Title: Predicting gene expression responses to environment in *Arabidopsis thaliana* using natural variation in DNA sequence

Authors: Margarita Takou[1], Emily S Bellis[1,2,] & Jesse R Lasky[1]

1 Pennsylvania State University, University Park, 16802, USA

2 Department of Computer Science, Arkansas State University, Jonesboro AR

Abstract:

The evolution of gene expression responses are a critical component of adaptation to variable environments. Predicting how DNA sequence influences expression is challenging because the genotype to phenotype map is not well resolved for *cis* regulatory elements, transcription factor binding, regulatory interactions, and epigenetic features, not to mention how these factors respond to environment. We tested if flexible machine learning models could learn some of the underlying *cis*-regulatory genotype to phenotype map. We tested this approach using cold-responsive transcriptome profiles in 5 diverse *Arabidopsis thaliana* accessions. We first tested for evidence that *cis* regulation plays a role in environmental response, finding 14 and 15 motifs that were significantly enriched within the up- and down-stream regions of cold-responsive differentially regulated genes (DEGs). We next applied convolutional neural networks (CNNs), which learn *de novo cis*-regulatory motifs in DNA sequences to predict expression response to environment. We found that CNNs predicted differential expression with moderate accuracy, with evidence that predictions were hindered by biological complexity of regulation and the large potential regulatory code. Overall, DEGs between specific environments can be predicted

based on variation in *cis*-regulatory sequences, although more information needs to be incorporated and better models may be required.

Key words: low frequency variants, gene expression prediction, machine learning, regulatory elements, evolution, genotype to phenotype map

Introduction

Adaptation is often a complex process that occurs via selection on traits and the underlying genetic mechanisms (Kawecki and Ebert 2004). The genetic basis of ecologically important quantitative traits often involves mutations with effects on gene expression. In variable environments, the evolution of gene expression responses to environment (expression plasticity) are likely a critical component of adaptation (Romero et al. 2012). These processes are often studied using evidence from variation in DNA sequence and mRNA abundance. For instance, the incorporation of transcriptomic information into genome wide association studies (GWAS) can help identify genes controlling a specific phenotype (Yang et al. 2022). Association or linkage mapping approaches can be used to map expression quantitative trait loci (eQTL), and allele-specific expression can determine the presence of *cis* or *trans* acting eQTL (Josephs et al. 2015; Josephs et al. 2020; Mack et al. 2023). However, the integration of sequence and expression data in evolutionary quantitative genomics is hindered by gaps in our understanding of how DNA sequence influences expression. In particular, understanding the genetic basis of expression plasticity might be beneficial to applied biology given the importance of predicting organismal responses to environmental changes(Keagy et al. 2023).

Gene expression responses to the environment differ among genes in the genome and among different natural genotypes due to DNA sequence variation. Differences in gene expression among genotypes may be driven by sequence differences at other loci in the genome, which regulate expression of a given gene (*trans* effects). At the same time, differences among genes and genotypes may arise from differences in sequence proximal to the gene that determine how its expression is regulated (*cis* effects). Researchers have documented cases where mutations in genes involved in environmental sensing (Huo et al. 2016; Mateos et al. 2017) or in environmentally responsive genes that bind DNA sequences (Adrian et al. 2010; Thomashow 2010) have downstream effects on expression of other genes. Additionally, researchers have characterized how some *cis* mutations influence DNA binding by environmentally responsive transcription factors, ultimately altering environmental responsive expression (Wittkopp et al. 2004; Cubillos et al. 2014; Lovell et al. 2016). Transcription factors that tend to show expression responses to environment have been documented, and the DNA sequence motifs they bind to have been roughly characterized (Schmitz et al. 2021). However, across the genome there are many transcription factors that often respond to a single environmental stressor and interact, and this knowledge is currently too limited to predict how genetic variation in DNA sequences influence genome-wide expression responses to environment.

The large number of *cis*-regulatory motifs allows the investigation of genomic and evolutionary properties, which are associated with their distribution across the genome (de Meaux 2006; Erwin and Davidson 2009; Meaux 2018). Genome-wide expression profiling in association mapping or linkage mapping populations as well as allele-specific expression studies have revealed that *cis* expression quantitative trait loci (*cis*-eQTL) are abundantly segregating within

species. *Cis*-regulatory mutations have been of particular importance for evolutionary biology because some have suggested they may be less likely to be deleterious than mutations in transcription factors, which are expected to be more pleiotropic (Wray et al. 2003; Brown and Kelly 2022). Relatedly, the possibility of easier dissection of the genotype to phenotype map for *cis*-eQTL (versus *trans*-acting mutations likely influencing multiple genes' expression) combined with the abundance of *cis*-eQTL may be an opportunity to build a more integrative understanding of quantitative trait genetics.

Using machine learning methods in fields such as genomics and population genetics can help navigate increasingly bigger datasets and reveal complex patterns (Korfmann et al. 2023). Especially in gene regulation and evolution, deep learning approaches can have an advantage over traditional methods in decoding enhancer grammar of gene regulation, as these models can learn complex *cis*-regulatory rules in a precise manner not biased by current knowledge (Raicu et al. 2023) and can outperform more traditional methods, like clustering by kmer or linear regression in predicting gene expression (Chen et al. 2016). For example, training a machine learning model on the imputed *cis* haplotypic information of RNA expression, Giri et al. (2021) managed to exclude the impact of *trans* effects on the gene expression variation. The prediction was less variable than using SNPs and had an increased accuracy for predicting within population variance (Giri et al. 2021). Moore et al. (2021) showed how known regulatory sites and transcription factors (TFs) interactions can be used in machine learning to predict clusters of gene co-expression related to stress response (Moore et al. 2021). There is increasing evidence that genetic diversity on the *cis* elements of plants are connected to selection (Marand et al. 2023). Using upstream and downstream genetic coding information improves the prediction of

the role of *cis* regulatory elements in differential gene expression, in contrast to using only upstream regions in maize (Benoit 2021), especially when more than one genotype was included in the analysis (Zhou et al. 2021).

In *Arabidopsis thaliana,* whose importance in studying evolution and ecology patterns is growing over the years (Takou et al. 2019), the conserved noncoding sequences are mostly concentrated close to the coding region and impact gene regulation (Yocca et al. 2021), as *cis*-regulatory variants are usually in linkage with the expressed transcript (de Meaux 2006; He et al. 2016; Steige et al. 2017). Even though there are conserved gene reactions between species and populations, the increased genetic diversity within populations or species causes phenotypic variance. Genes, whose expression is affected by genotype x environment interactions (GxE), tend to show allele frequency correlations with climate and have elevated genetic variation in stress responsive transcription factor binding sites (Lasky et al. 2014). Known TF binding site motifs of *A. thaliana* have been used to identify putative binding sites and then train a deep model for predicting *cis* changes in tomato (Akagi et al. 2022).

Our aim was to use genetic variation among genes within individuals and among genotypes in the regions upstream and downstream of genes to predict expression response to environment in *A. thaliana.* We aimed to predict in a both a more general sense than other studies by using the sequences instead of known TFBS, and in a more restrictive sense by classifying gene expression in specific environments within one species. For the analysis we used a published dataset on differentially expressed genes (DEGs) during cold acclimation of diverse *A. thaliana* genotypes (Hannah et al. 2006). The authors had identified approximately 1,500 up regulated DEGs and down regulated DEGs out of around 20,000 genes for multiple accessions as a response to cold

acclimation and stress (Table S1). Here, we focus on the five *A. thaliana* accessions *Can*, *Col-0*, *Cvi*, *Ler* and *Rsch*, for which the pseudogenomes from whole genome resequencing data are published (1001 Genomes Consortium 2016). We included multiple genotypes in the analysis, so as to capture some of the natural genetic diversity and potential GxE responses.

We first tested whether there are specific sequence motifs enriched along those regions of the differentially expressed genes that would suggest *cis* regulation drives some aspect of cold expression response. We determined if these motifs were related to known stress-responsive transcription factors (TFs). Then, we tested whether several naïve deep learning models could predict expression response, without a priori designation of important motifs, by using only the sequences upstream and downstream of genes.

Material and Methods

Transcriptome and genome data

We used the publicly available dataset of differentially gene expression during cold acclimation by (Hannah et al. 2006). In brief, the authors used multiple accessions of *A. thaliana* to test their cold acclimation potential. They found that gene expression responds to cold acclimation, indicating regulatory pathways being involved in their control. Out of the accessions used in the original study, we used the five accessions, *Col-0, Ler, Cvi, Can* and *Rsch* (Table S1) whose pseudogenomes are available in https://1001genomes.org/.

Searching for sequence motifs using STREME

Transcription factors bind DNA in a sequence specific manner, but not all sequences are known nor is the exact relationship between sequence and binding affinity known. For this reason we began with a *de novo* search for sequence motifs in the upstream and downstream regions of the coding sequence, as those often contain such signals. We used samtools faidx v1.10 (Li et al. 2009) to extract the 1kbp upstream and downstream regions of each gene present in the dataset, including the first and last part of the coding region as TFBS often bind there. Specifically, we extracted the 500bp upstream and 500bp downstream of the gene's start codon for the upstream regulatory coding region, and the 500bp upstream and 500bp downstream of the gene's end codon for the downstream regulatory coding region. We then used the s STREME (Bailey 2021) to calculate significantly enriched ungapped motifs within the set of up- and down- regulated DEGs in relation to their presents in non-DEGs. For both groups we used as a control a randomly down-sampled set of the non-DEGs and stopped the search for motifs when there were discovered 3 consecutive motifs with corrected for multiple testing p-value (e-value) larger than 0.05.

Because sequence motifs may have a variety of causes and functions, and not all are likely TFBS, we next sought to determine whether motifs discovered *de novo* by STREME specifically correspond to known transcription factor binding sites. We searched for significant overlaps of at least 5bp between the discovered enriched motifs and the JASPAR 2022 CORE (Sandelin et al 2004) plant transcription factor binding sites database with MEME suite's online tool TOMTOM (Gupta et al. 2007).

Random forest analysis

To determine whether known transcription factor binding sites regulate the response to cold, we attempted to predict DEGs based on their presence or absence in the regulatory regions. We used a dataset (Takou et al. 2022) of TF putative binding sites upstream of each gene, which were predicted from *A. thaliana* using JASPAR2018 Bioconductor package (JASPAR) and then matched to the aligned regions of *A. thaliana*. We treated the count of each TF binding site per gene as a separate feature and additionally determined the total number of TF binding sites per gene. Taken together, these descriptors of genomic variation provided a total of 413 features, of which we could assess the predictive impact on gene regulation.

We trained a random forest model for *Col-0* using the ranger library in R (Wright and Ziegler 2017) with up-regulated DEGs in response to cold and non-DEGs as variable and all compiled features of gene regulation as predictor variables. We set the number of trees of the forest to 500, the number of variables to possibly split at each node (mtry) to 200, and the impurity mode to impurity_corrected. From the trained random forest, we extracted the relative importance of each predictor variable as well as the out-of-bag prediction error, which is informative about the predictive power of the model. Finally, we estimated a *p* value for each predictor variable using the "Altmann" method with 100 permutations. We repeated the above analysis for all down-regulated DEGs. We estimated the prAUC in R library using the function prauc and extracted its significance by permuting 100 times the gene identity for each set of variables.

Training Convolutional Neural Networks

We employed a more naive machine learning approach to determine whether we can predict DEG using regulatory regions. We trained convolutional neural networks (CNNs) to identify a

model that would predict the whether each gene was differentially expressed in the stress conditions relative to the control with high accuracy. We prepared the dataset by separating the genes into a training and test set for each accession, incorporating information about the gene families that they belonged to (Washburn et al. 2019). Because sequences may carry spurious associations with expression that are due to shared evolutionary history, we followed the method of Washburn et al. (2019) and split genes into gene families and then tested predictions in gene families not used for training.

In the end, we used 68,979 (80%) genes as the train set and 17,475 (20%) as the test set out of the total 86,454 genes across all five genotypes. The proportion of up- and down-regulated DEGs to the non DEGs was approximately 1:11 in both sets (Table S1). For each gene, we used samtools faidx v1.10 (Li et al. 2009) to extract the 1.5kbp upstream and downstream regions of each gene from the published pseudogenomes, using custom scripts. Specifically, we extracted the 1,000 upstream and 500bp downstream of the gene's start codon for the upstream regulatory coding region, and the 500bp upstream and 1,000bp downstream of the gene's end codon for the downstream regulatory coding region. Those sequences were converted into one hot encoding of each nucleotide in python v3.6, to be used as input for the training and testing of the CNN models. We then created the labels needed for binary classification of DEG during training by scoring the up-regulated DEGs as 0 and the non-DEGs as 1. As the number of up-regulated DEGs was much smaller than non-DEGs, we oversampled with replacement the up-regulated DEGs to a ratio of 1:1 between the two groups. This process was repeated for training models to classify genes as down-regulated DEGs and non-DEGs.

We used the prepared inputs and labels to perform a grid search of tuning parameters, as we wanted to identify both the models with the best prediction of our dataset, as well as how different tuning parameters influence model accuracy using the regulatory regions of multiple accessions. The grid search was performed on a total of 1,344 combinations of hyperparameters (Washburn et al. 2019) by training each model for 50 epochs using the python library tensorflow-gpu v2 (Abadi et al.). The tuning parameters included: the number of convolutional layers, the filter of each convolutional layer, the width of each convolutional layer, the pool width and stride, the number of dense layers and their units, as well as the dropout rate. A list of the specific values used to create all the different combinations is provided in Table S2. For each trained model, we extracted the loss, accuracy, and the probability-based accuracy under the curve (prAUC) of the training, validation and testing of each model to compare the accuracy of the models. Moreover, we estimated the proportion of correct predictions of each class within the test set so that we could assess the potential overfitting of each model to predicting one class or the other. We repeated this process for all down-regulated DEGs.

