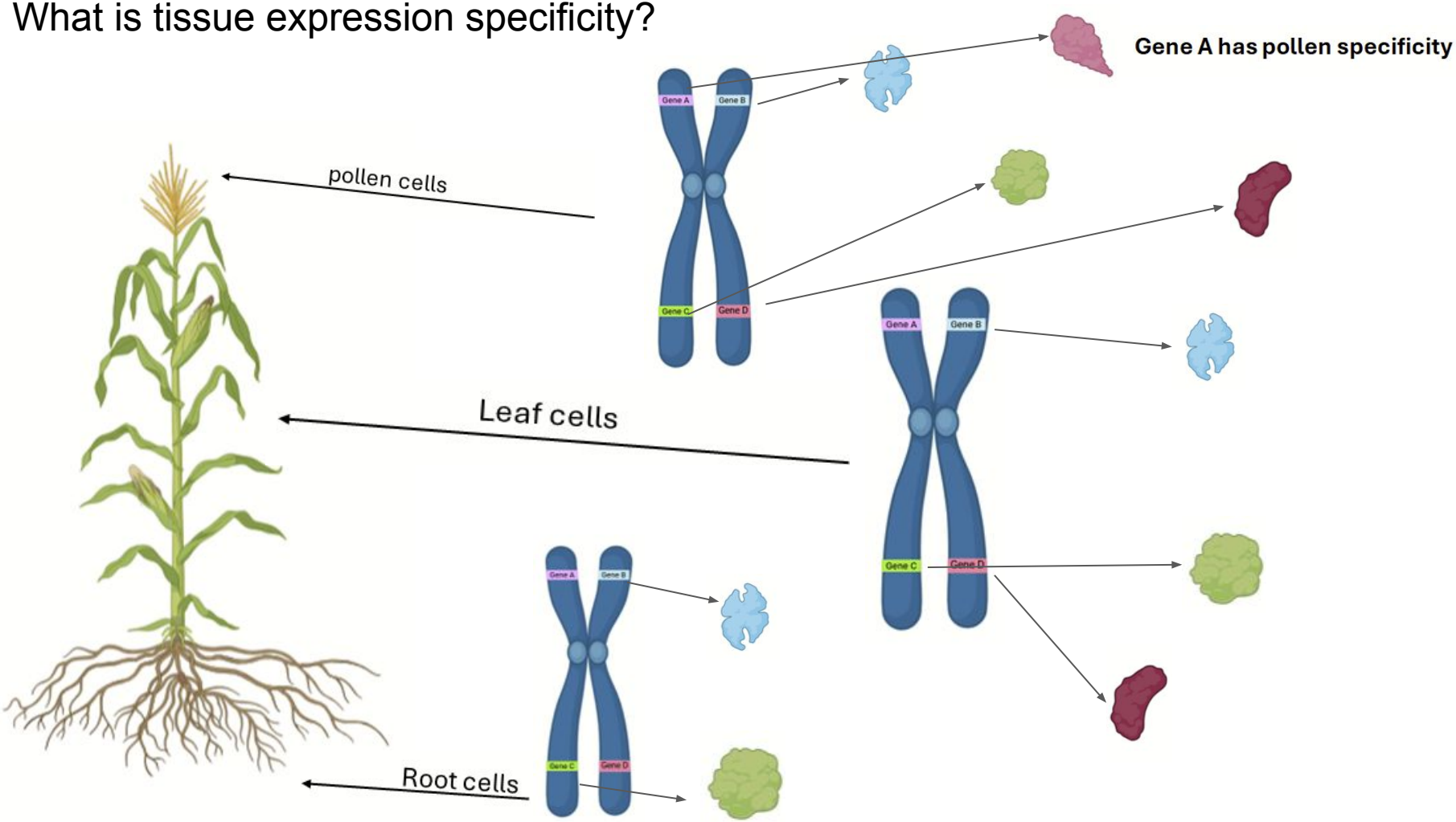
Predicting Maize Pollen Expression Specificity

