SAND2024-02899C

Sandia National Laboratories

Exceptional service in the national interest

APPLIED MACHINE LEARNING FOR ELUCIDATING RELATIONSHIPS BETWEEN EPIGENOMIC REGULATORY RULES AND GENE EXPRESSION

BETWEEN FUNGAL SPECIES ACROSS PHYLOGENETIC DISTANCES

Presented by: Laura Weinstock, PhD

Sandia National Laboratories

March 13, 2024

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

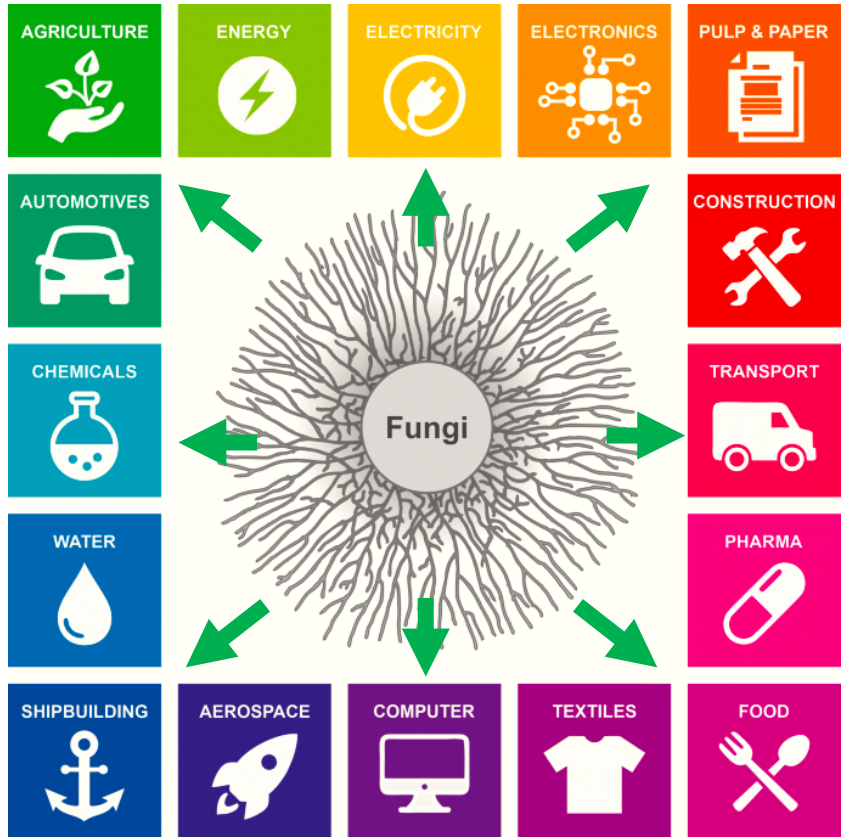
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-02899C

BACKGROUND: FUNGI IN BIOMANUFACTURING

- Engineered fungi are promising chassis for future sustainable biomanufacturing and bioproduction.
- Yeast alone (Nicholas Money, 2018):
  - 5% of USA's GDP (>$900B)
  - 3% of USA's workforce (>5M)
- Reliable regulation of the functionality in diverse fungi at scale remains a significant challenge to commercialization.

Improving biomanufacturing outcomes



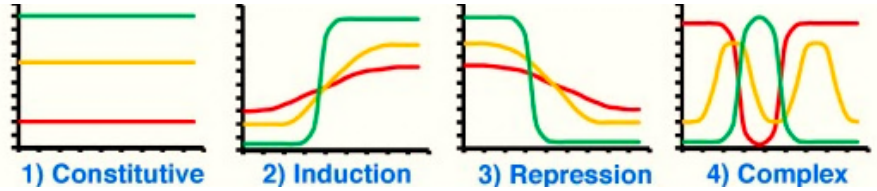
Meyer *et al.* (2020).

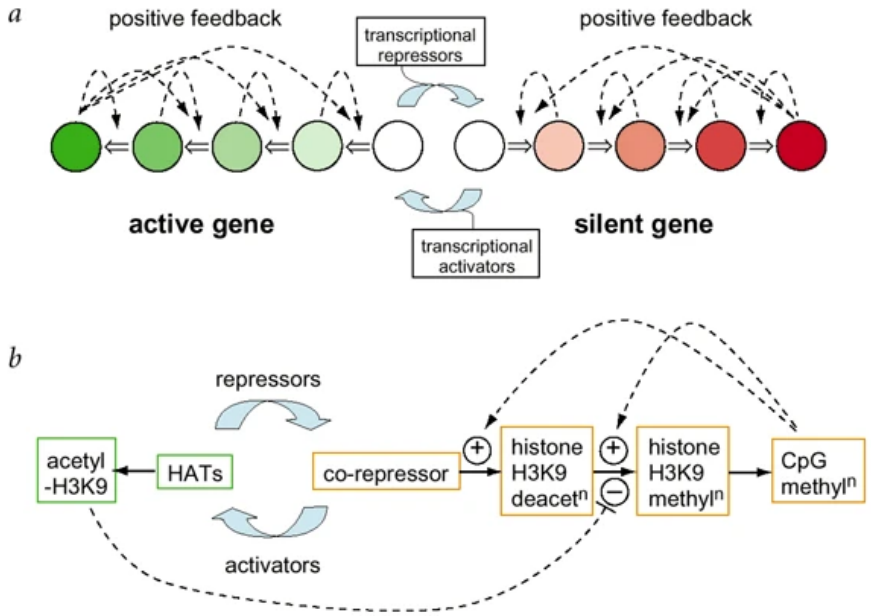
BACKGROUND: REGULATING BIOPRODUCTION

- Biomanufacturing requires dynamic, fine-tuned multicellular control
- Getting the exact desired properties of fungal chassis is challenging, and it is uneconomical to seek new chassis to overcome every limitation.
- Precise modulation of chassis attributes will support efficient biomanufacturing
- Epigenetic modifications play a crucial role in regulating gene expression and provide finer-resolution control of expression than genetic modifications

A systems problem needing a systems-level solution

In reality, you need more than on and off switches





Jaenisch & Bird. (2003). *Nat Gen.*

DEFINING THE PROBLEM FOR MODEL DEVELOPMENT

Can machine learning be used to determine relationships between epigenomics and gene expression in fungi?

- Fungal biomanufacturing requires: the right genes, available metabolic flux, controlled regulatory cross-talk:
  - Must find the right ‘knobs and levers’ of broad gene regulation for process optimization
  - Use of predictive models to assess process controllability, supplement with targeted experimentation

1

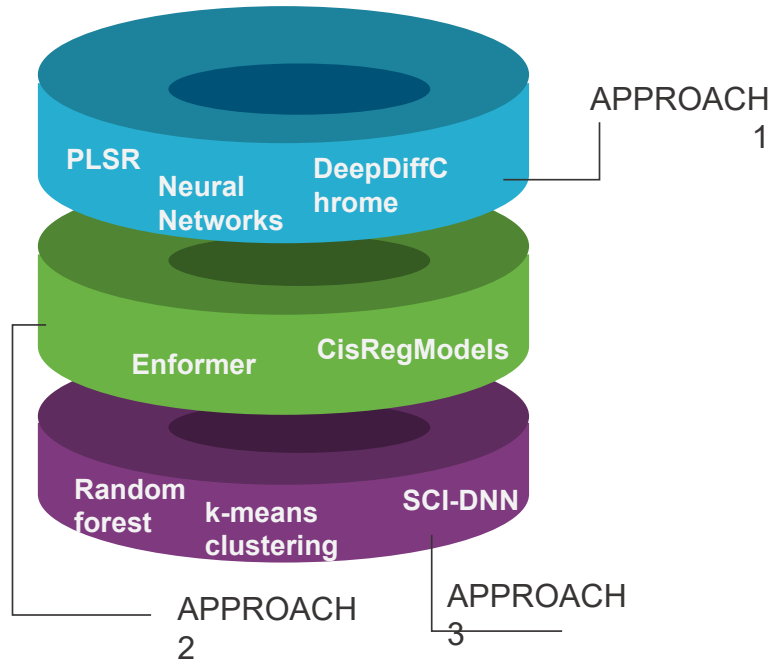
EPIGENETIC MODIFICATION TO PREDICT RNA EXPRESSION
Uses peri-gene (TSS +/- n bp) histone modifications and DNA methylation data to predict or regress against RNA expression.

