Predicting gene expression by XGboosting and Shapley

Completed for the Certificate in Scientific Computation
Spring 2025

Kimin Wu Nguyen
BSA Applied Mathematics & BS Computational Biology
College of Natural Science
The University of Texas, Austin

Supervising Faculty's Signature

Supervising Faculty's Full Name: Nhat Ho
Title: Assistant Professor
Department: Statistics and Data Science

Abstracts:

As climate change poses a challenge to agricultural productivity and food security, understanding plant stress adaptation mechanisms becomes increasingly significant. While the current computational biology methods are effective, they are often complex in processes as well as require high computational resources, thereby being less accessible for general application. In this case, our research explores the high density of DNA-binding proteins and conserved motifs within the Arabidopsis genome to establish how variations in these regulatory factors influence transcriptional activity. This study investigates the prediction of gene expression in Arabidopsis through the application of XGBoosting and Shapley values, focusing on the role of transcription factors in mediating stress responses to environmental factors, such as drought tolerance . With the use of advanced computational techniques, such as a grid search for hyperparameter tuning in XGBoost, we analyze the contribution of various cis-regulatory elements to drought tolerance. Our research reveals essential motifs with a significant contribution to drought response prediction, as well as synergistic interactions among them. The outcome of this work can inform engineering of climate-tolerant crops, an urgent need for adaptation to increasing climate variability. Despite promising insights, we acknowledge limitations in sample size and experimental confirmation. This research aims to simplify the process of attaining stress tolerance in crops, which may cut down the computational burden and assist in fostering sustainable agriculture in the face of a changing climate.

Introduction:

Unlike animals, plants are immobile and therefore cannot relocate to other environments. Instead, they have evolved mechanisms to react to different environmental stresses, ranging from drought to high-humidity environments. These are largely mediated through changes in gene expression. Arabidopsis, and likely most plants, contain an enormously large number of DNA-binding proteins that may be functioning as transcription factors (TFs). More than 3,000 genes are thought to be engaged in transcription, over half of which are expected to code for TFs. This accounts for greater than 5% of the Arabidopsis genome, approximately double the rate encountered in yeast and animal genomes. These transcription factors bind to specific cis-acting regulatory elements (CAREs) within the DNA and regulate the initiation of transcription(Rombauts et al., 2003).

Cis-regulatory elements are DNA sequences that serve as binding sites for specific transcription factors (TFs).They play a vital role in regulating the gene expression precisely facilitating the recruitment of the basal transcription machinery(Hernadez-Garcia and Finer, 2014). Transcription factors control gene expression during transcription by recognizing and binding to specific DNA sequences (Mitsis et al., 2020). The interaction strength between transcription factors and their respective cis-regulatory elements is primarily influenced by the sequence features of these elements. In eukaryotes, gene expression regulation is intricately complex, often requiring the orchestrated activity of multiple transcription factors. This combinatorial regulation is a conserved mechanism across various organisms, providing distinct advantages such as the integration of diverse environmental signals and the ability to achieve regulatory flexibility using a limited repertoire of transcription factors.Therefore,

predicting the DNA-binding motifs of the transcription factors is an important component of the functional analyses of transcription factors.

Despite extensive research into the principles and mechanisms of TF binding, identifying functional motifs and quantifying their regulatory importance remains a significant challenge. One simple method to discover the motifs is by linear model. In 2003, Conlon introduced the motifs regressor model by combining matrix-based motif finding techniques (like MDSCAN) with regression analysis to identify motifs that are significantly correlated with gene expression changes. This method employs linear regression and stepwise regression to filter out insignificant motifs and to determine which combinations of motifs collectively contribute to gene expression variability. The algorithm are summarized as:

$$Y_g = \alpha + \sum_{m=1}^{M} \beta_m S_{mg} + \varepsilon_g$$

- Y_g is the expression level of gene g.
- α is the intercept term of the regression model.
- β_m represents the regression coefficients for each motif m.
- S_{mg} indicates the presence (or abundance) of motif m in the upstream sequence of gene g.
- M is the total number of motifs being considered in the model.
- ε_g is the error term for gene g.

