SAND2024-02718C

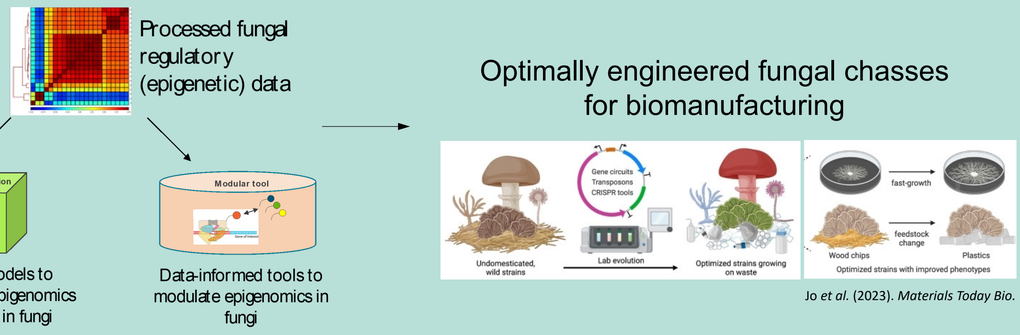
Sandia National Laboratories

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

BETWEEN FUNGAL SPECIES ACROSS PHYLOGENETIC DISTANCES

Laura Weinstock, Jenna Schambach, Anna Fisher, Cameron Kunstadt, Elizabeth Koning, Wittney Mays, and Raga Krishnakumar
Sandia National Laboratories

Abstract



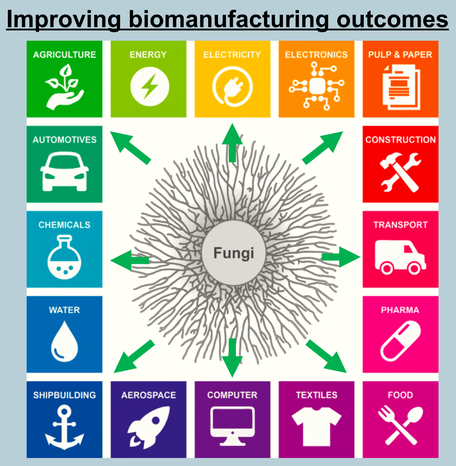
Jo *et al.* (2023). *Materials Today Bio.*

Can machine learning be used to determine relationships between epigenomics and gene expression in fungi, and can these predictions be validated experimentally?

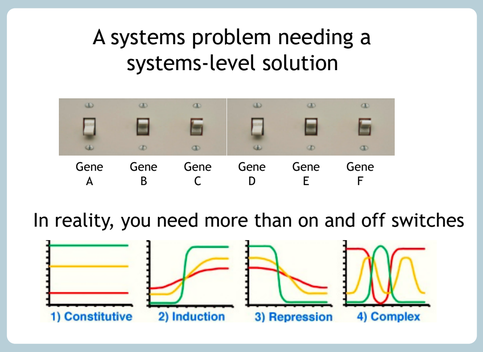
Background

- Engineered fungi are promising chassis for future sustainable biomanufacturing and bioproduction
- Reliable functional regulation of diverse fungi at scale remains a significant challenge to commercialization
- Optimized chassis require knowledge of conserved and variable controls of gene regulation across fungal species
  - Fungal biomanufacturing requires dynamic control of genes, metabolic flux, and regulatory cross talk
  - Modeling can help reduce the number of experimental conditions that need to be tested.

