


Expression-based machine learning models for predicting plant tissue identity

Sourabh Palande¹ | Jeremy Arsenault² | Patricia Basurto-Lozada³ | Andrew Bleich⁴ |
 Brianna N. I. Brown⁴ | Sophia F. Buysse^{4,5,6} | Noelle A. Connors⁷ |
 Sikta Das Adhikari^{1,8} | Kara C. Dobson^{5,9} | Francisco Xavier Guerra-Castillo^{10,11} |
 Maria F. Guerrero-Carrillo¹² | Sophia Harlow⁷ | Héctor Herrera-Orozco^{13,14} |
 Asia T. Hightower^{4,5} | Paulo Izquierdo¹⁵ | MacKenzie Jacobs^{16,17} |
 Nicholas A. Johnson^{5,18} | Wendy Leuenberger^{5,9} | Alessandro Lopez-Hernandez^{3,19} |
 Alicia Luckie-Duque¹² | Camila Martínez-Avila²⁰ | Eddy J. Mendoza-Galindo¹² |
 David Cruz Plancarte²¹ | Jenny M. Schuster^{17,22} | Harry Shomer² |
 Sidney C. Sitar^{15,23,24} | Anne K. Steensma^{4,17,25} | Joanne Elise Thomson^{17,22} |
 Damián Villaseñor-Amador²⁶ | Robin Waterman^{4,5,6} | Brandon M. Webster⁴ |
 Madison Whyte¹⁵ | Sofía Zorilla-Azcué²⁷ | Beronda L. Montgomery²⁸ |
 Aman Y. Husbands²⁹ | Arjun Krishnan³⁰ | Sarah Percival¹ | Elizabeth Munch^{1,31} |
 Robert VanBuren^{7,32} | Daniel H. Chitwood^{1,7}  | Alejandra Rougon-Cardoso^{12,33}

Correspondence

Alejandra Rougon-Cardoso, Escuela Nacional de Estudios Superiores Unidad León, Universidad Nacional Autónoma de México, León, Guanajuato, Mexico.
 Email: arougon@enes.unam.mx

Daniel H. Chitwood, Michigan State University, East Lansing, Michigan, USA.
 Email: dhchitwood@gmail.com

Abstract

Premise: The selection of *Arabidopsis* as a model organism played a pivotal role in advancing genomic science. The competing frameworks to select an agricultural- or ecological-based model species were rejected, in favor of building knowledge in a species that would facilitate genome-enabled research.

Methods: Here, we examine the ability of models based on *Arabidopsis* gene expression data to predict tissue identity in other flowering plants. Comparing different machine learning algorithms, models trained and tested on *Arabidopsis* data achieved near perfect precision and recall values, whereas when tissue identity is predicted across the flowering plants using models trained on *Arabidopsis* data, precision values range from 0.69 to 0.74 and recall from 0.54 to 0.64.

Results: The identity of belowground tissue can be predicted more accurately than other tissue types, and the ability to predict tissue identity is not correlated with phylogenetic distance from *Arabidopsis*. *k*-nearest neighbors is the most successful algorithm, suggesting that gene expression signatures, rather than marker genes, are more valuable to create models for tissue and cell type prediction in plants.

Discussion: Our data-driven results highlight that the assertion that knowledge from *Arabidopsis* is translatable to other plants is not always true. Considering the current

For affiliations refer to page 11.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

landscape of abundant sequencing data, we should reevaluate the scientific emphasis on *Arabidopsis* and prioritize plant diversity.

KEYWORDS

Arabidopsis, flowering plants, gene expression, machine learning, model species, tissue identity

Resumen

Premisa: La selección de *Arabidopsis* como organismo modelo desempeñó un papel fundamental en el avance de la ciencia genómica. Se descartaron los marcos de referencia que proponían seleccionar una especie modelo basada en criterios agrícolas o ecológicos, en favor de profundizar en el conocimiento de una especie que promueve la investigación enfocada en el genoma.

Métodos: Aquí, examinamos la capacidad de los modelos basados en datos de expresión génica de *Arabidopsis* para predecir la identidad del tejido en otras plantas con flores. Comparando diferentes algoritmos de aprendizaje automático, los modelos entrenados y probados con datos de *Arabidopsis* alcanzaron valores de precisión y recuperación casi perfectos. De manera contrastante, cuando se predice la identidad del tejido en todas las plantas con flores utilizando modelos entrenados con datos de *Arabidopsis*, los valores de precisión oscilan entre 0,69 y 0,74 y los de recuperación entre 0,54 y 0,64.

Resultados: La identidad del tejido subterráneo puede predecirse con mayor exactitud que otros tipos de tejido, y la capacidad de predecir la identidad del tejido no está correlacionada con la distancia filogenética de *Arabidopsis*. El algoritmo *k*-nearest neighbors es el más exitoso y sugiere que las firmas de expresión génica, más que los genes marcadores, son más valiosas para crear modelos en plantas de predicción de tejidos y de tipos celulares.

Discusión: Nuestros resultados sustentados en datos demuestran que no siempre se cumple la afirmación de que el conocimiento de *Arabidopsis* es traducible a otras plantas. Teniendo en cuenta el panorama actual de abundantes datos de secuenciación, deberíamos reevaluar el énfasis científico en *Arabidopsis* y priorizar la diversidad vegetal.