While “MOTIF REGRESSOR” has successfully identified known motifs in simple eukaryote organisms like Saccharomyces cerevisiae (yeast), it failed to apply on more complicated organism. For instance, we built the linear regression to explain gene expression (response) in our arabidopsis thaliana's gene expression by using 35 motif-based predictors (M_ variables), where each M_ represents the count of a specific DNA motif identified in the regulatory region of the gene. These motifs were treated as independent variables to assess their potential influence on gene expression patterns. The overall model is statistically significant, with an F-statistic of 1.619 (p = 0.04778), indicating that, collectively, the motif counts have a moderate effect on predicting gene expression. The model explains approximately 47.4% of the variance in expression levels (R-squared = 0.4736), but the adjusted R-squared drops to 0.1811, suggesting that many included motifs

```
Call:
lm(formula = response ~ ., data = model)

Residuals:
    Min      1Q  Median      3Q     Max
-177.18  -57.04   -8.88   30.19  363.37

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    83.17398   34.98797   2.377 0.020493 *
M_1_CTTCTTCTTCTTCTT_counts      18.14789   17.76389   1.022 0.310869
M_10_CGGCGATGGMG_counts        31.74804   18.79205   1.689 0.096079 .
M_11_GGMGGAGGAG_counts          6.92436   17.19573   0.403 0.688547
M_12_AAAAAAAAAAAAAAA_counts     1.92176   26.46314   0.073 0.942339
M_13_CCACCACCACCACCA_counts    -2.11073   16.55184  -0.128 0.898933
M_14_GCAGCAGCAGC_counts       -16.10805   24.50140  -0.657 0.513297
M_15_AAAAGGAAAAAAAAA_counts    -8.68073   14.93845  -0.581 0.563246
M_16_ATATATATATATATAT_counts    7.56568   22.29967   0.339 0.735532
M_17_AGAGAAGAGAGAAR_counts    -11.72602   11.21542  -1.046 0.299775
M_18_WGAYGATGA_counts          23.62181   12.64805   1.868 0.066467 .
M_19_CGGCGAG_counts            10.17020   23.22500   0.438 0.662957
M_2_AAAACCCTAA_counts          54.49564   15.80693   3.448 0.001013 **
M_20_GATGAYGASG_counts        -21.86513   19.45170  -1.124 0.265247
M_21_GAAGCAGWAG_counts        -27.17483   28.46209  -0.955 0.343342
M_22_HTATATATAD_counts        -43.93909   32.22227  -1.364 0.177538
M_23_AAAATAAAATAAAAT_counts     4.24906   17.78893   0.239 0.811990
M_24_AACAAAAAAAAAA_counts       4.09764   18.00526   0.228 0.820709
M_25_GAAGAAGA_counts          -24.54536   16.83504  -1.458 0.149810
M_26_AGGCCCAWTAA_counts         9.56338   14.57570   0.656 0.514138
M_27_CTASTASTAG_counts        185.22107   48.93318   3.785 0.000345 ***
M_28_ATTAAAAAAAAAA_counts     -13.31054   14.51657  -0.917 0.362681
M_29_CAAAACCAAAACC_counts      22.80064   17.57502   1.297 0.199248
M_3_MCGSCGSCGK_counts         -29.58141   29.02669  -1.019 0.312048
M_30_AGTWGAAGWAGYWGH_counts    35.85806   23.46925   1.528 0.131549
M_32_AATTAATTAAATTT_counts      3.04244   16.16092   0.188 0.851279
M_33_AACAAAACAAAA_counts       -1.06229   14.68254  -0.072 0.942552
M_34_ATTATTATTATATT_counts    -12.06022   19.38370  -0.622 0.536067
M_35_ATAAATAAATAAAA_counts      7.91576   15.69212   0.504 0.615711
M_36_ACACACACACAAMCA_counts     0.08597   15.66002   0.005 0.995637
M_4_AGAGAGAGAGAGAGR_counts      3.95997   11.94065   0.332 0.741263
M_5_GAAGAAGAAGA_counts          2.00836   15.70989   0.128 0.898682
M_6_GAAGAAGAGAGAGR_counts     -12.01822   14.07989  -0.854 0.396575
M_7_CATCATCATCATC_counts      -19.36200   15.56302  -1.244 0.218071
```

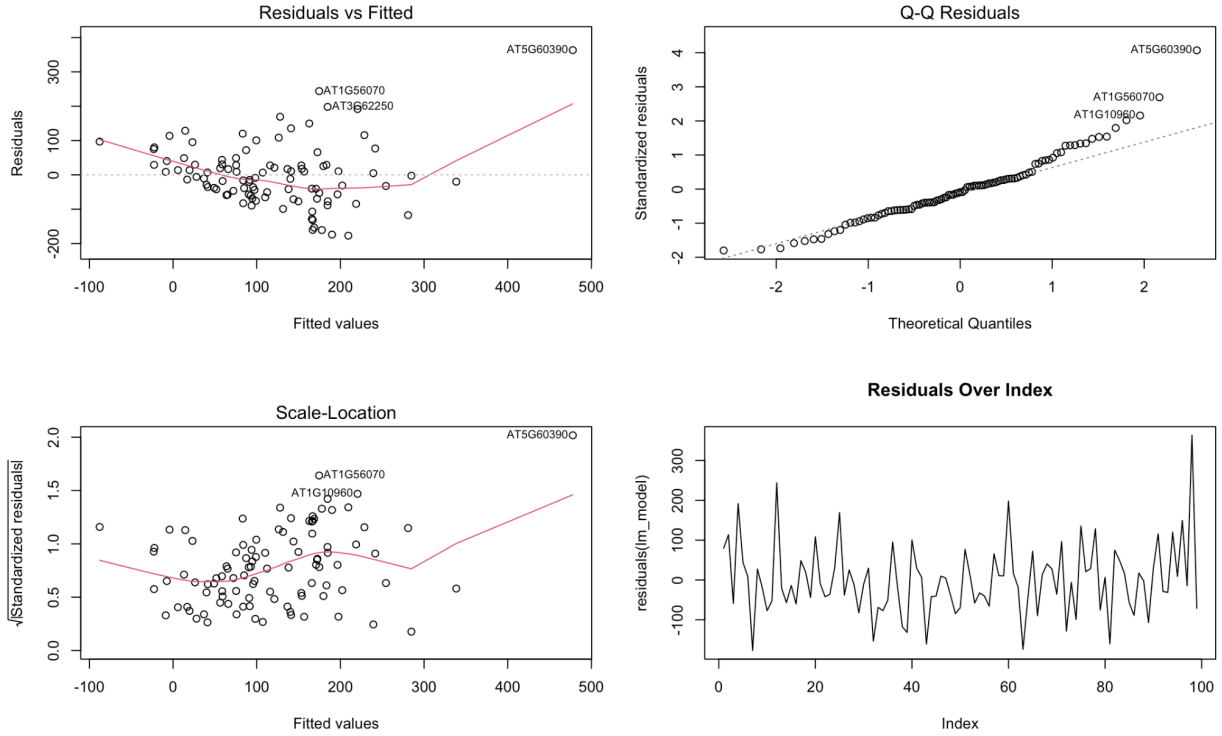
```
M_8_CAACAACAACAA_counts        7.41705   19.57365   0.379 0.706014
M_9_TCGCCGGAGAWDN_counts      -6.63657   17.71572  -0.375 0.709205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.4 on 63 degrees of freedom
Multiple R-squared:  0.4736,    Adjusted R-squared:  0.1811
F-statistic: 1.619 on 35 and 63 DF,  p-value: 0.04778
```

may not meaningfully contribute to the model and could be introducing noise or overfitting.

Among these motifs, M_2_AAAACCCTAA_counts and M_27_CTASTASTAG_counts stand out as statistically significant predictors ($p = 0.00101$ and $p = 0.00035$, respectively), each having a strong positive association with gene expression. This suggests that the presence of these motifs may play a critical regulatory role. Additionally, M_10_CGGCGATGGMG_counts and M_18_WGAYGATGA_counts show marginal significance ($p < 0.1$), indicating that they may also be biologically relevant, though further validation would be necessary. In contrast, the majority of the other motif predictors have high p-values ($p > 0.1$), suggesting limited or no individual contribution to the gene expression levels in this model.

