# Exploring the utility of regulatory network-based machine learning for gene expression prediction in maize

Taylor H. Ferebee[1*], Edward S. Buckler[2,3,4]

**1** Department of Computational Biology, Cornell University, Ithaca, NY, USA 14853
**2** United States Department of Agriculture, Ithaca, NY, USA 14853
**3** Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA 14853
**4** Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

* Email: tf259@cornell.edu (TF)

## Abstract

Genomic selection and gene editing in crops could be enhanced by multi-species, mechanistic models predicting effects of changes in gene regulation. Current expression abundance prediction models require extensive computational resources, hard-to-measure species-specific training data, and often fail to incorporate data from multiple species. We hypothesize that gene expression prediction models that harness the regulatory network structure of *Arabidopsis thaliana* transcription factor-target gene interactions will improve on the present maize models. To this end, we collect 147 *Oryza sativa* and 99 *Sorghum bicolor* gene expression assays and assign them to maize family-based orthologous groups. Using three popular graph-based machine learning frameworks, including a shallow graph convolutional autoencoder, a deep graph convolutional autoencoder, and the inductive GraphSage strategy, we encode an *Arabidopsis thaliana* integrated gene regulatory network (iGRN) structure and TF gene expression values to predict gene expression both within and between species. We then evaluate the network methods against a partial least-squares baseline. We find that the baseline gives the best predictions within species, with Spearman correlations averaging between 0.74 and 0.78. The graph autoencoder methods were more variable with correlations between -0.1 and 0.65. In particular, the GraphSage and deep autoencoders performed the worst, and the shallow autoencoders performed the best. In the most challenging prediction context, where predictions were in new species and on genes that were not seen, we found that the shallow graph autoencoder framework averaged around 0.65. Unlike initial thoughts about preserved network structure improving gene expression predictions, this study shows that within-species predictions only need simple models, such as partial least squares, to capture expression variations. In cross-species predictions, the best model is often a more complex strategy utilizing regulatory network structure and other studies' expressions.

## Introduction

In recent years, gene expression prediction models have been introduced as a computational tool to identify key regulatory elements that influence the effects of disease variants [1] or describe the nature of regulatory grammars [2–4]. Because of the ability of the plant genetics community to study variations of complex traits, understand the evolutionary processes behind these traits, and target sites for gene editing, expression prediction holds a lot of research potential. To gather the necessary resolution to explore these concepts, researchers generally focus on the systematic role of regulatory machinery that facilitates variation in gene expression over space and time [5,6]. Gene regulation requires elaborate interplay between linked (*cis*) and distant (*trans*) regulatory elements. In the *cis*-case, DNA sequence-specific transcription factors bind to the DNA to modulate nearby gene expression In addition

to transcription factors, untranslated regions of RNA modulate gene expression through length variation and through offering variants of mRNA from the same gene [7]. In plants, mutations in these *cis*-regulatory regions contribute to domestication [8, 9], adaptation [10, 11], and, in some cases, are the target of allele-generative pursuits in next-generation breeding approaches [12, 13].

Across plant species, the relationship between genes and transcription factors has been one of the main focus for understanding the transcriptomic patterning. For example, just under 50% of orthologous genes between *Arabidopsis thaliana* and *Oryza sativa*, species that diverged about 150 million years ago [14], show similar transcriptomic responses while under abiotic stress [15, 16]. Further, the different responses were due to changes in regulatory elements, such as transcription factor expression. It has also been shown that tissue-specific transcription factor activity induces patterning in gene target specificity [17]. Due to the multidimensional relationship between tissue, experimental conditions, and species, few studies have attempted to model gene expression patterns between species.

Gene regulatory networks (GRNs) represent the emergence of transcriptomic patterning by providing a graph representation of the TF to target gene relationships. GRNs have been crucial to understanding not only the interactions between genes, but also the coordinated targeting of specific genes by multiple transcription factors [18, 19]. Assays, such as CHIP-seq and ATAC-seq, inform the majority of the transcription factor-gene relationships in these networks, and with sufficient experimental data, these networks can also integrate the 3-D chromatin structure and dynamics to reveal complex, time-sensitive, and species-specific players in gene regulation [20–23]. Unfortunately, these combinations of experiments can be expensive, hard to measure, and computationally difficult to utilize.

There have been several recent successes in overcoming the limitations of predicting gene expressions or other large-scale datasets using deep learning. Generally, deep learning refers to computational methods that aim to learn a hierarchical representation of data by functionally relating the data in many layers [24]. In plant data, these models have learned to predict chromatin state from sequence [25], identify and classify key stress-responsive genes [26], and detect seasonal changes across fields [27]. In addition to these successes, many state-of-the-art models take in not only numerical data but also the structure of the relationships between the data. Traditionally, non-Euclidean graphical structures represent these complex, multidimensional relationships. The models, graph neural networks, use the same general machinery that many deep learning methods use, but take special care to ensure that the representation of the data is preserved through each abstraction. Graph neural networks have shown particular success in exploring disease and pest relationships in *Arabidopsis thaliana*, inferring protein-protein interactions key to biological functions [28], and finally, inferring gene expression in mice, given co-expression and gene interaction networks [29].

Graph neural networks solve problems typically associated with complex, non-Euclidean dimension reduction, image extraction, and link or node prediction. Strategies to solve these problems can be categorized into spectral and spatial-based approaches. Spectral approaches, such as the common graph convolutional network (GCN), focus primarily on the graph spectrum, or the set of eigenvectors and eigenvalues, associated with the adjacency matrix of the graph [30]. In these cases, the graph spectrum is used to map its structure to a Euclidean space for which convolutional operations can be performed similarly to convolutional neural networks. Spatial approaches work at the node and neighboring node-edge level to aggregate information across nodes. Due to this aggregation, many special methods are amenable to large networks. For example, GraphSage [31], aggregates node data in order to create low dimensional representations of unseen data. This aggregation to embedding step makes GraphSage a natural choice for large biological studies, and researchers use the method to explore large-scale predictions for large-scale protein-protein interactions [32] and drug-disease relationships [33].

However, to create the models, care must be taken to preserve the broad evolutionary relationships. Plant systems, because of genome duplication and adaptive responses to selection pressures, generally have larger expansion of transcription factor families than animal ones [34]. These expansions, however, often act in parallel, which preserves researchers' aim of accurate and precise global quantification of TF-gene relationships. In *Arabidopsis thaliana*, researchers mined thousands of experimental and computational data to find the most likely interactions between genes and transcription factors [35]. Their goal was to use machine learning machinery to infer a network that was biologically supported and

could be used to functionally validate biological hypotheses. The resulting network identified key regulators of reactive oxygen stress response [35]. In a cross-species comparison of gene regulatory networks, researchers inferred stress-responsive sub-networks that are evolutionarily conserved between *Marchantia polymorpha* and *Arabidopsis thaliana*, allowing for new hypotheses and methods that can provide insight into divergent species' gene regulatory mechanisms [36]. In this work, we use popular network embedding-based deep learning methods to predict gene expression in maize experiments. We design the machine learning task around the embedding of a regulatory network structure from the model species, *Arabidopsis thaliana*. Upon training the model on *Zea mays*, *Arabidopsis thaliana*, and *Sorghum bicolor*, we observe improved correlations between predictions and observed values relative to previous regression and machine learning models. Finally, we examine the characteristics of the regulatory network that may affect the prediction of gene expression.