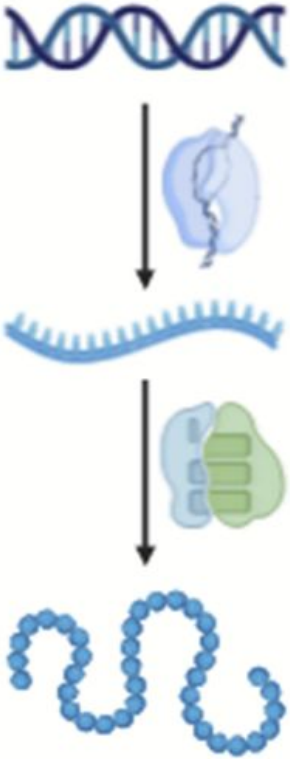
Group 3

What is tissue expression specificity?



Motivation: pollen specificity is important for predicting pollen fitness

Central Dogma of Molecular Biology



DNA → "Low fidelity" – cheaper, easy to get, more available

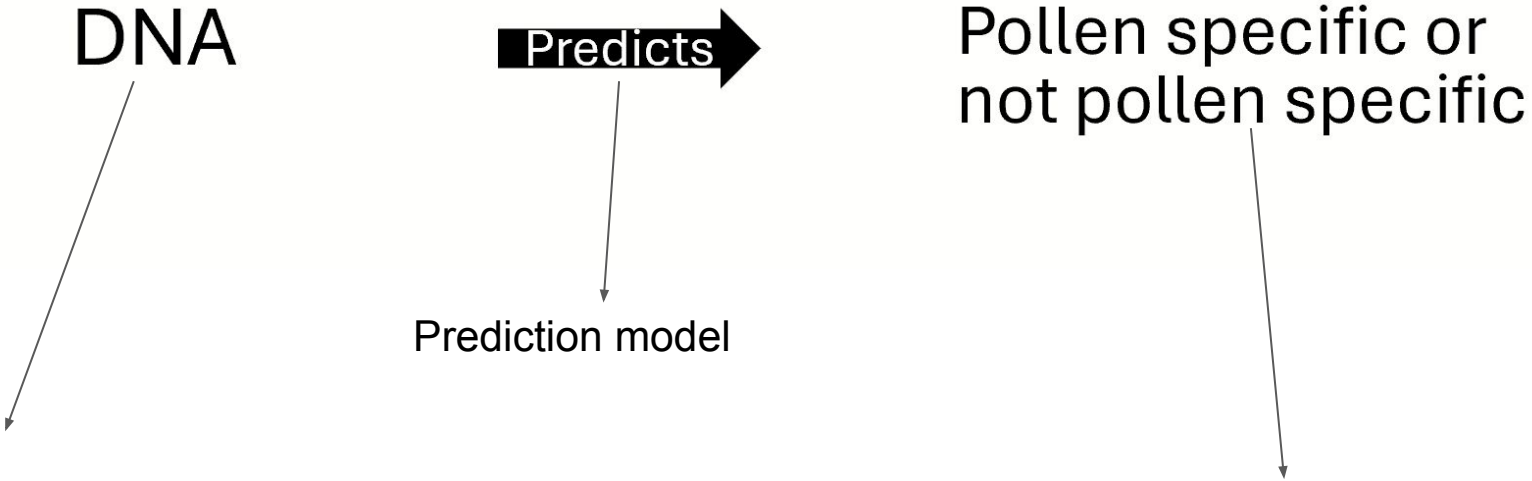
Can we make a separate model to use DNA information to predict pollen specificity?

RNA → Approximates → Pollen specificity

Pollen specificity → Predicts → "Good" pollen / Or / "Bad" pollen

Protein → "High fidelity" – expensive, harder to get, less available

Our proposed modeling



Information about a gene (e.g. kmer, amino acid composition, subcellular localization etc.)

