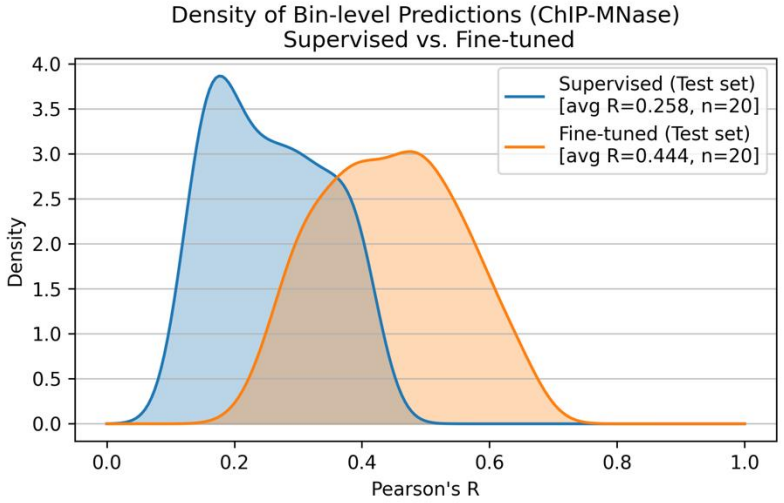
Track-level prediction evaluation

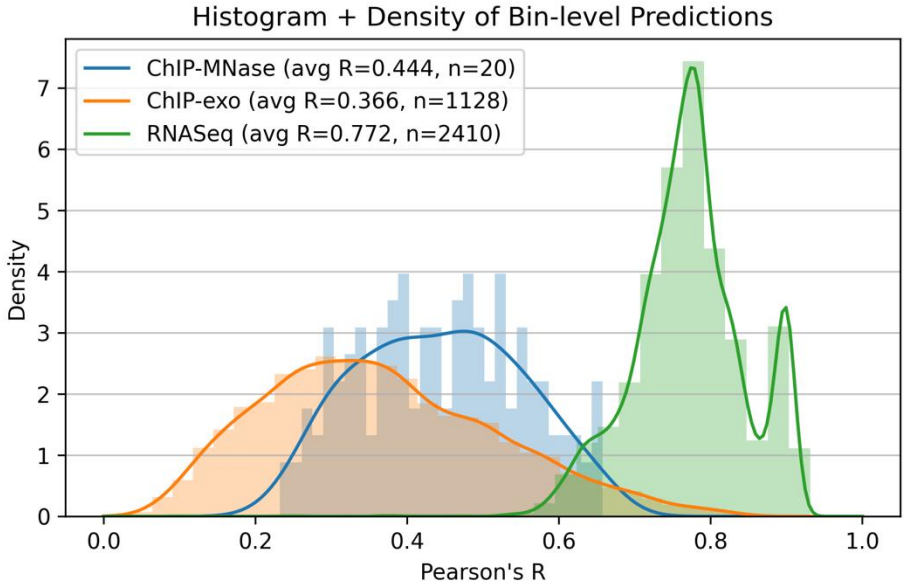
Fine-tuned vs Scratch-Trained (Test set)

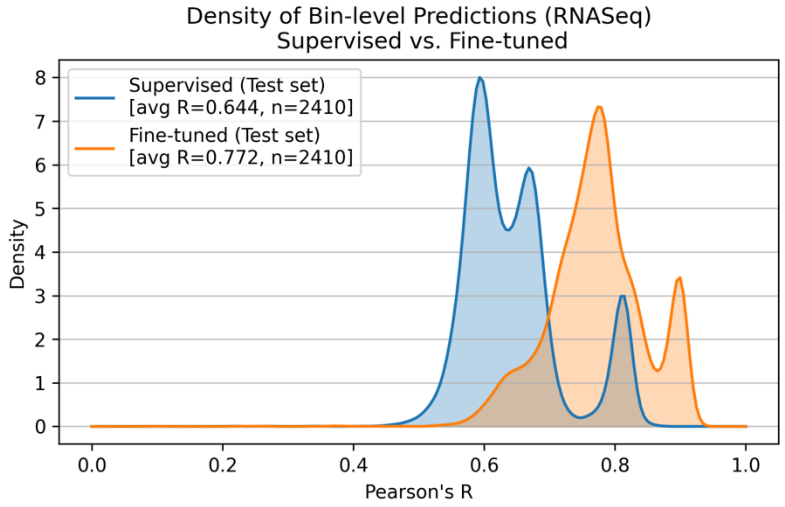


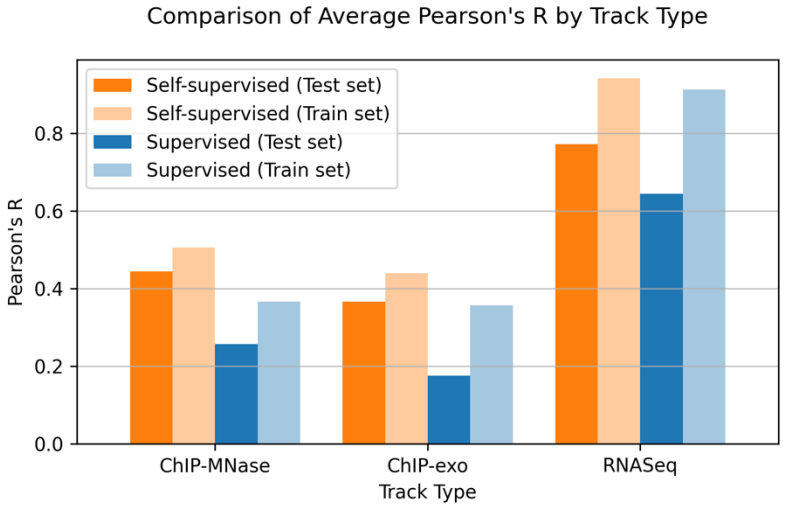
Fine-tune vs Scratch-Trained (Test set)