However, we found that many of the conditions for linear regression were violated. Based on the diagnostic plots, several assumptions show moderate to severe violations that may compromise the model's validity. The Residuals vs Fitted plot reveals a curved trend, indicating potential non-linearity in the relationship between the predictors and the response variable. This suggests that the linear model may not adequately capture the underlying structure of the data. The Q-Q plot indicates that while most residuals are approximately normally distributed, deviations in the upper tail highlight the presence of mild outliers, which could be influential. The Scale-Location plot shows increasing residual spread at higher fitted values, suggesting heteroscedasticity, or non-constant variance. This can lead to inefficient estimates and underestimated standard errors. On the other hand, the Residuals over Index plot shows no clear trend, supporting the assumption of independent residuals.

Although the model exhibits some predictive value, the violations of these assumptions suggest that a linear model may be inappropriate for capturing non-linear patterns. Using linear regression in such cases risks misrepresenting the true relationships in the data. Furthermore, multicollinearity, which is often unavoidable in biological systems due to interconnected molecular pathways, does not necessarily reduce predictive power but can severely distort interpretability. High multicollinearity inflates standard errors, increasing p-values and reducing t-statistics, which in turn masks the significance and individual contribution of predictors. This can lead to misleading conclusions regarding the role of specific features in predicting the response variable (Vijay, 2018).



Here we defined a new machine learning methods to predict the cis-elements in arabidopsis thaliana by using XGboosting and Shapley to understand the contribution of each cis-regulator for prediction

Data Reprocessing:

After performing DSeq2, we visualized the differential expression analysis of 7,976 genes where each point is for one gene, after filtering out all of the non statistical genes by MA plot (fig.1). The x-axis represents the mean normalized count that is the average expression level of each gene, while the y-axis is the log2 fold change that represents relative expression difference between two groups. In this plot, gray dots are genes that are not strongly differentially expressed, and blue dots are genes with statistically significant changes (adjusted p-value < 0.05). Triangles at the top and bottom edges are genes with extreme log2 fold changes that are outside the range of the plot. Most genes cluster around a log2 fold change of zero, meaning that their expression levels are not changing between conditions. However, one large subgroup of genes, particularly those with reduced mean expression, have significant upregulation or downregulation.

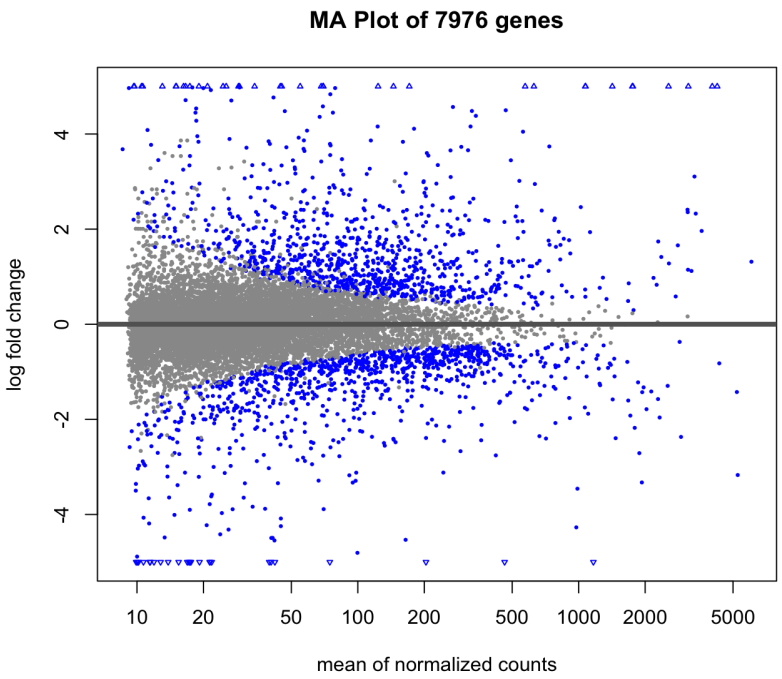


Figure 1: MA Plot of Differential Gene Expression. The plot shows the relationship between the mean of normalized counts (x-axis) and the log2 fold change in gene expression (y-axis) for 7,976 genes. Gray dots represent non-significant genes, while blue dots indicate genes with statistically significant differential expression (adjusted p-value < 0.05). Triangle-shape at the top and bottom edges mark genes with log2 fold changes beyond the plotting range.

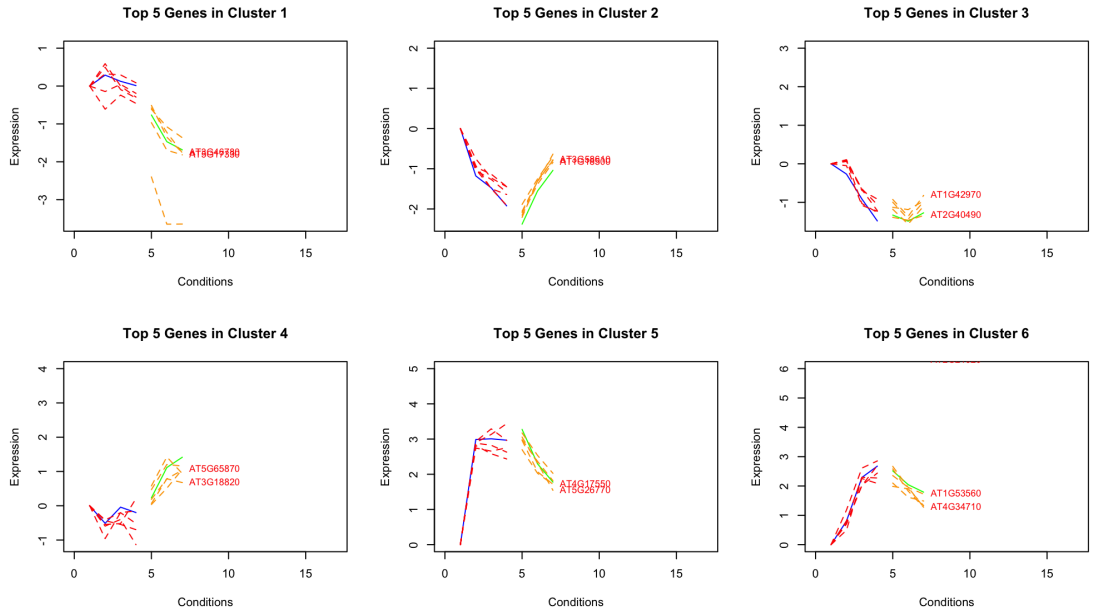
After calculating gene expression changes (log_2 fold change), we applied K-means clustering to group genes that shared similar expression patterns across drought and recovery conditions. We used the elbow method to determine the optimal number of clusters, which indicated that 6 clusters best captured the variation in the data. For each of these 6 clusters, we selected the top 5 representative genes—typically based on either expression level or centrality to the cluster—and plotted their expression trajectories over time or conditions (fig 3a).