Historically, plant biology has focused on inferring genetic, molecular, physiological, and ecological mechanisms. Conventionally, through quantifying phenomena and applying statistics, hypotheses are tested and decisions regarding the most likely scenarios are determined. New technologies and computational approaches have caused a shift from hypothesis-to data-driven research (Mazzocchi, 2015). Moreover, plant biology has embraced the inclusion of machine learning methods in addition to traditional statistical approaches (Ij, 2018). The combination of a deluge of data and new computational methods has allowed for predictive, rather than inferential, methods. Both statistics and machine learning can be used for inference and prediction, but machine learning methods more often classify and predict based on class labels rather than inferring the statistical parameters of a population. In plant biology, such predictive approaches underlie the frameworks of phenotyping (Coppens et al., 2017), precision agriculture (Zhang et al., 2002), genomic prediction (Crossa et al., 2014), linking transcriptomic profiles to phenotype (Azodi et al., 2020), and protein structure determination (Jumper et al., 2021). Just as inferential statistics has its limitations, the robustness and ability to extrapolate predictive models are also constrained by the empirical context from which the data originates. Although data-driven research is

slowly becoming more theoretical and predictive (Hogeweg, 2011), the creation of universal plant models is hindered by their overwhelming diversity. Not only is the phylogenetic diversity among flowering plants immense (The Angiosperm Phylogeny Group et al., 2016), but plants are exceptionally responsive to their environments (Sultan, 2000) and have evolved symbiotic interactions with and defense mechanisms against innumerable microbes (Mitchell et al., 2006). Furthermore, technical variability in data acquisition makes it difficult to exploit the huge amount of expression data archived in databases. The number of ways we sample molecular profiles from plant tissues and the interaction effects that arise between phylogenetically diverse species with environments, stresses, and biotic interactions are countless and prevent extrapolating results between studies.

Due to the clear advantages of studying a single model species, the early days of the genomics era tended to overlook the importance of prioritizing plant diversity. The candidates considered for the first sequenced genome were either easily transformable (e.g., species within Solanaceae; Knapp et al., 2004) or were already used for genetics (e.g., maize; Strable and Scanlon, 2009), but biodiversity was never considered (Meyerowitz, 2001). Reasons for choosing *Arabidopsis* as the first sequenced plant genome (*Arabidopsis* Genome

Initiative, 2000) include ease of transformation (Clough and Bent, 1998), its small genome (Bennett et al., 2003), life history traits that allow for genetics through crossing, and short generation times (Meyerowitz, 1987). The justification for initially sequencing the genome of a single model species was that such focus would allow unprecedented molecular discoveries that could be translated to other species and improve our understanding of all plants (Bevan and Walsh, 2005). The strategy to focus on a single model species was successful, and *Arabidopsis* is the most cited plant in the past 20 years, even surpassing key crops and all other plant species (Marks et al., 2023). Our molecular knowledge in plants was purposefully constructed to focus on *Arabidopsis* over crops and plant genetic diversity. However, such a choice has little relevance in a changing climate with dwindling natural resources and vanishing biodiversity that have become the most pressing concerns of our time. The cultural dynamics, dominated by the Global North, that influenced the choice of *Arabidopsis* as the first sequenced genome are reflected in the subsequently sequenced plant genomes. Plants native or endemic to land outside the Global North or first described by Indigenous cultures and territories have been sequenced by outside colonial powers (Marks et al., 2021; Dwyer et al., 2022). While sequencing *Arabidopsis* has certainly expanded our knowledge of molecular processes, this intense focus has limited our understanding of other species, raising the question: To what extent can the insights from *Arabidopsis* be extrapolated to the rest of flowering plants?

In the 20 years since the release of the *Arabidopsis* genome sequence (Arabidopsis Genome Initiative, 2000), the number of sequenced plant genomes has dramatically increased (Michael and Jackson, 2013; Li and Harkess, 2018; Marks et al., 2021), leading to a greater understanding of the evolutionary origin and genetic mechanisms underlying numerous traits across the green lineage. Next-generation sequencing, for example, has enabled unprecedented surveys of genome-scale features across species, tissue types, environments, and interactions between plants with abiotic and biotic factors. There are currently over 300,000 public gene expression datasets spanning thousands of diverse plant species (Lim et al., 2022). Cross-species comparisons of gene expression across plants have usually been limited by the number of species analyzed (Proost and Mutwil, 2018) or their sampling breadth. Most studies have generated datasets from scratch (Julca et al., 2021) instead of leveraging public repositories. Databases and datasets curating and making vast numbers of gene expression profiles and their associated metadata have been created. For example, an *Arabidopsis* RNA-Seq database compiles 20,068 publicly available *Arabidopsis* RNA-Seq libraries (Zhang et al., 2020), and the Plant Public RNA-seq Database has ~45,000 maize, rice, wheat, soybean, and cotton samples (Yu et al., 2022). Previously, a dataset of 2671 publicly available gene expression profiles from 54 flowering plant species across seven developmental tissue types and nine stresses had been curated (Palande et al., 2023). More than 20 years after the release of the *Arabidopsis* genome, we have accumulated enough data across plants to ask

unprecedented questions, and we have the computational tools that permit comparative approaches to analyze such massive amounts of data.