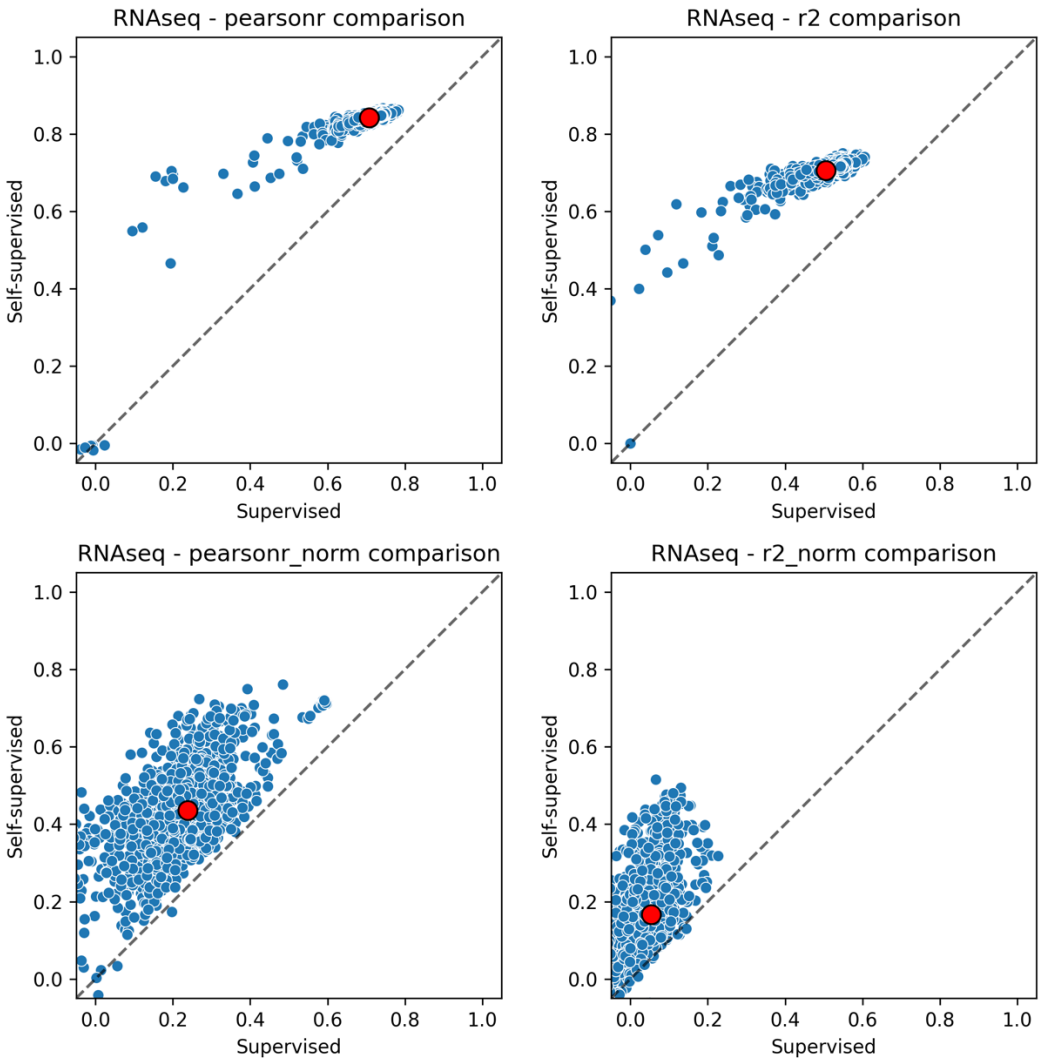




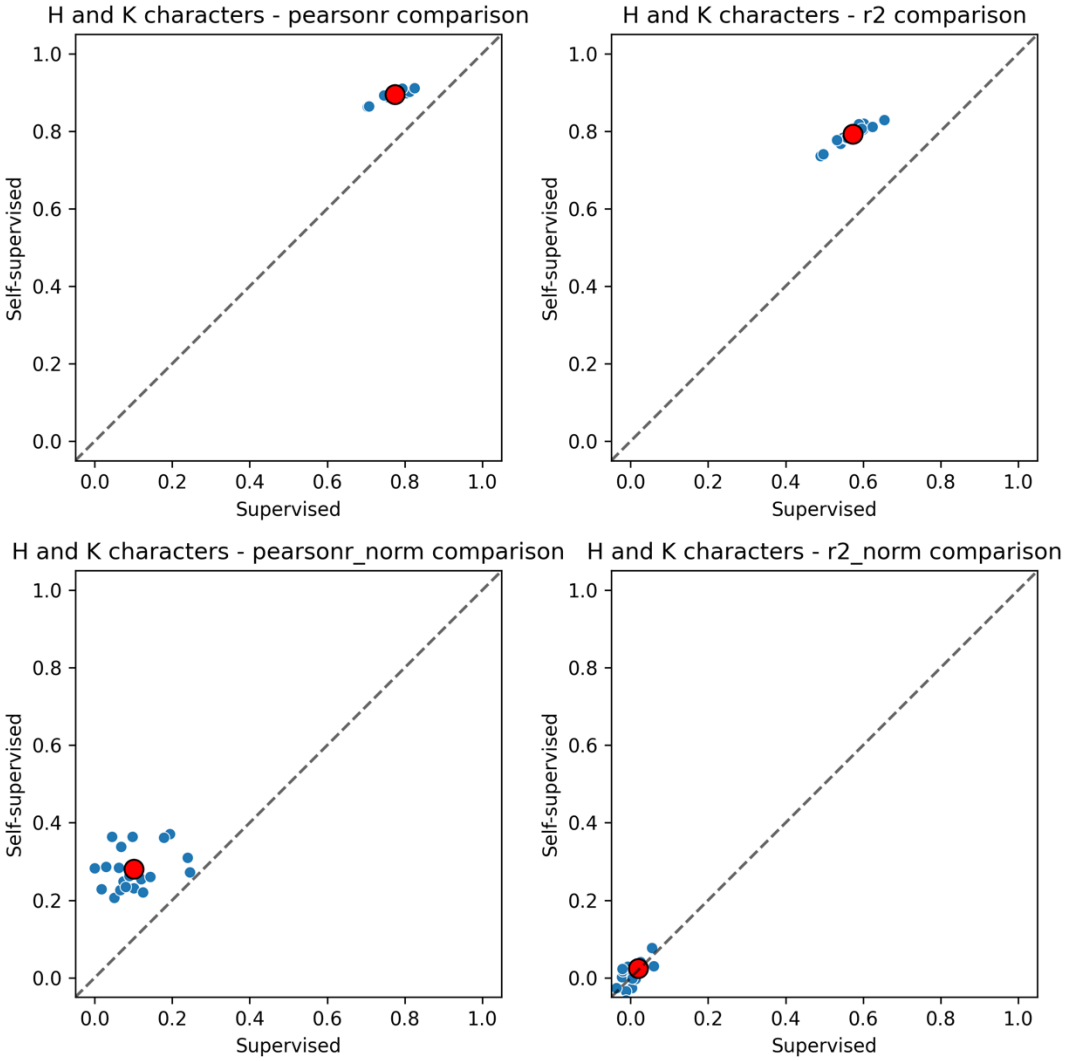