# Materials and methods

## Raw data and preprocessing

Gene expression data for *Zea mays*, *Sorghum bicolor,* and *Oryza sativa* was downloaded and processed as in [37]. In summary, datasets were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), trimmed and quality checked with FastQC and Sickle, and aligned to Maize V4, Sorghum 3.1.1, and *Oryza sativa* V7 genomes with HISAT2 [37], and read counts were normalized to transcripts per million with Stringtie. Identifications of transcription factors for *Zea mays*, *Sorghum bicolor*, and *Oryza sativa* were downloaded from the Plant Transcription Factor Database website [38]. Orthologous groups based on gene families were recovered from the PLAZA 4.5 database [39]. The Arabidopsis integrated gene regulatory network (iGRN) was downloaded from the supplementary of [35]. The handling of the data sets was done with custom R scripts that utilize the data.table and dplyr packages [40] were used to process the data sets. All data, including code to generate input formats, are enclosed in the code repository on Zenodo (doi: 10.5281/zenodo.7194372). Supplemental Figure 1 shows the flow chart of the data preprocessing.

### Orthogroup assignments

For each species, gene expression matrices were transformed into orthogroup matrices by cross-referencing the PLAZA gene family-based orthogroups [39]. To create an orthogroup expression matrix, the mean orthogroup expression per experiment was taken across replicates. Expression levels were log-transformed and normalized by the standard deviation of an orthogroup's expression across experiments. The Arabidopsis iGRN nodes were similarly transformed into orthogroups. The nodes and their corresponding edges were collapsed to create a network with orthogroups as the nodes rather than the original Arabidopsis genes. The resulting network contained 756 transcription factors and 18771 target genes. From this list, an adjacency matrix was created to represent the network's connections.

### Model splitting

Training and testing splits were made at the orthogroup and experimental levels. In the case we predict expression on orthogroups that the model has seen, we perform an 80:20 train-test split across experiments. To ensure there is minimal experiment-experiment leakage, we completely mask these values during the training process. In the case we predict expression on genes that have not been seen by the model, we do a two part procedure for training. We first perform an 80:20 split on the experiments. Then, we perform another 80:20 split on the orthogroups. The final predictions are completed on the unseen genes within the test experiments.

## Model development

Expression abundance predictions were made with two architectures, structural embedding with random forest (graph autoencoder models) [33] and a partial least squares model. All models structural embedding models were constructed using Python 3 with the Pytorch Geometric [41] and Scikit-learn libraries. Models were tested in 4 scenarios: 1) within-species prediction of the unseen expression profile when all genes have been seen by the model, 2) prediction in maize from a single species when all genes have been seen by the model, 3) prediction of unseen expression profile when all genes have not been seen by the model, and 4) prediction of maize expression profiles from all species.

Expression predictions were done with two architectures: structural embedding with random forest prediction layer (graph autoencoder models) and a baseline partial least squares model. We tested three main architectures for the GAE-a shallow graph convolutions, deep graph convolutions, and Graph Sage. The random forest layer was constructed with $\lfloor \frac{n}{3} \rfloor$ decision trees, as suggested by [42]. All structural embedding models were constructed in Python 3 with the Pytorch Geometric library [41]. Predictions were made within each related species and into maize (target species). We tested both the PLS and structural embedding random forest architectures on a new test set containing genes which the model has been trained on. In structural embedding models, we also tested in situations where the predicted genes' expression profiles are masked during training (Supplementary Figure 1).

### Partial least squares baseline

To establish a baseline performance, we use a partial least squares model. In this case, consider $\mathbf{X}_{nonTF}$ to be a gene expression matrix consisting of orthogroups not associated with transcription factors. Similarly, let $\mathbf{X}_{TF}$ be a gene expression matrix of only transcription factor orthogroups. We want to predict each column of the non-transcription factor, $\mathbf{X}_{nonTF}$, from a combination of the columns of $\mathbf{X}_{TF}$. Partial least squares transforms the space of predictors in $\mathbf{X}_{TF}$ to orthogonal variables, which we then use for prediction (Fig1). We chose to use the PLS2 algorithm which both $\mathbf{Y}$'s and $\mathbf{X}_{TF}$s variable spaces are simultaneously modeled. Here, the underlying model, with components $T$ and $U$, loading $P$ and $Q$, and error terms $E$ and $F$ becomes,

$$\mathbf{X}_{TF} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

$$\mathbf{X}_{nonTF} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2}$$

We aim to maximize the covariance of $T$ and $U$ so that we can perform a regression such that

$$\tilde{\mathbf{X}}_{nonTF} = \mathbf{X}_{TF}\tilde{\beta} + \beta_0 \tag{3}$$

To find the optimal number of components cross-validation was performed, resulting in choosing 3 components.

### Graph autoencoder with random forest model

To create a low-dimensional representation of the expression and network data, we use a standard non-variational graph autoencoder, as seen in [29]. In this method, we consider input expression $\mathbf{X}_{tot}$ as a matrix with $N$ genes and $Q$ experiments and a graph $G$, as the Arabidopsis orthogroup iGRN. The encoder block of the autoencoder maps the regulatory network and gene expression features into a latent embedding (Fig 1). Then, in order to determine the validity of the embedding, the model decodes this embedding back into the input space and the cross-entropy loss of the reconstruction is measured. The predicted expression values are obtained using this latent representation $\mathbf{Z}$ such that

$$\tilde{\mathbf{X}}_{Tot} = f(\mathbf{Z}) \tag{4}$$

where $\tilde{\mathbf{X}}_{tot}$ is the predicted matrix of the input expression values and f denotes a random forest with a max depth scaled by $\lfloor Q \rfloor$ with $\lfloor \rfloor$ denoting the floor operation.

To establish performance within this framework, we consider the use of two major encoding schemas, induction-based and convolution-based methods. For the induction-based method, we use the GraphSAGE layer available in PyTorch Geometric [41]. The GraphSAGE algorithm works by propagating hidden representations of the nodes (genes) to their neighbors [31]. Here, by stochastic descent, the combination of weight matrices and the mean aggregation function is tuned so that genes with similar behavior have similar representations within the resulting embedding [31]. Thus, a node, $i$, representation, $h$, for a neighborhood size of $k$ is given by,

$$h_i^k = W_1(h_i^{k-1}) + W_2 aggregate_{j \in N(i) \cup i}(h_j^{k-1}) \tag{5}$$

where the *aggregate* function is given by mean pooling and $W_1$ and $W_2$ are trainable weight matrices.