Prior to applying the XGBoost model for evaluation of drought tolerance, we did t-distributed Stochastic Neighbor Embedding (t-SNE) first to visualize the overall structure of our data based on gene expression. As shown in the left-hand panel of Figure 3b, samples were color-coded according to six clusters of gene expression defined by the top five genes for each cluster. Visualizing the projection using t-SNE indicated considerable overlap among these clusters, which implies the gene expression profiles between the clusters were not distinguishable enough in the low dimension. Noting this overlap raised concerns for us to evaluate classification via downstream analysis via the original definition of the clusters, particularly separating drought tolerant from drought intolerant gene expression profiles.

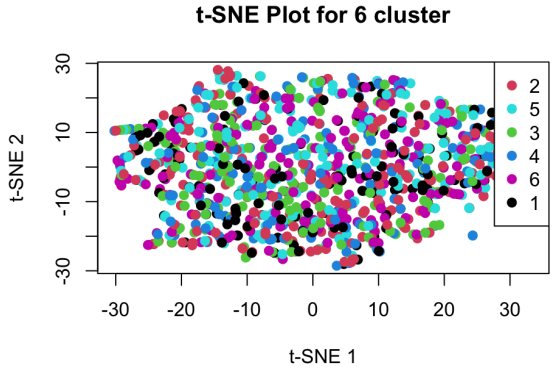
To improve the prediction, we re-evaluated the expression trajectories and observed that Clusters 5 and 6 exhibited highly similar patterns under drought and recovery conditions(figure 2a). Therefore, we decided to merge these two clusters into a

single group, effectively reducing redundancy and simplifying the class structure. When we re-ran the t-SNE analysis after this adjustment, we observed a clearer separation between drought-tolerant (green) and drought-intolerant (red) genes in the right panel of Figure 2c. This separation suggests that combining expression clusters with similar temporal profiles can enhance the biological relevance of the clustering and improve the feature structure for predictive modeling. This restructured representation was then used as input for the XGBoost model to classify drought tolerance status

A)



B)



C)

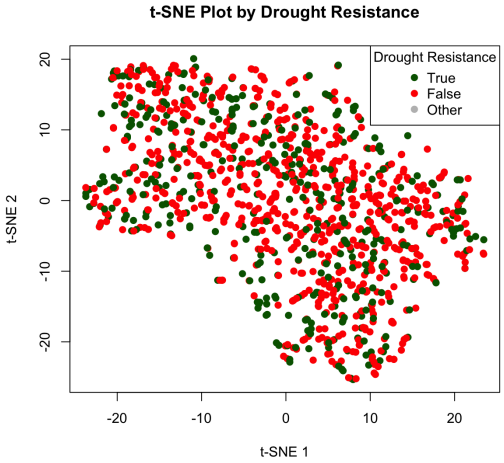


Figure 2: Gene expression pattern & t-SNE Visualization of Samples Based on Gene Expression Clusters and Drought Resistance.

B) A t-SNE plot of samples derived from the expression patterns of the top 5 genes in each of the six gene expression clusters. Each color represents a distinct cluster, reflecting transcriptional similarity across conditions. C) The panel overlays drought resistance phenotypes (True = green, False = red, Other = gray) on the same t-SNE coordinates.

Considering Natural Evolutions

Next, we mapped our gene list to various accession assemblies. Genes can exhibit different promoters across accessions due to the specificity of gene regulation, which ensures that genes are activated at the appropriate times and levels. Variations in promoter sequences facilitate the recruitment of diverse combinations of transcription factors, leading to differential gene expression in response to environmental conditions or specific cellular signals. This variable promoter architecture is important for multicellular organisms to manage intricate biological processes such as cell differentiation and stress responses. As illustrated by Ushijima (2017), the use of alternative promoters is common in response to environmental stimuli, serving as a mechanism to enhance transcriptional regulation amidst varying developmental and environmental contexts.

Instead of identifying motifs from these genes using MEME (Multiple EM for Motif Elicitation), which uncovers motifs in unaligned biological sequences through a probabilistic method grounded in the Expectation-Maximization (EM) algorithm. Meme establishes a position-specific probability matrix (PSPM) by randomly choosing subsequences (w-mers) from the input to serve as initial motif seeds. We chose to consider the role of natural selection in relation to these genes. Indeed natural selection is important for motif discovery and helps distinguish biologically meaningful motifs from those arising by random chance. We chose motifs that were obtained from the Catalog of Inferred Sequence Binding Preferences (CisBP v2.00) database (Weirauch et al., 2014), specifically for Arabidopsis thaliana. For each transcription factor (TF), we selected one motif following CisBP precedence rules prioritizing direct motifs, then inferred motifs with highest DBD similarity, and further resolved ties using CisBP's “Motif_Type” attribute.

These conserved motifs likely enhance an organism's fitness, allowing them to persist over evolutionary time. Ignoring the influence of selection could lead to the identification of misleading patterns that, while statistically significant, lack biological relevance. By incorporating measures of natural selection, like conservation scores or phylogenetic footprinting, we can improve the biological relevance and predictive power of the computational models used for motif detection (Elnitski et al., 2006). These important motifs can be considered as Cis-elemt which are short DNA sequences that are typically found in the regulatory regions of genes

Determining drought regions:

Next, we aimed to assess the drought severity associated with each accession. Given that the 1001 Genomes Project provides the geographical coordinates for each accession, determining the drought conditions can be challenging due to variability from year to year. To address this, we used the 168-year average (from 1850 to 2018) of the Palmer Drought Severity Index (PDSI), obtained from the UCAR Climate Data Guide. The PDSI is a widely recognized metric that incorporates precipitation, temperature, and soil moisture to evaluate long-term drought severity (Dai, 2011).