Here, building upon large, curated databases of *Arabidopsis* (Zhang et al., 2020) and flowering plant gene expression profiles (Palande et al., 2023), we examine how predictive *Arabidopsis* is as a model species relative to the rest of the flowering plants and to what degree we can apply our knowledge from model organisms to diverse plant species. Dimension reduction through principal component analysis (PCA) reveals that biotic stress response and tissue type are primary, orthogonal sources of structure in gene expression data from *Arabidopsis*, and while angiosperm data projected onto this space retain some structure, the regions occupied between tissue types become less distinct. We next compare the performance of different machine learning models. The *k*-nearest neighbors (kNN) method yields precision and recall values of up to 0.99 using models trained and tested on *Arabidopsis* data. Model performance decreases significantly, with higher precision than recall values, when data from across flowering plants are tested using models trained on *Arabidopsis* data. Belowground tissue is more separated from and predictable than other tissue types, and phylogenetic distance from *Arabidopsis* does not appear to influence prediction rates. We end with a discussion of the implications of our results for the current structure of the plant science community, acknowledging that the past focus on *Arabidopsis* as a model organism based on decisions made decades ago was effective at that time; however, we now advocate for a shift in approach due to changing circumstances, particularly in light of the pressing issue of biodiversity loss. We argue for a more decentralized and inclusive research framework that better encompasses the diversity of plants and the human cultures that represent them, adapting to current environmental and scientific challenges.

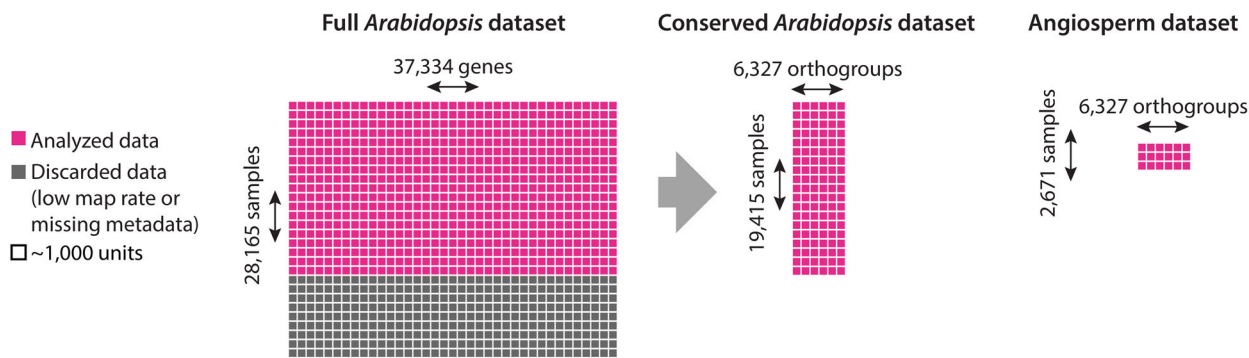
METHODS

The code necessary to reproduce the results presented here is available on GitHub (<https://github.com/PlantsAndPython/arabidopsis-gene-expression>; see Data Availability Statement). The gene expression data are not included in the repository due to the large file size; these are available at Dryad (Chitwood and Palande, 2024; <https://datadryad.org/stash/dataset/doi:10.5061/dryad.4b8gthtn7>). To reproduce the analysis presented in this paper, first clone the GitHub repository, then download the dataset from Dryad and deposit it in the “data” directory of the cloned repository. The code assumes that the data files are available in the directory.

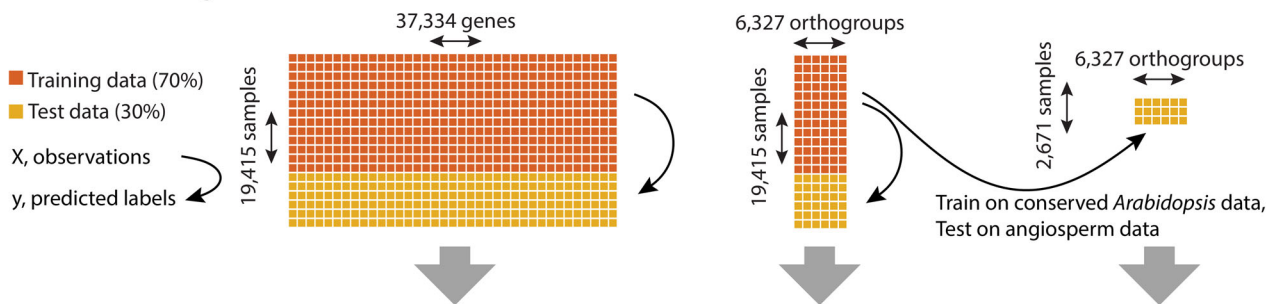
Datasets

We used two curated databases in this analysis (Figure 1A). The first contained 28,165 *Arabidopsis* gene expression profiles across 37,334 genes (Zhang et al., 2020). The second