Improving biomanufacturing outcomes



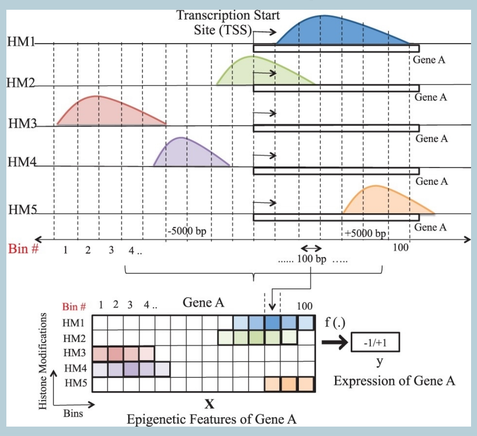
Meyer *et al.* (2020). *Microb Cell Fact.*



- Yeast value alone (Nicholas Money, 2018): 5% of USA's GDP (>$900B) & 3% of USA's workforce (>5M)
- Epigenetic modifications are critical in regulating gene expression
  - Provide finer-resolution control of expression than genetic modifications
- There remains limited understanding of how similar epigenetic modifications control gene expression across fungal species

Previous work: CNN for Epigenetic Prediction of Gene Expression

- Recent efforts have sought to understand the relationship between gene sequence epigenetic modifications, and gene expression within fungal species through the use of machine learning and deep learning (ML/DL) methods
- The discovery of conserved epigenetic modification design rules that control gene expression would greatly improve the efficiency of engineering across diverse fungal species

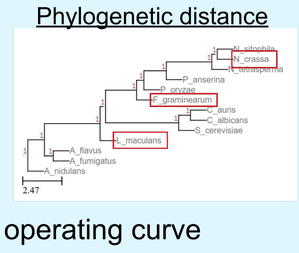


Singh *et al.* (2016) *Bioinformatics.*

Results

- Species used in analysis span genetic distances
  - Tree developed by custom phylogenetic algorithm

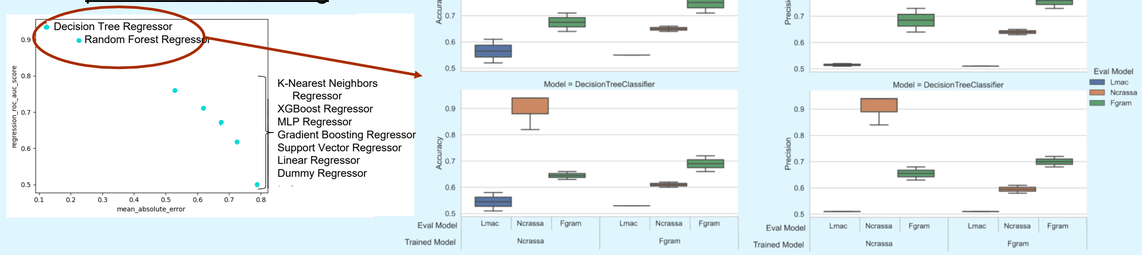
Phylogenetic distance



Shallow model results

- Tree-based models outperform other shallow models on mean absolute error (MAE) and area under the receiver operating curve (AUROC) metrics with *N. crassa* data
- Cross-species predictions using shallow learning models have below 60% accuracy

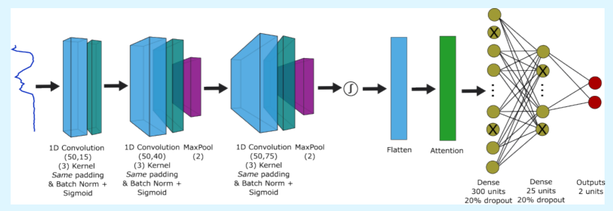
Intra-species shallow model performance screening