2

DNA SEQUENCE TO PREDICT EPIGENETIC MODIFICATIONS
Uses native DNA sequence to predict or regress against epigenetic modification binding.

3

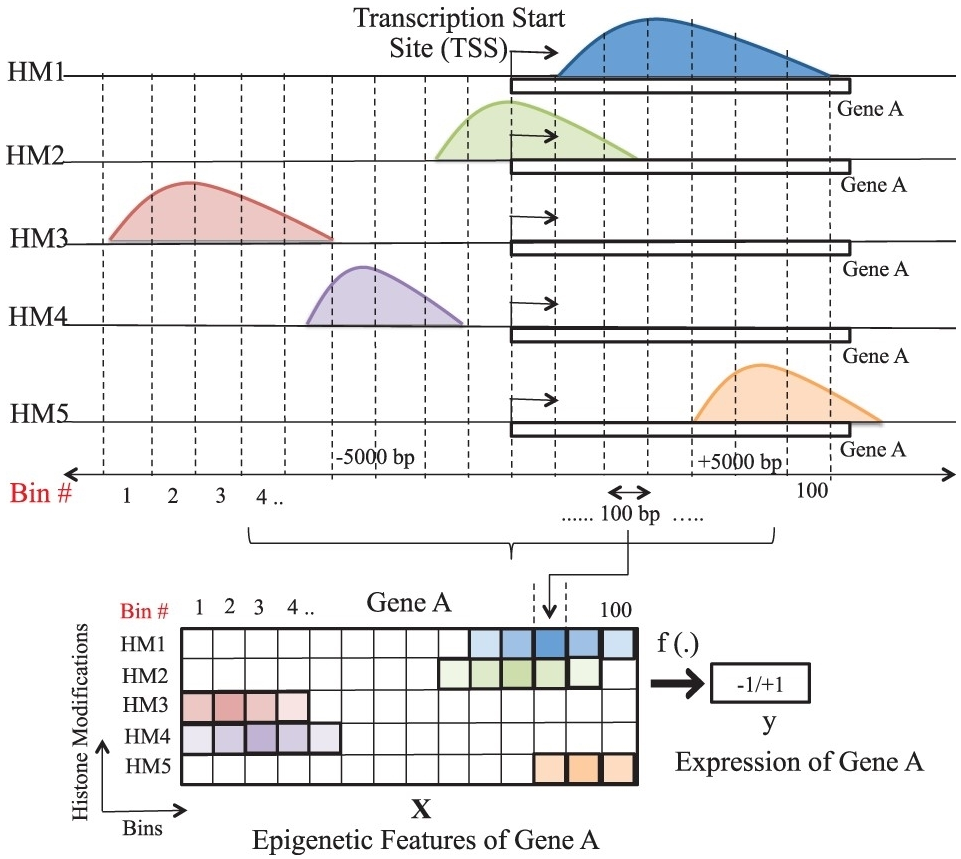
UNSUPERVISED LEARNING OF EPIGENETIC MODIFICATIONS AND/OR DNA SEQUENCE
Learn patterns in the epigenetic modification or DNA sequence data without expression information.



PREVIOUS EFFORTS USING MACHINE LEARNING TO UNDERSTAND GENE REGULATION

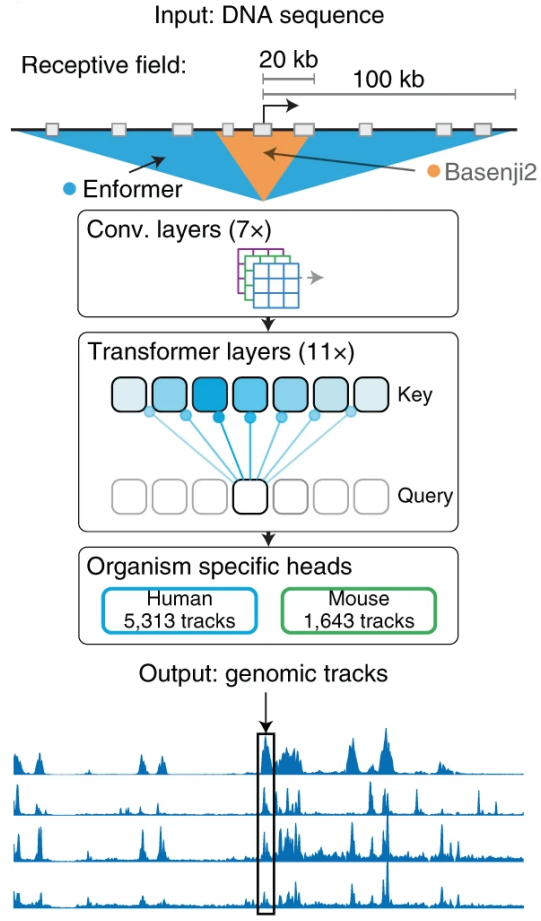
Convolution and attention are powerful tools to combine for extracting features from sequence and epigenetic signal information

DeepChrome: binary classification of gene expression using binned histone modification expression profiles



Singh *et al.* (2016) *Bioinformatics.*

Enformer: sequence-to-sequence prediction using convolutional and transformer layers