The convolutional framework takes into account the high order of structure in neighborhoods around nodes through the use of convolutional layers [43]. Hidden representations of those high-order structures are then merged together to form each node's hidden state [43]. A single layer can be seen as

$$h_i^k = \sum_{j \in N(i) \cup i} \frac{1}{\sqrt{deg(i)} * \sqrt{deg(j)}} . (W h_j^{k-1}) \tag{6}$$

where $W$ is a trainable weight matrix, $deg(i)$ and $deg(j)$ are the number of neighbors of node $i$ and $j$ respectively [43]. For the shallow autoencoder, we choose to use a single convolutional layer with hidden dimension of size 32. For the deep autoencoder, we choose three convolutional layers with a hidden dimension size of 32.

## Model evaluation

For each model task, we will use the Spearman rank correlation to compare the expression level of the observed and predicted orthogroup. Spearman's is given by:

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)}, \sigma_{R(y)}} \tag{7}$$

where $Cov(R(x), R(y))$ represents the covariance of the x and y ranks and $\sigma_{R(x)}$,$\sigma_{R(y)}$ represents the standard deviation of x $R(x)$ and $R(y)$, respectively.

### Network statistics

To understand the predictions with respect to structural embedding, we looked at three key network statistics: Degree centrality [30], Kleinberg Authority [30], and Harmonic centrality [44]. Degree centrality directly measures the total amount of edges connected to an individual node [30]. That is, for node $i$ in the graph $G$, degree centrality is given by the degree of $i$:

$$c_d = d_i(G) \tag{8}$$

The harmonic centrality measure extends the degree of centrality by taking into account the distance between node i and all other nodes. To mitigate the effects of extreme high and low distances, the harmonic centrality calculation aggregates the inverse distances between nodes [30]. That is, for two nodes $i, j$:

$$c_h = \sum_{i \neq j} \frac{1}{dist(x_i, x_j)} \tag{9}$$

Finally, to examine the hub-like nature of each node, we calculated the Kleinberg Authority metric. Here, we take the first eigenvector of $\mathbf{A}^T\mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix of graph $G$. The eigenvalues of this vector gives a measure of how much information a node has (authority) and how well-connected the node is to other informative nodes (hubness) [44].

# Results

## Prediction data and model parameters

To predict the abundances of gene expression between experiments, we collected RNA-seq from *Sorghum bicolor*, *Oryza sativa*, and *Zea mays* and an *Arabidopsis thaliana* gene regulatory network. For the gene regulatory network, we took advantage of the *Arabidopsis thaliana* integrated gene regulatory network (iGRN) [35]. This network consists of 1,491 transcription factors and 31,393 target genes (about 27,000 protein-coding and a mixture of long noncoding RNAs and replicated transposable elements ), resulting in a total of 1.9 million interactions weighted by the strength of a biologically supported interaction [35]. In order to reduce the number of parameters model estimates, we thresholded the network edges to only include an edge if its weight is at least 0.9. This gave a resulting network of about 250,000 interactions. For gene expression abundances, we used log2-transformed gene expression levels (in transcripts per million) that were collected in *Sorghum bicolor*, *Oryza sativa*, and *Zea mays*. We collapsed genes into orthogroups to allow for cross species comparisons. Table 1 describes a summary of the number of orthogroups, experiments, and transcription factors for each species.

To set up a baseline for comparison with graph networks, we predicted non-transcription factor expression abundance from transcription factor expression using a partial least squares regression (Fig1). For the graph neural networks, we predicted expression under a graph autoencoder framework (Fig 1) with three different strategies for encoding the graph into a latent space. The first two strategies, deep and shallow convolutions, utilize graph convolutional layers, and the third strategy, GraphSAGE, uses an inductive convolutional layer that finds a mean representation of a node based on the information of nodes in its neighborhood. The representation of these models is then passed through a random forest model to predict gene expression profiles under different conditions. The models were tested in four scenarios: 1) prediction within species of unseen nonTF expression profiles when all genes have been seen by the model, 2) prediction in maize from a single species when all genes have been seen by the model, 3) prediction of unseen expression profile when all genes have not been seen by the model, and 4) prediction of maize expression profiles from all species.

**Table 1. Summary of Input Data for Gene Expression Prediction Models.**

| Species | Number of Transcription Factors | Number of Orthogroups | Number of Experiments |
|---|---|---|---|
| *Zea Mays* | 3336 | 18449 | 454 |
| *Sorghum bicolor* | 2665 | 16961 | 99 |
| *Oryza sativa* | 2409 | 20538 | 147 |

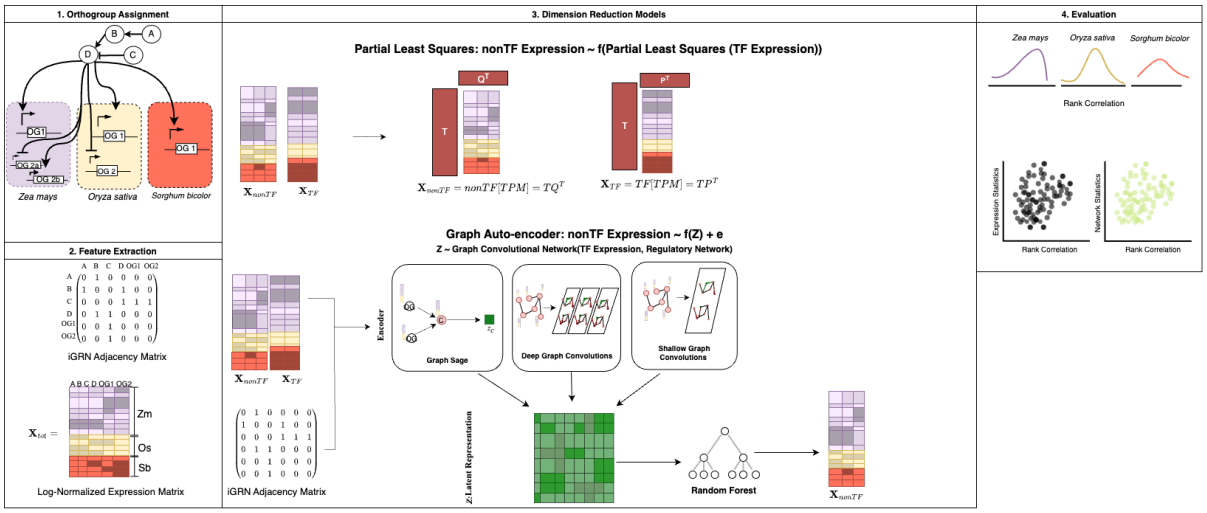



**Fig 1. Schematic of Cross Species Prediction Model** Flowchart of predicting cross-species gene expression in maize. First, we assign orthogroups to each of the species, *Zea mays*, and *Sorghum bicolor*. Then, we gather species' gene expression data and the Arabidopsis integrated gene regulatory network(iGRN). The network is converted to an adjacency matrix, and species expression data is log-normalized and transformed into an orthogroup-by-condition expression matrix. Next we establish a baseline prediction model using partial least squares. We then construct predictions from three different encoding strategies for the encoding of the expression and network– Shallow Graph Convolutions, Deep Graph Convolutions, and Graph Sage. The resulting latent representation is then used in a random forest model in order to predict unseen expressions across various conditions. Finally, we evaluate the models by examining the Spearman correlation between observed and predicted expression profiles. To determine the drivers of predictions, we also examine expression statistics versus model performance and network statistics versus model performance relationships.