Determining factors influencing accuracy of CNNs

To determine the degree to which predictions relied on upstream sequence versus both up and downstream sequences, we repeated the above analysis described for the regulatory regions of all accessions by using only the upstream regions.

Because genetic variation among genotypes in expression can be driven by *trans*-effects (from other loci far from the gene), we sought to control for some of this variation. In particular, we focused on known loss of function variation at large effect CBF transcription factors that are

involved in adaptation to different temperatures (Oakley et al. 2014; Lee et al. 2024). The CBFs, or *C-REPEAT BINDING FACTORs* are a set of 3 transcription factors that regulate cold acclimation and segregate for loss of function variation at high frequency in *Arabidopsis thaliana*, with loss of function variants associated with warmer climates (Fowler et al. 2005; Oakley et al. 2014). Hannah et al. (2006) previously showed with our dataset that intact CBF function explained a substantial portion of variation in cold response transcriptomes and Des Marais et al. (2017) showed that genetically variable cold responses were enriched in central locations of coexpression networks.

To determine the role that genetic variation among genotypes, versus only among genes within a genotype, we trained models with only *Col-0* genes as the set. We repeated the training while also oversampling the dataset to be the same size as for all five accessions combined to control for underpowered training due to the smaller dataset.

To assess whether models may have been hindered by factors such as model architecture or data complexity we introduced a “spike” of 5bp in the upstream region of up-regulated DEGs. This way we created a positive control that can help validate the capability of the models to recognize the up regulated DEG group. We replaced the first 5bp with the sequence of AGAGA.

Evaluating and interpreting model training

We compared the area under the precision-recall curve (prAUC), cluster accuracy, as well as the difference of the two classes accuracy with a Kolmogorov Smirnov test using the function ‘ks.test’ in R v4.1.2. To see which hyperparameters had a significant influence on those CNN performance metrics of the spiked and non-spiked dataset, we ran a generalized linear mixed

model with all the hyperparameters as fixed effect and the models' loss as random effect. We dropped from the model all the fixed effects that had no significant impact on the prAUC. Thusly, the final model included the number of convolutional layers, dense layers, pool stride and dense units as fixed effects. We tested the impact of those four hyperparameters in the original complete dataset by running another generalized linear mixed model. Generalized linear mixed models were performed using the package lme4 (Bates et al. 2015) and *p* values were extracted using the function Anova from the R package car (Fox and Weisberg 2019).

For each type of CCN training we selected the best fitting model. We selected models that had a prAUC value on the test set of at least 0.7, and that they did not overfit into predicting one class of DEGs. To do so, we estimated the absolute difference of the accuracy of predicting each class in the dataset. We used the same criteria for all cases.

Once we had selected the model parameters that comprise the architecture which best predicted the DEG class, we evaluated the significance of the result and excluded the possibility that is the result of learning from spurious sequences by permuting the label identity of each gene 100 times and re-train using the same model architecture. We estimated the *p* value by counting the cases that had better performance based on both the prAUC and the class accuracy difference and divided by 100.

All results were visualized in R v4.1.2 (R Core Team 2018).

Results

The presence of enriched motifs in the up regulated DE gene sequences suggests their potential contribution to environmental response.

We searched for motifs enriched along the putative regulatory regions of the up-regulated DEGs in comparison to the non-DEGs to assess the possibility of patterns in the regulatory regions associating with the differential expression status of each gene. We used STREME (Bailey 2021) to search for de novo enriched motifs within the upstream and downstream regions of all genes of interest (Figure 1a). When we searched for enriched motifs only within the sequences of up-regulated DEGs for the reference genotype, *Col-0,* we discovered 9 motifs. In contrast, we identified 14 motifs significantly enriched within the up-regulated DEGs of all 5 accessions, suggesting greater power with more genotypes ($p < 0.05$). The median discovered motif length was 10bp (Table S3) and the majority of the discovered motifs were significantly enriched in high frequency in our dataset. Specifically, five of the fourteen discovered motifs were present in more than 98% of the DEGs' regulatory regions tested, even though a few rare ones with less than 1% frequency were also identified (Figure 1b). When we looked at which genotypes each discovered motif was significantly enriched for at least once, the average number of motifs per up-regulated DEG was similar between the genotypes (approximately 6 motifs per gene per genotype; Figure 1c).

There was a similar number of enriched motifs among the down-regulated DEGs. In total 15 motifs were discovered that were significantly enriched in all genotypes ($p < 0.05$; Table S4). Five of those motifs were present in more than 99% of tested DEG sequences for enrichment (Figure 1d), with an average of 6 discovered motifs per tested DEG sequence (Figure 1e).

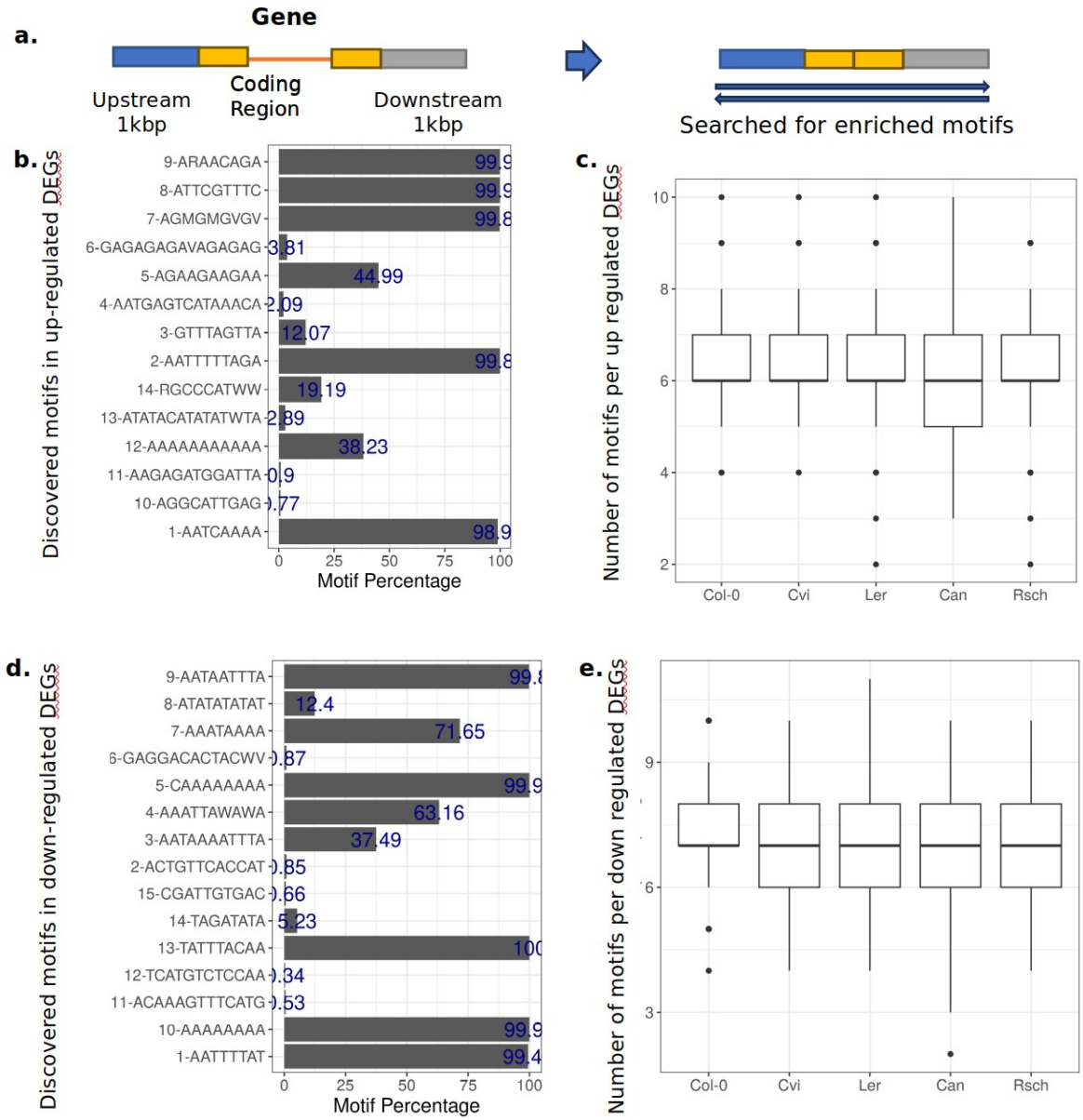

Figure 1: Significantly enriched motifs within the regulatory regions of the up- and down-regulated DEGs. **a.** For each up- or down- regulated gene across all five *A. thaliana* accessions, we extracted the 1kbp upstream (blue block) and downstream regions (grey block). We also extracted within the first 500bp after the start codon and 500bp before the end codon (yellow

blocks) to include in the motif search. **b,d.** The discovered motifs in the upstream and downstream regions of up- or down- regulated DEGs and how frequent they are in the DEG sequences with STREME. **c,e** The distribution of discovered motifs in the upstream region of up-DEGs or down- DEGs per accession.

When we cross referenced all the motif sequences we identified as enriched in up- and down-regulated DEGs in a database of experimentally defined TFBSs (Sandelin et al. 2004), the discovered motifs had a significant overlap with 135 and 143 known TFBS, respectively ($p < 0.05$). The most common TFBS across the discovered motifs were members of the DOF family, with CDF5, DOF1.7 and DOF5.1 binding sites being present in 3 and 4 of the total discovered motifs in the up-regulated and down-regulated DEGs respectively. The most common discovered sequences overlapping with the members of the DOF family were ARAACAGA in up-regulated and CAAAAAAAA in down-regulated DEGs. Members of the DOF family of transcription factors have been characterized in being involved in cold response(Yanagisawa 2016). Moreover, *PISTILLATA* (PI), member of the MADS-box factors family MIKC, and the C2H2 zinc finger factor SGR5 binding sites, were present in 3 and 4 of the discovered motifs in the up -regulated DEGs, respectively. There was an overlap of the most common TFBS across the discovered motifs of the up-regulated DEGs and the down-regulated DEGs (Table 1). Therefore, within the promoter regions of the up- and down-regulated DEGs, there are motifs present that could be potentially informative for classifying the gene expression.

Table 1: The top known transcription factor binding sites (TFBS) that significantly overlap with the enriched discovered motifs along the upstream and downstream regions of up and down-regulated regions of up and down DEGs.

TFBS	Number of discovered motifs	Gene	Family
MA1159.1	4	SGR5	C2H2 zinc finger factors (class)
MA1268.1	3	CDF5	DOF
MA0556.1	3	DOF1.7	DOF
MA0558.1	3	FLC	MIKC
MA0559.1	3	(PISTILLATA) PI	MIKC
MA0940.1	3	AP1	MADS box factors (AP1)
MA1085.1	3	WRKY40	WRKY (class)
MA1281.1	3	DOF5.1	DOF
MA1367.1	3	AT1G76870	Trihelix
MA1380.1	3	TCX6	CPP
MA1182.1	5	RVE8	Myb-related
MA1184.1	5	RVE1	Myb-related
MA1380.1	5	TCX6	CPP
MA1277.1	4	DOF1.7	DOF
MA1281.1	4	DOF5.1	DOF
MA1268.1	4	CDF5	DOF
MA1278	4	DOF3.4	DOF

.1			
MA1279.1	4	DOF1.5	DOF
MA1190.1	4	RVE5	Myb-related
MA0933.1	4	AHL20	HMGA factors
MA1267.1	4	DOF5.8	DOF

Known transcription factor binding sites within *Col-0* do not accurately predict the expression response to cold using random forests.

Given the many known and documented transcription factor binding site motifs that exist for different transcription factors, many of which are known to be involved in cold response, and given that we found TFBS motifs in the cold DEGs, we asked if these TFBS could predict genome-wide expression response to cold. We used a dataset of known TFBS in putative promoters 1kbp upstream of genes in the *A. thaliana* accession *Col-0* (Takou et al. 2022) to test whether they can predict DEGs using random forests classification trees (Figure S1). For this analysis, we also split the genes based on up-regulated and down-regulated and compared each group to all non-DEGs. Even though both random forests models correctly predicted the gene regulation status for approximately 80% of the genes, the prAUC of each of those models was low. The probability AUC-ROC (prAUC) for predicting up-regulated DEGs was 0.5013 (permutation based p = 0.17) and the prAUC for predicting down-regulated DEGs was 0.499 (permutation based p = 0.22). Therefore, RF models based on known TFBS involved in temperature related reactions were not able to predict DEGs with high accuracy within this dataset. This, taken together with the various discovered motifs within the upstream and

downstream regions of the genes, indicates that a method that incorporates more genetic information within the code might improve the prediction of regulatory status better within specific conditions.