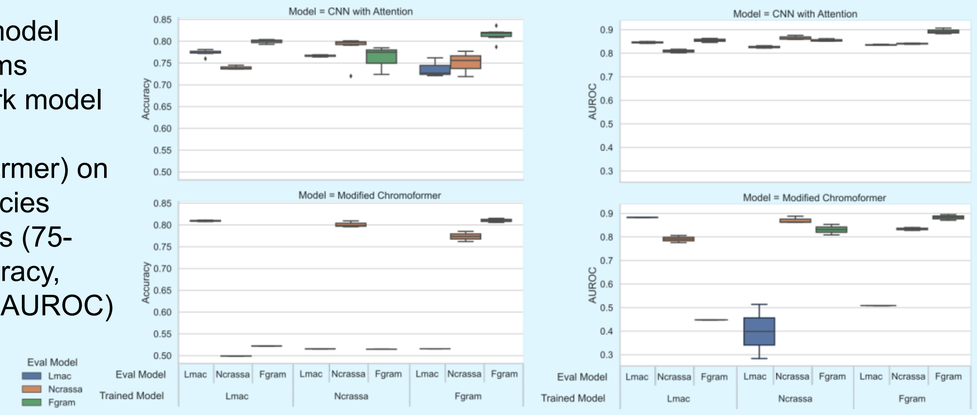


Legend: *Lmac = Leptosphaeria maculans; Ncrassa = Neurospora crassa; Fgram = Fusarium graminearum*

Deep learning model results

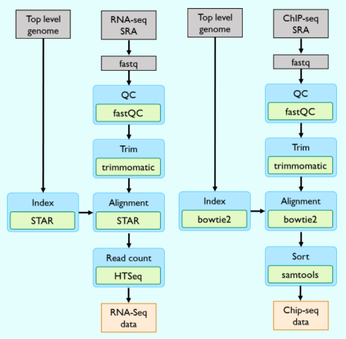
- Custom deep learning model architecture stacks three 1D CNN layers, a self-attention layer, and two fully-connected layers
- Model output is 2-class prediction of gene expression
- Custom model outperforms benchmark model (modified Chromoformer) on cross-species predictions (75-80% accuracy, higher on AUROC)





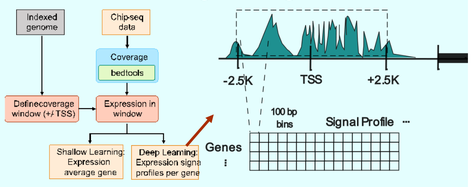
Methods

Data processing pipelines
(Mined available data)

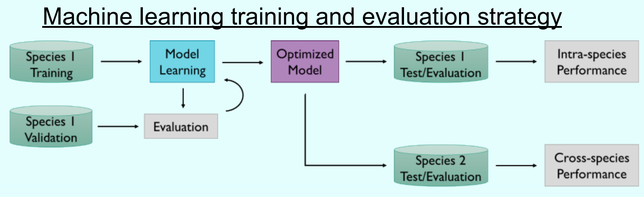


- Focus on using epigenetic modification as predictor for gene expression
- Explored both regression and classification models

Feature engineering



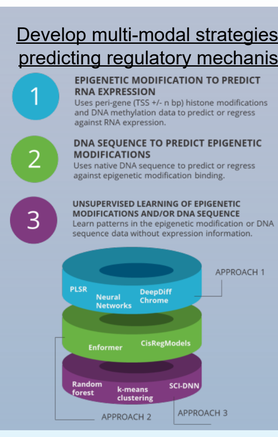
Machine learning training and evaluation strategy



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.
- There is therefore likely some conservation of epigenetic mechanisms across related species
- Limitations in predictive capacity may be due to underlying biological constraints or limited data availability.
- Highly modular gene expression through understanding and predictive modeling of regulatory epigenetic modification rules is possible through machine learning and will support sustainable and secure bioenergy and biomanufacturing goals.

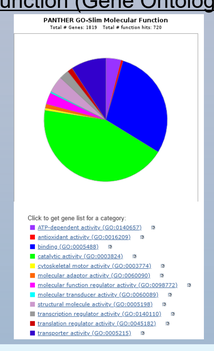
Develop multi-modal strategies for predicting regulatory mechanisms



Future Work

- Building more robust predictive models by adding data from published and in-house sequencing work
- Continue exploration into adaptability of tools across more species
- Identify biomanufacturing-pertinent scenarios to deploy our machine learning technology
- Identify and experimentally validate epigenetic modification profiles that regulate gene-of-interest expression
- Develop model-informed tools to modulate epigenomics in fungi

Explore Link Between Predictability and Gene Function (Gene Ontology)



U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

SAND2024-02718C

Sandia National Laboratories