Ideally, each accession would have a perfectly matched PDSI value based on its exact latitude and longitude. However, due to limitations in the spatial resolution of climate data and to ensure more robust environmental representation, we introduced a buffer—matching accessions to the nearest available PDSI grid point within a 50-mile radius. This approach offers flexibility while still preserving the environmental relevance of each accession. To reduce the influence of extreme values, we applied the Interquartile Range (IQR) method to remove outliers, retaining only the central 75% of the data. This method is a standard statistical practice to ensure robust analysis by mitigating the impact of anomalies (Penn State, n.d.).

We then introduced a standardized metric, Drought_norm, based on the assumption that the underlying distribution of PDSI values approximates a normal distribution—an assumption supported by our data. Normalization in this way is common in climatological studies, as it enables clearer identification of drought anomalies and classification into meaningful categories (Dai, 2011). We defined thresholds at ±1 standard deviation: accessions with values below -1 (left tail) were classified as originating from “Drought regions,” while those above +1 (right tail) were categorized as from “Wet regions.”

The geographic distribution of categorized drought severity across Arabidopsis accessions is visualized in Figure 3. Each point represents an accession colored by its assigned drought category, including “Highly Wet,” “Moderate Wet,” “Somewhat Wet,” “Normal,” “Somewhat Dried,” and “Moderate Dried.” This mapping highlights the regional variation in environmental conditions among accessions.

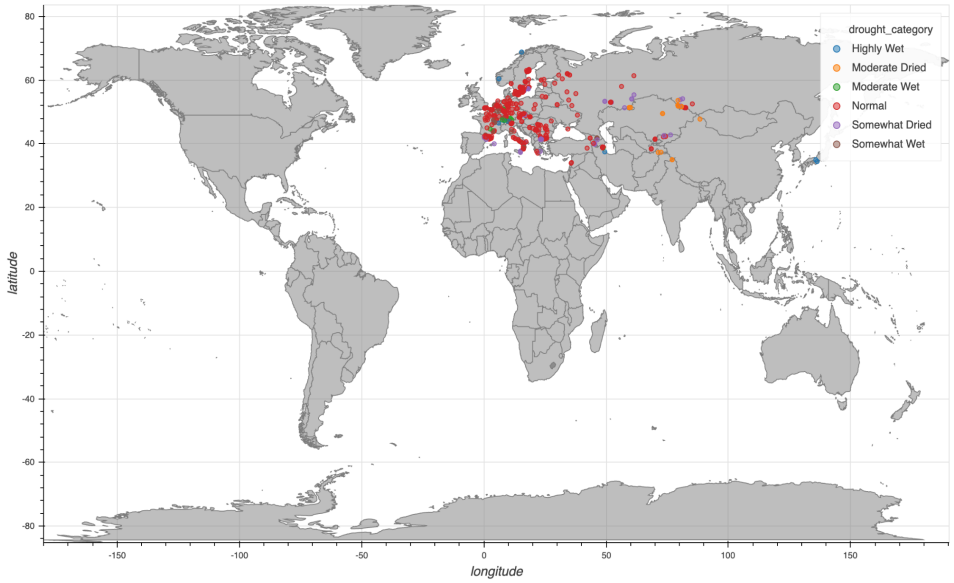


Figure 3. Geographic distribution of Arabidopsis accessions overlaid with categorized drought severity. Each point represents an accession colored by its assigned drought category based on the normalized Palmer Drought Severity Index (Drought_norm). Categories include "Highly Wet", "Moderate Wet", "Somewhat Wet", "Normal", "Somewhat Dried", and "Moderate Dried". Accessions were matched to PDSI grid points within a 50-mile radius of their reported coordinates (Interactive)

Result

XGboost hypertuning

Before starting the classification task, we tuned our XGBoost model through a comprehensive grid search with cross-validation. This process explored combinations of key hyperparameters, including learning rate (eta), maximum tree depth (max_depth), and subsampling ratio (subsample), while keeping colsample_bytree fixed at 0.8. Model performance was evaluated using cross-validated ROC AUC across increasing boosting iterations (figure 4). Lines in the figure represent different tree depths (max_depth ranging from 1 to 5), allowing us to observe the impact of tree complexity on performance. Models trained with a lower learning rate (eta = 0.3) and moderate subsample ratios (0.625–0.875) consistently achieved strong and stable AUC values. In particular, models with max_depth = 2 and 3 provided the best balance between model complexity and generalization, making them suitable candidates for downstream classification of drought-tolerant and drought-intolerant accessions.