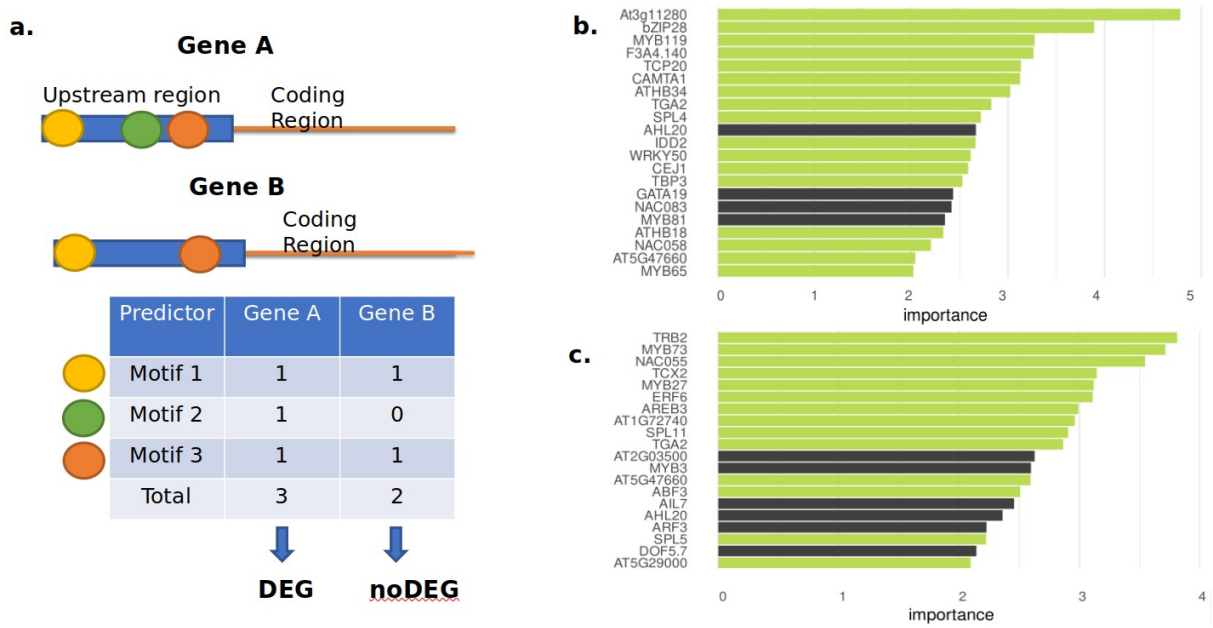

Figure S1: Known transcription factor binding sites of *Col-0* poorly predict the expression response to cold, with PRAUC = 0.503. **a.** We used a database of known TFBS upstream of *Col-0* to predict whether a gene is DEG or none DEG. Each TFBS was scored as being present (1) or not present (0) in each feature. **b.** The random forest features with the highest relative importance within the up regulated DEGs within *Col-0*. **c.** *T*he random forest features with the highest relative importance for predicting down-regulated DEGs within *Col-0*. In green and black it is marked if a gene's relative importance is significant or not significant, respectively. The *p* value was estimated based on 100 permutations of the dataset.

Predicting gene expression regulation based on regulatory regions using naive methods is possible

We next investigated whether a currently commonly used machine learning method could learn *cis* regulatory motifs simply from sequence and DE data alone. If *de novo* motifs can be identified in DE genes, even though known TFBS were not good predictors of expression in the RF model, we might obtain better predictions learning *de novo* important motifs. To this end, we trained convolutional neural networks (CNNs) that incorporate the upstream and downstream genetic sequence as input layer via one hot encoding (Figure 2a; see Material and Methods). We trained 1,344 models with different combinations of the hidden layer's parameters to find the model architecture that best predicts the genes DE (Table S2). We trained the models using the information from all 5 accessions that had available pseudogenomes to predict up- or down-regulated DEGs separately.

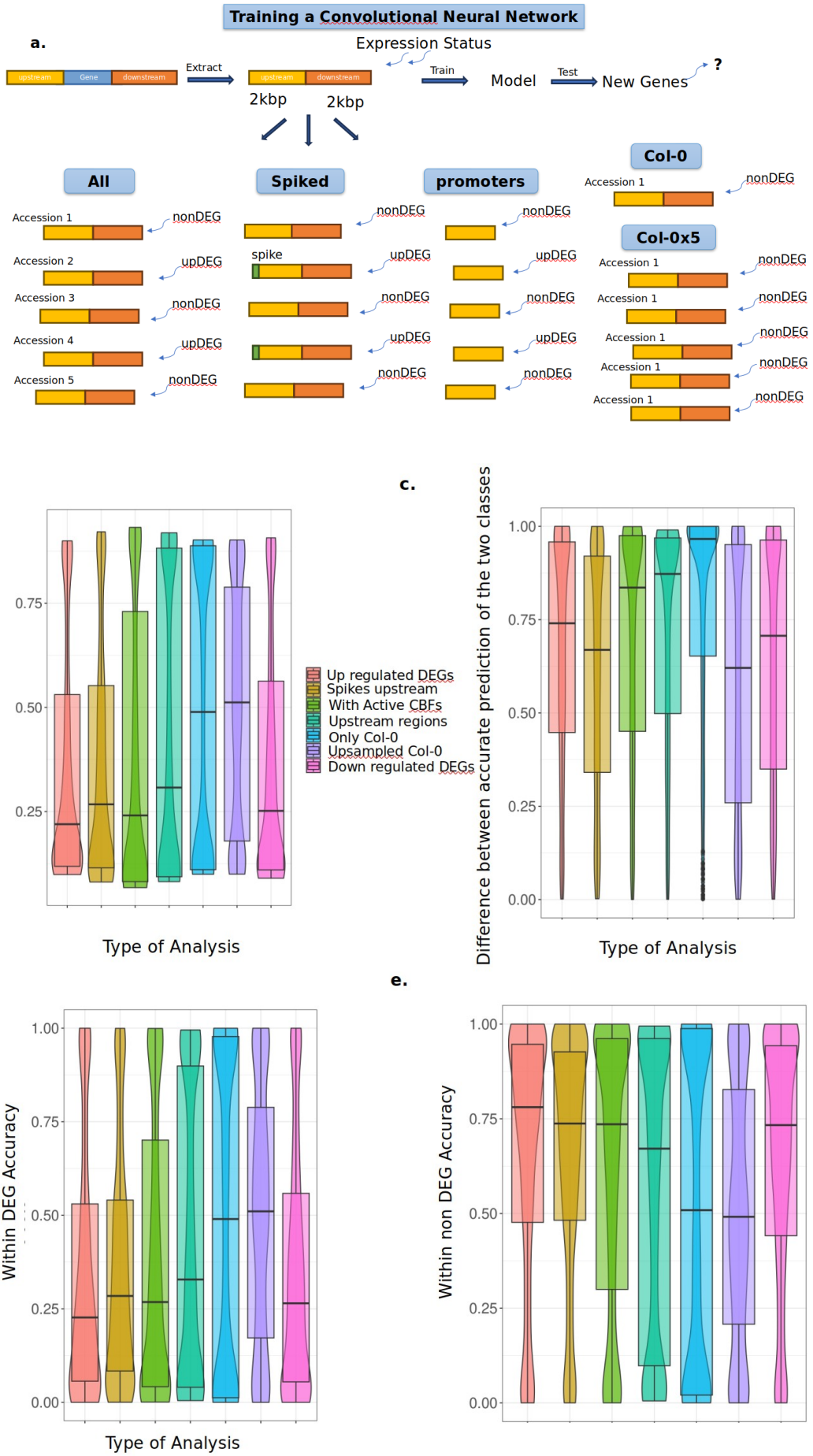
Training a Convolutional Neural Network
a.
Expression Status
upstream
Gene
downstream
Extract
Train
Model
Test
New Genes
?
2kbp
2kbp
All
Spiked
promoters
Col-0
Col-0x5
Accession 1
Accession 2
Accession 3
Accession 4
Accession 5
nonDEG
upDEG
spike
b.
prAUC of test set
Type of Analysis
Up regulated DEGs
Spikes upstream
With Active CBFs
Upstream regions
Only Col-0
Upsampled Col-0
Down regulated DEGs
c.
Difference between accurate prediction of the two classes
Type of Analysis
d.
Within DEG Accuracy
Type of Analysis
e.
Within non DEG Accuracy

Figure 2: Training Convolutional Neural Networks to predict up-regulated or down-regulated DEGs between specific environments. a. Description of the different type of input data used to train CNNs to predict up- or down-regulated DEGs vs non DEGs. The analysis "up-regulated DEGs" included the five *A.thaliana* accessions and both upstream and downstream regions for up-regulated DEGs. The analysis "spiked upstream" included training on up-regulated DEGs with the first 5bp being replaced consistently with the same sequence. "With active CBFs" refers to the training of models using the two accessions with all *3 CBF* genes active. The analysis "upstream regions" included only the upstream regions of the up-regulated DEGs as input. "Only *Col-0"* was used to train models, as well as when it was five times oversampled ("Col-0x5"). Finally, the results under the category "down regulated DEGs" show the outcome of training in all **b.** prAUC values of the test set in the different analysis. c. The difference in accurately predicting the up/down-regulated DEGs to the non DEGs in all different analysis. d. The within up/down-regulated DEGs prediction accuracy of the test set in all analysis. e. The within non-DEGs prediction accuracy of the test set in all analysis.

The tested CNN model architectures have had a wide variety of accuracy in predicting the DEG of each gene (Table S5). We extracted the prAUC of each model during training and moreover, we estimated the prAUC of the test set for each trained CNN model. The median and mean prAUC values during training for predicting up-regulated DEGs vs non-DEGs were 0.5013 and 0.5017, respectively, indicating that many tuning parameter settings did not lead to substantial learning. Note that the test set is not over sampled, but it is the original ratio which is mostly non-DEGs. Overall, the median prAUC in the test set of the models was 0.229, with a mean of 0.354 (Figure S2). We identified 15.6% of the models with prAUC in the test set values above 0.80.

To gain insight into whether tuning parameters had any consistent effects on model learning, we used a generalized linear mixed on the prAUC of the test set using the first layer's filters as random effect and the number of convolutional and dense layers, with pool stride and dropout rate as fixed effects. Both the number of convolutional and dense layers had a significant

negative impact on the prAUC of the test set, with p = 0.000754 and p = 3.9708e-16, respectively, suggesting more complex models were more poorly trained. Additionally, the pool stride (p = 2.918e-10), which indicates the number of base pairs that are considered together in each filter, and the dropout rate, or how many model nodes are used each time to predict, had a significant negative impact of the prAUC of the test set. However, none of the interactions between the fixed effects were significant, suggesting little importance of specific combinations of the tuning parameters on learning.

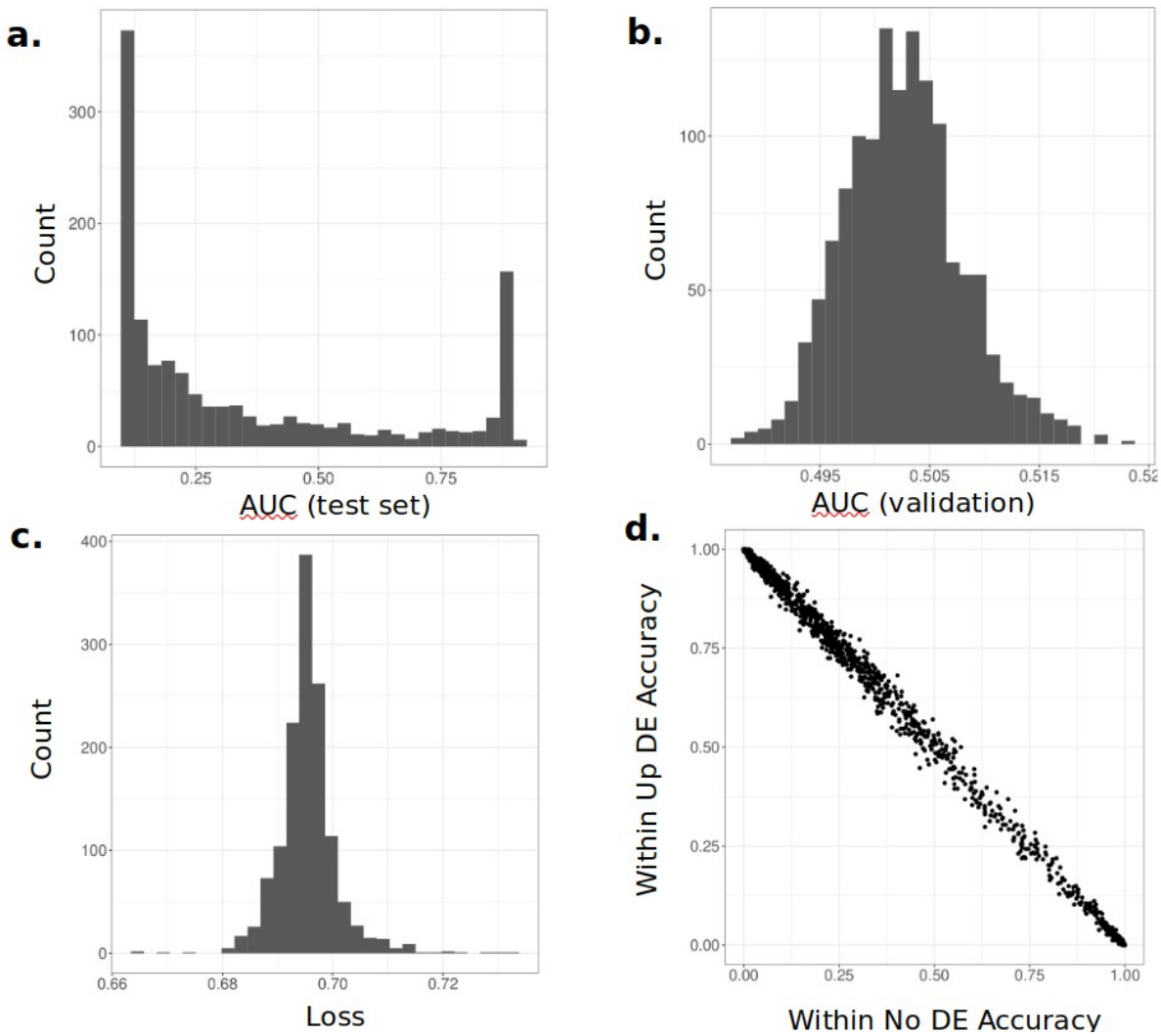

Figure S2: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs non DEGs. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class in the test set, The distributions of summary statistics for all tested CNN architectures are given.

The trained models for predicting down-regulated DEG or non-DEGs (Figure S3) were very similar with the results observed for the up-regulated DEGs vs non-DEGs. The median prAUC of the models during training was 0.5009 and the mean prAUC was 0.5012. The median prAUC of the test set for the down-regulated DEGs vs non DEGs was 0.253 and the mean 0.363 (Table S6). The same hyperparameters as for predicting up-regulated DEGs were identified to have a significant negative effect with *p-values* of 1.511e-07, 0.0168, 4.068e-08 and 0.0485 for number of convolutional layers, number of dense layers, pool stride and dropout respectively.