## Partial least squares baseline performance

Before using the gene regulatory network model, we performed a partial least squares regression with three components. The partial least squares model does not use any information from the gene regulatory network's structure. For each species, we predicted the expression profile of 1434 genes across 16 experiments and compared those to the observed abundances. To assess the performance of the model we decided to use Spearman's rank correlation. *S. bicolor* performed the best at a correlation of mean 0.80, but *Z. mays* ($\rho = 0.77$) and *O. sativa* ($\rho = 0.76$) correlated well. Figure 3a gives the distribution of the mean correlations experiments with respect to each species.

For the next baseline, we wanted to see how well each species does to naively predict maize ortholog expression. To do this, we separated all of the maize ortholog expressions from all of the other species. Next, we used each species' ortholog profiles to predict the corresponding ortholog's profile. While overall lower than maize just predicting within itself, we found that the average Spearman correlations were still comparable to within species predictions. When we used *S. bicolor* to predict maize, we get correlation of 0.77. Similarly, when we used *O. sativa* to predict maize expressions we got correlation of 0.74, we used wanted to look at how well PLS does when we are using *O. sativa* and *S. bicolor* orthologs to predict maize expression abundance. For the cases when we used *O. sativa* and *S. bicolor*, we find that the mean Spearman correlations are 0.74 and 0.78, respectively. These results show that regardless of species combination, partial least square models tend to estimate gene expression similarly between and within species.

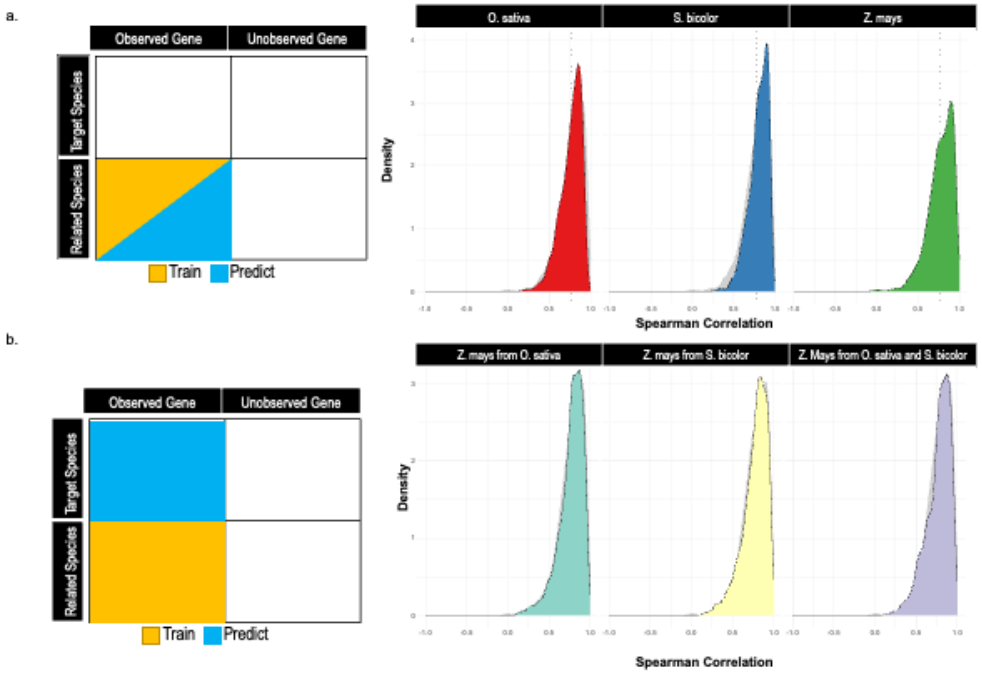



**Fig 2. Baseline partial least squares prediction performance of within-species and between-species in terms of Spearman correlation coefficient** Density plot of Spearman correlation between predicted and observed gene expression values across each species. (a) Distribution of the mean Spearman correlations for orthogroup expression predictions within each species. (b) Distribution of mean Spearman correlations for maize orthogroup expression predictions from *O. sativa*, *S.bicolor*, and the combination of *O. sativa* and *S.bicolor*.

## Graph autoencoder model performance not comparable for within-species predictions

To test the hypothesis that graph structure would improve baseline predictions, we implemented a graph autoencoder (GAE) using three different encoder architectures. We trained each model within *Oryza sativa*, *Sorghum bicolor*, and *Zea mays* to predict expression profiles within each species. For each species. the training sets were 75% of each species' orthogroups. To ensure, no data leakage, we masked the 25% test set during the training process. We ran each model for a total of 200 epochs with a learning rate of 0.001. The first method, Deep GAE, performed poorly in all species with mean Spearman correlations of 0.116, 0.179, 0.117 for *Oryza sativa*, *Sorghum bicolor*,and *Zea mays*, respectively. The inductive method, GraphSage, also did not perform well within species with mean correlations of 0.054, 0.052, and -0.015. The best performing method, the shallow graph autoencoder, gave mean predictions of 0.46 in *Oryza sativa*, 0.56 in *Sorghum bicolor* and 0.61 in *Zea mays*. Figure 3 shows the distributions of the Spearman correlations of the predictions.

## Shallow graph autoencoders outperform complex strategies in cross-species expression predictions

We next predicted in unseen maize gene expressions using the best performing strategy-the shallow graph autoencoder. We used the shallow graph convolutional autoencoder in three models covering *Z. mays* + *O. sativa*, *Z. mays* + *S. bicolor*, and *Z. mays* + *O. sativa* + *S. bicolor*. Again, we used the transcription factor expressions as input to the model to predict the unobserved maize gene expression profiles. Each model predicts 1435 unobserved maize genes across 217 experiments.In all three contexts, the shallow graph allowed the predictions of these unseen genes in maize with $\rho$ of 0.42 to 0.47 (Fig 4).