A Data curation, cleaning, and preparation



B Create training and test sets



C Classification models

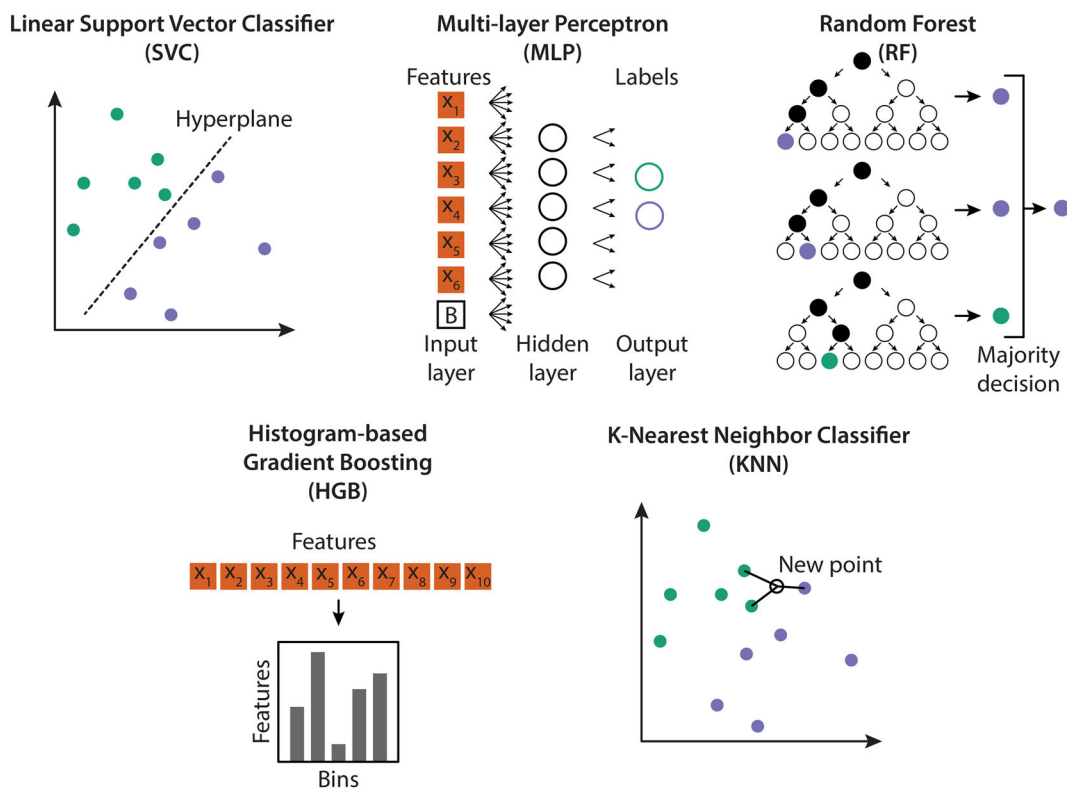


FIGURE 1 (See caption on next page).

contained 2671 flowering plant expression profiles across 6327 orthogroups (Palande et al., 2023). We originally classified samples from both databases into 23 tissue types: “anther,” “carpel,” “cotyledon,” “flower,” “hypocotyl,” “inflorescence,” “internode,” “leaf,” “other,” “petal,” “petiole,” “pistil,” “reproductive-other,” “root,” “root cell,” “seed,” “seedling,” “sepal,” “shoot,” “stamen,” “stigma,” “vasculature,” or “whole plant.” Although we discuss model performance using the detailed tissue designations above, due to large differences in sample size between these categories, for our main analysis we aggregated tissue designations into four tissue type labels: “aboveground,” “belowground,” “whole plant,” or “other.” The categories are purposefully encompassing and were chosen to facilitate accurate assignment across the broad categories of experimental data we analyzed, focusing on aboveground and belowground tissue identity as one of the simplest cases to test tissue predictability. Samples for which tissue identity could not be determined from their description were discarded, as they were incompatible with our machine learning pipeline. Additionally, we discarded low-quality samples, which we measured by unique mapped rate, or the number of uniquely mapping reads divided by the total number of reads. After removing samples with missing metadata and samples with low unique mapped rate (<75%), the *Arabidopsis* database was left with 19,415 samples. A conserved *Arabidopsis* database was also constructed by keeping only the genes mapped to the orthogroups from the flowering plant database. The conserved *Arabidopsis* database contained the same number of samples, but with much smaller expression profiles across only the 6327 orthogroups shared with the angiosperm dataset.

Classification models

Classification is a common machine learning task in which, given data points belonging to two or more classes, the goal is to *learn* a function that best differentiates between points from different classes. Then, given a new data point, the function can be used to decide which class the point belongs to. The classifier function can be learned in many ways, leading to various types of machine learning models. For each classifier model in this study, we employed the following modeling methods:

Linear support vector classifier (SVC): In linear classification, each point is viewed as a vector in k -dimensional

space (Cortes and Vapnik, 1995), where k is the number of desired groups to predict. The goal is to find $(k - 1)$ -dimensional hyperplanes that separate the points belonging to different classes. For example, if we wish to predict which of two groups ($k = 2$) samples belong to, then in a two-dimensional space, we find one $(k - 1 = 1)$ hyperplane to divide the space and separate the points of the different classes (Figure 1C). There are many possible choices for hyperplanes that can classify the points. A reasonable choice is to find the ones that maximize the separation between points from different classes. These are known as maximum-margin hyperplanes. Geometrically, the maximum-margin hyperplanes are defined by the points that lie closest to them; therefore, such points are called support vectors.

Multi-layer perceptron (MLP): The SVC model (see above) assumes that the classes are linearly separable, which may not be true. MLPs are a class of artificial neural networks (Haykin, 1998) with three or more layers of “perceptrons” with non-linear activation. An MLP consists of an input and an output layer, with one or more hidden layers of neurons. As is conventional for MLP model parameterization, we experimented with one and two hidden-layer MLPs (see Figure 1C) and used rectified linear unit (ReLU) activation in all cases to optimize the prediction of our classifier. In ReLU, a neuron’s activation is the weighted sum of its inputs if the sum is non-negative, and zero otherwise. Even with this simple non-linear activation function, MLPs can outperform the linear SVC model.