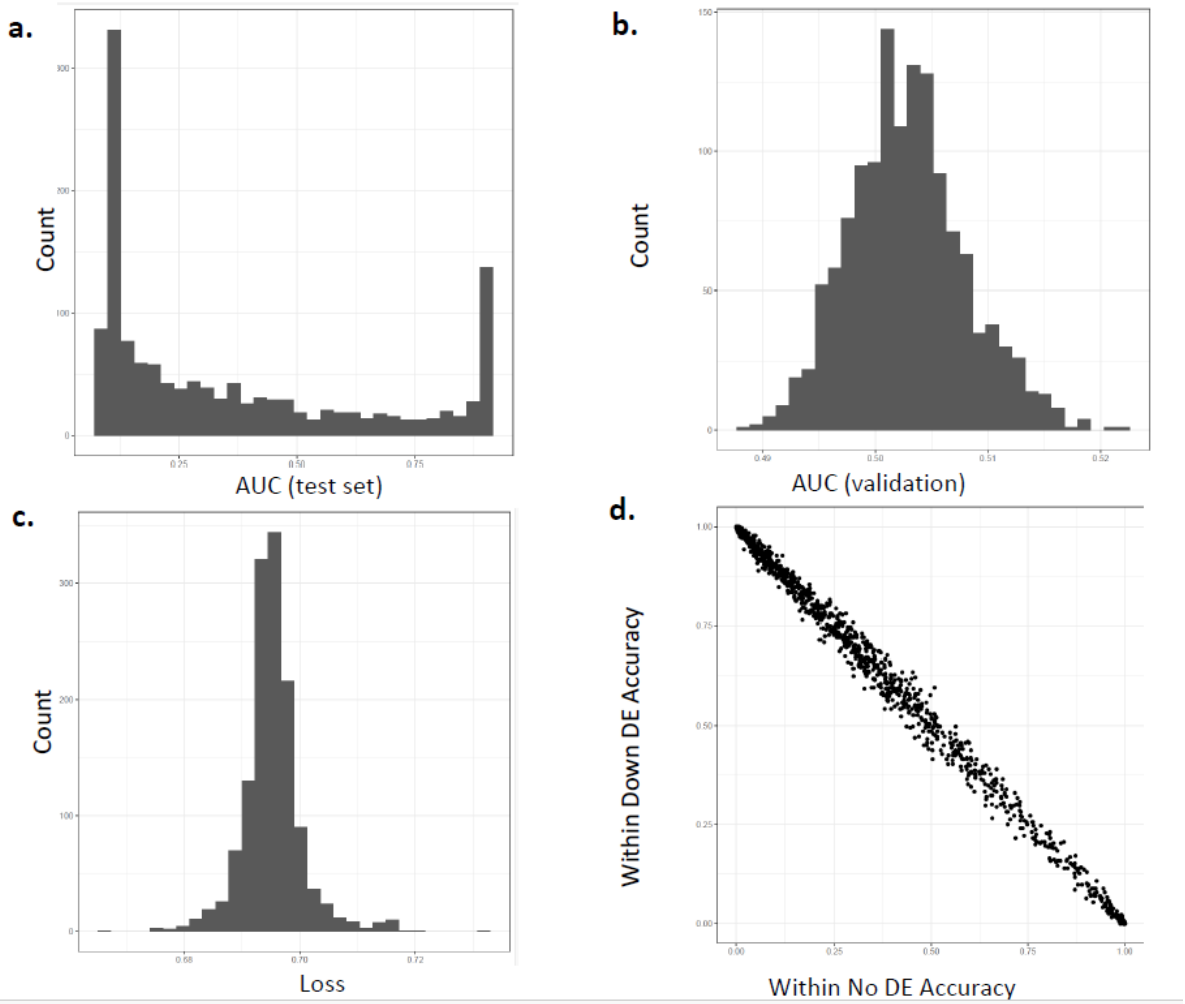

Figure S3: Summary statistics of all models trained during the grid search for predicting down-regulated DEGs vs non DEGs. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class in the test set, The distributions of summary statistics for all tested CNN architectures are given.

We identified the best models that could predict the up-regulated DEGs. For that purpose, we selected a model that had prAUC in the test set above 0.7. Moreover, in order to understand the models' potential to predict each class separately we estimated for each class a prediction accuracy as the number of correct predictions of the class divided by the total number of the genes in the class tested. So, after this prAUC cutoff, we picked a model whose prediction

accuracy difference was the lowest. The best model for predicting up regulated DEGs had a prAUC in the test set of 0.701, and prediction accuracy in the test set was 0.368 for up regulated and 0.693 for non DEG (Table 2). The class prediction accuracy difference was 0.325. The best model for predicting up-regulated DEGs had 2 convolutional layers and 2 dense layers, a dropout rate of 0.25 and pool stride 4. Using the same criteria, we were able to identify the best model in predicting the down-regulated DEGs. The prAUC of the test set for this model is 0.7128 and the difference in the class prediction accuracy is 0.319, with per class accuracy being 0.365 for down-regulated DEGs and 0.665 for non-DEGs. The model had 3 convolutional and 2 dense layers, pool stride of 8 and dropout rate of 0.25.

We assessed whether these models performed better than randomness by permuting the labels and training the selecting model 100 times (Figure S4). For each permutation we estimated the prAUC of the test set and the prediction accuracy between the two predicted classes. The significantly higher prAUC for the up-regulated DEGs ($p < 2.2e-16$) and down-regulated DEGs ($p < 2.2e-16$) than the permuted dataset indicate that our picked models perform better than randomly picking genes.

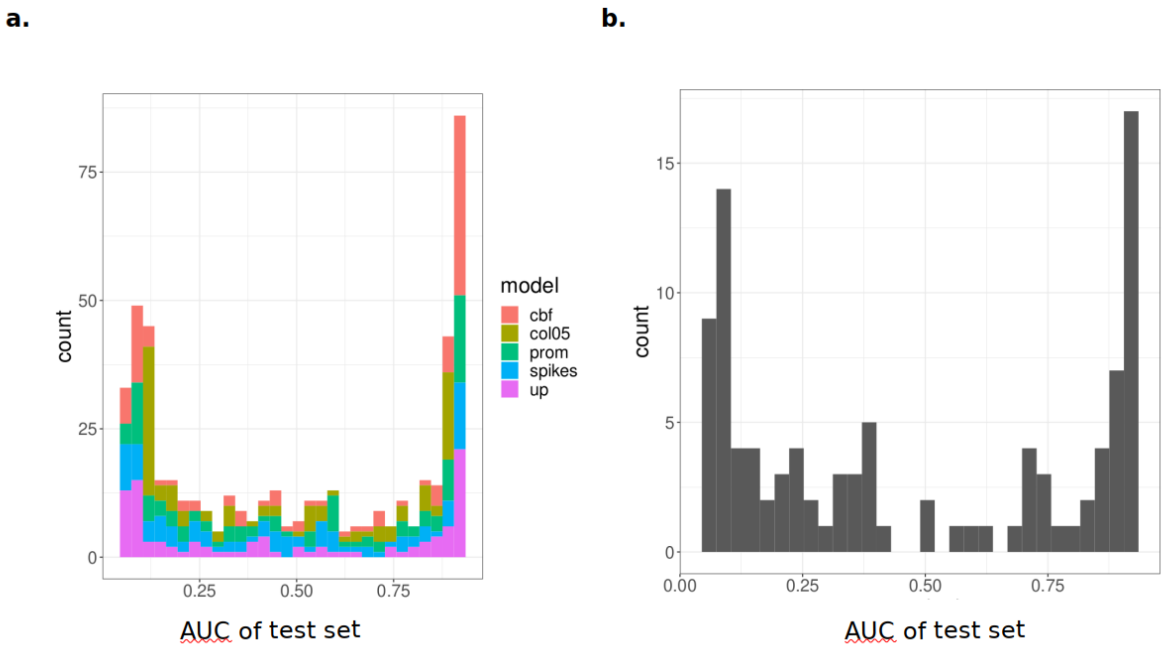

Figure S4: Distribution of the prAUC in the test set for the permuted datasets of a) up-regulated DEGs (up) and b) down-regulated prAUC,.For the up-regulated DEGs we permuted the labels for all the subsequent analysis, specifically for when only the 2 accessions with all 3 CBFs are active (cbf), for up-sampled *Col-0* dataset (col05), for only the upstream regulatory regions included in the analysis (prom) and so when a spike as positive control was incorporated (spikes).

Table 2: The best model for each training performed are given. For each type of analysis run (Figure 2), we selected the best performing model based on the prAUC value of the test set and the the difference in accurately predicting the DEGs vs non DEGs. For each model, the model parameters, the training, validation and testing prAUC, loss and accuracy of predicting both DEG classes is given.

	Up regulated DEGs	Down regulated DEGs	Spiked Up DEGs	Upstream regions of up DEGs	Up regulated DEGs of only *Col-0*	UpDEGs of upsampled *Col-0*	Genotypes with active all *CBF* genes.
1st Conv. filters	128	64	128	64	128	128	128
2nd Conv. filters	128	64	None	128	None	128	128
3rd Conv filters	None	128	None	128	None	None	64
Conv width	4	4	4	4	4	8	8
Pool width	4	4	8	4	4	4	4
Pool stride	4	8	4	4	8	4	8
Dropout	0.25	0.25	0.25	0.1	0.1	0.1	0.25
1st Dense layerunits	64	128	128	128	128	128	64
2nd Dense units	128	None	128	None	128	128	None
Number of conv. layers	2	3	1	3	1	2	3
Number of dense layers	3	2	3	2	3	3	2
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
loss	0.692332	0.6912516	0.6912327	0.6917669	0.6909633	0.6922191	0.6915633
accuracy	0.6213274	0.682407	0.6737391	0.6661143	0.6927858	0.6368272	0.6728636
accuracy of DEGs	0.3684211	0.3650794	0.4802343	0.3394294	0.3479532	0.3684211	0.3411514
Accuracy of non DEGs	0.6936842	0.6849123	0.6657669	0.6717727	0.6975439	0.6807018	0.681660371 08139

prAUC test set	0.7008184	0.7128399	0.7040513	0.7004735	0.7207444	0.7034359	0.7108139
prAUC training	0.4996175	0.5023556	0.5229402	0.4994888	0.4971713	0.5053142	0.4998409
PrAUC of validation set	0.500912	0.5024606	0.5784325	0.5094903	0.4851478	0.5178313	0.0.50604
Loss of the validation set	0.6931664	0.693195	0.6922947	0.6930951	0.693379	0.6930395	0.693965
Epoch of best model	37	50	49	29	49	48	49
Difference in accuracy of DEG classes	0.3252632	0.3198329	0.1855327	0.3323433	0.3495906	0.3122807	0.3405089

Number of discovered motifs can impact the correct prediction of up-regulated DEGs

We investigated the profile of the genes that are accurately predicted within the up regulated DEGs by the selected model with the highest achieved performance. Out of the 4,564 up DEGs across all five genotypes in the test set, 592 were accurately predicted. The correctly predicted genes had a significantly higher (wilcoxon rank sum test p = 0.018) average number of discovered motifs in their regulatory regions (median of 12 motifs) by the STREME analysis than the genes that were incorrectly predicted (mean of 10; Figure S5a). Interestingly, when we repeated this analysis for the down-regulated DEGs, we did not detect a different accumulation of TFBS within the correctly predicted genes. Both the correctly and wrongly predicted genes had an median of 11 motifs per gene and the accumulation of the TFBS upstream and downstream of the regions was not significantly different (wilcox sum rank test p = 0.9446; Figure S5b). Moreover, the gene expression level of the genes that were correctly predicted vs

not correctly predicted was not significantly different (based on Kolmogorov Smirnov test $p > 0.05$) for both down-regulated DEGs and up-regulated DEGs.

Up until this point, the discovered motif diversity, the CNN training results, and even the best model's behavior were very similar for up-regulated DEGs versus down-regulated DEGs. For the rest of the analyses, we focused on training CNN models for predicting up regulated DEGs versus non DEG, as we do not expect that altering the input of the CNNs training will alter the observed pattern between the two regulatory groups.

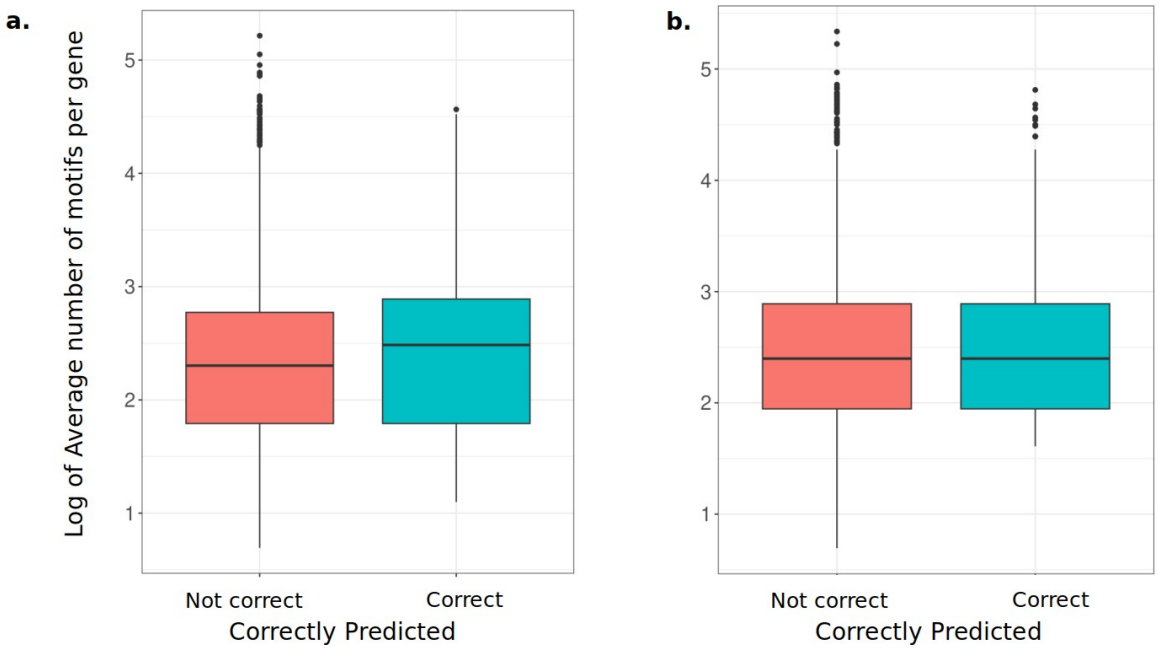

Figure S5: The average number of motifs per gene for the correctly predicted and not correctly predicted DEGs by the best model predicting a. up-regulated DEGs or b. down regulated DEGs..