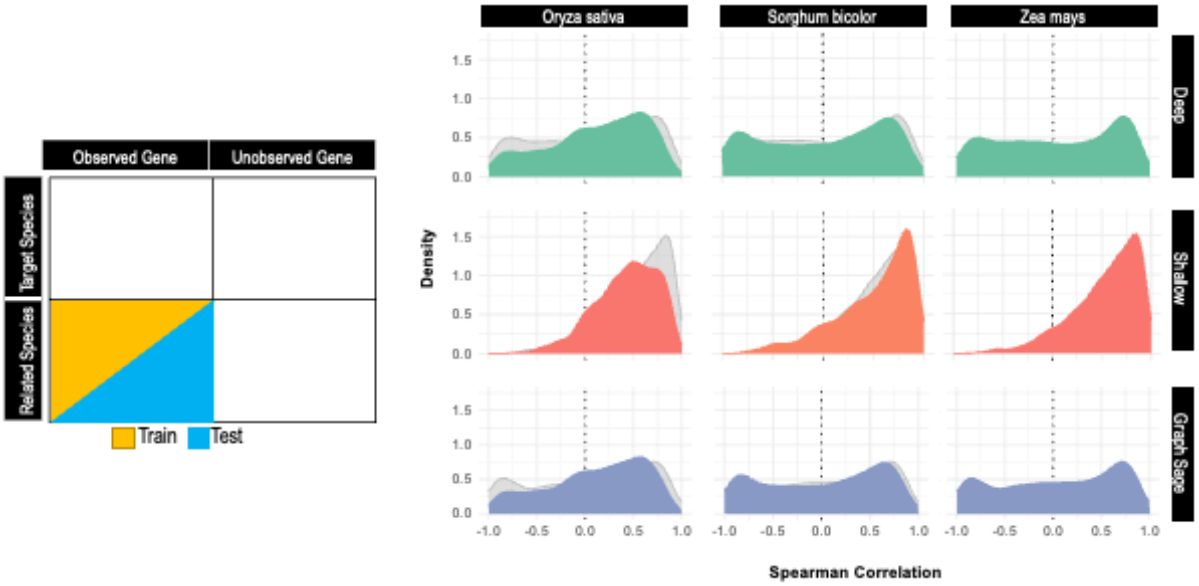



**Fig 3. Prediction performance of within-species autoencoder models in terms of the Spearman correlation coefficient** Density plot of Spearman correlation between predicted and observed gene expression values across each species. Three strategies for mapping to latent spaces, Shallow Graph Convolutions, Deep Convolutions, and Graph Sage were utilized across *Oryza sativa*, *Sorghum bicolor*, and *Zea mays*.

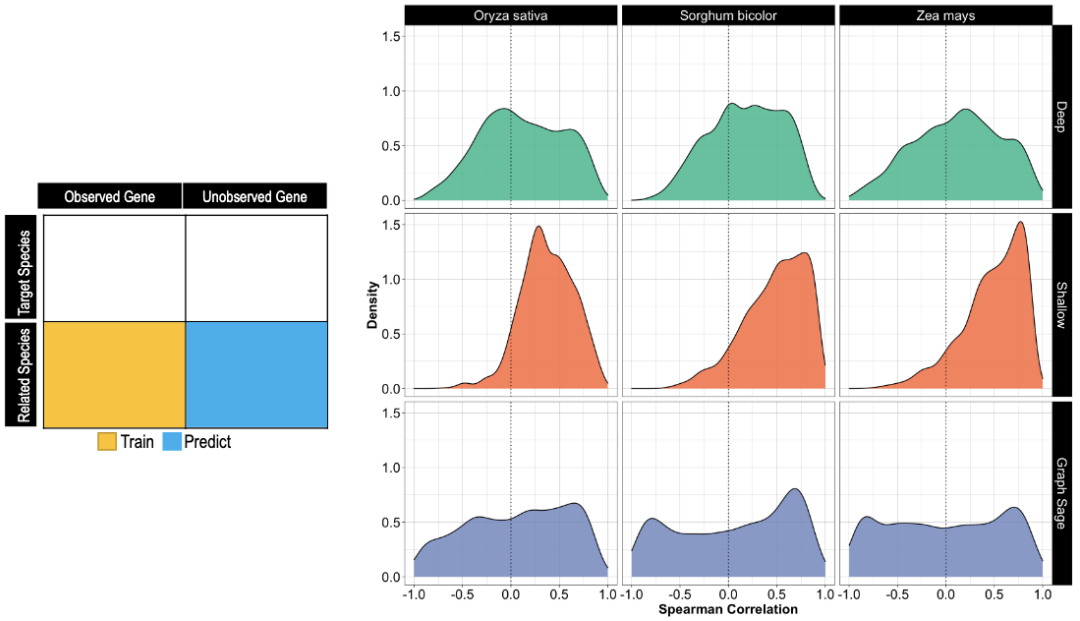



**Fig 4. Distribution of the predictive performance of the shallow graph convolutional autoencoder for the abundance of maize expression** Density plot of predictive performance Spearman Correlation of graph convolutional autoencoder. Predictive performance (x-axis) is given as the Spearman Correlation between observed and predicted orthogroup expressions. Plots are colored according to the species at hand. Dashed, vertical lines represent the mean of each distribution.

## Graph autoencoders allow for predictions in never-before-seen samples

To test the utility of using a graph autoencoder for gene expression predictions, we predicted gene expression in genes which were previously unobserved in the training. To do this, we first made sure that any genes the model could possibly predict would be included are accounted for in the iGRN. Next, we used the three autoencoder strategies to predict unobserved genes in masked experiments within each species. We trained each model for 2000 epochs with a learning rate of 0.001. Just as with the within species model, we note that the GraphSage and Deep autoencoder strategies performed the worst. The deep autoencoder had within-species prediction performances for *O. sativa*, *S. bicolor*, and *Z.mays* of -0.10, 0.31, and 0.11, respectively. GraphSage had correlations of 0.01, 0.02, and 0.002 for *O. sativa*, *S. bicolor*, and *Z.mays*, respectively. The shallow autoencoder, however, performed surprisingly well compared to the observed gene, within-species baselines with correlations of 0.47 - 0.53 (Fig 5).

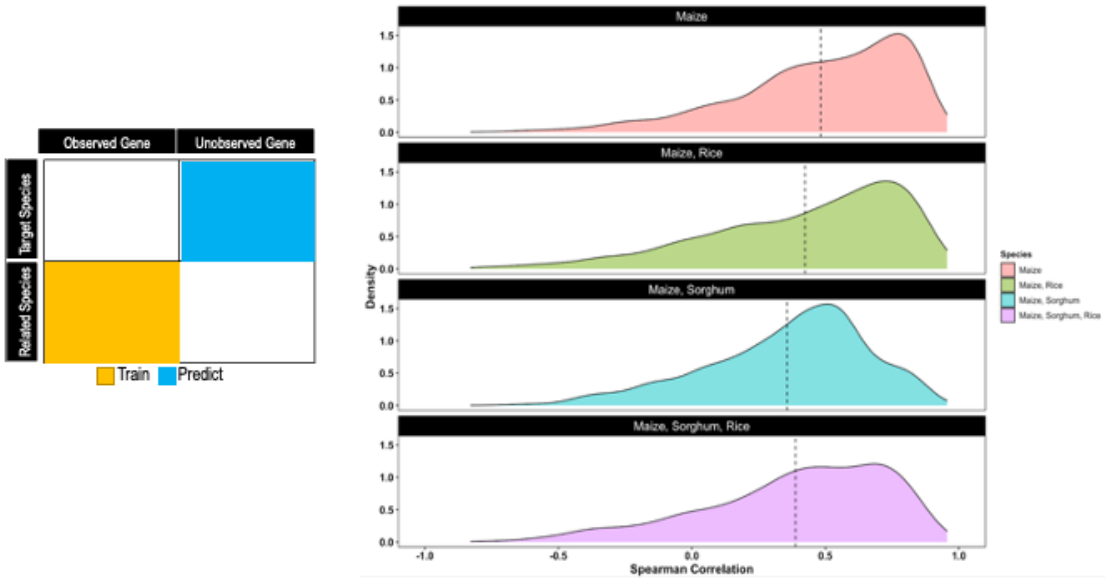



**Fig 5. Distribution of predictive performance of shallow graph convolutional autoencoder for maize expression abundance** Density plot of predictive performance Spearman Correlation of graph convolutional autoencoder. Predictive performance (x-axis) is given as the Spearman Correlation between observed and predicted orthogroup expressions. Plots are colored according to the species at hand. Dashed, vertical lines represent the mean of each distribution.