Binary true or false. True if a gene is pollen specific protein expression false otherwise

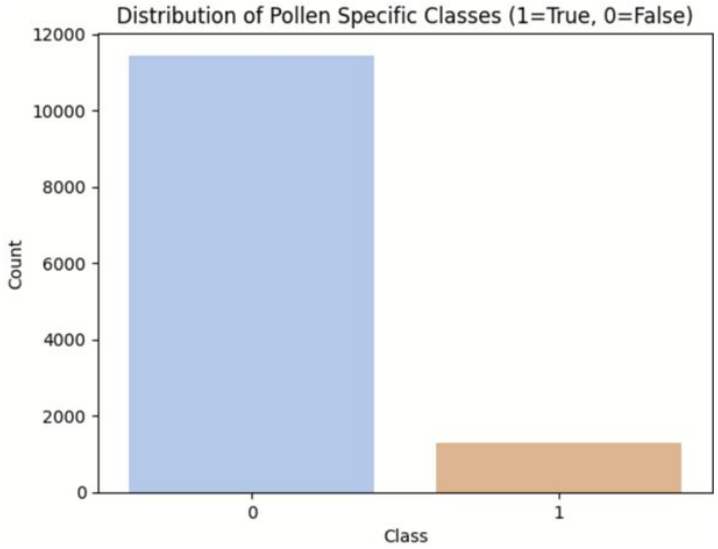
Our Data

Input:

12725 data points (rows)
2626 dimension (columns)

Response Variable:

11,443 Non-Pollen Specific
1282 Pollen Specific

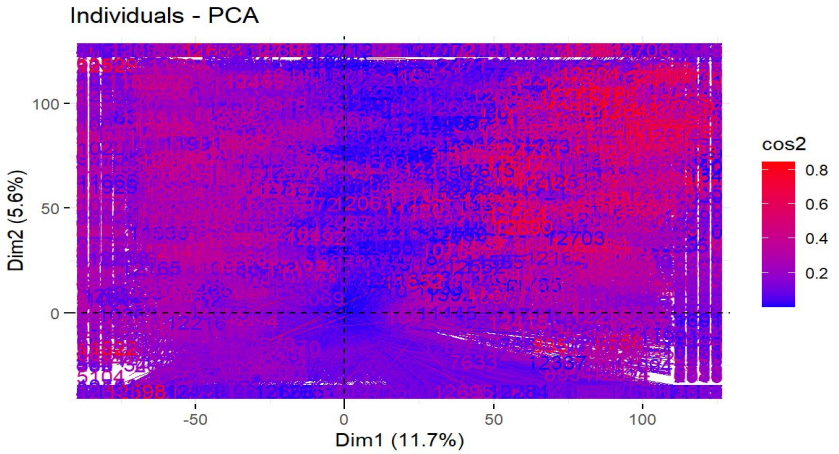


Gene ID	Kmer ACT	CTDC G1	…	PseDNC Xc1 AC	Pollen Specific
Gene 1	0.0434	0.2234	…	0.0672	0
Gene 2	0.0783	0.6673	…	0.0932	0
…	…	…	…	…	…
Gene 12725	0.0033	0.0021		0.4402	1

PCA Analysis: Dimensionality Reduction in Gene Features

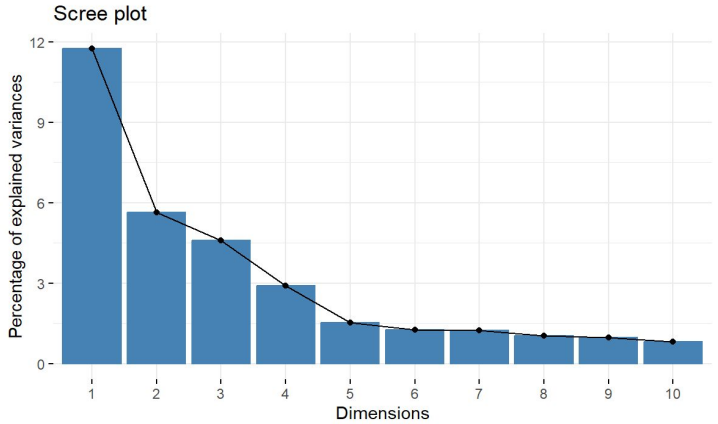
PCA Factor Map Shows No Clear Clusters:

- The **PCA plot does not show distinct separation** between pollen-specific and non-pollen-specific genes.
- This suggests that **PCA alone may not be enough for classification**, and **additional features or nonlinear techniques may be needed**.