Random forest (RF): Random forests (Ho, 1995) perform classification by constructing an ensemble of decision trees. Each decision tree outputs a class label for the given sample, and the output of the RF is the class label predicted by the majority of the trees. In a decision tree, each internal node is labeled by an input feature, and the leaf nodes are labeled by the class labels. Starting from the root node, the input set is recursively partitioned into children nodes using the input feature associated with the node. The recursion ends when all data points in the node belong to the same class, or when some pre-specified termination criteria, such as maximum depth of the tree, are met. Which feature to split the data on at each level is determined using information criteria such as Gini impurity or entropy that measure how consistent the subsets are with respect to the class labels after the split.

Histogram-based gradient boosting (HGB): Gradient boosting (Mason et al., 1999) is another class of methods that uses a large ensemble of decision trees. In HGB, the

FIGURE 1 Experimental design visualization. (A) Data curation, cleaning, and preparation. The original *Arabidopsis* dataset consists of 28,165 samples with 37,334 gene expression features. Samples with low unique map rates or missing metadata were discarded (gray), yielding the full *Arabidopsis* dataset with 19,415 samples (magenta). The conserved *Arabidopsis* dataset consists of the same 19,415 samples as the full dataset while the angiosperm dataset has only 2671 samples. Both the conserved *Arabidopsis* and angiosperm datasets consist of 6327 gene expression features that represent conserved orthogroups shared by both datasets. (B) Training and test set creation. Classification models were fitted using 70% of the full and conserved *Arabidopsis* datasets and testing on the remaining 30%. The angiosperm dataset was used as a test set on models created from the conserved *Arabidopsis* dataset. (C) Classification models. Once the training and test sets have been created, classification models and prediction can be run. The five models we test are linear support vector classifier (SVC), multi-layer perceptron (MLP), random forest (RF), histogram-based gradient boosting (HGB), and k -nearest neighbor classifier (kNN), which are described in the Methods section.

real-valued input features are first discretized into a few (typically 256) bins using histograms. This allows the training algorithm to run much more efficiently and construct a much larger ensemble of decision trees to support the classification.

k-nearest neighbors (*k*NN) classifier: In *k*NN classifiers (Cover and Hart, 1967), class labels are assigned based on a majority vote of the *k*-nearest training points. The distance metric and the number of neighbors are specified by the user. In our experiments, correlation distance between the expression profiles was used to train the *k*NN classifier.

Experimental design

To establish the utility of gene expression profiles in predicting tissue type, we trained supervised machine learning models (by tuning hyperparameters using a Bayesian optimization procedure, see below) to classify the *Arabidopsis* data by tissue types (Table 1). The database was split into training and test sets (70%:30% split; Figure 1B). An arbitrary percentage of random samples is used to train a model, and the remainder are used to test its performance. The 70%:30% split for the training and test sets is standard and worked for our purposes here, but any proportion could work. To ensure comparability, all five models were trained and tested on the same training and test sets. Next, we wanted to examine how predictive *Arabidopsis* is to the rest of the flowering plants (Table 2). To test this, we used a set of conserved *Arabidopsis* transcripts with orthogroups across angiosperms, split into training and test sets (70%:30% split) as before. The same five machine learning models were trained on the conserved *Arabidopsis* training set. The performance of these models was first tested on the conserved gene *Arabidopsis* test set to make sure that the models were still able to predict the tissue types with a significantly smaller number of features. We then used the same models to classify the angiosperm data to test how well they extrapolate to species other than *Arabidopsis*. Each machine learning model employed in our experiments requires additional hyperparameters that need to be tuned to optimize model performance. We used the Bayesian optimization procedure implemented in the Hyperopt package in Python (Bergstra et al., 2013). Briefly, by evaluating an objective function (e.g., model accuracy), a Bayesian probability model can be built that uses past parameter search values to inform the selection of the next parameter values to evaluate and arrive at optimized parameter values. To gain insights into the functional annotation and enrichment of our gene list, we performed a Gene Ontology (GO) term (Ashburner et al., 2000) analysis using the DAVID Functional Annotation Clustering tool (version 2021; <http://david.ncifcrf.gov>) (Huang et al., 2009). Each principal component (PC) is calculated as a linear combination of input features. The weight attributed to each feature that defines a PC is known as a loading. Each gene expression feature thus has a single loading value, allowing

TABLE 1 Classification performance of models trained on the full *Arabidopsis* dataset.

Model	Precision	Recall	F1 score
SVC	0.765131	0.80103	0.777531
MLP	0.843599	0.844979	0.832854
RF	0.845664	0.826609	0.833746
HGB	0.976665	0.976481	0.976319
<i>k</i> NN	0.98921	0.989185	0.989193

Note: HGB = histogram-based gradient boosting; *k*NN = *k*-nearest neighbors; MLP = multi-layer perceptron; RF = random forest; SVC = support vector classifier.

TABLE 2 Classification performance of models trained on the conserved *Arabidopsis* dataset and tested on conserved *Arabidopsis* or angiosperm datasets.