RNA-Seq

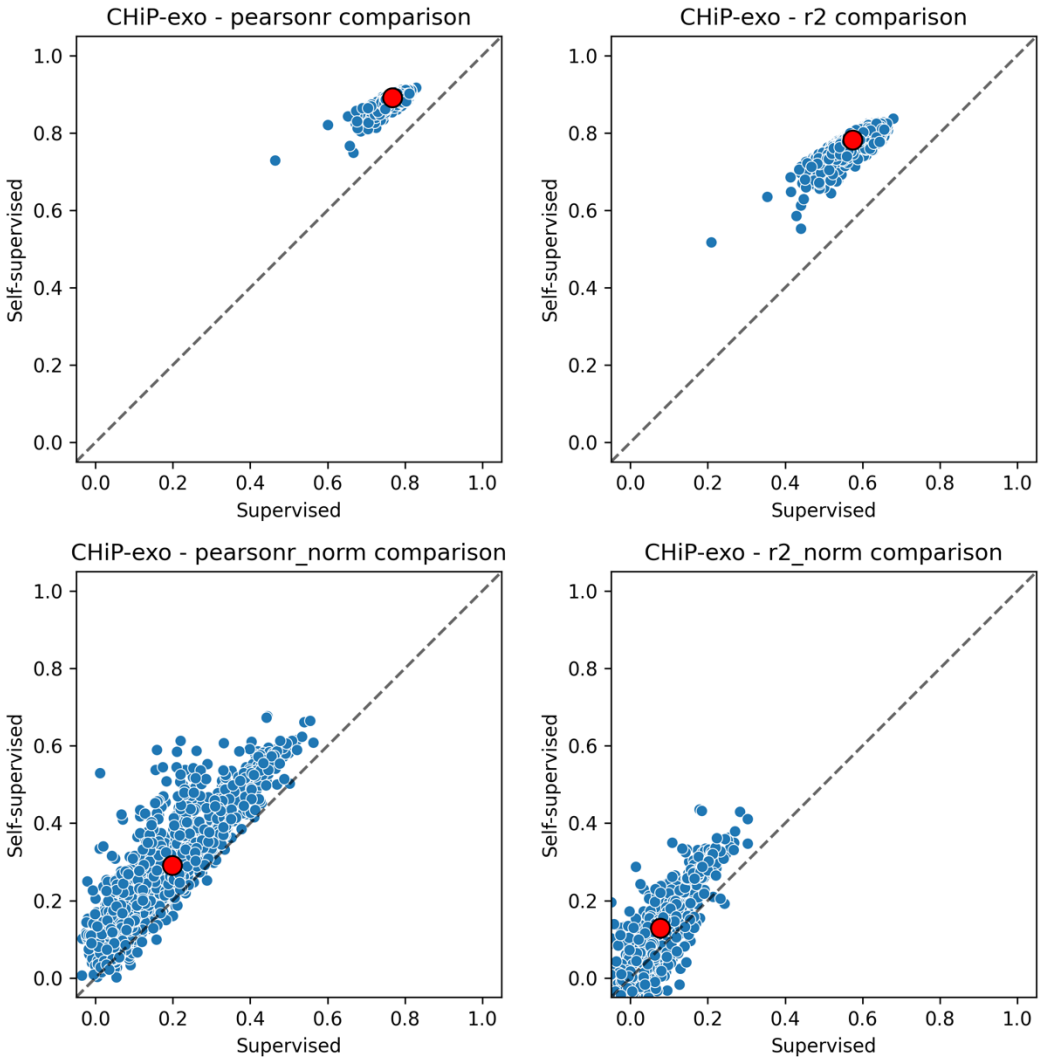


Histone Marks