Variance Explained is Low (~11.7% for PC1):

- The first two PC's only explains **~17% of the variance**, which is **not enough for strong classification**.
- This indicates that **other methods like feature selection (LASSO) or nonlinear models (Random Forest, XGBoost) may work better**.



Understanding Gene Clusters Using t-SNE

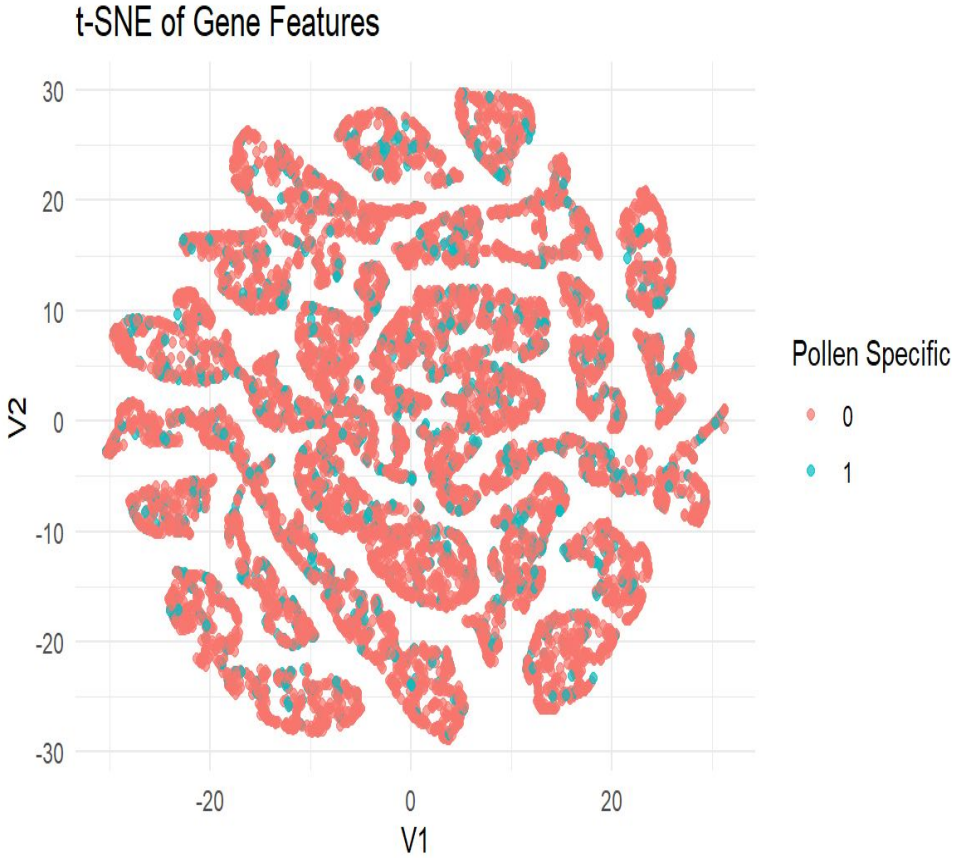
- **What This Plot Shows:**
  Each dot represents a **gene**, colored by **pollen-specific (cyan) or non-pollen-specific (red)** status.
- **Key Observations:**
  Genes form clusters, suggesting **underlying feature patterns**.
- **Overlap exists between pollen-specific and non-pollen-specific genes**, meaning classification is **not straightforward**.

Why t-SNE?

- Captures **nonlinear structures** in genomic data.
- Helps **visualize complex relationships** that PCA might miss.

Implications for Machine Learning:

- **No clear separation** → More **advanced models (Lasso, Random Forest, XGBoost) needed**.
- Feature engineering may help improve classification.