## Predictive performance is not strongly driven by network structure or variation in expression

Our final hypothesis was that structure of the Arabidopsis iGRN and the intensity of gene expression drives model prediction accuracy. In order to test the hypothesis, we plotted the maize cross-species prediction accuracy against network descriptive statistics and expression statistics. (Figure 4a) and expression statistics (Figure 5). In the case of the network statistics, there was little to no negative correlation between the measure of Log10-centrality ($\rho = 0.051$) and Kleinberg Authority ($\rho = 0.055$) and harmonic centrality ($\rho = 0.022$). These findings indicate that the influence of network structure on the model is minimal. For the expression variation statistics, we use Spearman correlation to compare overall gene expression mean, median, and variance with predictive performance. Just as with network

statistics, we find there is a very slightly negative relationship between mean ($\rho = -0.028$), median ($\rho = -0.027$), and variance ($\rho = -0.0058$), but we note that these values are not significant.

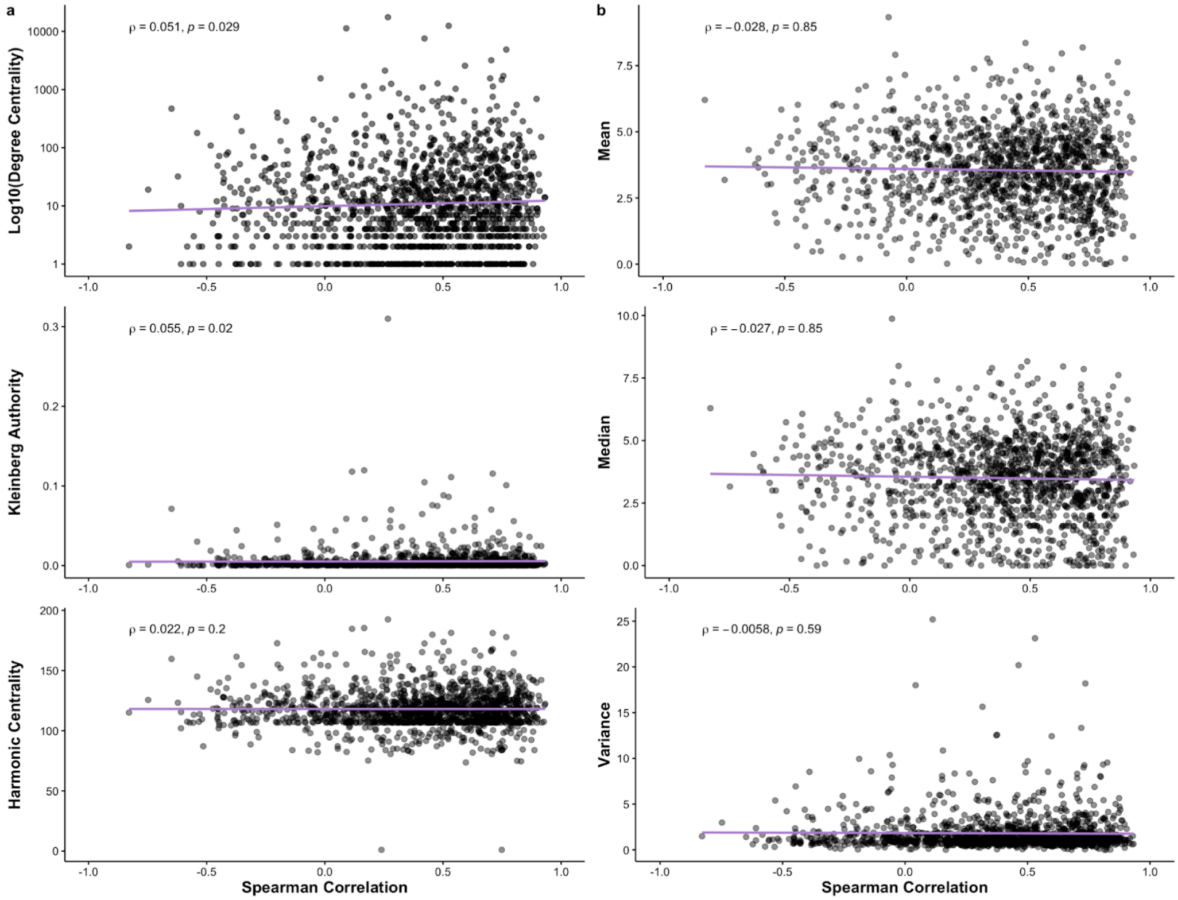



**Fig 6. Relationships between shallow graph convolutional autoencoder performances and key descriptive statistics** (a) Relationship between three common network measures, Degree Centrality, Kleinberg Authority, and Harmonic Centrality, with the Spearman correlations for predicting maize expression abundance with a shallow graph autoencoder. (b) Relationship between descriptive statistics about the expression data, mean, median, and variance, with the Spearman correlations for predicting maize expression abundance with a shallow graph autoencoder.

# Discussion

Gene expression inference across species, experiments, and time require effective technological frameworks, carefully curated data, and powerful mathematical frameworks. In traditional gene expression inference frameworks, such as the LINCS linear regression-based platform, there is little focus on the non-linearity of gene expression. We present a way to utilize the gene regulatory network's nonlinear connections in a prediction framework that allows for the prediction of gene expression values for unobserved genes in unobserved experiments and is comparable to within-species predictions. Our results suggest that gene expression prediction accounting for discrete biological network structure can be used to achieve moderate-accuracy target gene expression values.

We note that there is a large imbalance between the accuracy of predictions in the within-species models. In *Sorghum bicolor*, this discrepancy is likely to be from a few genes which are highly variable.

In *Zea mays* and in *Oryza sativa*, our hypothesis is that this discrepancy is likely due to the different types of tissue classes included in the dataset. There is evidence that the transcription factor background can be attributed to the determination of biological roles such as tissue specificity [45], gene function [17], and cellular behavior [46]. In addition to the overall differences per species, there is a wide range of differences between each of the model types.