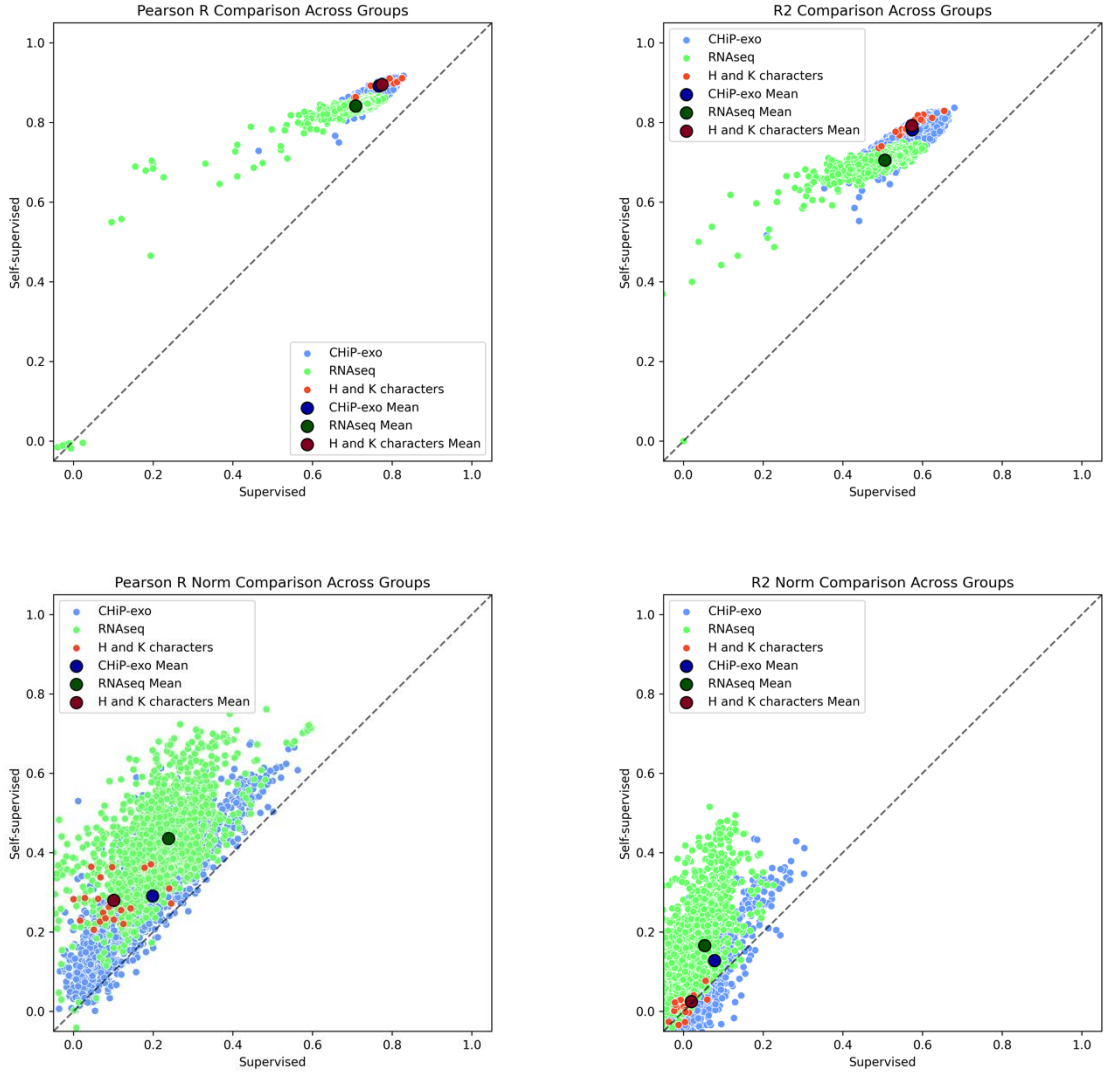


Average results across 8 folds. Each dot is a track.