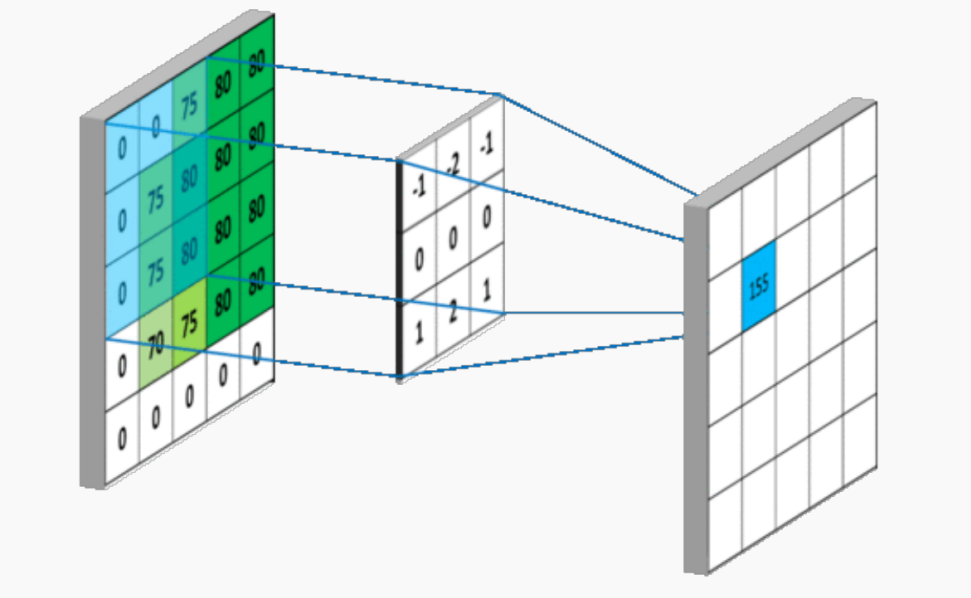


Avsec *et al.* (2021). *Nat Methods.*

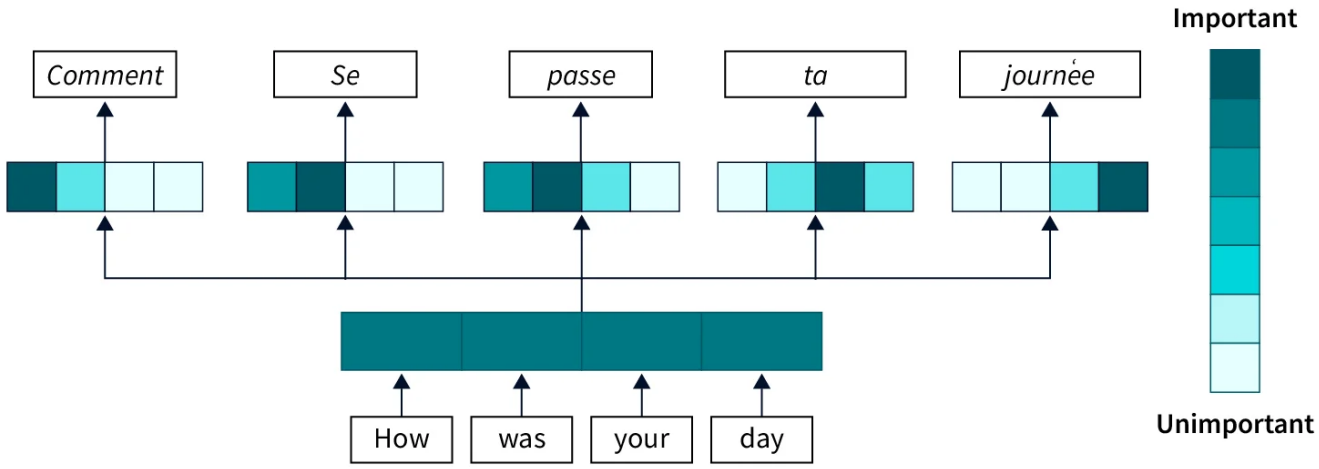
OVERVIEW OF KEY MACHINE LEARNING LAYERS

Convolution finds array of features, attention finds key features

Convolutional layers

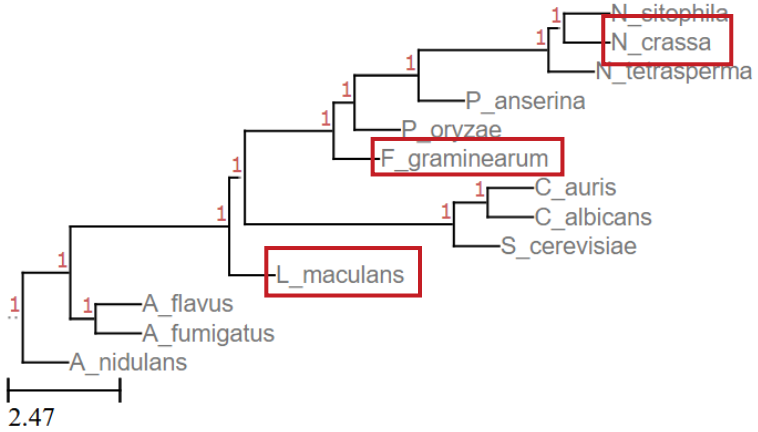


Attention layers

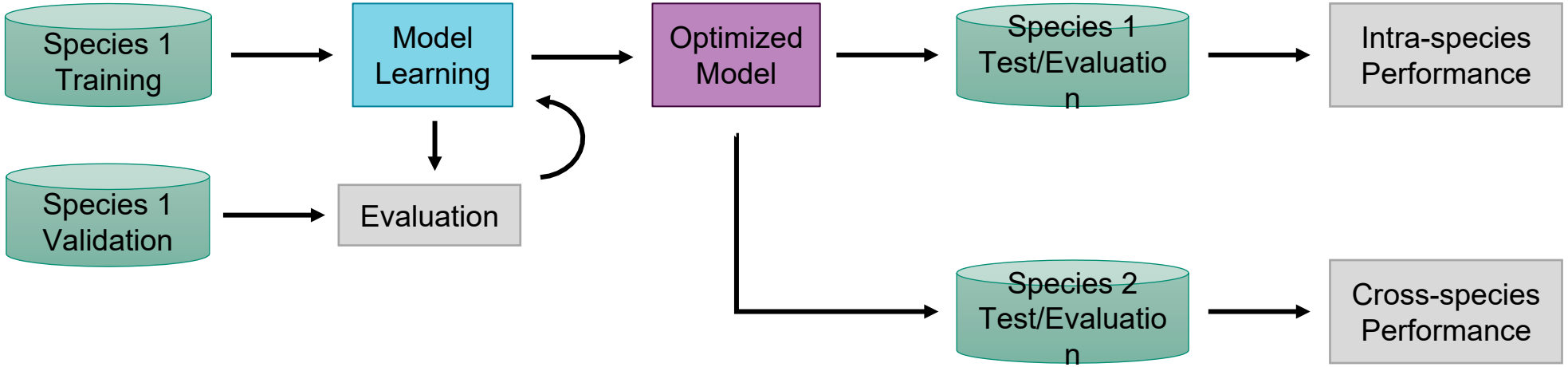


MODEL TRAINING AND EVALUATION STRATEGY

- Species selected based on phylogenetic distance and overlapping epigenetic modification data availability

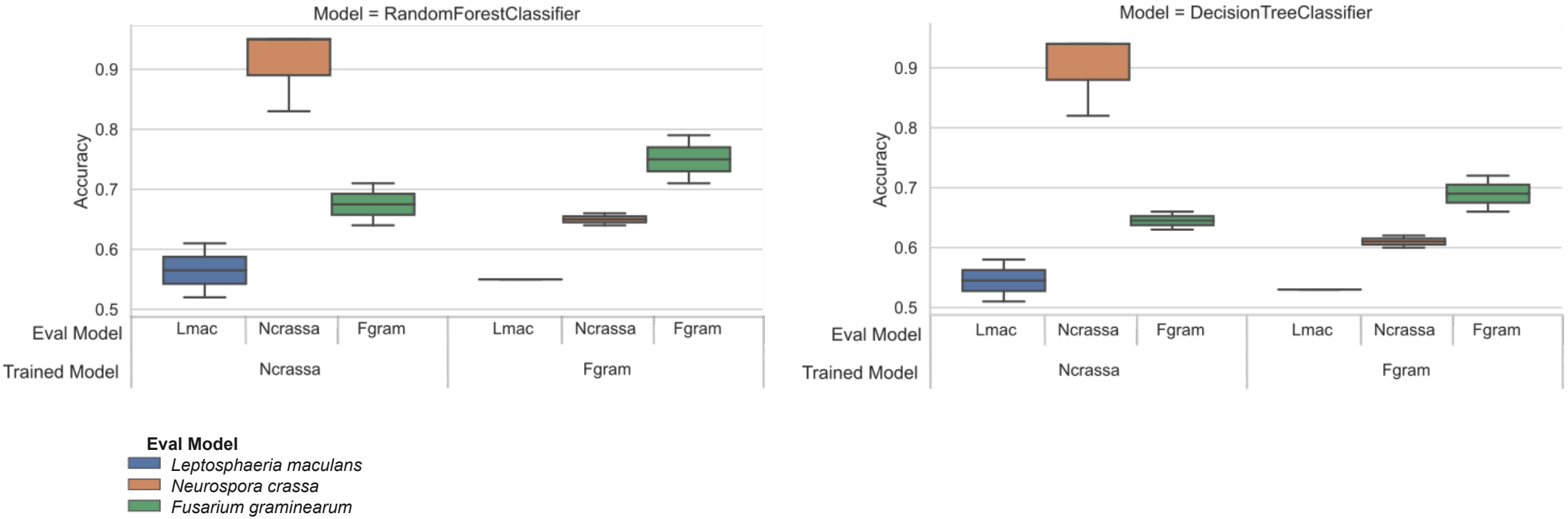


- Training and testing data → epigenetic modification data
- Predicted values → RNA expression