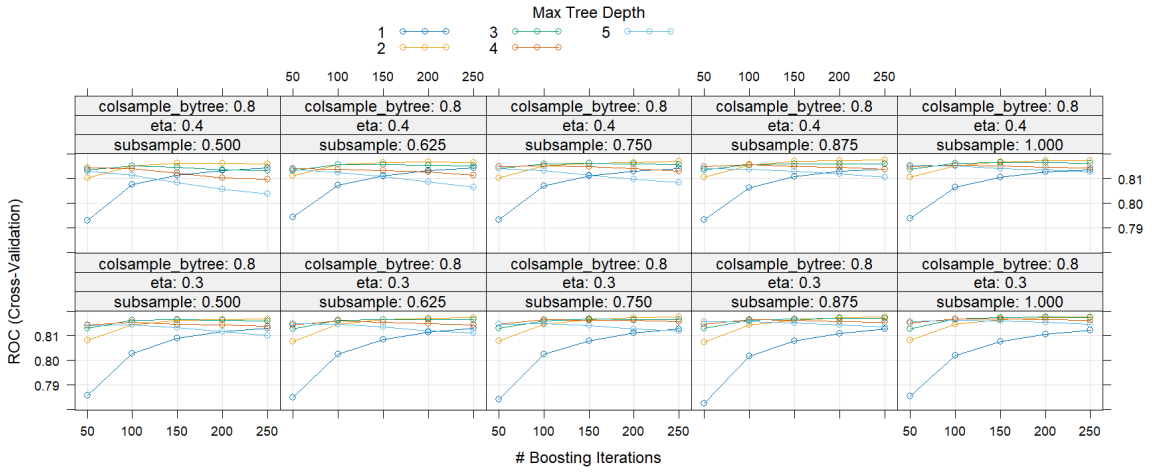


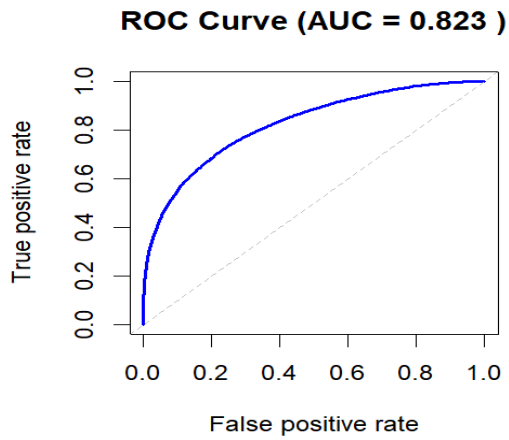
Figure 4: Cross-validation ROC performance across XGBoost hyperparameter combinations. Grid search was conducted across different combinations of learning rate (eta), subsample ratio, and maximum tree depth to optimize model performance by combination of eta and subsample, with a fixed colsample_bytree = 0.8.

Model Performance Evaluation

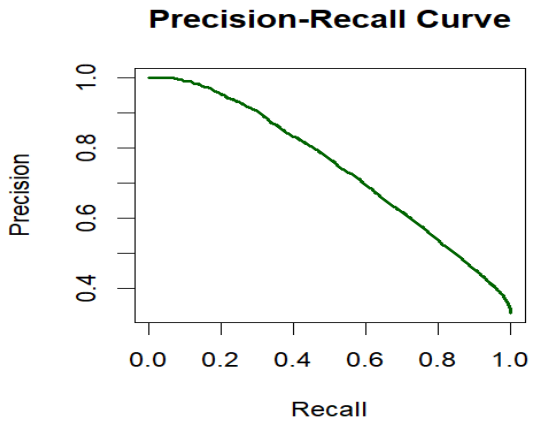
After hypertuning the XGBoost model, we evaluated its performance to classify drought-tolerant and drought-intolerant accessions using the test dataset. The final model had an overall classification accuracy of 78%, indicating stable classification performance. To further test the model, we examined the ROC curve(fig 5a), which showed superb discriminatory power between the two classes, with an AUC of 0.823(fig 5b), indicating high sensitivity and specificity at all thresholds. The Precision-Recall curve revealed the strength of the model in the presence of class imbalance with high precision even at low recall values. We also compared class-wise performance using precision and F1-score measures(fig 5c). Class 0 (non-tolerant accessions) revealed high precision of 0.852 and good F1-score of 0.794, reflecting the confidence and accuracy of the model in predicting non-tolerant samples. In contrast, Class 1 (drought-tolerant accessions) yielded a moderate precision of 0.587 and F1-score of

0.654, indicating the model's ability to identify tolerant accessions successfully, though with some room for improvement in terms of not picking up false positives. This result is expected, because our dataset is slightly imbalance (~33% Drought-Tolerant class)

A)



B)



C)

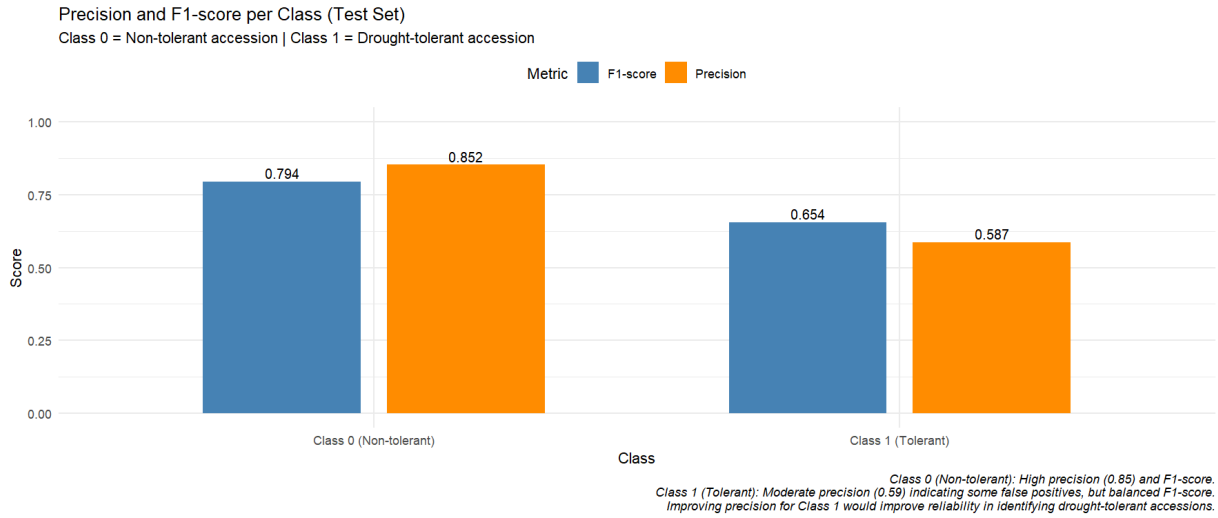
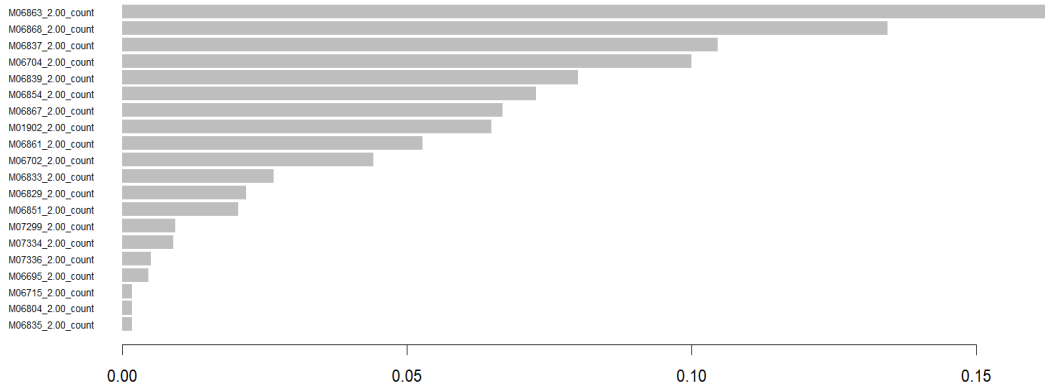