Model	Test set	Precision	Recall	F1 score
SVC	<i>Arabidopsis</i>	0.740855	0.778026	0.754276
	Angiosperm	0.695691	0.576189	0.591683
MLP	<i>Arabidopsis</i>	0.822682	0.828155	0.824351
	Angiosperm	0.734603	0.547361	0.611767
RF	<i>Arabidopsis</i>	0.862941	0.864721	0.861927
	Angiosperm	0.747272	0.569075	0.622122
HGB	<i>Arabidopsis</i>	0.971034	0.970987	0.970574
	Angiosperm	0.741902	0.567952	0.640741
<i>k</i> NN	<i>Arabidopsis</i>	0.987804	0.987811	0.987803
	Angiosperm	0.733478	0.643205	0.663313

Note: HGB = histogram-based gradient boosting; *k*NN = *k*-nearest neighbors; MLP = multi-layer perceptron; RF = random forest; SVC = support vector classifier.

us to determine which genes most contribute to a PC. We filtered the 200 genes with the most positive and negative PC1 loading values. The annotation was performed using The Arabidopsis Information Resource (TAIR) IDs (<https://www.arabidopsis.org/>; Reiser et al., 2024) and selecting GO terms from levels 3 and 4 of the molecular function and biological process categories.

RESULTS

Dimension reduction and alignment between *Arabidopsis* and angiosperm gene expression datasets

Although inferential statistics is sensitive to imbalances between factor levels, predictive methods are less so, as long as there is sufficient sampling of the features of the smallest class. Although we originally classified samples into 23 tissue types (see Methods section) and we compare model results of this classification with the main results (see

below), out of an abundance of caution, we categorized samples into bins with adequate sampling of the aboveground, belowground, whole plant, and other labels. A PCA performed on the full dataset of 19,415 *Arabidopsis* RNA-Seq samples shows a clear separation by the four tissue type labels (Figure 2A). The aboveground, belowground, and other tissue types are well separated from each other, although the belowground tissue has the least overlap with other tissues. The whole plant tissue type, composed of different combinations of the other tissues, is not well separated, as we would expect. The separation of tissues occurs along a gradient defined by PC2, demonstrating that tissue type is not the primary source of variance in the data. Rather, a small proportion of samples are distributed across PC1 in an additive, orthogonal manner, preserving the separation of tissue types defined by PC2. To investigate the underlying cause responsible for the primary source of variation in the data, we performed GO enrichment on genes with the most extreme PC1 loading values that are

most responsible for defining PC1. In the full *Arabidopsis* dataset (Figure 2A), high PC1 values, which include a small number of samples that contribute to a disproportionate amount of variance in the data, are defined by high expression of genes associated with response to the biotic stress and oxidative damage GO terms (Appendix S1). Low PC1 values, which include a majority of samples across tissues and which we assume arise from plants grown under regular conditions associated with less stress, are defined by high expression of genes with GO terms associated with biosynthesis, biogenesis, and cell growth. Remarkably, in the full *Arabidopsis* dataset, negative PC1 loading values are enriched for glucosinolate biosynthesis and other metabolic processes (false discovery rate <0.05).

From these large-scale datasets, we developed a predictive model to test if tissue type could be inferred from expression patterns alone and if this *Arabidopsis*-trained model could be transferred to other flowering plants. We previously created a set of 6327 low-copy orthogroups that