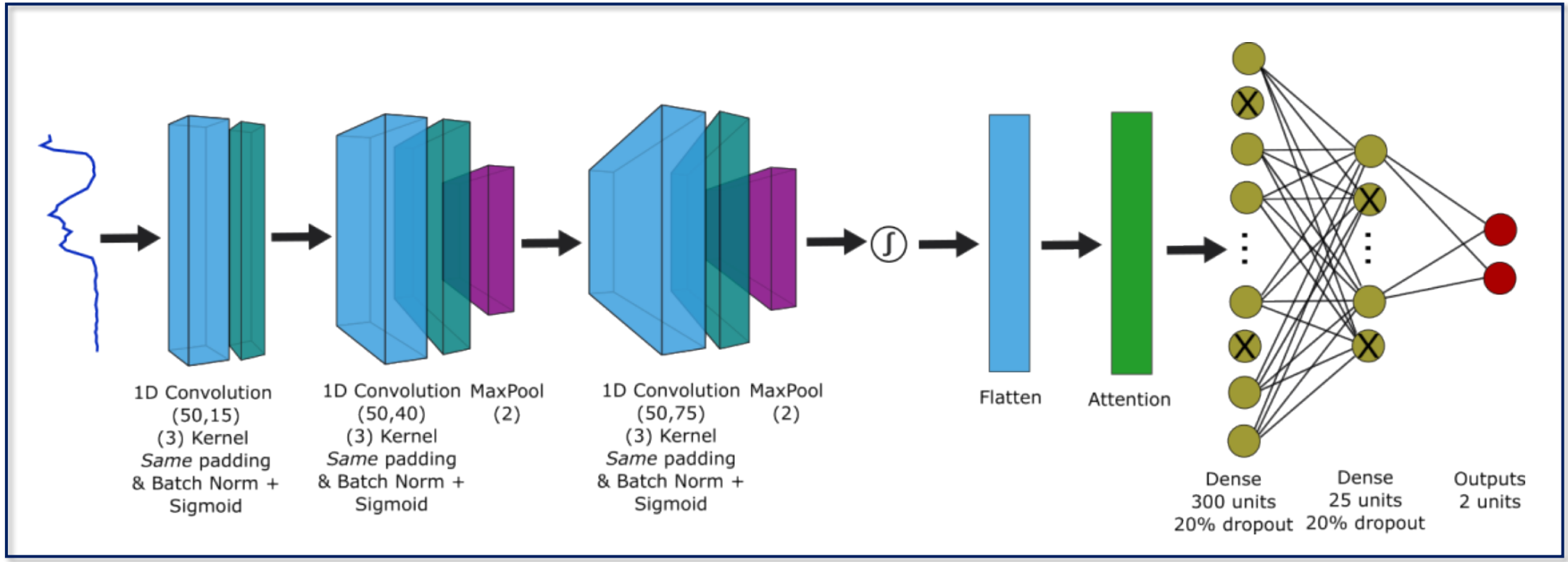
SHALLOW MODEL RESULTS: CROSS-SPECIES EVALUATION

Candidate tree-based models do not maintain performance in cross-species prediction.



DEEP LEARNING MODEL ARCHITECTURE

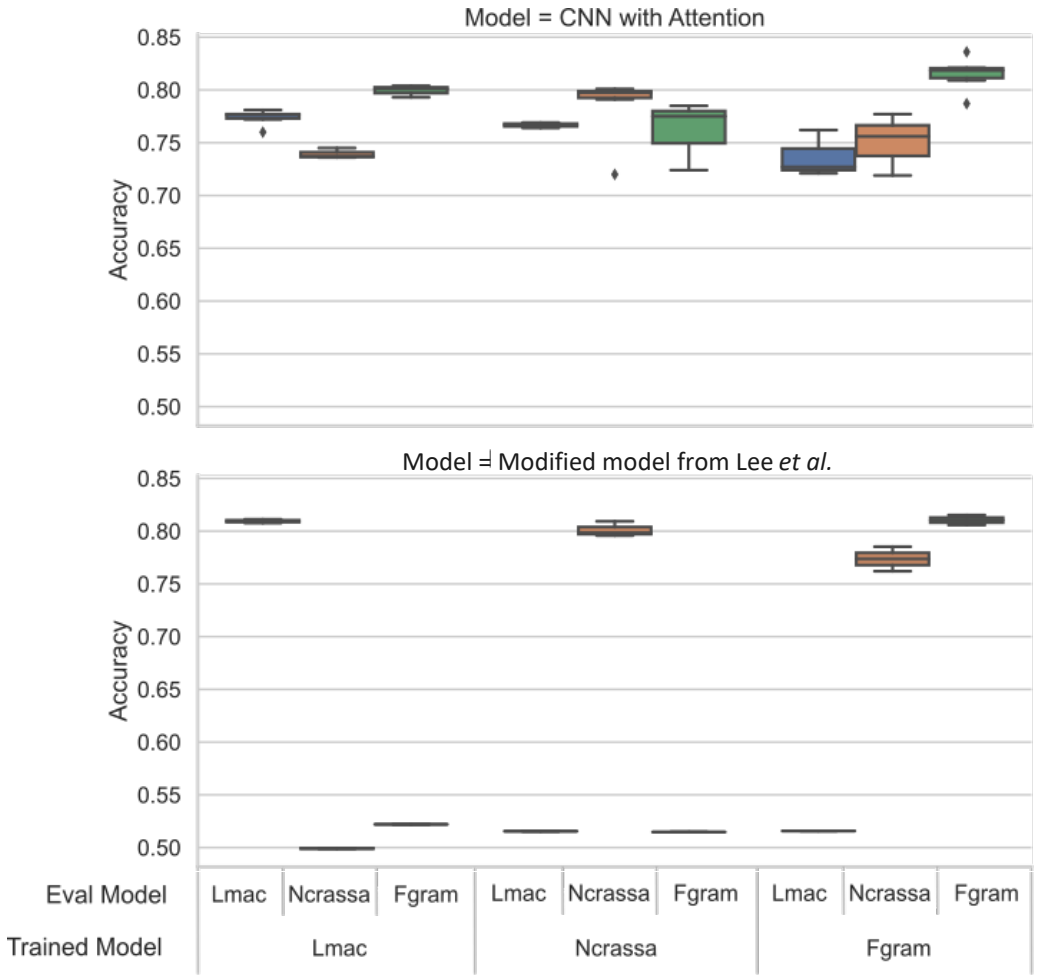
Custom model with Convolution with MaxPooling, Attention, and Fully Connected layers uses epigenetic signal profile to predict 2-class gene expression