Logistic Regression + L1 Regularization

Logistic Regression: Models probability of a class $y \in \{0, 1\}$ as: $P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$

Where $\sigma(z) = \frac{1}{1+e^{-z}}$

Loss function (negative Log-Likelihood):

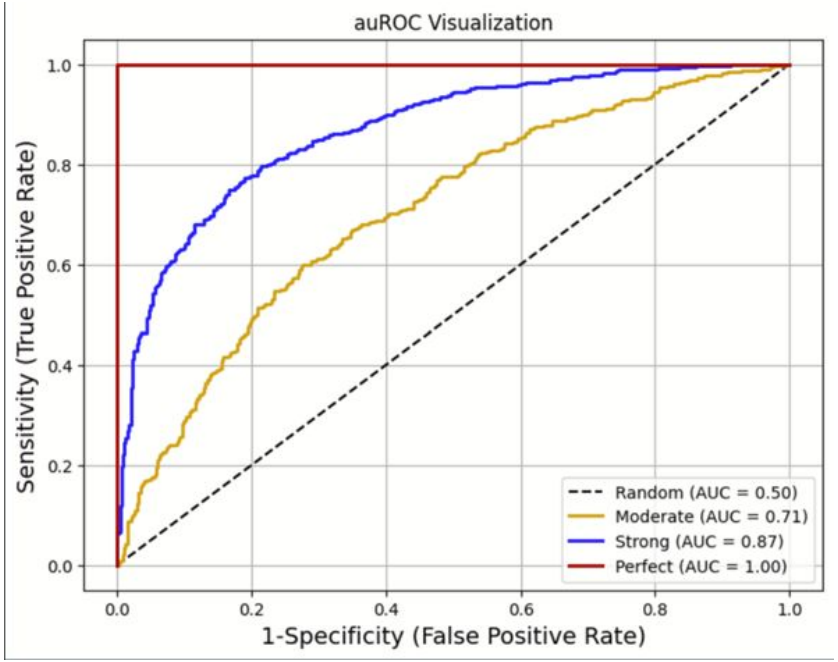
$$\mathcal{L}(\mathbf{w}, b) = -\sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

Add L1 regularization to achieve sparsity:

$$\mathcal{L}_{\text{reg}}(\mathbf{w}, b) = \mathcal{L}(\mathbf{w}, b) + \lambda \sum_{j=1}^{d} |w_j|$$

What is auROC?