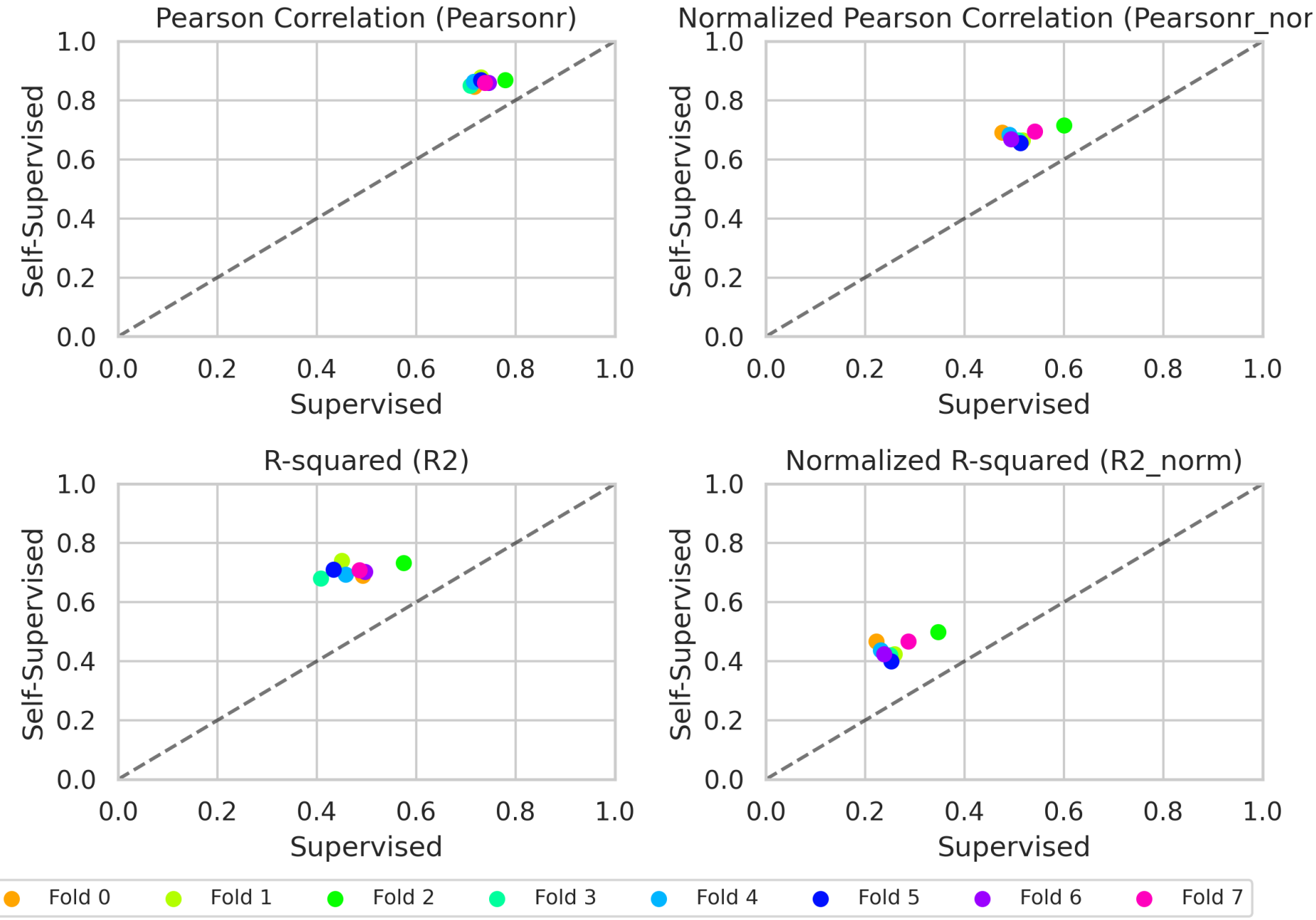
CHiP-exo



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



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