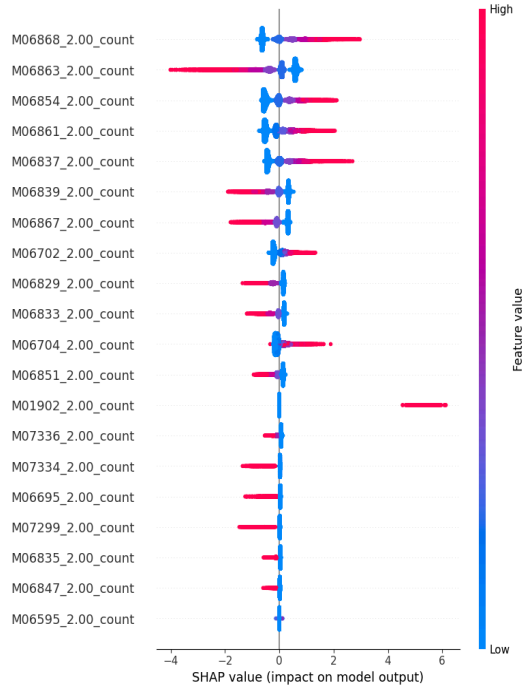
Figure 5: Model performance evaluation and class-wise precision/F1 analysis. *(A) The ROC curve shows strong discrimination between Drought- tolerance class ande Drought-intolerance class, with an AUC of 0.823, indicating high sensitivity and specificity at all thresholds.(B) The Precision-Recall curve shows the ability of the model to maintain high precision for low recall values, which shows good performance under class imbalance.(C) Bar plot of precision and F1-score by class on the test set. Class 0 (non-tolerant accessions) has high accurate on predictions (0.852) and strong F1-score (0.794), while Class 1 (drought-tolerant accessions) has moderate precision (0.587) and F1-score (0.654).*

Important Cis-Elements

A)



B)



C)

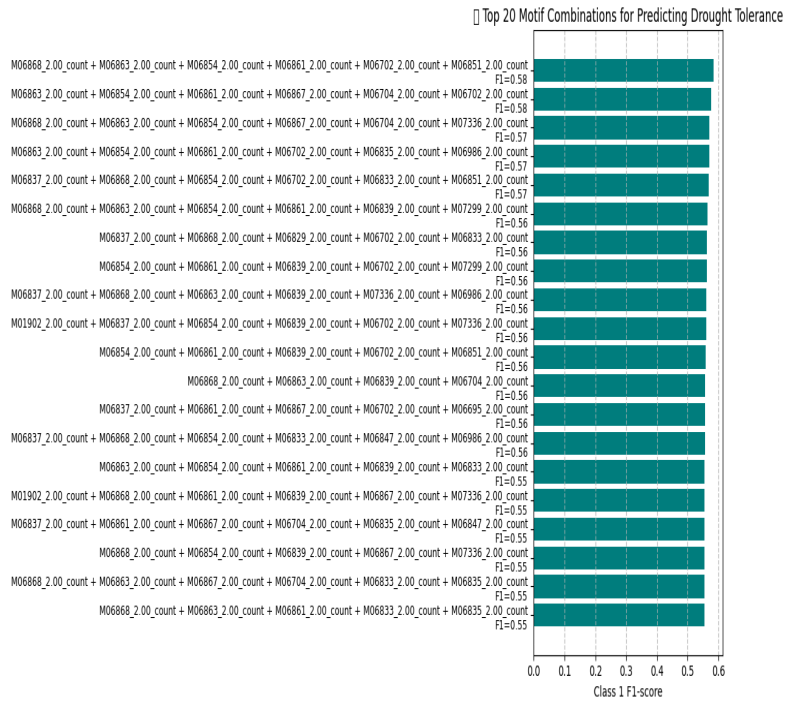


Figure 6: Top Cis Elements and Top Cis Elements Combinations for Predicting Drought Tolerance

a) Feature importance values represent the contribution of each cis elements to the overall model performance, with higher values indicating greater influence. Cis elements at the top, particularly M06863_2.00_count and M06868_2.00_count) SHAP summary plot highlighting the top 20 motif features and their contributions to the model's prediction of drought tolerance. Each point represents a sample, colored by the motif count (blue = low, red = high), and positioned by its SHAP value (influence on model output). Motifs with consistently high SHAP values are strong predictors of class 1

(drought tolerant). c) top 20 motif (cis-element) combinations associated with accurate prediction of the drought tolerance group, ranked by F1-score. These combinations reflect synergistic patterns of motif presence that help differentiate drought-tolerant genes (class 1) from non-tolerant ones, underscoring the regulatory power of motif interactions in gene expression under stress.

Conclusion

Our integrative analysis combining SHAP interpretation and motif interaction modeling reveals that both individual motifs and their combinations are critical for predicting drought tolerance. SHAP values identified several high-impact cis elements such as M06863_2.00_count, M06868_2.00_count, and M06837_2.00_count—as key contributors to model predictions, indicating their strong association with drought response regulation.

Furthermore, the top performing cis elements combinations, consistently included these high-impact motifs and achieved F1-scores up to 0.58. Notably, the combination of M06868_2.00_count + M06863_2.00_count + M06854_2.00_count + M06861_2.00_count + M06702_2.00_count + M06851_2.00_count was among the highest-ranked, suggesting that co-occurrence and interaction of multiple motifs enhances the model's ability to discriminate drought-tolerant genes from others. These results highlight the regulatory importance of motif synergy, providing new insights into the combinatorial logic of cis-regulatory elements under drought stress. This knowledge can inform downstream experimental validation and may serve as a basis for engineering stress-resilient crops through targeted manipulation of promoter architectures.