auROC = a metric for indicating how well a model separate two classes

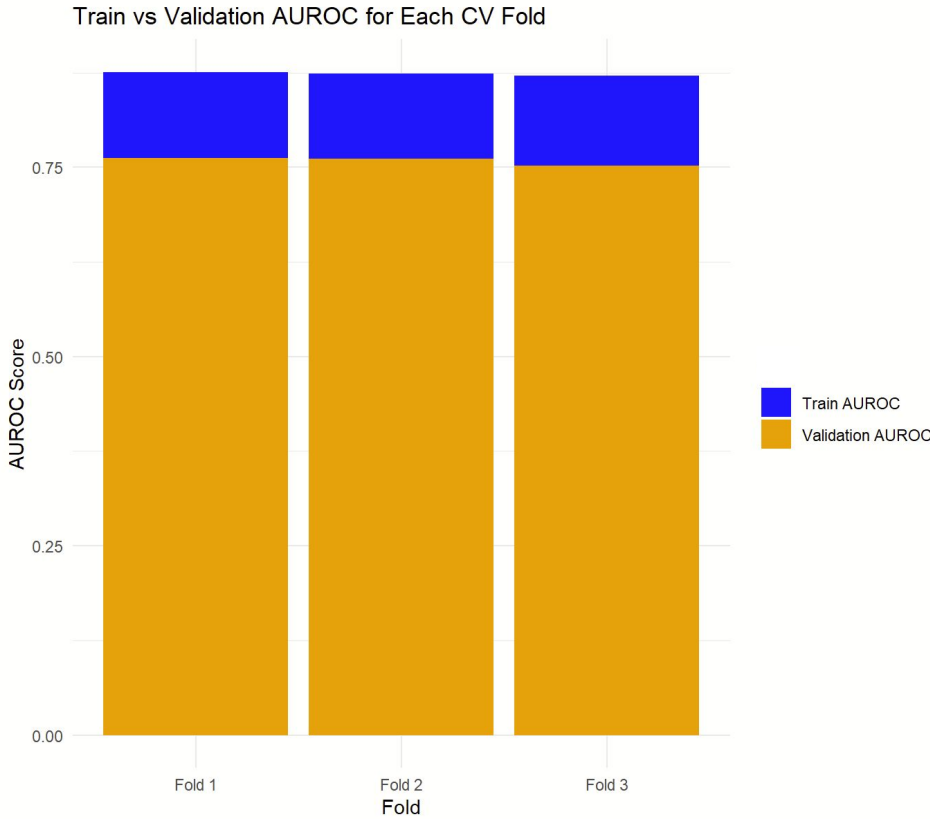


How much approx. information gain in the model (auROC)?

50% = none
60% = small
70% = moderate
80% = strong
90% = very strong
100% = perfect model

Model results

Figure 1: Train/test auROC for 3 fold cross validation



Train AUROC (Blue) is consistently high (~0.8+)

- This suggests the model **performs well on training data**.
- The model is capturing patterns in the gene features that help distinguish pollen-specific genes.

Validation AUROC (Brown/Gold) is slightly lower (~0.75)

- This means the model is still performing reasonably well on unseen validation data.
- Since the validation AUROC is **not too far below the train AUROC**, the model has **moderate generalization ability**.

Key Note:

- The AUROC scores for all three folds are fairly similar.

The model achieves AUROC ≈ 0.75 in validation,
meaning it is better than random (0.5) but still has room for improvement.

A **small train-validation gap** suggests the model is **not heavily overfitting**, which is a positive sign.

- **X-Axis (False Positive Rate - FPR):** Measures how often the model incorrectly classifies non-pollen-specific genes as pollen-specific.
- **Y-Axis (True Positive Rate - TPR):** Measures how often the model correctly classifies pollen-specific genes.
- The **diagonal dashed line (y = x)** represents random guessing (AUROC = 0.5). A model performing at this level has no predictive power.

Key Observations

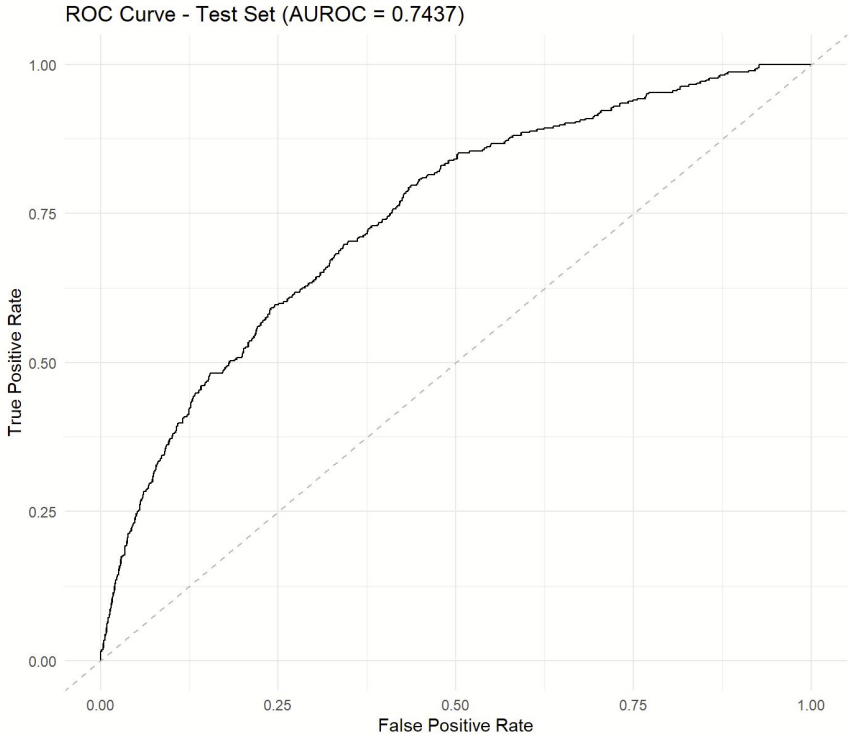
AUROC = 0.7437 (~0.74) : This means the model is **better than random (0.5) but not perfect (1.0)**.