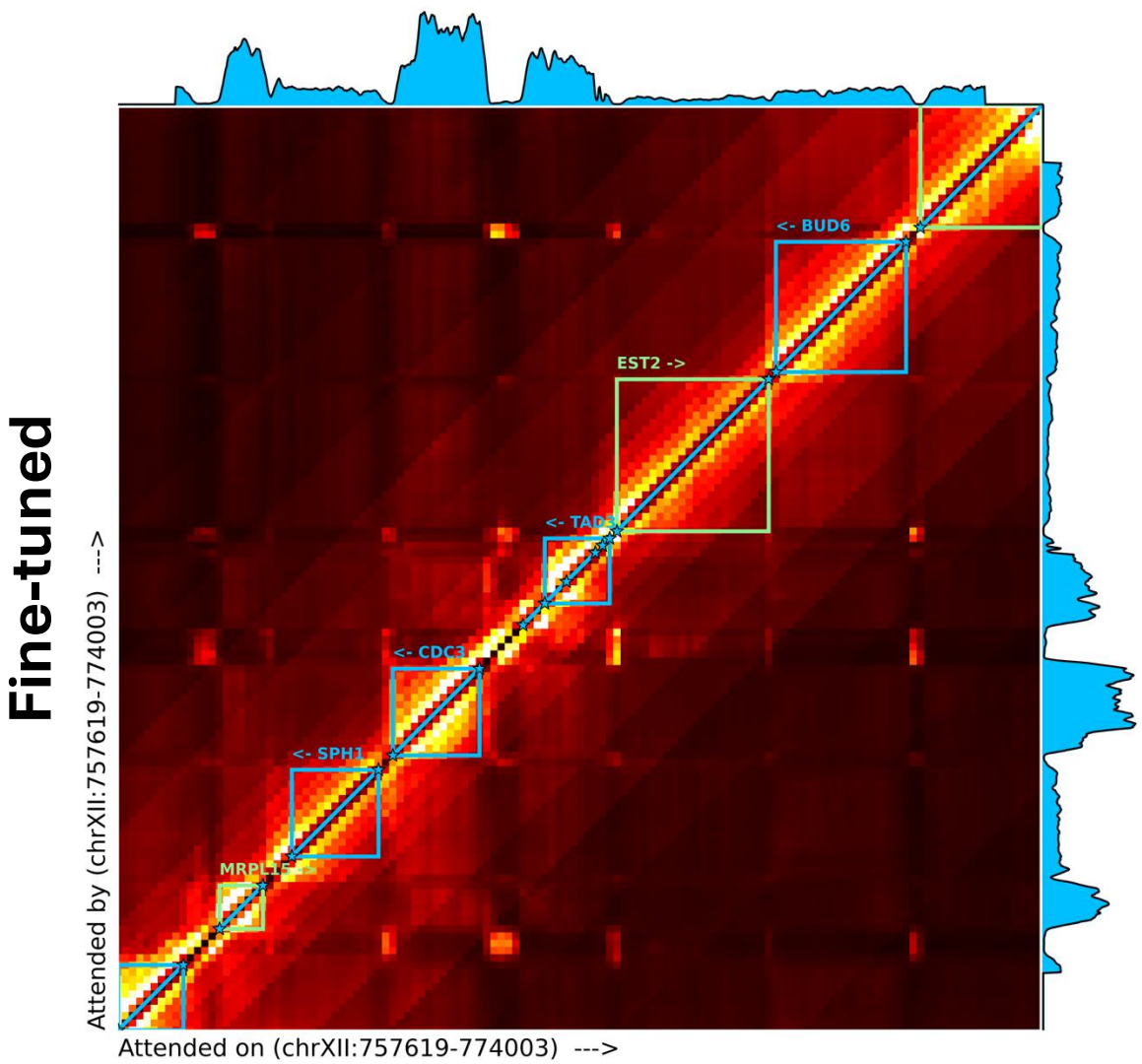
Average results across tracks. Each dot is a fold

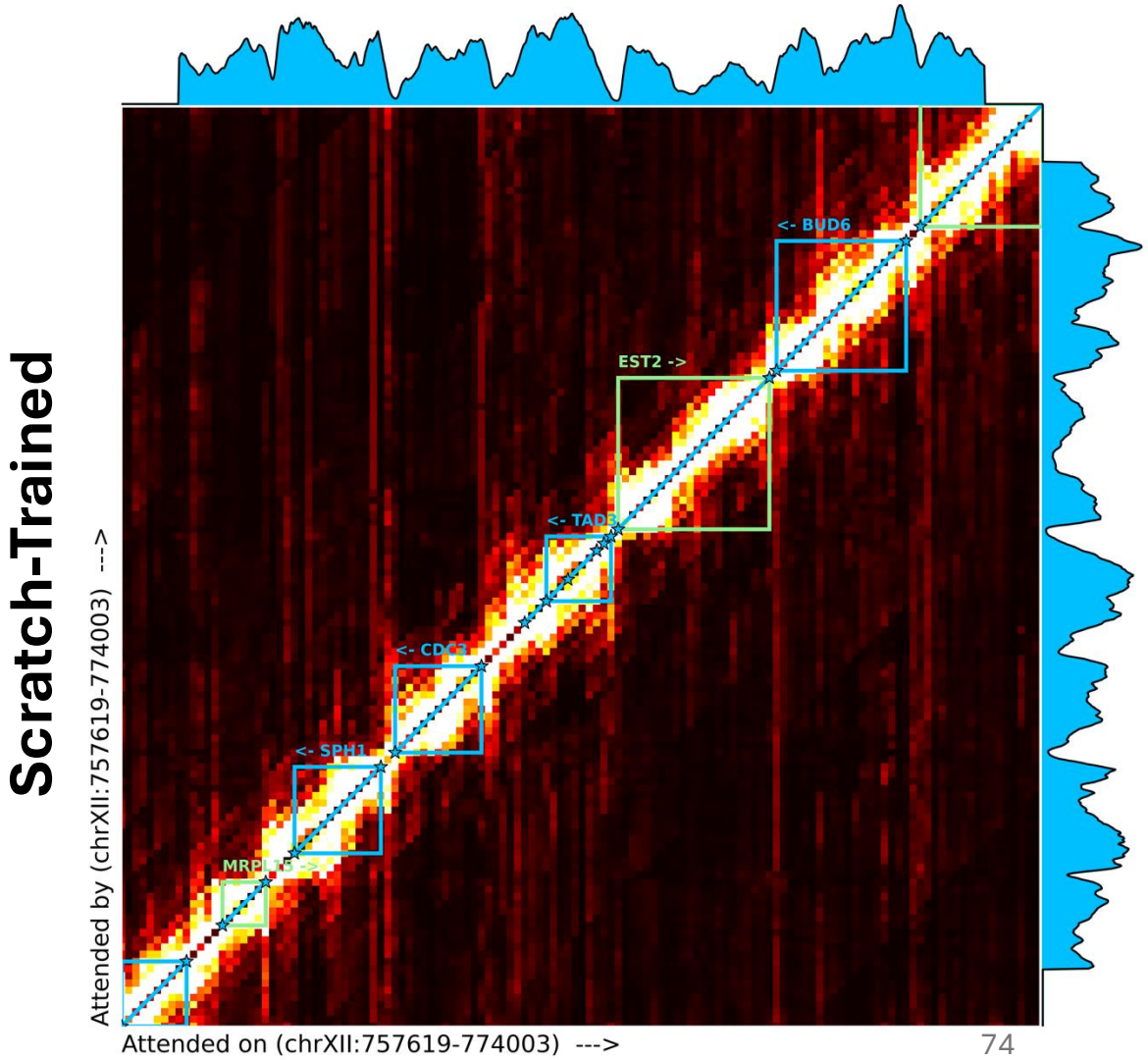
Scratch-trained model vs.