However, some limitations should be acknowledged. The sample size is relatively small, which may limit the generalizability of the model. Additionally, while computational methods provide strong insights into regulatory patterns, these predictions currently lack experimental validation. Future work should include wet-lab experiments to confirm the functional roles of the identified motifs and their combinations, and expand the dataset to improve model robustness and biological relevance.

Appendix: Computer Code and Data Availability

Due to the extensive length of the computer code developed for this research, the full source code is included as an appendix in this document. To ensure accessibility and facilitate reproducibility, the code has also been made available in a public repository on GitHub (or similar platform):

- **Accession_XGboosting.py**: Python script for drought tolerance prediction using the XGBoost algorithm.
- **DSeq2.R**: R script for differential expression analysis on TPM-filtered genes.
- **Motifs matching by Fimo and MCAST.R**: R script for motif detection using FIMO and MCAST tools.
- **Visualization for clustering.R**: R script for generating visualizations of clustering results.
- **motif occur.R**: R script for analyzing motif occurrences across accessions.
- geographical & rainfall.R: R script integrating geographical location and rainfall data for accessions.
- **TPM10.txt**: Original gene expression dataset containing TPM (Transcripts Per Million) values, filtered at a threshold >10.
- **gene_motifs dataset.csv**: Merged dataset linking genes with essential motifs.
- **17_accession_drought_category and.csv**: Classification of 17 accessions by drought response categories.
- **xgboost_predictions_drought_tolerance.csv**: Output of XGBoost model predictions for drought tolerance.
- **Accession Distribution.html**: HTML visualization displaying drought response distributions among accessions.
- **README.md**: Detailed documentation and instructions for reproducing the full analysis pipeline.

  All materials are available at: [GitHub Repository Link]

Acknowledgements

I would like to express my deepest gratitude to Dr. Nhat Ho, my project supervisor, whose expertise in statistics, machine learning, and optimization was invaluable throughout this project. His consistent guidance, especially in model selection and hyperparameter tuning, played a critical role in achieving the best possible results.

This project would not have been possible without the support of Dr. Hong Qiao, an associate professor in molecular biosciences. Her deep knowledge of biology—particularly in the context of Arabidopsis thaliana—was essential in bridging the gap between computational methods and biological interpretation.

I am incredibly thankful for their mentorship and encouragement throughout this interdisciplinary work.

Reflection

Throughout this project, I explored the intersection of computational biology and machine learning by building a predictive model for drought tolerance in Arabidopsis thaliana. I developed an XGBoost-based classifier that used motif occurrences as features and integrated environmental drought indices. I also applied SHAP analysis to identify the most influential cis-regulatory elements and uncovered synergistic motif combinations that strongly contributed to drought response predictions. From initial gene expression clustering with k-means, to dimensionality reduction via t-SNE, and through iterative model refinement—including hyperparameter tuning and class restructuring—I was able to produce a model with both biological insight and predictive power.

This hands-on experience revealed key challenges in applying machine learning to biological systems. The data was highly dimensional and the features—cis-elements—were often highly correlated, making it difficult to isolate individual contributions without introducing noise. I learned that without experimentally validated data, even powerful models like XGBoost can struggle to find consistent patterns. Additionally, interpreting the outputs meaningfully required deep biological knowledge, underscoring the importance of domain expertise when bridging data science and life sciences.

Another insight was the impact of class structure on model performance. Initially, the use of six tightly clustered expression classes led to low predictive performance, with models plateauing at about 16% accuracy. After consolidating clusters that shared similar expression dynamics, performance improved significantly—highlighting the importance of biologically informed data simplification. It improved model accuracy and reinforced the idea that statistical sophistication alone is insufficient without thoughtful, context-aware design.

Overall, this project strengthened my skills in scientific computing and machine learning, deepening my appreciation for the biological complexity behind gene regulation. It taught me the value of iteration, interdisciplinary thinking, and the careful in

Works References

Conlon, E. M., Liu, X. S., Lieb, J. D., & Liu, J. S. (2003). Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis. Proceedings of the National Academy of Sciences of the United States of America, 100(6), 3339–3344. http://www.jstor.org/stable/3139361

Dai, Aiguo & National Center for Atmospheric Research Staff (Eds). Last modified 2023-08-19 "The Climate Data Guide: Palmer Drought Severity Index (PDSI).” Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi on 2025-04-14.

Hernandez-Garcia, C. M., & Finer, J. J. (2014). Identification and validation of promoters and cis-acting regulatory elements. Plant science : an international journal of experimental plant biology, 217-218, 109–119. https://doi.org/10.1016/j.plantsci.2013.12.007

National Center for Atmospheric Research (NCAR). (n.d.). Palmer Drought Severity Index (PDSI). UCAR Climate Data Guide. Retrieved from https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi

Penn State. (n.d.). Lesson 3.2: Outliers. Penn State University, STAT 200: Elementary Statistics. Retrieved from https://online.stat.psu.edu/stat200/lesson/3/3.2

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P., & van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant physiology, 132(3), 1162–1176. https://doi.org/10.1104/pp.102.017715

Vijay, A. K. (2021, March 18). Linear regression assumptions, violation of assumptions, and rectification. Medium. https://medium.com/@akshivijaykk/linear-regression-assumptions-violation-of-assumtions-rectification-81e3a144bc74

Ushijima, T., Hanada, K., Gotoh, E., Yamori, W., Kodama, Y., Tanaka, H., Kusano, M., Fukushima, A., Nakabayashi, R., Nishizawa, T., Ohnishi, M., Okamoto, M., Yamamoto, T., Iba, K., Yamaguchi, J., & Yamamoto, Y. Y. (2017). Light controls protein localization through phytochrome-mediated alternative promoter selection. Cell, 171(6), 1316–1325.e12. https://doi.org/10.1016/j.cell.2017.10.018

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., ... & Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158(6), 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009