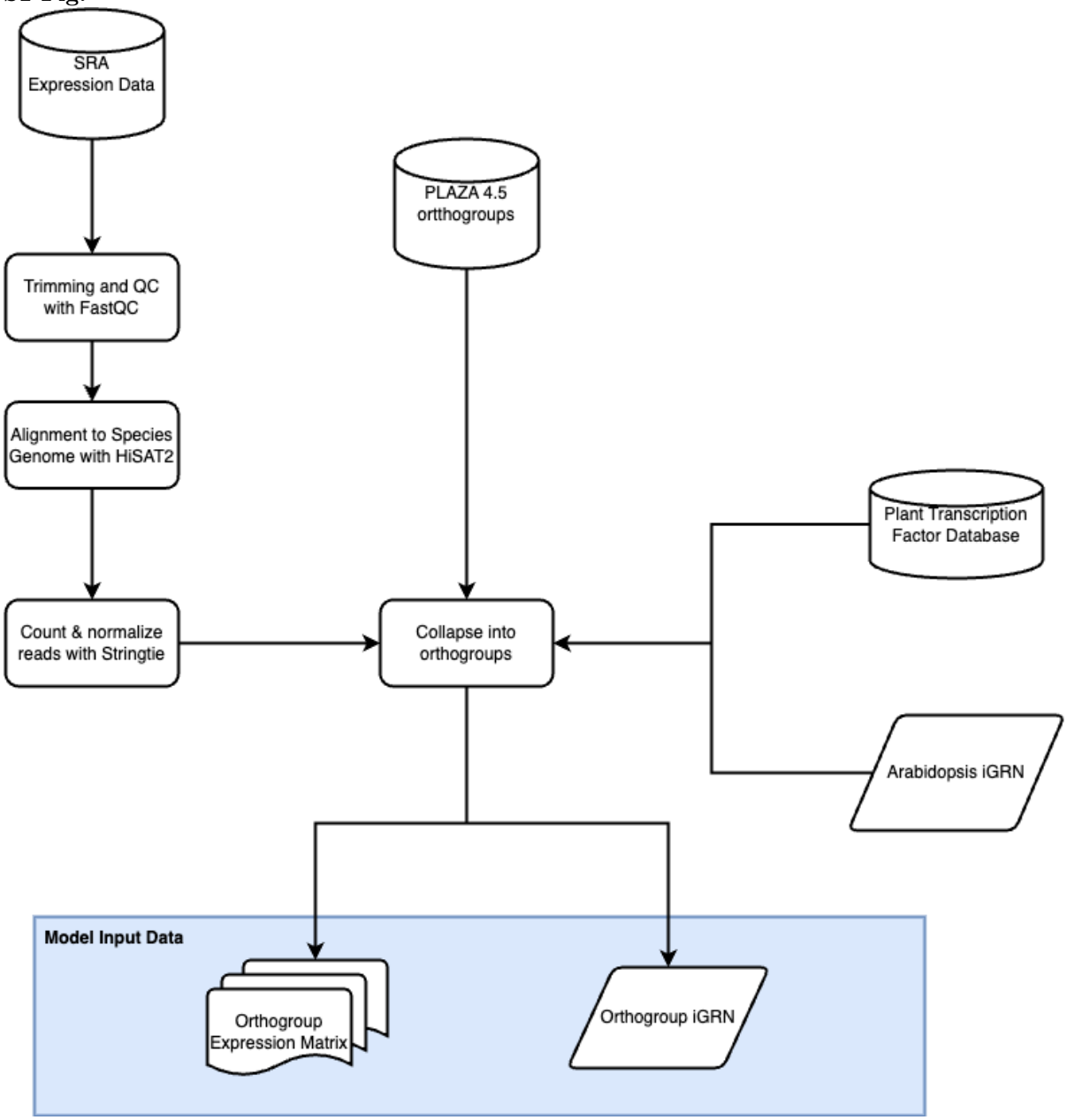
One of the main results of the study is that the shallow convolution operator consistently outperforming the more complex models. We hypothesize this is due to a combination of the network structure and the ability for simpler model to generalize on new data. For the network case, we decided to take the most representative edges in the Arabidopsis iGRN. Deeper GNNs and GraphSAGE models inform their embedding projections by taking in information from surrounding nodes. By paring down the network at hand, these models put heavy emphasis on nodes that are already known to have a large bias toward genes that are known and expressed. This means that the 2nd and 3rd order relationships that may be in the larger, less thresholded network are largely unseen. This can lead to under-accounting for certain relationships and over-accounting for relationships that actually involve more connections. The relationships at question, however, must generalize across species and expression pattern. As seen in genomic prediction and language modeling, the less complex model often performs better than the more complex models [47].

One of the key problems in this analysis is that we only use one data modality, RNA-seq. The landscape of gene expression is rather complex and requires not only understanding the non-linearity of regulator-target relationships, but also understanding the interactions between regulatory elements that cannot be identified exclusively in assays such as RNA-seq.This type of modeling has been done with models such as sequence-specific models which use wide ranges of regulatory information to guide the model's training and testing. In other models, such as GC-merge [48], cis-regulatory information from Chip-Seq integrates expression measurements with a deep graph convolutional framework much like the one we utilized. In the future, prediction across species using more types of readily-accessible data could provide a more biologically informed way to power these sophisticated deep-learning models.

# Figures

# Supporting information

**S1 Fig.**




**Flowchart for pre-processing of model input data.**

# Acknowledgments

We thank Travis Wrightsman and Michelle Stitzer for their suggestions in the preparation of this manuscript.

# References


1. Zrimec J, Buric F, Kokina M, Garcia V, Zelezniak A. Learning the Regulatory Code of Gene Expression. Frontiers in Molecular Biosciences. 2021;8:673363. doi:10.3389/fmolb.2021.673363.

2. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Reports. 2020;31(7):107663. doi:10.1016/j.celrep.2020.107663.

3. Akagi T, Masuda K, Kuwada E, Takeshita K, Kawakatsu T, Ariizumi T, et al. Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. The Plant Cell. 2022;34(6):2174–2187. doi:10.1093/plcell/koac079.

4. Kelley DR. Cross-species regulatory sequence activity prediction. PLOS Computational Biology. 2020;16(7):e1008050. doi:10.1371/journal.pcbi.1008050.

5. Muhammad D, Schmittling S, Williams C, Long TA. More than meets the eye: Emergent properties of transcription factors networks in Arabidopsis. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2017;1860(1):64–74. doi:10.1016/j.bbagrm.2016.07.017.

6. Grah R, Friedlander T. The relation between crosstalk and gene regulation form revisited. PLOS Computational Biology. 2020;16(2):e1007642. doi:10.1371/journal.pcbi.1007642.

7. Srivastava AK, Lu Y, Zinta G, Lang Z, Zhu JK. UTR-dependent control of gene expression in plants. Trends in Plant Science. 2018;23(3):248–259. doi:10.1016/j.tplants.2017.11.003.

8. Lemmon ZH, Bukowski R, Sun Q, Doebley JF. The role of cis regulatory evolution in maize domestication. PLoS Genetics. 2014;10(11). doi:10.1371/journal.pgen.1004745.

9. Lang Z, Wills DM, Lemmon ZH, Shannon LM, Bukowski R, Wu Y, et al. Defining the role of prolamin-box binding FACTOR1 gene during maize domestication. Journal of Heredity. 2014;105(4):576–582. doi:10.1093/jhered/esu019.

10. Lasky JR, Des Marais DL, Lowry DB, Povolotskaya I, McKay JK, Richards JH, et al. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in Arabidopsis thaliana. Molecular Biology and Evolution. 2014;31(9):2283–2296. doi:10.1093/molbev/msu170.