CNNs imperfectly identify a simple artificial signal within the regulatory regions.

Because many tuning parameter values resulted in models that learned little (training prAUC near 0.5), we sought to assess potential sources of constraints on these models. We investigated whether the tested model architectures can learn a simple motif within the sequences that perfectly predicts DEGs, a motif we refer to as a "spike". If these spiked models are perfectly able to predict DEGs in this setting, it suggests that the limitations we observed with real data arise from a lack of such good, simple predictors in the sequence. If models have limited learning in this setting, it suggests that the model architecture is limited in its ability to find signals in the data were using, perhaps due to the complexity and abundance of sequence data (2,000 bp per gene).

To test these hypotheses, we included a spike sequence of 5bp in the place of the first 5bp of all up-regulated DEGs (Figure 2a). The spike's sequence (AAGGG) does not overlap with the sequence of any of the discovered enriched sequences with the up DEGs. The expectation was that the accuracy of the models would increase as they have a perfect sequence to use for predicting the up-DEG group. During training the prAUC values had median of 0.5018 and mean of 0.5046. The median prAUC in the test set of the spiked models was 0.36, with a mean of 0.266, which was on average higher than when trained without the spikes (Table S7). The distributions of the prAUC values in the test set between the spiked and non-spiked models were significantly different (Kolmogorov Smirnoff test, $p < 2.2e\text{-}16$; Figure 2b). Generally, the spiked models had fewer overfitted models than the non-spiked models for both up-regulated and non-DEG classes (Figure 2c; Figure S6). In total 188 spiked models with prAUC values of the test set above 0.8 were detected and 561 models with prAUC values of the test set below 0.2, and thus,

in total, 55.7% of the models tested are overfitted to one of the two classes. In contrast, among the non-spiked models, 61.5% of the models are overfitted.

When we used the same criteria to pick the best performing models for the spiked data (permutation $p < 2.2e-16$) as for the non-spiked data, we identify a model that predicts the two classes better than the non-spiked best model. prAUC of the test set (0.704) is only marginally higher by 0.03. In contrast, the difference in the prediction accuracy of the two classes in the test set is 0.185, which is 1.75 times lower than the best model predicting non-spiked up-regulated DEGs. This improvement is due to the higher accurate prediction of the up-regulated DEGs (0.48) in the spiked best model than the non-spiked best model (0.36).

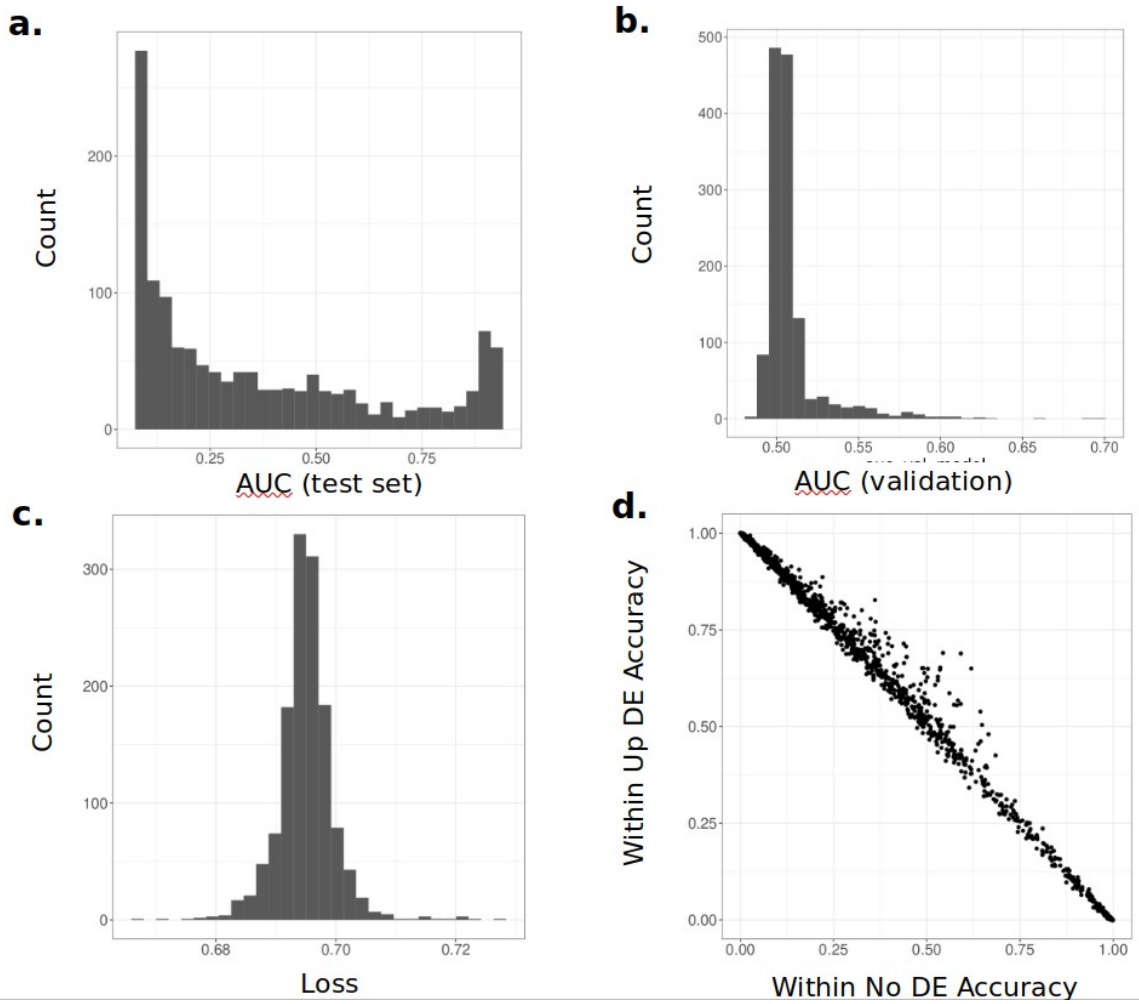

Figure S6: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs no DEGs, while including a spiked region at the start of the up-regulated genes' putative promoters. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class.

There is no significant correlation between the prAUC values in the test set of the spiked and non-spiked models under the same values of tuning parameters (p=0.96, ρ =0.0012), indicating that the inclusion of the spikes did not improve models with good tuning parameters for the real data, but instead were best captured with distinct tuning parameters (Figure 3). When we further

investigated the different hyperparameters' impact on the prediction accuracy of each class, we noticed that the spiked models with 1 convolutional layer had higher accuracy and prAUC in the test set and simultaneously smaller difference between the accurate prediction of both classes than the models with 2 and 3 convolutional layers. (Figure 3a). The fact that the models with less convolutional layers performing better that with more of them was also true for the spiked models. A generalized mixed model including the number of convolutional layers, dense layers, pool stride and dense units as fixed effects indicated that the number of convolutional layers had a significant effect on the difference of the prediction accuracy of the two classes (p = 8.091e-10). Fewer layers were predicting more accurately, as the model's estimates were negative. This observed pattern is not as extreme as for the non-spiked models (p = 0.01023; Figure 3a). Other parameters that were found to have a significant negative impact on the spiked models' difference in the prediction accuracy of the two classes were the number of dense layers (p = 9.036e-05), the pool stride (p = 0.002137) , the first dense layer's units (p=0.048), as well as the interaction of the number of convolutional and dense layers (p = 0.00983). All these factors, except for the interaction of the number of dense and convolutional layers, had a significant effect on difference between the accuracy of the two models for the non-spiked models. Specifically, the number of dense layers, the pool stride, and the number of convolutional layers had p values of 6.191e-06 and 0.0102 respectively (Figure 3b-c). The estimates for all parameters were negative in both spiked and non-spiked data, which means that generally models with fewer layers were able to predict with higher per class accuracy in the test set. This result suggests that although different specific sets of tuning parameters for good for spiked data versus the non-spiked, the general effect of tuning parameters, and specifically model complexity, was consistent. Therefore, a consistent signal within the up-regulated DEGs can

improve the model performance, indicating that the regulatory complexity among genotypes is quite high, that we lack information in the sequence data for prediction, or that the DEG response is highly stochastic.

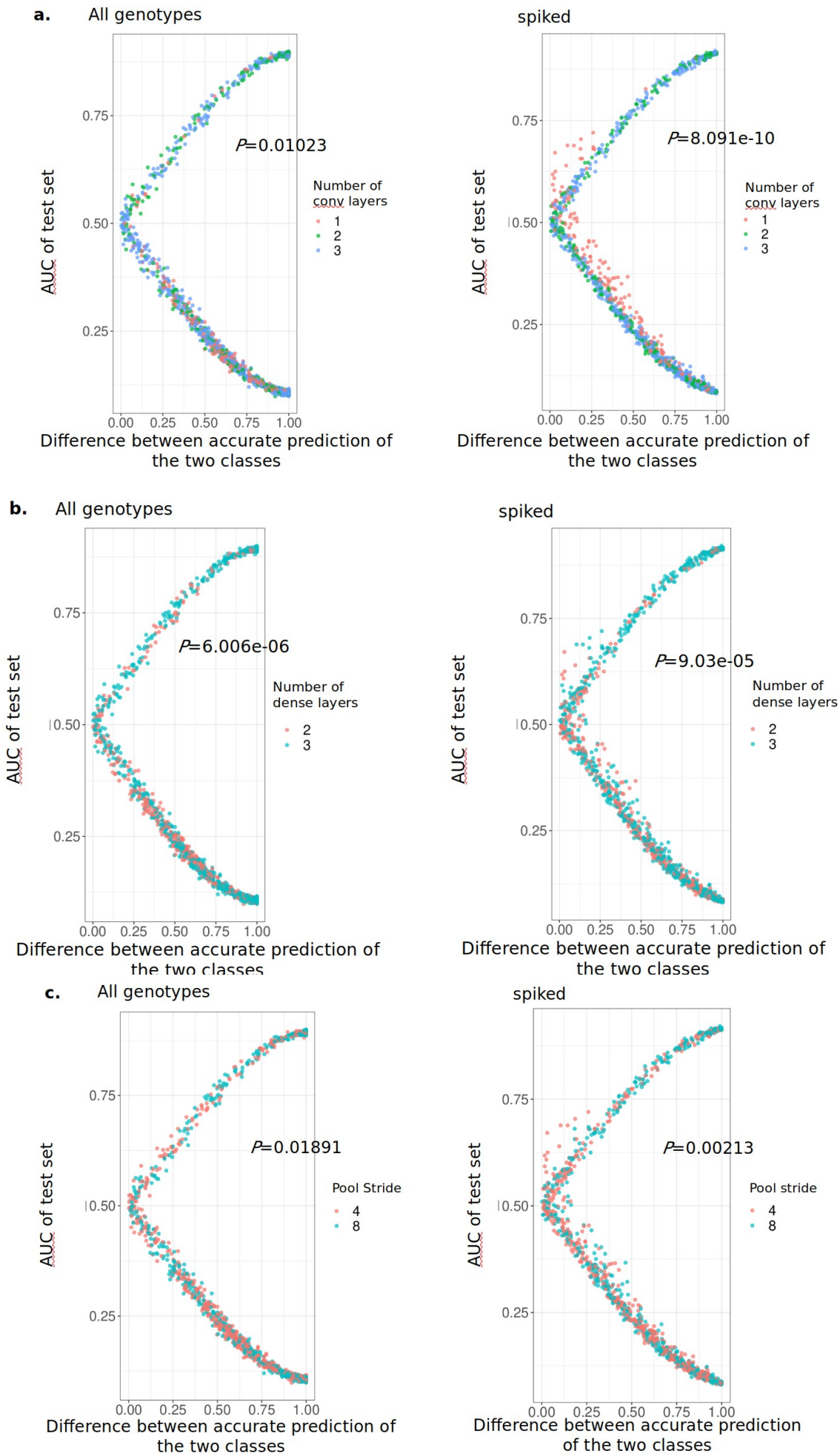
a.
All genotypes
spiked
P=0.01023
P=8.091e-10
Number of conv layers
1
2
3
AUC of test set
Difference between accurate prediction of the two classes
b.
All genotypes
spiked
P=6.006e-06
P=9.03e-05
Number of dense layers
2
3
c.
All genotypes
spiked
P=0.01891
P=0.00213
Pool Stride
Pool stride
4
8
0.00
0.25
0.50
0.75
1.00

Figure 3: Simple spiked models can accurately predict an introduced spike of 5bp at the beginning of the up-regulated DEGs promoter region. Comparison of the prAUC in the test set and the difference between the accurate prediction of the two classes for non- spiked (left) and spiked (right) input data. Comparisons are done for a. number of convolutional layers, b. number of dense layers and c. pool stride.

Predicting only among genotypes sharing alleles for a major *trans* eQTL does not improve the prediction of the up-regulated DEGs.