A value of **0.74** suggests that when randomly selecting a pollen-specific and non-pollen-specific gene, the model correctly ranks them ~74% of the time.
The **ROC curve rises above the diagonal**, indicating that the model has predictive ability.

The steep initial increase means the model captures a significant number of **true positives** at lower false positive rates.

The curve flattens towards the top right, meaning additional correct classifications come at the cost of more false positives.

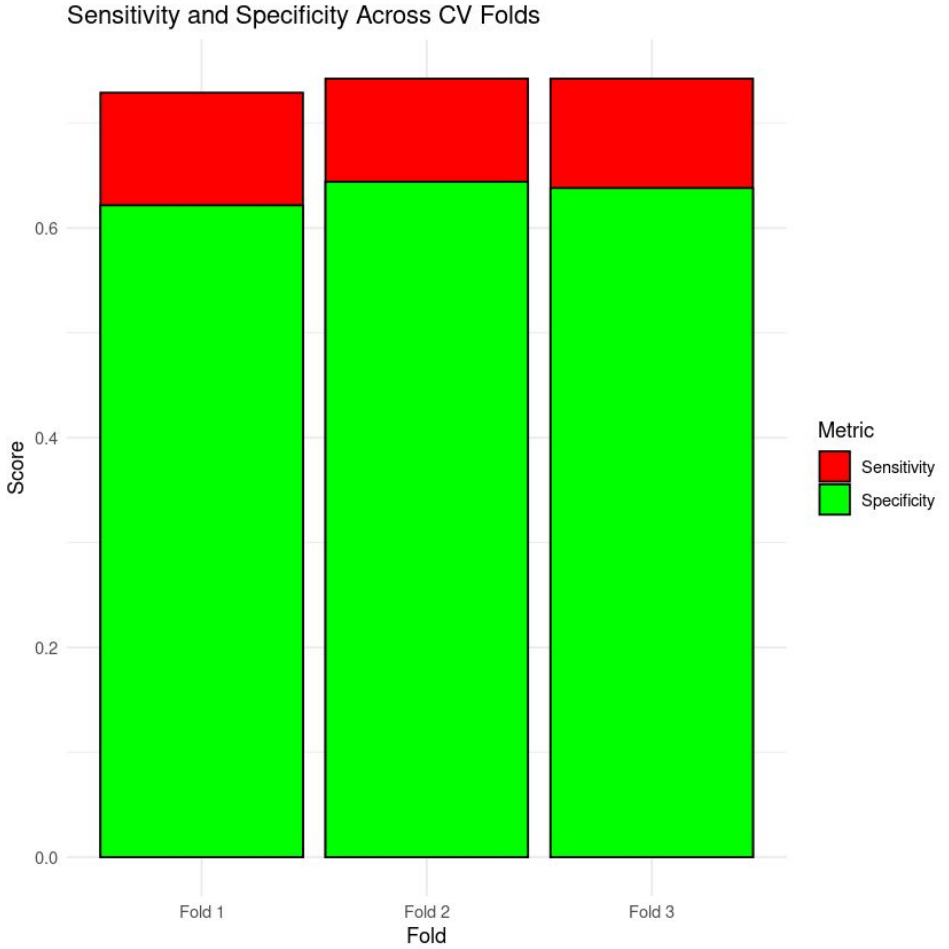
Figure 2: Held out test set auROC



Model results

Sensitivity and Specificity for all three cross-validation (CV) folds.