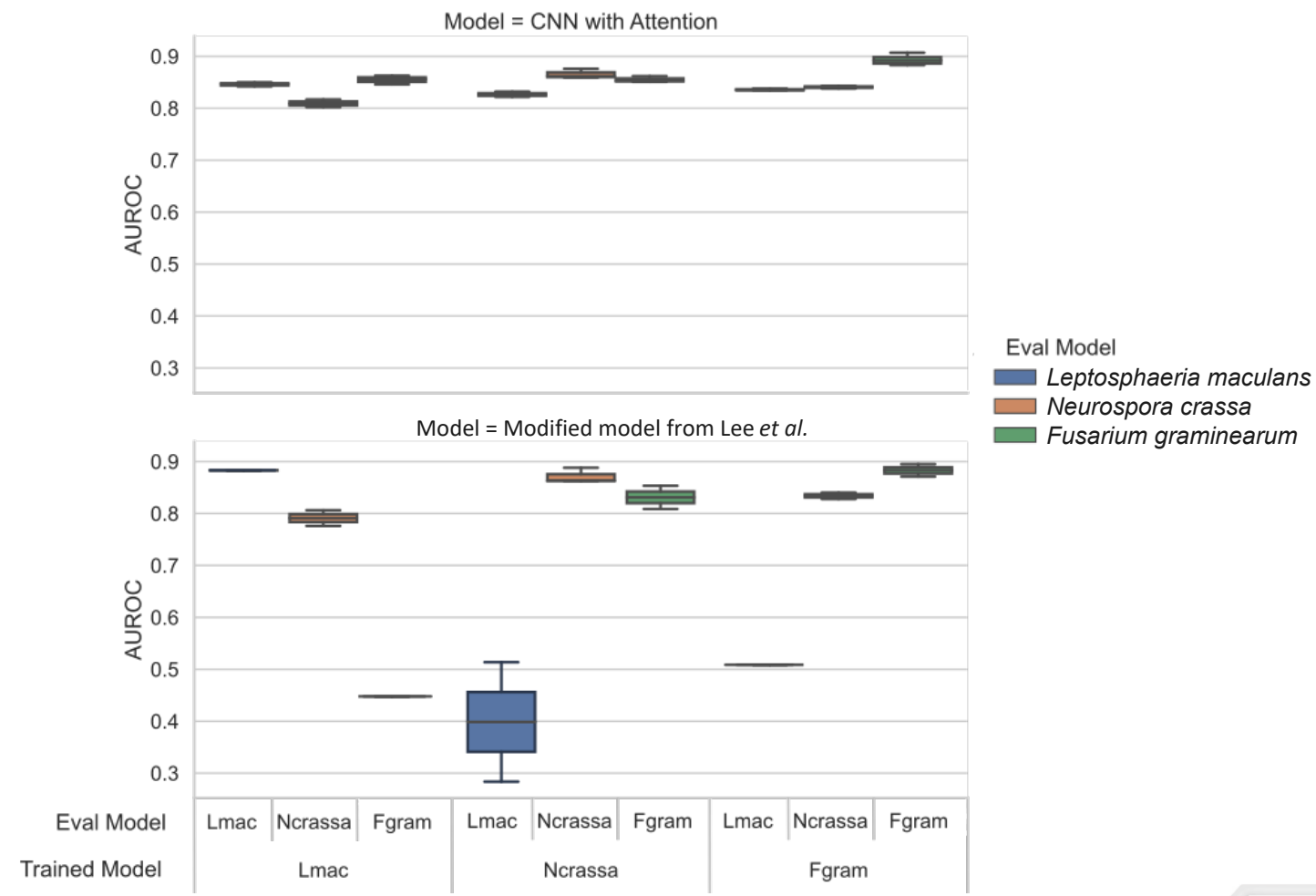


- Based on previous models for similar purposes in other species

DEEP LEARNING MODEL RESULTS

Custom CNN with Attention model outperforms benchmark models on cross-species prediction performance.

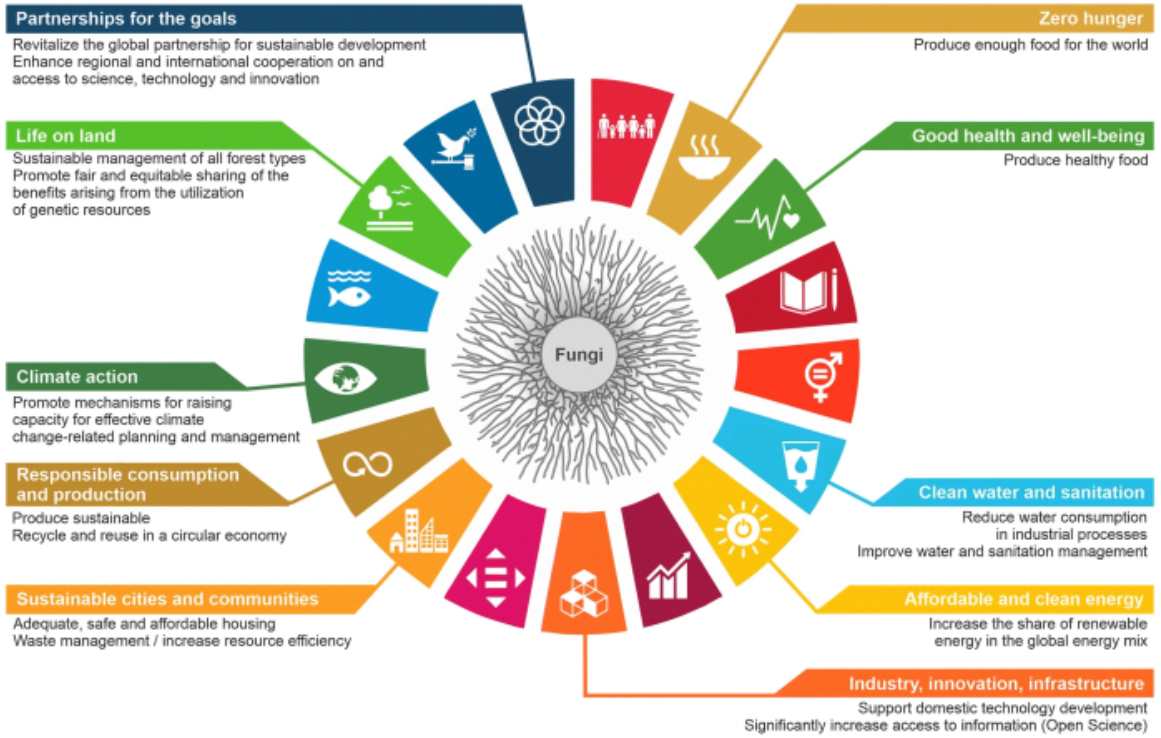




CONCLUSIONS

- Shallow models have limited ability to predict intra-species gene expression based on average epigenetic modification expression signals.
- Custom CNN + Attention model predicts intra- and cross-species gene expression based on epigenetic modifications signal profiles.
- We can use conservation of epigenetic mechanisms across related species to predict functional outcomes
- Limitations in predictive capacity may be due to underlying biological constraints or limited data availability.
- Highly modular gene expression enabled by predictive ML modeling of regulatory epigenetic modification rules is possible
- Support sustainable and secure bioenergy and biomanufacturing goals, and understanding of fungal disease.

Supporting sustainable circular bioeconomy goals



Meyer *et al.* (2020).

FUTURE DIRECTIONS

- Additional sequence, chromatin structure, and unsupervised prediction strategies
- Identify modification features that drive prediction
- Experimentally validate predicted modifications that regulate gene expression
- Develop modular epigenome editing tool for testing across fungi

1

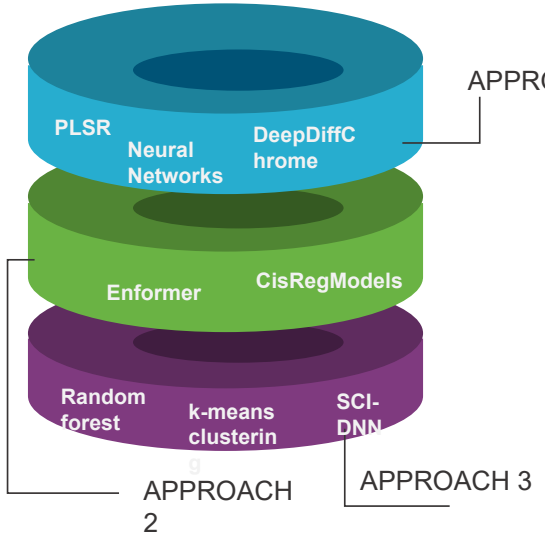
EPIGENETIC MODIFICATION TO PREDICT RNA EXPRESSION
Uses peri-gene (TSS +/- n bp) histone modifications and DNA methylation data to predict or regress against RNA expression.