Fine-tuned Fungal LM:

Attention maps

Fine-tuned vs Scratch-Trained





Scratch-trained model vs.

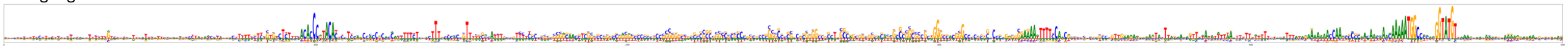
Fine-tuned Fungal LM:

Motif usage

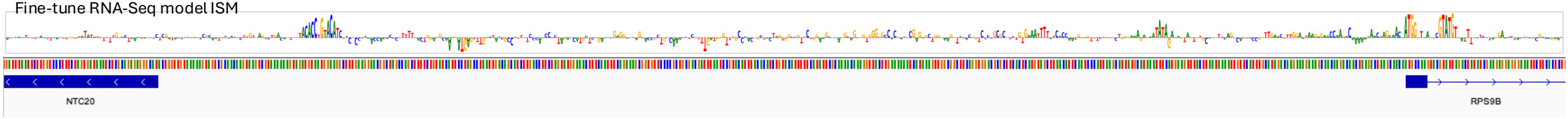
Fine-tuned vs Scratch-Trained

RPS9B

Language model



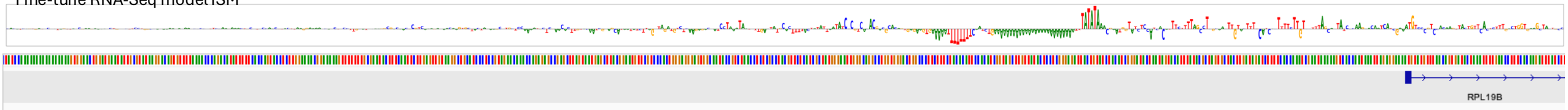
Fine-tune RNA-Seq model ISM



RPL4A

Language model

Fine-tune RNA-Seq model ISM



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

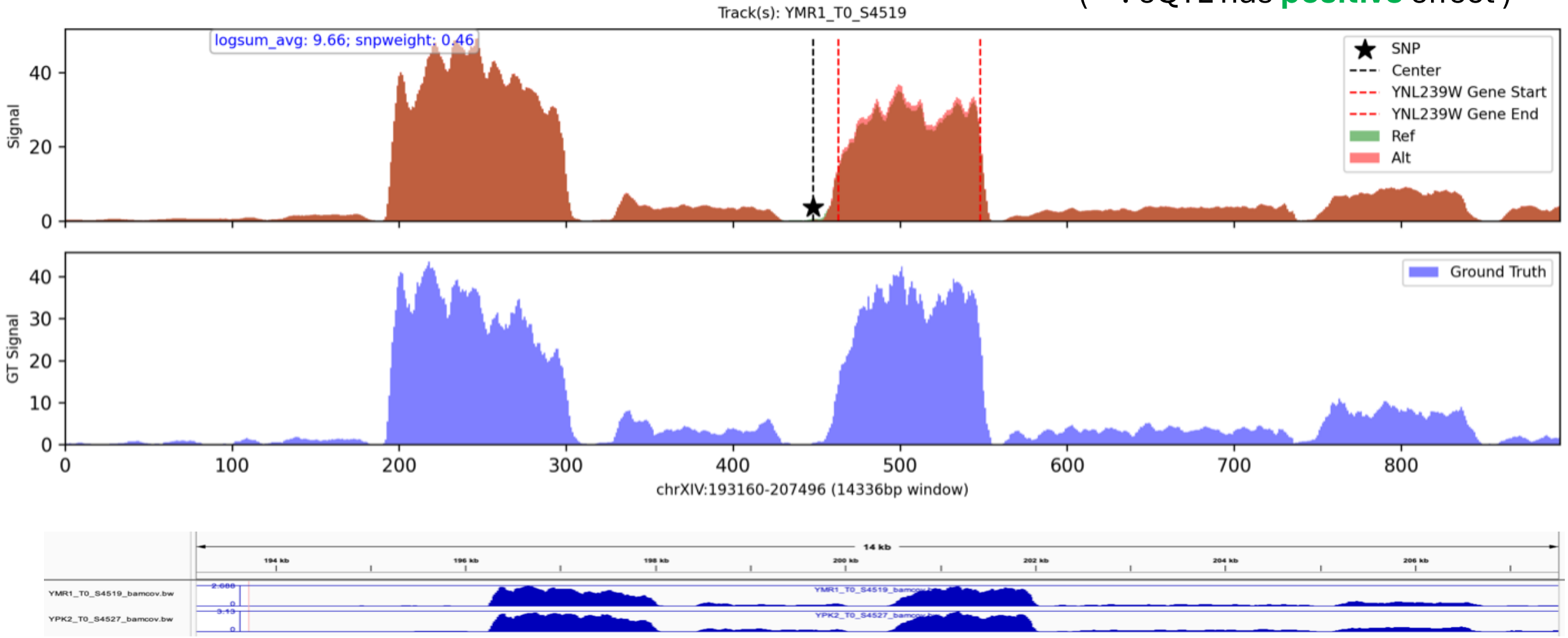
1. Assessing the variant effects on eQTLs and negatively selected eQTLs
2. Predicting the influence of distal regulatory elements (i.e. enhancers) on gene expression.
3. MPRA mutation effect prediction