- **Sensitivity (Red):** Measures the proportion of correctly identified pollen-specific genes.
- **Specificity (Green):** Measures the proportion of correctly identified non-pollen-specific genes.
- **Results show that specificity is consistently high across all CV folds, while sensitivity is lower, suggesting that the model identifies non-pollen-specific genes more accurately than pollen-specific genes.**



Model Results

Confusion Matrix

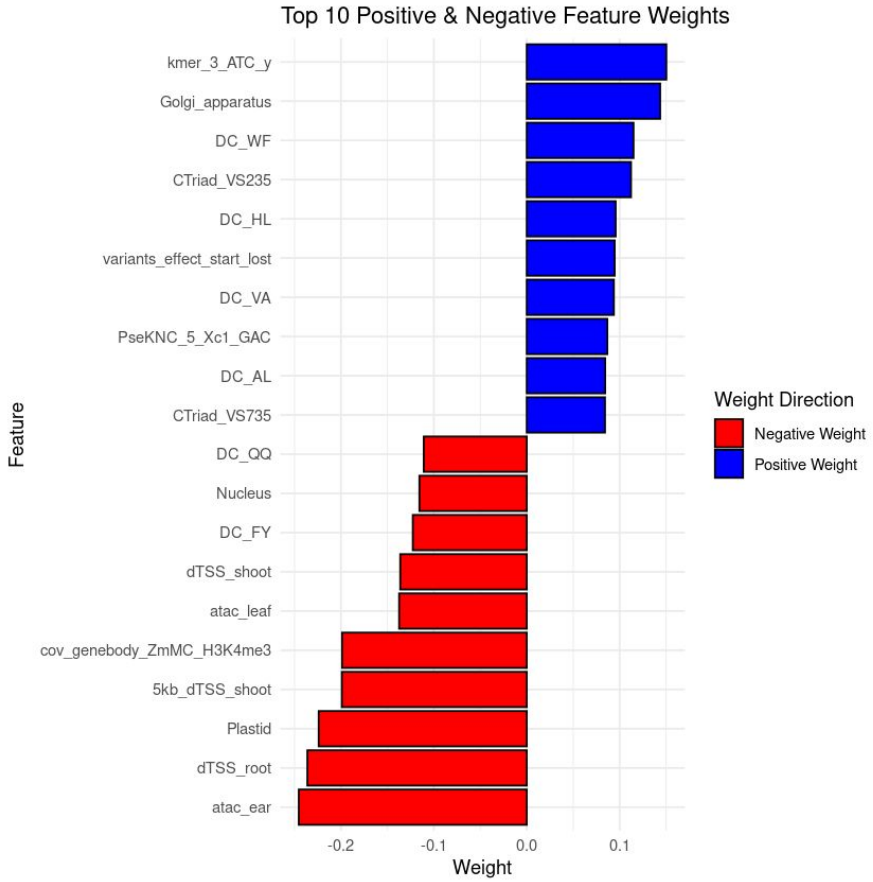
- **True Positives (TP = 301):** Correctly predicted pollen-specific genes.
- **False Positives (FP = 1339):** Non-pollen-specific genes incorrectly classified as pollen-specific.
- **False Negatives (FN = 84):** Pollen-specific genes incorrectly classified as non-pollen-specific.
- **True Negatives (TN = 2094):** Correctly predicted non-pollen-specific genes.

	Actual positive	Actual negative
Predicted positive	301 (TP)	1339 (FP)
Predicted Negative	84 (FN)	2094 (TN)

Model Results

Figure 5: Feature weights

336 weighted feature after L1



Conclusion

Expression specificity is difficult to predict from genome sequence

- PCA helps reduce dimensionality but does not show clear separation between classes.
- PC1 and PC2 explain only ~17% of the variance, indicating that more features or alternative methods are needed.
- LASSO regression selected 336 important features, improving interpretability and reducing overfitting.
- Final AUROC (~74.37%) indicates the model performs better than random but has room for improvement.
- Further refinement would be needed to improve model accuracy. Feature engineering? Class imbalance – maybe too many starting dimension

Thank you!

Data: https://mfs.maizegdb.org/featurebase