Since the results with the sequence spike indicate some data complexity limited our CNN predictions, we investigated whether we could simplify the prediction task by homogenizing known large-effect *trans* mutations (CBFs). We trained models predicting up-regulated DEGs vs non DEGs on *Can* and *Cvi*, which are the two accessions that have functional copies of all three major well-studied cold-response *CBF* transcription factors, including *CBF2*. These transcription factors show variation in functionality among genotypes, and in the other natural genotypes tested at least some copies are non-functional. The median and mean prAUC during training was 0.5014 and 0.5016 respectively, while the prAUC of the test set had mean value of 0.38 and median 0.268 (Figure S7; Table S8). The difference in the accurate prediction of the up-regulated DEGs and non DEGs within the test set has a median of 0.835 and mean 0.701, which is on average higher than the difference in prediction accuracy of the models applied to all genotypes. This difference is reflected in the percentage of model architectures, many of which are overfitted for predicting the upregulated DEGs; 45.8%. of the models in contrast to the 22.5% overfitted in predicting the non DEGs. Overall, those models did not perform better than the original models including all the *A. thaliana* accessions. The best model trained on only the two accessions with all copies of *CBF* are functional had prAUC in the test set of 0.7 and difference in predicting accuracy of the two DEG groups of 0.31. The model was performing better than a

trained model (p < 2.2e-16) on a random dataset, based on 100 permutations of label identity (Table 2). This suggests that a major *trans* acting expression variant did not hinder predictability in our bigger set of diverse genotypes here.

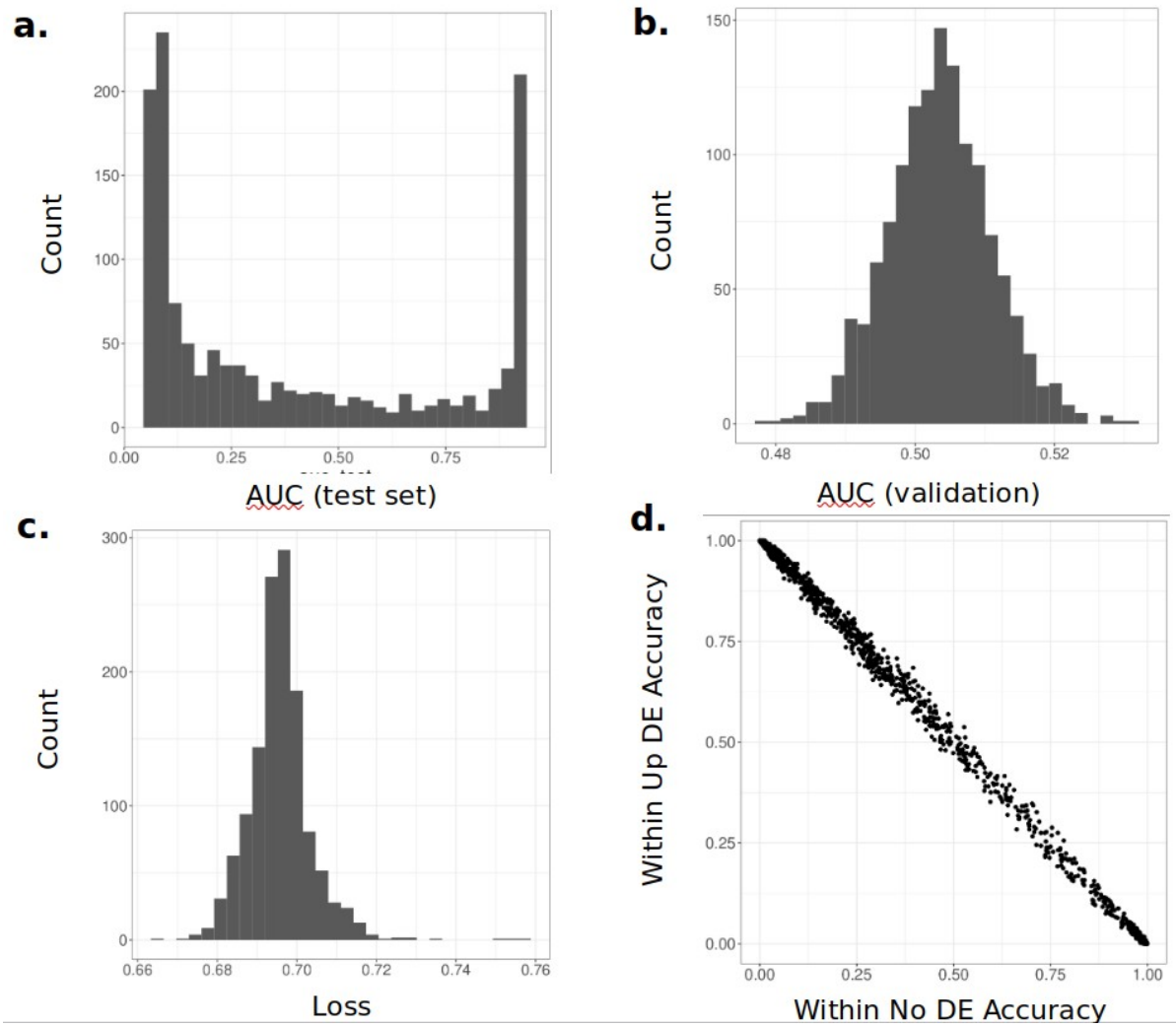

Figure S7: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs non DEGs, while using the two accessions with all 3 *CBF* genes functional. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class, The distributions of summary statistics for all tested CNN architectures are given.

Including both upstream and downstream regions improves the prediction of up-regulated DEG class

As often the *cis* elements are mostly identified in the upstream regions (Lenhard et al. 2012), we tested whether including the downstream coding region has limited the prediction accuracy, perhaps by incorporating noise. Noise might be introduced by incorporating long sequences that do not usually have a lot of TFBS, such as the downstream regions. Therefore, we retrained all the models by including only the 2kbp upstream regions for all 5 accessions, compared to the 1 kb up and 1kb downstream of the coding region in the approach presented above (Figure 2a). The median prAUC of the training set was 0.5 and the median prAUC (0.307) in the test set of the models trained only using the 2kb upstream sequences was slightly higher than including both the 1kb upstream and 1kb downstream regions, by 0.048 (Figure S8; Table S9). The distribution of the prAUC results in the test set of the two analysis was again significantly different (KS test $p < 2.2e\text{-}16$). The within class prediction accuracy of up-regulated DEGs (0.671) in the test set was higher than for non-DEGs (0.328). Moreover, 29.8% of the models had prAUC values in the test set above 0.8. The same pattern was observed for models that had low prAUC in the test set. Approximately 41% of the models had prAUC in the test set below 0.2, indicating overfitting to predicting all genes as up-regulated DEGs, to a total of 71.1% of the models overfitting to either class. In comparison, including both upstream and downstream regions leads to overfitting of 61.5% of the models. (Figure 2c, Figure 2d). Therefore, excluding the downstream regions from the analysis leads to more cases when the models get fixed in recognizing only one class.

The best model in predicting up-regulated DEGs and non DEGs trained only on putative promoter regions had 3 convolutional layers, 3 dense layers and a dropout of 0.1 (Table 2). The prAUC in the test set was 0.7004 and the difference in accurately predicting up-regulated DEGs and no DEGs is 0.332, which is a similar performance to the best model learning using both upstream and downstream regions (Table 2). Its performance is better than models with the same architecture trained on random input (p < 2.2e-16). Therefore, including only the 2kb upstream regulatory regions of the genes during training does not significantly alter the predictive ability compared to including 1kb upstream and 1kb downstream.

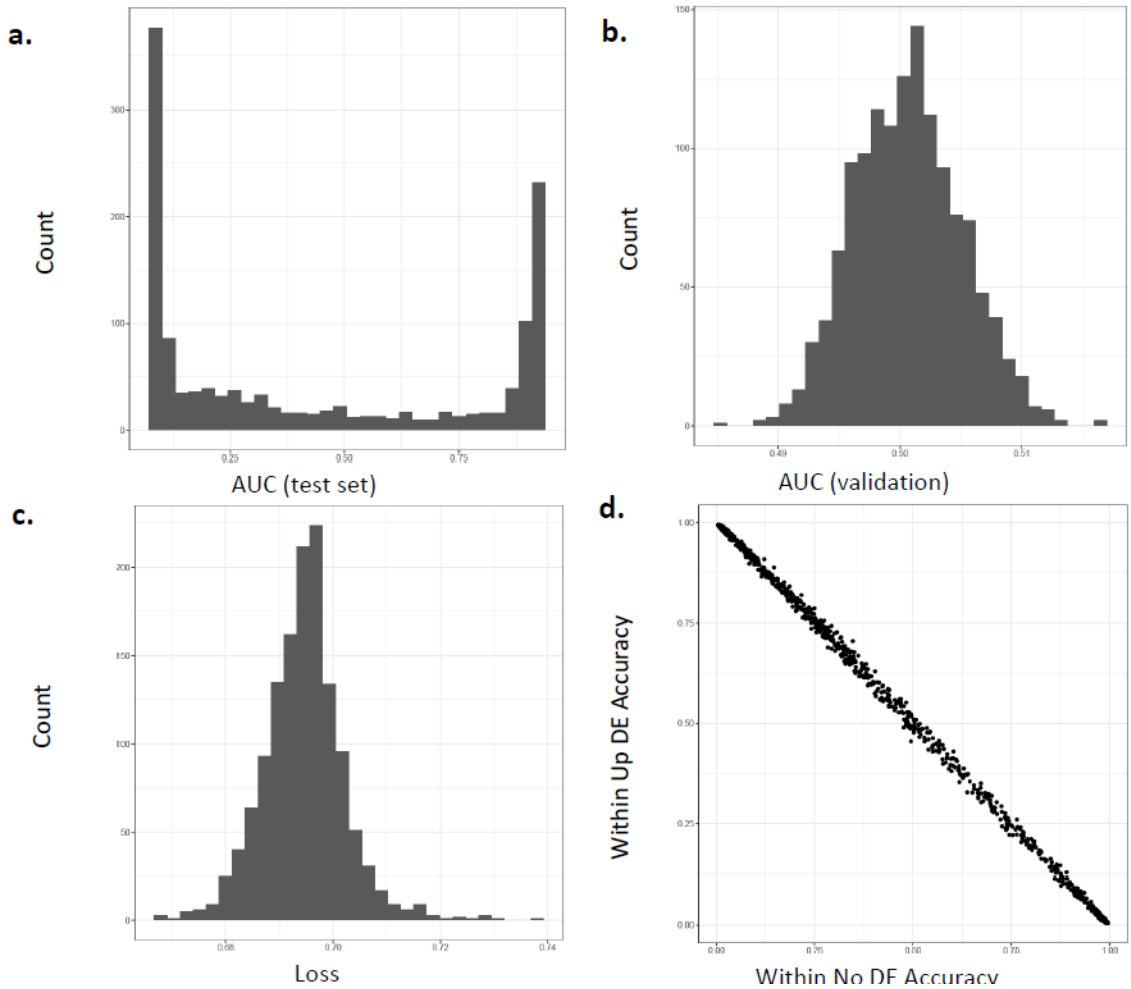

Figure S8: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs non DEGs, while including only up-regulated genes' putative promoters. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class, The distributions of summary statistics for all tested CNN architectures are given.

Genetic diversity improves accuracy of differentially expressed genes.

Genetic diversity is an important dimension of natural populations. Incorporating multiple genotypes in the STREME analysis has had a positive impact into discovering more enriched motifs. If we consider the introduced spike in the upstream region of up-regulated DEGs as a large effect mutations due to its 100% frequency within the group, we also showed that the majority of the information that the models learn is based on small effect mutations. Small effect mutations can be better identified when incorporating more genetic diversity into any kind of analysis, as the less frequent variants are easier to identify. We finally compared how effective CNNs were when including multiple genotypes in the training set, by training all the different model architectures including only the most commonly studied *A. thaliana* accession, *Col-0* (Figure 2a). We repeated this twice, once that had the original number of genes in *Col-0* and once that we up sampled the *Col-0* genes enough times to have the same size of training set as for the complete set of the genotypes. In both cases, we oversampled the up regulated DEGs to a 1:1 ratio with the no DEGs. The training of the non-oversampled and oversampled sets wielded median prAUC values of 0.5004 and 0.501 respectively.