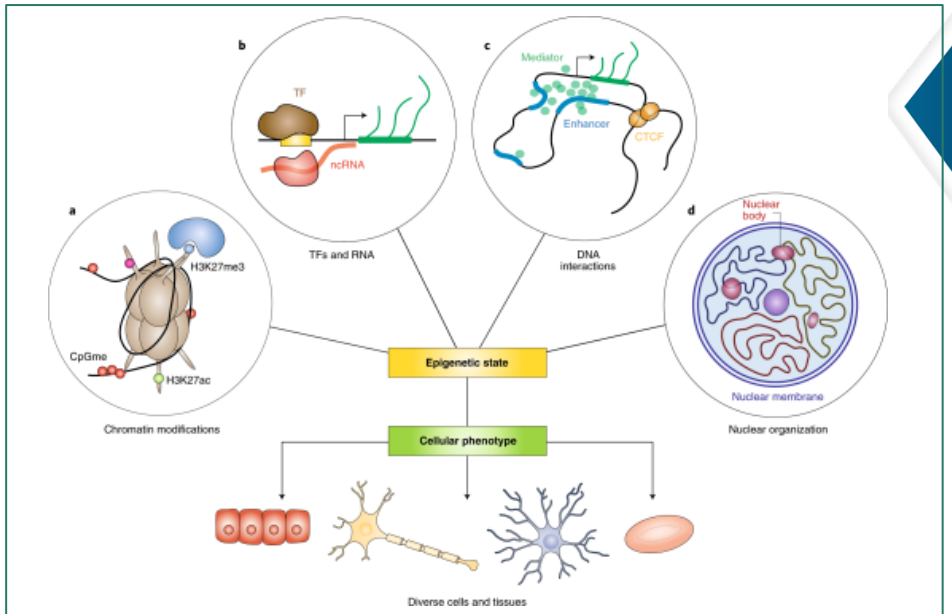
2

DNA SEQUENCE TO PREDICT EPIGENETIC MODIFICATIONS
Uses native DNA sequence to predict or regress against epigenetic modification binding.

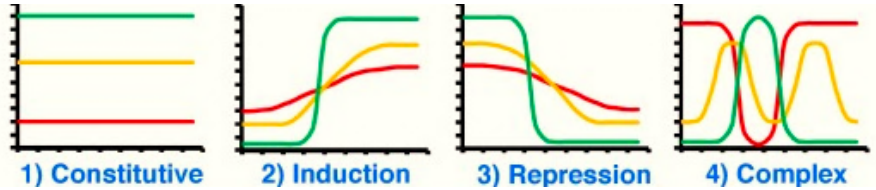
3

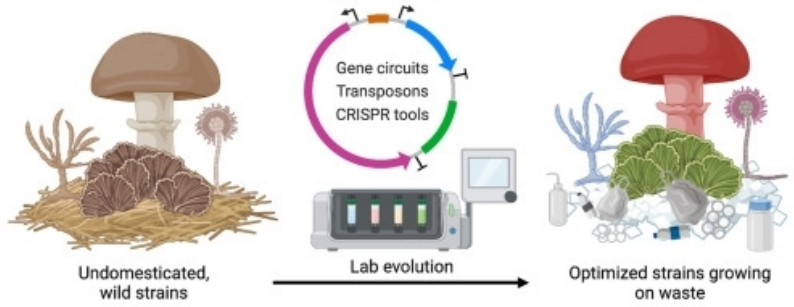
UNSUPERVISED LEARNING OF EPIGENETIC MODIFICATIONS AND/OR DNA SEQUENCE
Learn patterns in the epigenetic modification or DNA sequence data without expression information.