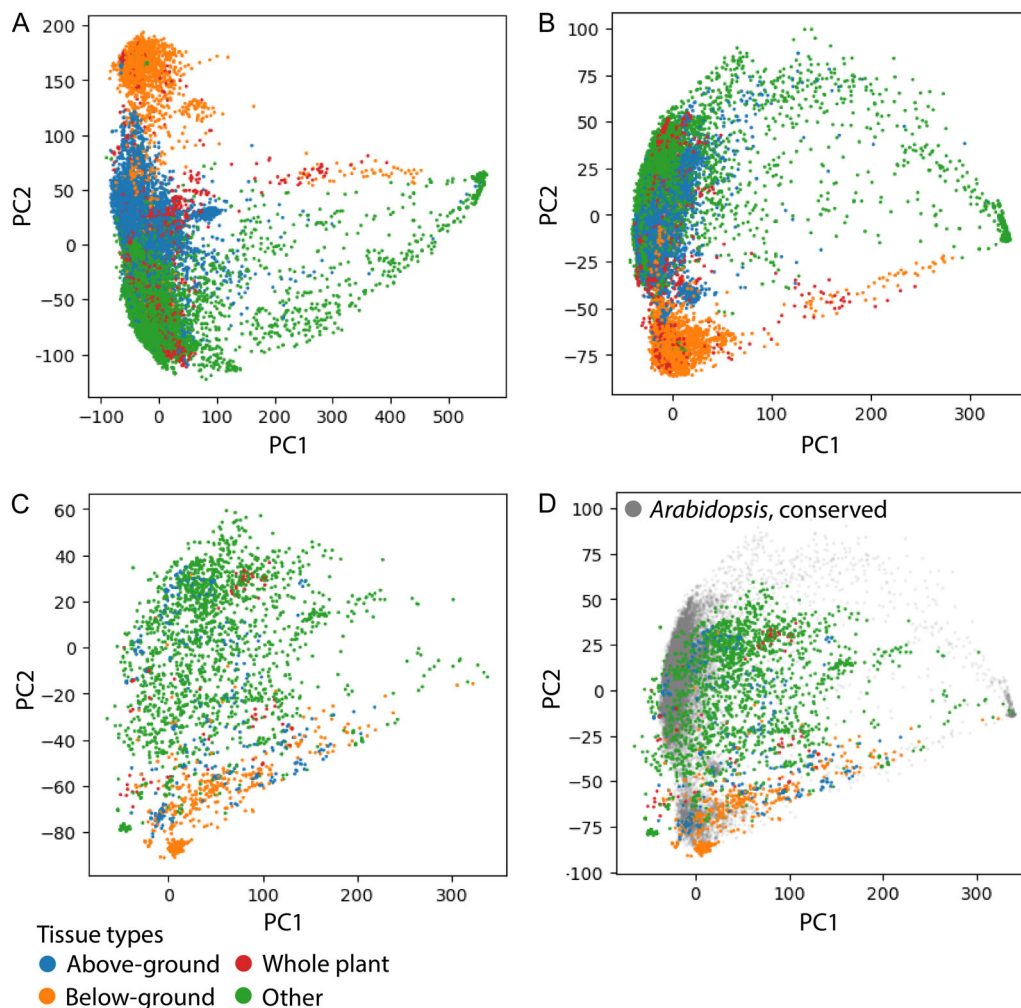


FIGURE 2 Principal component analysis (PCA) of gene expression profiles. PCAs with gene expression profiles colored by aboveground (blue), belowground (orange), whole plant (red), and other (green) tissue types for (A) the full *Arabidopsis* dataset, (B) the conserved *Arabidopsis* dataset, (C) the angiosperm dataset projected onto the conserved *Arabidopsis* PCA from (B), and (D) the same as (C), but with conserved *Arabidopsis* gene expression profiles in the background (transparent gray).

are deeply conserved across flowering plants (Palande et al., 2023) and used a set of 6327 *Arabidopsis* genes corresponding to these orthogroups for all downstream analyses. A PCA performed on this subset of 6327 conserved flowering plant genes shows mostly the same structure as the analysis with all *Arabidopsis* genes included (Figure 2B). However, while the belowground tissue type remains distinct from the rest of the data, the aboveground tissue type overlaps more with the whole plant and other tissue types. Note that whether a PC is positive or negative is arbitrary, which explains the “flip” of PC2 values relative to the full set of *Arabidopsis* genes. An analysis of the enriched GO terms for PC1 loading values from the conserved gene PCA reveals that high PC1 values are associated with biotic responses, but also with anther- and pollen-related GO terms (Appendix S1). Low PC1 values are associated overwhelmingly with photosynthesis. Because the two datasets have corresponding orthogroup features, we are able to project the angiosperm dataset onto the PCA defined by the conserved gene *Arabidopsis* dataset (Figure 2C, D). While the overall structure defining the distributions of tissue types is maintained in the projected angiosperm data, there is substantial overlap between the aboveground and belowground tissue types. We conclude that indeed there is conservation of tissue-specific expression between *Arabidopsis* and the rest of the flowering plants, but as expected, the alignment of the underlying structures of gene expression patterns defining tissue type identity is not identical.

Predictive modeling of plant tissue from gene expression

We used supervised learning classifiers to test if gene expression profiles could predict tissue type in

Arabidopsis and if these *Arabidopsis*-trained models could be applied more broadly to flowering plants. We first split the *Arabidopsis* data into testing and training sets, with samples divided into four classes of aboveground, belowground, whole plant, or other, as described above. Models trained on *Arabidopsis* expression data and used to predict tissue type in *Arabidopsis*, whether the full or conserved gene datasets, achieved high precision and recall scores. The highest F1 scores (the harmonic mean of precision and recall) for the full and conserved datasets were achieved using a kNN algorithm (0.99 and 0.99, respectively; Tables 1 and 2) and the lowest using the SVC model (0.78 and 0.75). The HGB model also achieved high F1 scores (0.98 and 0.97), whereas the results for RF (0.83 and 0.86) and MLP (0.83 and 0.82) were intermediate. When used to predict *Arabidopsis* data, the precision and recall values for each model were similar to each other, indicating similar positive prediction value (precision, true positives divided by true positives and false positives) and sensitivity (recall, true positives divided by true positives and false negatives). The relative prediction rates of different tissue types to each other were equivalent for the full *Arabidopsis* dataset (Figure 3A). If we run the kNN model using the 23 tissue labels instead of four, similarly high prediction statistics are achieved both for the full *Arabidopsis* dataset (precision: 0.980999, recall: 0.980944, F1 score: 0.980830) and the conserved *Arabidopsis* dataset (precision: 0.975517, recall: 0.975279, F1 score: 0.974980).