11. de Meaux J. cis-regulatory variation in plant genomes and the impact of natural selection. American Journal of Botany. 2018;105(11):1788–1791. doi:10.1002/ajb2.1180.

12. Wang H, Li C, Zhang J, Wang J, Ma Y, Lian Y. A new LSTM-based gene expression prediction model: L-GEPM. Journal of Bioinformatics and Computational Biology. 2019;17(04):1950022. doi:10.1142/s0219720019500227.

13. Lemmon ZH, Reem NT, Dalrymple J, Soyk S, Swartwood KE, Rodriguez-Leal D, et al. Rapid improvement of domestication traits in an orphan crop by genome editing. Nature Plants. 2018;4(10):766–770. doi:10.1038/s41477-018-0259-x.

14. Freeling M. Grasses as a single genetic system. reassessment 2001. Plant Physiology. 2001;125(3):1191–1197. doi:10.1104/pp.125.3.1191.

15. Hartmann A, Berkowitz O, Whelan J, Narsai R. Cross-species transcriptomic analyses reveals common and opposite responses in Arabidopsis, rice and barley following oxidative stress and hormone treatment. BMC Plant Biology. 2022;22(1):62. doi:10.1186/s12870-021-03406-7.

16. Narsai R, Ivanova A, Ng S, Whelan J. Defining reference genes in Oryza sativa using organ, development, biotic and abiotic transcriptome datasets. BMC Plant Biology. 2010;10(1):56–56. doi:10.1186/1471-2229-10-56.

17. Fagny M, Kuijjer ML, Stam M, Joets J, Turc O, Rozière J, et al. Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. Frontiers in Genetics. 2021;11:606285. doi:10.3389/fgene.2020.606285.

18. Long TA, Brady SM, Benfey PN. Systems Approaches to Identifying Gene Regulatory Networks in Plants. Annual Review of Cell and Developmental Biology. 2008;24(1):81–103. doi:10.1146/annurev.cellbio.24.110707.175408.

19. Saint-Antoine MM, Singh A. Network Inference in Systems Biology: Recent Developments, Challenges, and Applications. arXiv. 2019;.

20. Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, et al. Integration of omic networks in a developmental atlas of maize. Science. 2016;353(6301):814–818. doi:10.1126/science.aag1125.

21. Patel N, Bush W. Modeling Transcriptional Regulation Using Gene Regulatory Networks Based on Multi-Omics Data Sources. 2020;doi:10.21203/rs.3.rs-112300/v1.

22. Matthews ML, Williams CM. Modeling Transcriptional Regulation, Methods and Protocols. Methods in Molecular Biology. 2021;2328:115–138. doi:10.1007/978-1-0716-1534-8_7.

23. Krouk G, Lingeman J, Colon AM, Coruzzi G, Shasha D. Gene regulatory networks in plants: learning causality from time and perturbation. Genome Biology. 2013;14(6):123. doi:10.1186/gb-2013-14-6-123.

24. Ba LJ, Caruana R. Do Deep Nets Really Need to be Deep? arXiv. 2013;.

25. Wrightsman T, Marand AP, Crisp PA, Springer NM, Buckler ES. Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. The Plant Genome. 2022; p. e20249. doi:10.1002/tpg2.20249.

26. Zhou W, Ji H. Genome-wide prediction of chromatin accessibility based on gene expression. Wiley Interdisciplinary Reviews: Computational Statistics. 2021;13(5). doi:10.1002/wics.1544.

27. Santos JPRd, Fernandes SB, Lozano R, Brown PJ, Buckler ES, Garcia AAF, et al. Novel Bayesian Networks for Genomic Prediction of Developmental Traits in Biomass Sorghum. bioRxiv. 2019; p. 677179. doi:10.1101/677179.

28. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Cho K, Vatanen T, et al. Structure-Based Function Prediction using Graph Convolutional Networks. bioRxiv. 2019; p. 786236. doi:10.1101/786236.

29. Hasibi R, Michoel T. A Graph Feature Auto-Encoder for the prediction of unobserved node features on biological networks. BMC Bioinformatics. 2021;22(1):525. doi:10.1186/s12859-021-04447-3.

30. Zhang J, Luo Y. Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network. Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017). 2017;doi:10.2991/msam-17.2017.68.

31. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. arXiv. 2017;.

32. Su XR, You ZH, Hu L, Huang YA, Wang Y, Yi HC. An Efficient Computational Model for Large-Scale Prediction of Protein-Protein Interactions Based on Accurate and Scalable Graph Embedding. Frontiers in genetics. 2020;12:635451. doi:10.3389/fgene.2021.635451.

33. Zhang XM, Liang L, Liu L, Tang MJ. Graph Neural Networks and Their Current Applications in Bioinformatics. Frontiers in Genetics. 2021;12:690049. doi:10.3389/fgene.2021.690049.

34. Shiu SH, Shih MC, Li WH. Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals. Plant Physiology. 2005;139(1):18–26. doi:10.1104/pp.105.065110.

35. Clercq ID, Velde JVd, Luo X, Liu L, Storme V, Bel MV, et al. Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. Nature Plants. 2021;7(4):500–513. doi:10.1038/s41477-021-00894-1.

36. Wu TY, Goh H, Azodi CB, Krishnamoorthi S, Liu MJ, Urano D. Evolutionarily conserved hierarchical gene regulatory networks for plant salt stress response. Nature Plants. 2021;7(6):787–799. doi:10.1038/s41477-021-00929-7.

37. Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, et al. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proceedings of the National Academy of Sciences. 2019;116(12):5542–5549. doi:10.1073/pnas.1814551116.

38. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Research. 2020;48(D1):D1104–D1113. doi:10.1093/nar/gkz1020.

39. Bel MV, Diels T, Vancaester E, Kreft L, Botzki A, Peer YVd, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Research. 2018;46(Database issue):D1190–D1196. doi:10.1093/nar/gkx1002.

40. Ross Z, Wickham H, Robinson D. Declutter your R workflow with tidy tools. 2017;doi:10.7287/peerj.preprints.3180.

41. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. arXiv. 2019;.

42. Helland IS. Partial Least Squares Regression and Statistical Models. Scandinavian Journal of Statistics. 1990;17(2):97–114.

43. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. CoRR. 2016;doi10.48550/arXiv.1706.02216.

44. Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM. 1999;46(5):604–632. doi:10.1145/324133.324140.

45. Huang J, Zheng J, Yuan H, McGinnis K. Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. BMC Plant Biology. 2018;18(1):111. doi:10.1186/s12870-018-1329-y.

46. Vachon G, Engelhorn J, Carles CC. Interactions between transcription factors and chromatin regulators in the control of flower development. Journal of Experimental Botany. 2018;69(10):2461–2471. doi:10.1093/jxb/ery079.

47. Runcie D, Cheng H. Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods. G3: Genes—Genomes—Genetics. 2019;9(11):3727–3741.

48. Bigness J, Loinaz X, Patel S, Larschan E, Singh R. Integrating Long-Range Regulatory Interactions to Predict Gene Expression Using Graph Convolutional Networks. Journal of Computational Biology. 2022;doi:10.1089/cmb.2021.0316.