Nakamura et al. (2021). *Nat Cell Bio.*





Jo *et al.* (2023). *Materials Today Bio.*

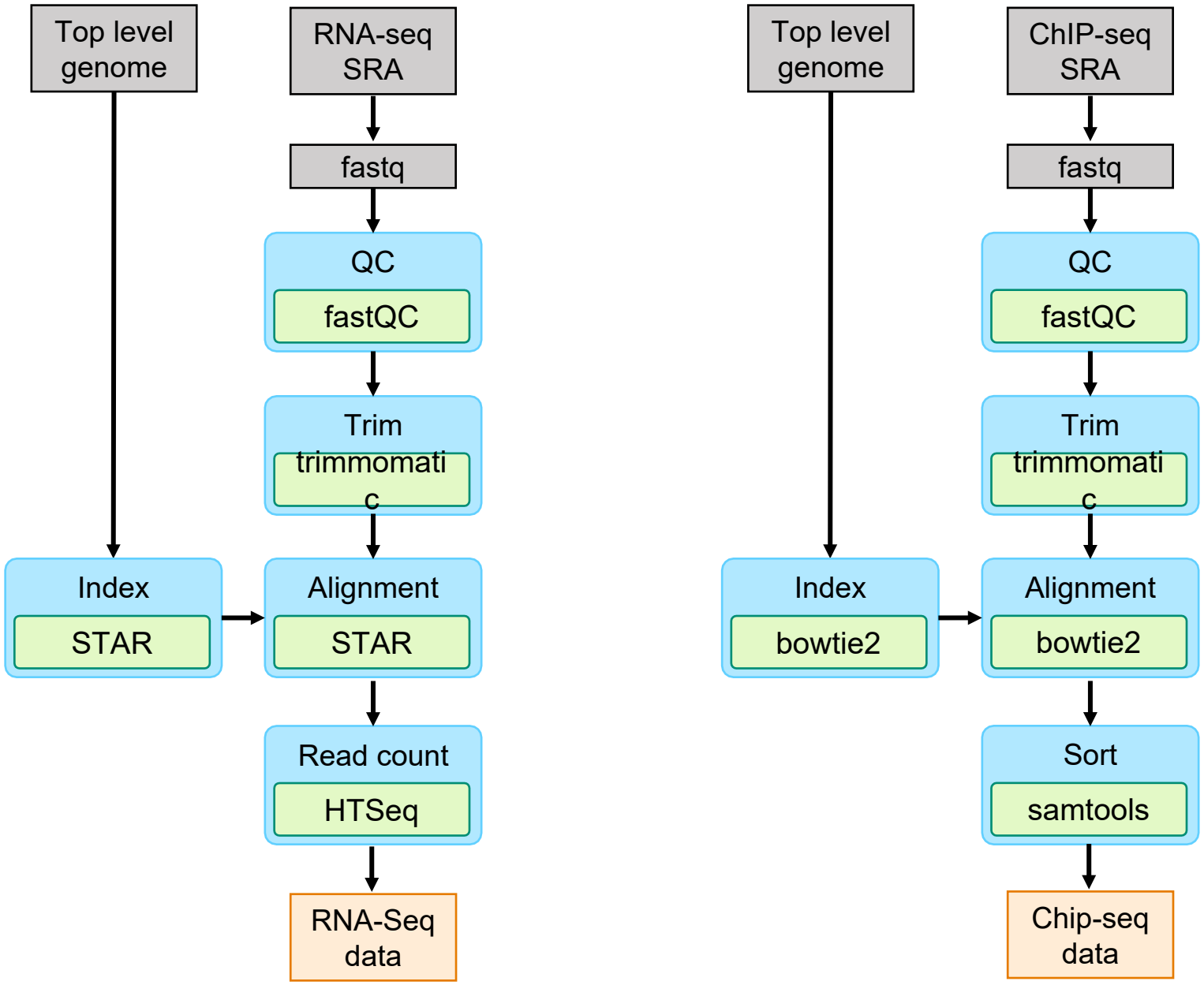
ACKNOWLEDGMENTS

- **PI: Raga Krishnakumar**

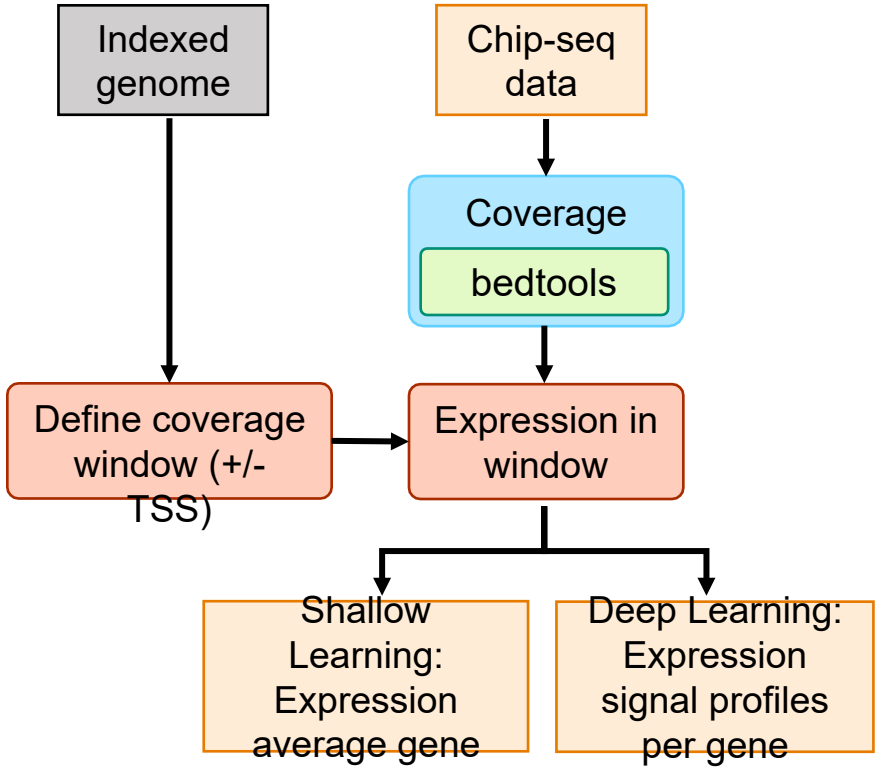
- Cameron Kundstadt
- Jenna Schambach
- Anna Fisher
- Matthew Hirakawa
- Elizabeth Koning
- Ethan Lee
- Wittney Mays
- Warren Davis
- Yooli Light
- Joshua Podlevsky

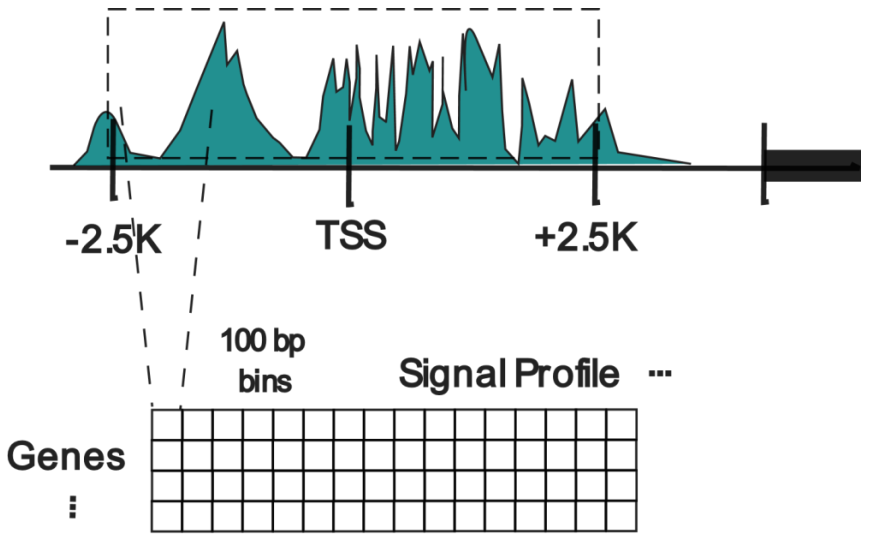
BACKUP

DATA PRE-PROCESSING PIPELINES



DATA PREPARATION





MINED EPIGENETIC DATA

Fgram matchup table

	Run	genotype	GEO Accession	Antibody	
1	SRR999608	WT	GSM1226424	Active Motif 39155	H3K27me3
4	SRR999611	WT	GSM1226427	abcam ab 9050	H3K36me3
6	SRR999613	WT	GSM1226429	Millipore 07-030	H3K4me2
10	SRR999617	WT	GSM1226433	abcam ab 8580	H3K4me3

N crassa matchup table

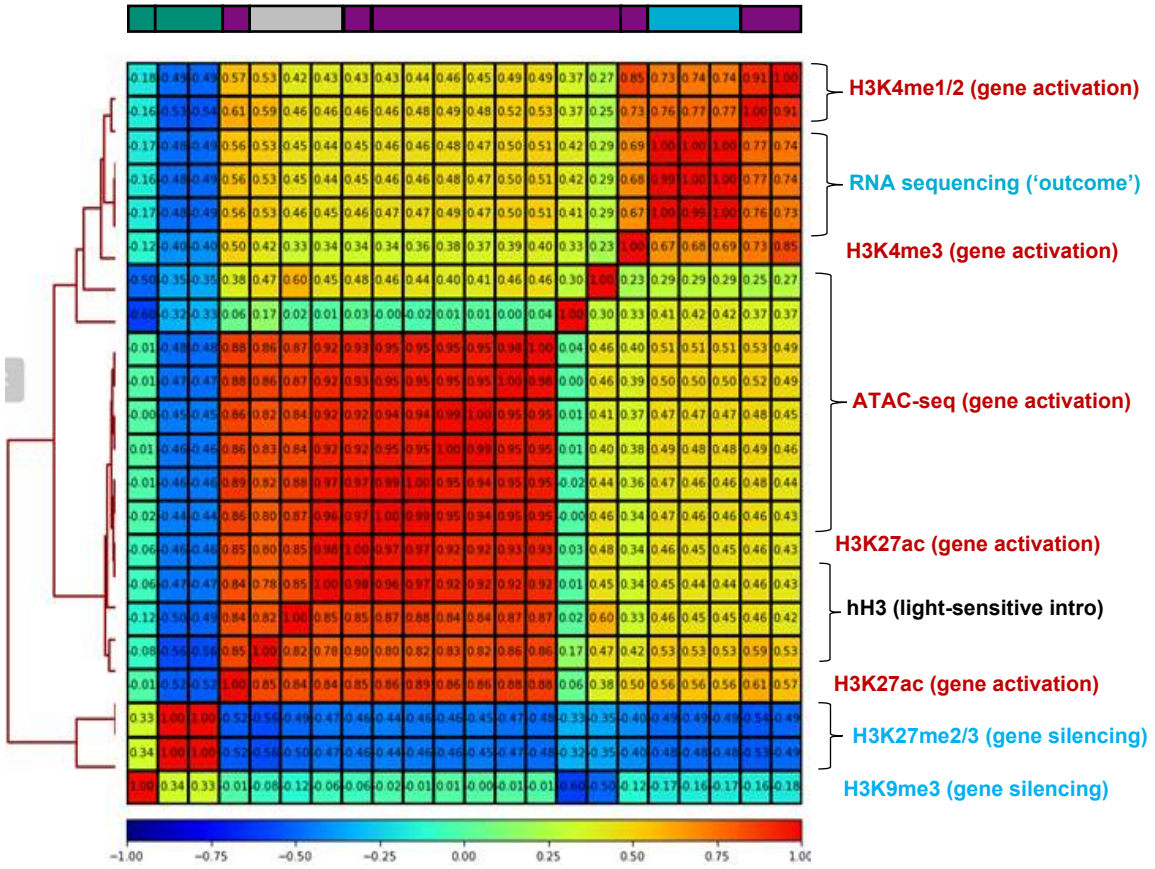
	Run	Assay Type	genotype	chip_antibody
9	SRR12229305	ChIP-Seq	WT	H3K27ac
10	SRR12229306	ChIP-Seq	WT	H3K4me1
11	SRR12229307	ChIP-Seq	WT	H3K4me2
12	SRR12229308	ChIP-Seq	WT	H3K4me3
13	SRR12229309	ChIP-Seq	WT	H3K36me3
14	SRR12229310	ChIP-Seq	WT	H3K9me3
18	SRR12229314	ChIP-Seq	WT	H3K27me2/3