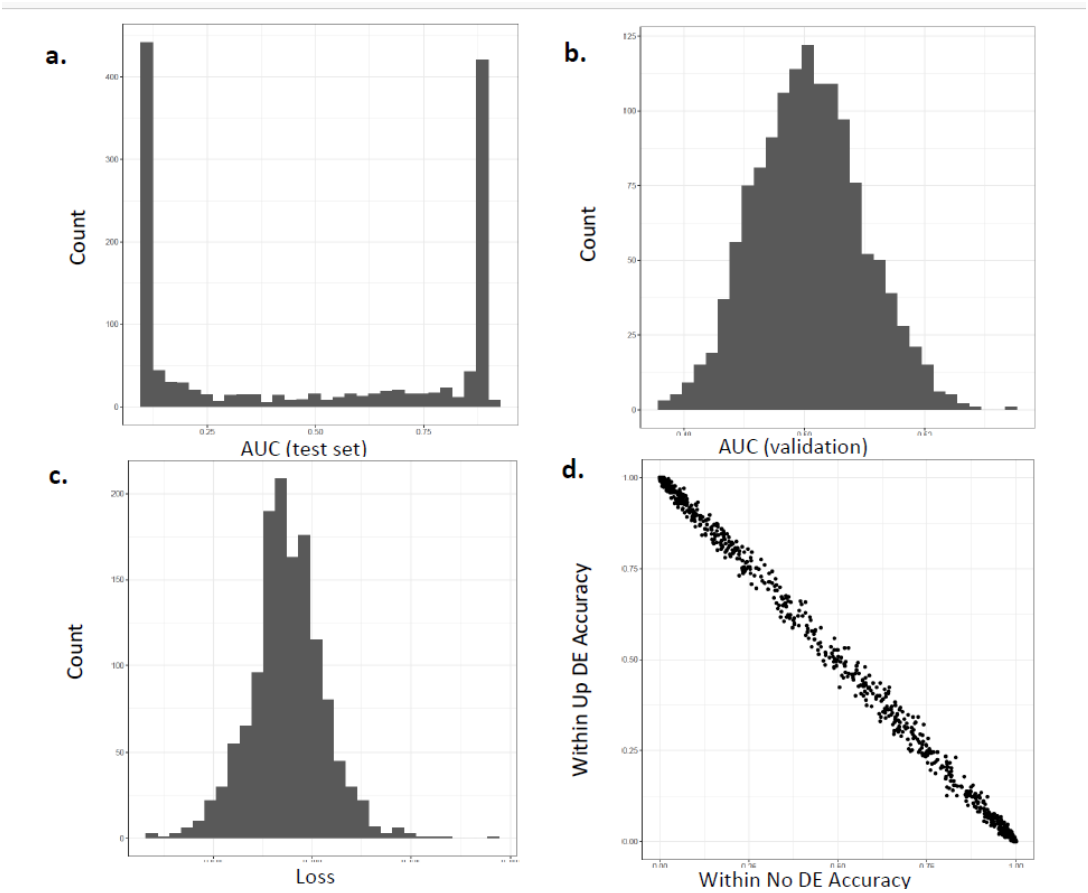

Figure S9: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs no DEGs, while including only *Col-0*.The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class, The distributions of summary statistics for all tested CNN architectures are given.

When compared to the models that have been trained on the *Col-0* dataset only, we see that the median AUC of the test set is higher, as it is at 0.488, for the multiple genotypes model (Figure 2b, Figure S9; Table S10). The accuracy was significantly impacted by the number of replicates, as the models trained on the up sampled *Col-0* set had median prAUC of the test 0.508 (KS test p = 8.8e-16; Table S11; Figure S10). During training, both datasets had a high proportion of overfitted models. Within the up sampled dataset, 50.3% of the prAUC values in the test set were either above 0.8 or below 0.2. Training on the up-sampled dataset of *Col-0* had fewer overfitted

models than training on the non-oversampled dataset of *Col-0*, in which 77.4% of the models were overfitting in either direction. The better performance of the oversampled dataset of *Col-0* is due to the more appropriate size of the dataset; the original *Col-0* dataset has only 34,002 data points when the up-regulated DEGs are oversampled to 1:1 ratio (Wei and Jr 2013; Buda et al. 2018). Based on this observation, below we compare only the models trained on the oversampled dataset of *Col-0* to the models trained on all five accessions.

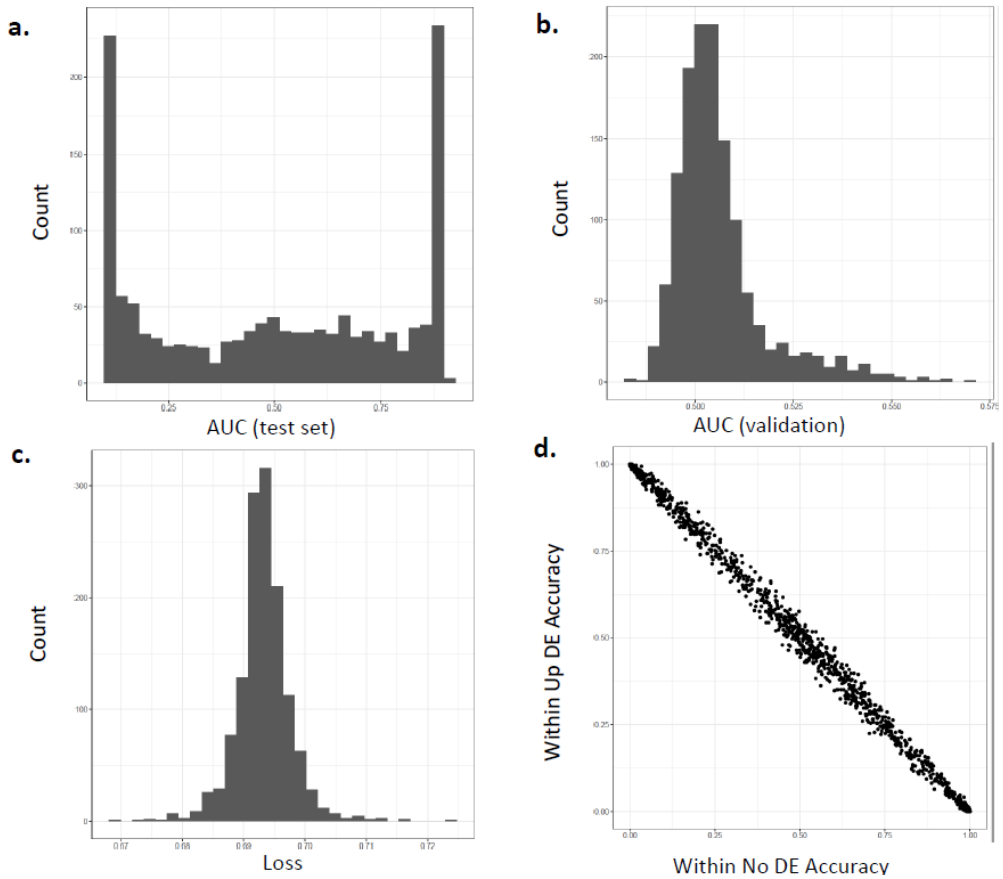

Figure S10: Summary statistics of all models trained during the grid search for predicting up-regulated DEGs vs no DEGs, while including only an up sampled set of *Col-0*. The models were evaluated by estimating the a. prAUC of the test set, b. the prAUC of the validation set during training, c. the loss of each model during training and the d. accuracy of correctly predicting each class, The distributions of summary statistics for all tested CNN architectures are given.

The models trained on the *Col-0* dataset have fewer overfitted models (50.3%) than the models trained on all five accessions (61.5%). The best model in predicting both up-regulated and non-DEGs on the *Col-0* dataset has similar architecture and performance statistics as the best model trained on all five accessions. The prAUC value in the test set trained on the oversampled *Col-0* was 0.7034 and the difference in the accurate prediction between the two classes was 0.3122. It performed better than when trained on a permuted dataset ($p < 2.2e-16$). Both the prAUC in the test set (0.7008) and the difference in accurate prediction of both classes (0.3252) indicate that training on all five accessions was performing not as well as the model trained on the *Col-0* dataset only. Therefore, including more accessions may either marginally increase the noise and not improve the accuracy of the models by a lot, or not be variable enough to capture more contrasts between alleles rather than genes.

Discussion

We used the genomic sequence of upstream and downstream regions of *Arabidopsis thaliana* genes to identify motifs related to the differential expression in cold and control conditions and also to predict the gene expression response to an abiotic stressor. Many *cis* regulatory elements have been documented (Wittkopp et al. 2004; He et al. 2016; Metzger et al. 2017) and they are often involved in gene evolution and adaptation to novel environments (Fraser et al. 2011; Josephs et al. 2015). Here, within both the up-regulated and down-regulated DEGs we found motifs of variable length that are enriched within them in comparison to the non-DEGs. The discovered motifs in *cis* regulatory regions significantly overlap with known transcription factor binding sites that are involved in the regulation of gene expression under cold stress (Hannah et al. 2006; Chawade et al. 2007; Yanagisawa 2016). Knowledge of specific transcription factor

binding sites can aid the prediction of gene expression under different types of abiotic and biotic stress (Li et al. 2019; Moore et al. 2021; Zhou et al. 2021; Smet et al. 2023). The five accessions used in this study are regulated by differential function of *CBF* and *ZAT12* regulons, leading to their variable acclimation capacity to cold (Hannah et al. 2006). The three *CBF* genes control the acclimation to cold conditions and reaction to freezing stress (Fowler and Thomashow 2002; Fowler et al. 2005). It has been shown that genetic diversity within the gene and its network is connected to adaptation to different temperatures and latitudes (Oakley et al. 2014; Mateos et al. 2017; Fagny and Austerlitz). We identified a frequent overlap with known TFBS, which are known to be involved in controlling responses to cold or freezing stress such as the RVEs (Kidokoro et al. 2023), members of the DOF family of transcription factors (Fowler et al. 2005; Fornara et al. 2015; Yanagisawa 2016; Corrales et al. 2017) and FLC (Kim et al. 2004). We also identified frequent binding sites for members of the DREAM complex, which are involved in repressing growth in response to DNA damage (Lang et al. 2021), as well as SGR5, which is differentially spliced under heat stress (Kim et al. 2016). This indicates the possibility to predict gene expression regulation based not only on factors with biological meaning, but also with genetic diversity, as the incorporation of more accessions than *Col-0*, yielded more discovered motif sequences.

However, as often the available information to the researcher is only DNA and mRNA sequence (and not experimentally verified transcription factor binding sites), we investigated the possibility of using the genetic code to predict gene expression. The need for this was also evident by trying to predict up regulated or down-regulated DEGs using random forest analysis. The random forest used known presence or absence of specific TFBS upstream and downstream

of the genes to predict DEGs. However, neither for up-regulated DEGs nor for down-regulated DEGs, the prediction accuracy was better than chance. In *A. thaliana*, using known transcription factor binding sites has not yielded very good prediction of differential gene expression (Moore et al. 2021), while the most enriched *cis* regulatory elements did not also predict the most differentially expressed genes in maize (Benoit 2021). Training CNNs, with the 2kbp upstream and downstream regions of the genes wielded better prediction accuracy. The model that had both high accuracy and had the least difference in predicting accuracy per class had prAUC in the test set of 0.7. This value is slightly lower than what has been reported in similar studies in other species, which identified models with prAUC between 75% to 85% (Washburn et al. 2019; Meng et al. 2021). In our case, higher prAUC values showed signs of overfitting; non DEGs, which comprise more than 80% of the dataset were then almost exclusively correctly predicted, inflating the accuracy and prAUC. Despite this, in all cases we were able to train models that could predict differential gene expression within conditions substantially better than chance.

Phenotypic responses to environment (plasticity) can be adaptive, which are ultimately fixed by populations in the specific environment (Savolainen et al. 2013). Phenotypic plasticity can also be genetically variable (i.e. GxE), with gene expression being an important example. Our hypothesis was that incorporating more genotypes in this study would improve the prediction accuracy as more genetic variation in *cis* sequences and expression would be incorporated during training. Incorporating genetic variability between species has been successfully used to train models with simple architecture to predict gene network relationships (Ferebee and Buckler 2023). Within maize, when more than one genotype was used for predicting gene expression, then the results were more accurate (Zhou et al. 2021). We did not observe the same pattern here;

in fact an oversampled dataset of *Col-0* upstream and downstream regions to the same size as the dataset including the multiple genotypes, yielded a model with higher prAUC and more comparable prediction of the two DEG classes. We do not believe that this is simply an issue of power due to the dataset being on the smaller side of what is required for machine learning. A potential contributing factor could be that the *cis* -regulatory genetic variation might not be of large enough frequency between genotypes but is stronger between genes. Additionally, the highest prAUC in the test set was achieved when the accessions with functional all 3 *CBF* genes were only included during training, suggesting *trans* effects in high frequency could limit our predictions across multiple genotypes. This pattern was confirmed when a high frequency motif was incorporated in the upstream regions of the up-regulated DEGs; the spiked dataset could be used to train models with higher accuracy. Taken together with the fact that more than half of the enriched discovered motifs within the up- and down-regulated DEGs are present in less than 50% of the genes across all genotypes and thus not very frequent, it indicates that the large number of *cis* regulatory variants across the genome in low frequency can decrease the accuracy in prediction of gene expression differences. The putative promoter region captures the *cis*-regulatory elements, which have been related to large impact to gene expression divergence and polygenic selection but are often each individually mutations of small effect (Wittkopp and Kalay 2012; Metzger et al. 2017). Incorporating a spike within the upstream region, which represents a single factor of large effect, supports this hypothesis. Our findings are consistent with machine learning predictions of protein folding, where models that accurately predict variation among genes perform worse on predicting effects of individual amino acid substitutions (Buel and Walters 2022), a task that was better suited to a distinct prediction approach (Cheng et al. 2023). During training, the models were able to detect the spike and increase the accuracy of

the up-regulated DEGs. Therefore, even though including multiple accessions can increase the accuracy of predicting GxE, care has to be taken to which accessions are selected for the training.

One explanation for the remaining variation in expression unpredicted by CNNs in this study is that these models overlook much of the biological complexity driving gene expression response to environments. Genes are expressed as a part of complicated regulatory networks that *cis* regulatory elements only partially capture. We have currently ignored epigenetic states and the sequence of the coding region. In a sorghum species, it has been more successful to predict DEGs as response to cold as models incorporated information about the coding region sequence and methylation information (Meng et al. 2021). Alongside specific transcription factor binding sites, their positions on the genome and the chromatin state have been useful to predict the transcription factor's target genes, with information that is also transferable with high accuracy between datasets (Rivière et al. 2022). Prediction accuracy of gene expression differences and the impact of *cis* variants were improved when large up- and downstream regions, as well as when open methylated regions were included in the analysis (Benoit 2021; Zhou et al. 2021). Trained bimodal CNNs based on chromatin state and DNA sequence for different cell types can differentially predict the transcription factor binding specificity, which performs better than neural networks based only on DNA sequence information (Srivastava et al. 2021). During training, therefore, using only *cis* regulatory regions to predict gene expression response to environment within *A. thaliana* can be used up to a certain degree without incorporating other data, especially when that biological information is missing and it is both complicated and expensive to acquire as would be the case in many non-model systems.