Predicting eQTLs

SNPWeight from GWAS **(+)**

Yeast LM prediction **(+)**

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

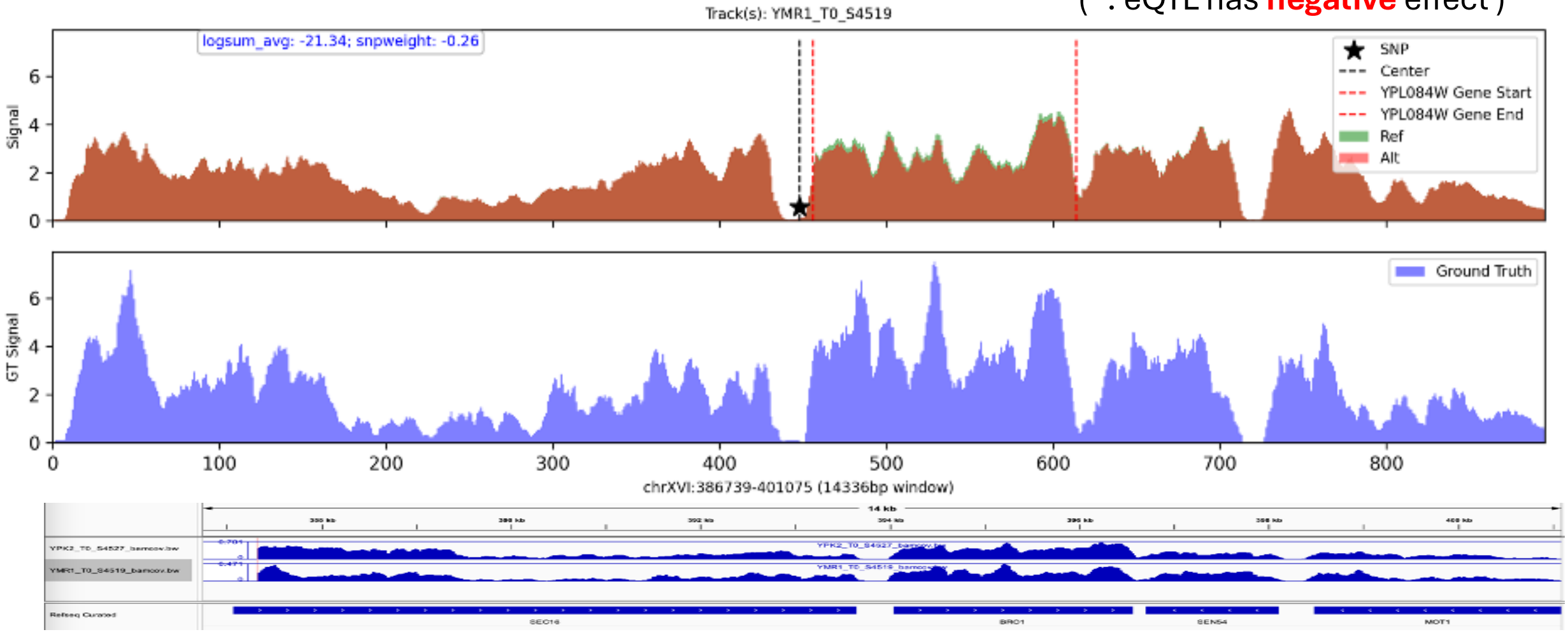


Predicting eQTLs

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!


Meet 5:07 PM • Wed, Aug 21
Calico
Calico
Join or start a meet
4:00 PM
Call a phone


Calico

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!!

Presenter Form

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 Mod	https://doi.org/10.48550/arXiv.2306.15794
11/19	Cristina Martin Linares	Machine-guided design of cell-type-targeting cis-r	https://doi.org/10.1038/s41586-024-08070-z
12/03	Eduarda Vaz	Effective gene expression prediction from sequen	https://doi.org/10.1038/s41592-021-01252-x
1/7	Gus Fridell	Applying interpretable machine learning in compu	https://doi.org/10.1038/s41592-024-02359-7
01/21	Stephen Hwang	Evolutionary-scale prediction of atomic-level prote	https://doi.org/0.1126/science.ade2574
02/04	Celine Hoh	A foundation model of transcription across huma	https://doi.org/10.1038/s41586-024-08391-z