HISTONE MODIFICATION FUNCTIONS

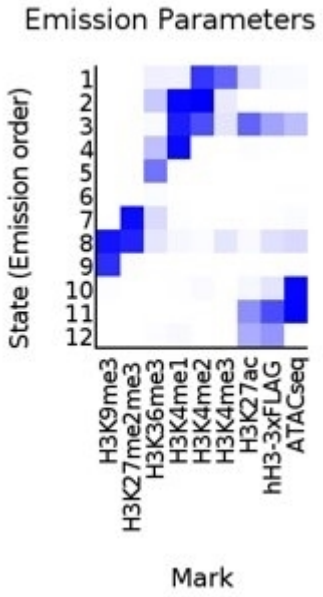
Histone/Position/Modification	Location	Effect	Enzyme
H3K4me2		Gene activation	Set1, MLL, Set7/9, SMYD3, LSD1, JAR1D1A
H3K4me3	5′ End of transcriptionally active genes	Gene activation	Set1, MLL, Set7/9, SMYD3, JAR1D1A
H3K9me	Euchromatin	Gene silencing	G9a; Suv91, StB1, PRD14, CLL8, GLP, Suv39h1, Suv39h2
H3K9me2	Euchromatin	Gene silencing	G9a; Suv91, StB1, PRD14, CLL8, GLP, Suv39h1, Suv39h2, JMJD2A
H3K9me3	Promoters and heterochromatin, Gene coding region	Gene silencing Gene activation	G9a; Suv91, StB1, PRD14, CLL8, GLP, Suv39h1, Suv39h2, JMJD2A
H3K27me1	Heterochromatin	Gene activation	
H3K27me2/3	Inactive-X chromosome, homeotic genes	Gene silencing	EZH2
H3K36me	Promoter	Not well characterized	JHDM1A
H3K36me2	Near double strand breaks, for repair	Gene silencing	NSD1, JMJD2A, JHDM1A
H3K36me3	3′ End of active genes. Marks exons.	Gene activation	JMJD2A
H3K79me2		Gene activation	Dot1L

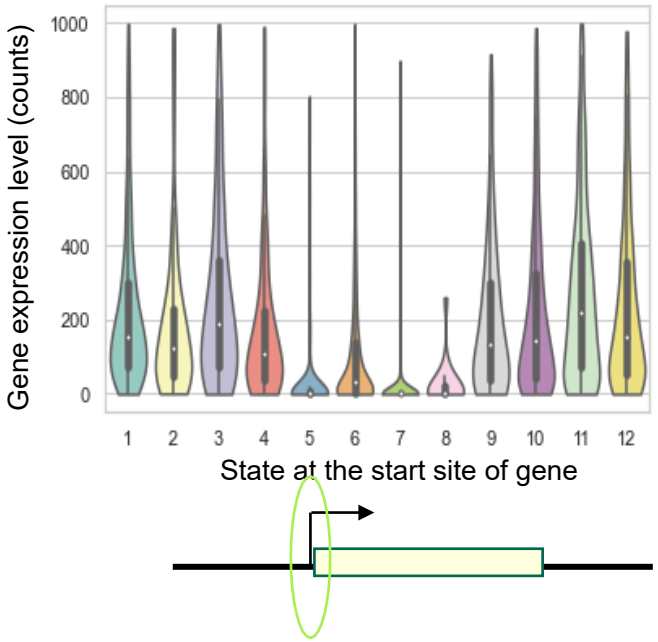
DATA VIZ TO CONFIRM EXPRESSION PATTERNS

deepTools to create a correlation clustergram



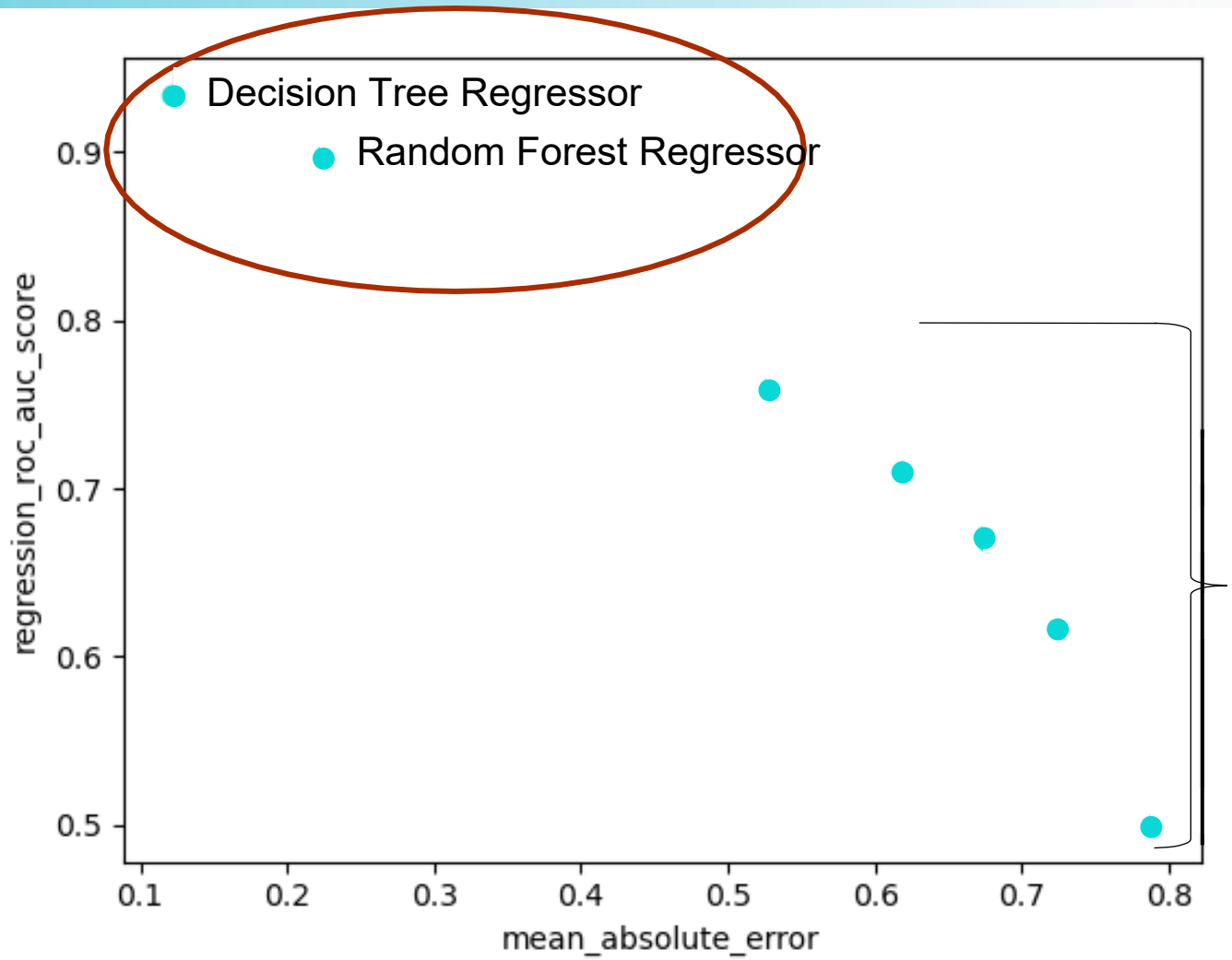
ChromHMM to observe combinatorial correlations





SHALLOW MODEL RESULTS: MODEL BATTERY SCREENING

Tree-based models outperform other shallow learning models on intra-species prediction using *N. crassa* data.



K-Nearest Neighbors Regressor
XGBoost Regressor
MLP Regressor
Gradient Boosting Regressor
Support Vector Regressor
Linear Regressor
Dummy Regressor

SHALLOW MODEL RESULTS: PRECISION