Interpretation of what might have been having an impact on the training could be difficult to understand. Even models with the best performance do not necessarily have the best interpretable structure of layers (Koo and Ploenzke 2021). Looking into the statistics of all the different models trained during each grid search can give insights into this. First, the distribution of the prAUC during testing can give clear indications of how easy it is to overfit to the model, and therefore how complex this training task can be. Despite these limitations, using machine learning methods and limited data can enable researchers to draw predictions and conclusions about reactions to specific environments within species, which we believe it would be useful for systems with less information about their (molecular) biology than the model species.

Acknowledgments

The Pennsylvania State University ACI cluster has provided us with computational support. We would like to thank S. Mahony and H. Dittberner for providing valuable feedback in running convolutional neural network analysis and M.G. Stetter, T.S. Winkler, A. Singh and J. Floret for feedback on the manuscript. This project was supported by NIH grant R35GM138300 to JRL.

References

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Adrian J, Farrona S, Reimer JJ, Albani MC, Coupland G, Turck F. 2010. cis-Regulatory Elements and Chromatin State Coordinately Control Temporal and Spatial Expression of FLOWERING LOCUS T in Arabidopsis. *Plant Cell* 22:1425–1440.

Akagi T, Masuda K, Kuwada E, Takeshita K, Kawakatsu T, Ariizumi T, Kubo Y, Ushijima K, Uchida S. 2022. Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. *Plant Cell*:koac079.

Anon. JASPAR - A database of transcription factor binding profiles. Available from: https://jaspar.genereg.net/

Bailey TL. 2021. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* 37:2834–2840.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67:1–48.

Benoit M. 2021. Hot ‘n cold: Applying the cis-regulatory code to predict heat and cold stress response in maize. *Plant Cell* [Internet]. Available from: https://doi.org/10.1093/plcell/koab269

Brown KE, Kelly JK. 2022. Genome-wide association mapping of transcriptome variation in Mimulus guttatus indicates differing patterns of selection on cis- versus trans-acting mutations. *Genetics* 220:iyab189.

Buda M, Maki A, Mazurowski MA. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* 106:249–259.

Buel GR, Walters KJ. 2022. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29:1–2.

Chawade A, Bräutigam M, Lindlöf A, Olsson O, Olsson B. 2007. Putative cold acclimation pathways in Arabidopsis thaliana identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors. *BMC Genomics* 8:304.

Chen Y, Li Y, Narayan R, Subramanian A, Xie X. 2016. Gene expression inference with deep learning. *Bioinformatics* 32:1832–1839.

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381:eadg7492.

Corrales A-R, Carrillo L, Lasierra P, Nebauer SG, Dominguez-Figueroa J, Renau-Morata B, Pollmann S, Granell A, Molina R-V, Vicente-Carbajosa J, et al. 2017. Multifaceted role of cycling DOF factor 3 (CDF3) in the regulation of flowering time and abiotic stress responses in Arabidopsis. *Plant Cell Environ.* 40:748–764.

Cubillos FA, Stegle O, Grondin C, Canut M, Tisné S, Gy I, Loudet O. 2014. Extensive cis-Regulatory Variation Robust to Environmental Perturbation in Arabidopsis. *Plant Cell* 26:4298–4310.

Erwin DH, Davidson EH. 2009. The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genet.* 10:141–148.

Fagny M, Austerlitz F. Understanding the adaptation of polygenic traits: the importance of gene regulatory networks. :22.

Ferebee TH, Buckler E. 2023. Exploring the utility of regulatory network-based machine learning for gene expression prediction in maize. :2023.05.11.540406. Available from: https://www.biorxiv.org/content/10.1101/2023.05.11.540406v1

Fornara F, de Montaigu A, Sánchez-Villarreal A, Takahashi Y, Ver Loren van Themaat E, Huettel B, Davis SJ, Coupland G. 2015. The GI–CDF module of Arabidopsis affects freezing tolerance and growth as well as flowering. *Plant J.* 81:695–706.

Fowler S, Thomashow MF. 2002. Arabidopsis Transcriptome Profiling Indicates That Multiple Regulatory Pathways Are Activated during Cold Acclimation in Addition to the CBF Cold Response Pathway[W]. *Plant Cell* 14:1675–1690.

Fowler SG, Cook D, Thomashow MF. 2005. Low Temperature Induction of Arabidopsis CBF1, 2, and 3 Is Gated by the Circadian Clock. *Plant Physiol.* 137:961–968.

Fox J, Weisberg S. 2019. An R Companion to Applied Regression. Third. Sage Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, Schadt EE. 2011. Systematic Detection of Polygenic cis-Regulatory Evolution. *PLOS Genet.* 7:e1002023.

Giri A, Khaipho-Burch M, Buckler ES, Ramstein GP. 2021. Haplotype associated RNA expression (HARE) improves prediction of complex traits in maize. *PLOS Genet.* 17:e1009568.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol.* 8:R24.

Hannah MA, Wiese D, Freund S, Fiehn O, Heyer AG, Hincha DK. 2006. Natural Genetic Variation of Freezing Tolerance in Arabidopsis. *Plant Physiol.* 142:98–112.

He F, Arce AL, Schmitz G, Koornneef M, Novikova P, Beyer A, de Meaux J. 2016. The Footprint of Polygenic Adaptation on Stress-Responsive *Cis* -Regulatory Divergence in the *Arabidopsis Genus*. *Mol. Biol. Evol.* 33:2088–2101.

Huo H, Wei S, Bradford KJ. 2016. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proc. Natl. Acad. Sci.* 113:E2199–E2206.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc. Natl. Acad. Sci.* 112:15390–15395.

Josephs EB, Lee YW, Wood CW, Schoen DJ, Wright SI, Stinchcombe JR. 2020. The Evolutionary Forces Shaping Cis- and Trans-Regulation of Gene Expression within a Population of Outcrossing Plants. *Mol. Biol. Evol.* 37:2386–2393.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. *Ecol. Lett.* 7:1225–1241.

Keagy J, Drummond CP, Gilbert KJ, Grozinger CM, Hamilton J, Hines HM, Lasky J, Logan CA, Sawers R, Wagner T. 2023. Landscape transcriptomics as a tool for addressing global change effects across diverse species. *Mol. Ecol. Resour.* [Internet] n/a. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13796

Kidokoro S, Konoura I, Soma F, Suzuki T, Miyakawa T, Tanokura M, Shinozaki K, Yamaguchi-Shinozaki K. 2023. Clock-regulated coactivators selectively control gene expression in response to different temperature stress conditions in Arabidopsis. *Proc. Natl. Acad. Sci.* 120:e2216183120.

Kim Hyoun-Joung, Hyun Y, Park J-Y, Park M-J, Park M-K, Kim MD, Kim Hye-Joung, Lee MH, Moon J, Lee I, et al. 2004. A genetic link between cold responses and flowering time through FVE in Arabidopsis thaliana. *Nat. Genet.* 36:167–171.

Kim J-Y, Ryu JY, Baek K, Park C-M. 2016. High temperature attenuates the gravitropism of inflorescence stems by inducing SHOOT GRAVITROPISM 5 alternative splicing in Arabidopsis. *New Phytol.* 209:265–279.

Koo PK, Ploenzke M. 2021. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat. Mach. Intell.* 3:258–266.

Korfmann K, Gaggiotti OE, Fumagalli M. 2023. Deep Learning in Population Genetics. *Genome Biol. Evol.* 15:evad008.

Lang L, Pettkó-Szandtner A, Elbaşı HT, Takatsuka H, Nomoto Y, Zaki A, Dorokhov S, Jaeger GD, Eeckhout D, Ito M, et al. 2021. The DREAM complex represses growth in response to DNA damage in Arabidopsis. *Life Sci. Alliance* [Internet] 4. Available from: https://www.life-science-alliance.org/content/4/12/e202101141

Lasky JR, Des Marais DL, Lowry DB, Povolotskaya I, McKay JK, Richards JH, Keitt TH, Juenger TE. 2014. Natural Variation in Abiotic Stress Responsive Gene Expression and Local Adaptation to Climate in Arabidopsis thaliana. *Mol. Biol. Evol.* 31:2283–2296.

Lee G, Sanderson BJ, Ellis TJ, Dilkes BP, McKay JK, Ågren J, Oakley CG. 2024. A large-effect fitness trade-off across environments is explained by a single mutation affecting cold acclimation. *Proc. Natl. Acad. Sci.* 121:e2317461121.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* 13:233–245.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li W, Yin Y, Quan X, Zhang H. 2019. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front. Genet.* [Internet] 10. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.01077

Lovell JT, Schwartz S, Lowry DB, Shakirov EV, Bonnette JE, Weng X, Wang M, Johnson J, Sreedasyam A, Plott C, et al. 2016. Drought responsive gene expression regulatory divergence between upland and lowland ecotypes of a perennial C4 grass. *Genome Res.* 26:510–518.

Mack KL, Square TA, Zhao B, Miller CT, Fraser HB. 2023. Evolution of spatial and temporal cis-regulatory divergence in sticklebacks. *Mol. Biol. Evol.*:msad034.

Marand AP, Eveland AL, Kaufmann K, Springer NM. 2023. cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. *Annu. Rev. Plant Biol.* 74:111–137.

Mateos JL, Tilmes V, Madrigal P, Severing E, Richter R, Rijkenberg CWM, Krajewski P, Coupland G. 2017. Divergence of regulatory networks governed by the orthologous transcrip-tion factors FLC and PEP1 in Brassicaceae species. *Proc. Natl. Acad. Sci. U. S. A.* 114:11037–11046.

de Meaux J. 2006. An adaptive path through jungle DNA. *Nat. Genet.* 38:506–507.

Meaux J de. 2018. Cis-regulatory variation in plant genomes and the impact of natural selection. *Am. J. Bot.* 105:1788–1791.

Meng X, Liang Z, Dai X, Zhang Y, Mahboub S, Ngu DW, Roston RL, Schnable JC. 2021. Predicting transcriptional responses to cold stress across plant species. *Proc. Natl. Acad. Sci.* 118:e2026330118.

Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biol. Evol.* 9:843–854.

Moore BM, Lee YS, Wang P, Azodi C, Grotewold E, Shiu S-H. 2021. Modeling temporal and hormonal regulation of plant transcriptional response to wounding. *Plant Cell*:koab287.

Oakley CG, Ågren J, Atchison RA, Schemske DW. 2014. QTL mapping of freezing tolerance: links to fitness and adaptive trade-offs. *Mol. Ecol.* 23:4304–4315.

R Core Team. 2018. R: A language and environment for statistical computing. Available from: http://www.R-project.org/

Raicu A-M, Fay JC, Rohner N, Zeitlinger J, Arnosti DN. 2023. Off the deep end: What can deep learning do for the gene expression field? *J. Biol. Chem.* 299:102760.

Rivière Q, Corso M, Ciortan M, Noël G, Verbruggen N, Defrance M. 2022. Exploiting Genomic Features to Improve the Prediction of Transcription Factor Binding Sites in Plants. *Plant Cell Physiol.*:pcac095.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* 13:505–516.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32:D91–D94.

Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. *Nat. Rev. Genet.* 14:807–820.

Schmitz RJ, Grotewold E, Stam M. 2021. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. *Plant Cell*:koab281.

Smet D, Opdebeeck H, Vandepoele K. 2023. Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. *Front. Plant Sci.* [Internet] 14. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2023.1212073

Srivastava D, Aydin B, Mazzoni EO, Mahony S. 2021. An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding. *Genome Biol.* 22.

Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. 2017. Genomic analysis reveals major determinants of *cis*- regulatory variation in *Capsella grandiflora*. *Proc. Natl. Acad. Sci.* 114:1087–1092.

Takou M, Balick DJ, Steige KA, Dittberner H, Göbel U, Schielzeth H, Meaux J de. 2022. Strength of stabilizing selection on the amino-acid sequence is associated with the amount of non-additive variance in gene expression. :2022.02.11.480164. Available from: https://www.biorxiv.org/content/10.1101/2022.02.11.480164v2

Takou M, Wieters B, Kopriva S, Coupland G, Linstädter A, De Meaux J. 2019. Linking genes with ecological strategies in Arabidopsis thaliana. *J. Exp. Bot.* 70:1141–1151.

Thomashow MF. 2010. Molecular Basis of Plant Cold Acclimation: Insights Gained from Studying the CBF Cold Response Pathway. *Plant Physiol.* 154:571–577.

Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H. 2019. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci.* 116:5542–5549.

Wei Q, Jr RLD. 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLOS ONE* 8:e67863.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 430:85–88.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13:59–69.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The Evolution of Transcriptional Regulation in Eukaryotes. *Mol. Biol. Evol.* 20:1377–1419.

Wright MN, Ziegler A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* 77:1–17.

Yanagisawa S. 2016. Chapter 12 - Structure, Function, and Evolution of the Dof Transcription Factor Family. In: Gonzalez DH, editor. Plant Transcription Factors. Boston: Academic Press. p. 183–197. Available from: https://www.sciencedirect.com/science/article/pii/B9780128008546000129

Yang Z, Xu G, Zhang Q, Obata T, Yang J. 2022. Genome-wide mediation analysis: an empirical study to connect phenotype with genotype via intermediate transcriptomic data in maize. *Genetics*:iyac057.

Yocca AE, Lu Z, Schmitz RJ, Freeling M, Edger PP. 2021. Evolution of Conserved Noncoding Sequences in Arabidopsis thaliana. *Mol. Biol. Evol.* 38:2692–2703.

Zhou P, Enders TA, Myers ZA, Magnusson E, Crisp PA, Noshay J, Gomez-Cano F, Liang Z, Grotewold E, Greenham K, et al. 2021. Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *Plant Cell* [Internet]. Available from: https://doi.org/10.1093/plcell/koab267

Data Accessibility

All scripts are available at the github repositories:

https://github.com/mtakou/cnns_DEGpromoters and https://github.com/em-bellis/CNN_GxE.

Conflict of Interest Statement

ESB is currently a full time employee of Avalo, a crop improvement company.