The projection of gene expression patterns from across flowering plants onto a PCA (calculated using Python scikit-learn functions PCA and StandardScaler to scale gene expression features [Pedregosa et al., 2011]) using a conserved set of genes from *Arabidopsis* shows considerable variability (Figure 2C, D). Using models trained on

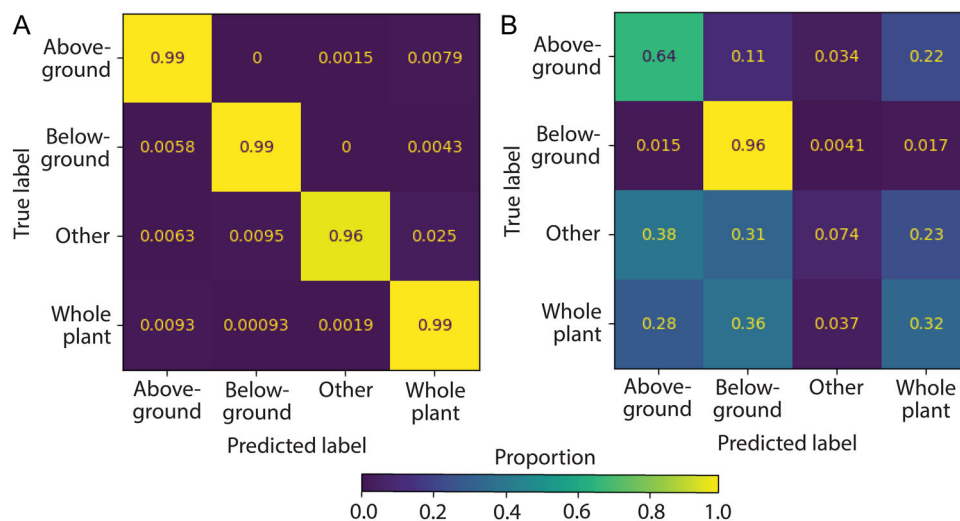


FIGURE 3 Confusion matrices using the kNN classifier. Confusion matrices showing the true label identity (vertical axis) and the proportion of samples assigned to predicted label identities (horizontal axis) for (A) the full *Arabidopsis* dataset and (B) the angiosperm dataset. Proportion indicated by viridis color scale (Garnier et al., 2024).

Arabidopsis data and tested on flowering plants, prediction rates are more similar to each other using different algorithms than *Arabidopsis* alone but perform much worse, and with higher precision than recall rates (Table 2). For the kNN, HGB, RF, MLP, and SVC methods, precision values were 0.73, 0.74, 0.75, 0.73, and 0.70, respectively, whereas the rates of recall were 0.64, 0.57, 0.57, 0.55, and 0.58. The precision values are uniformly around the same as the values of the worst-performing models using only *Arabidopsis* data and may reflect that, although a reasonable classifier can be constructed for *Arabidopsis* data, this is not the case when predicting flowering plant data using models trained on *Arabidopsis*. Although these rates are moderately high, they must be interpreted in the context of using only four tissue type labels. The relatively higher precision rates compared to recall indicate that when a sample is retrieved, there is a higher rate of the models calling a true positive (positive prediction value) compared to the fraction of relevant samples retrieved (sensitivity). The prediction rates across tissue types were not evenly distributed (Figure 3B). Belowground tissue was accurately classified, at a rate of 0.96, while aboveground tissue was only correctly predicted at a rate of 0.64. The tissue types other and whole plant were classified poorly (0.074 and 0.32, respectively), and almost no samples were predicted as the tissue type other, including samples classified as other. Although the prediction accuracy varies considerably across plant families (Figure 4), from around 0.4 to 0.8, we could not identify any phylogenetic signal or find any support that prediction of tissue identity is inversely correlated with the distance of a plant family from *Arabidopsis* in the Brassicaceae. If we run the kNN model predicting flowering plant data trained on *Arabidopsis* data using the 23 tissue labels instead of four, we achieve similarly poor prediction results (precision: 0.523399, recall: 0.515943, F1 score: 0.490203).

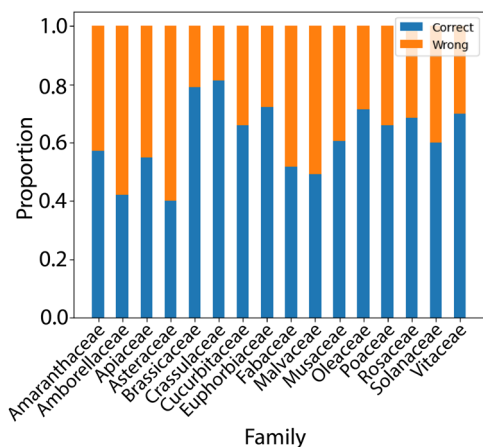


FIGURE 4 Prediction accuracy by plant family. Using the kNN classifier on the angiosperm dataset, the proportion of samples correctly (blue) and wrongly (orange) predicted from *Arabidopsis* data is shown as a stacked bar plot.

DISCUSSION

Arabidopsis-only models are highly accurate

Although we focus on tissue identity in this study, we note that the strongest source of variance (PC1) in publicly available *Arabidopsis* gene expression profiles is a signature associated with biotic defense (Appendix S1) and that it acts in an additive, orthogonal manner with respect to tissue type, which is the next strongest source of variance (PC2). Higher prediction rates are expected for the *Arabidopsis*-only models both because the same dataset is being used for training and testing, and because the data structure that separates the main factors being tested (i.e., aboveground and belowground tissues), as visualized in a PCA, is substantial (Figure 2A, B). From this perspective, it is perhaps not surprising that kNN is the best-performing algorithm, based on the overall distance-based proximity of gene expression profiles for each label to each other (Table 1). The other methods, which are based on decision trees or neural networks, focus on individual gene expression values as parameters, and thus fail to account for overall distance. The focus on individual gene expression values instead of the overall signature or profile is reminiscent of the molecular biology concept of “biomarkers” to indicate the tissue or stress from which a sample arises. The out-performance of kNN over other algorithms we tested may suggest that gene expression signatures (rather than individual gene expression values) are more valuable in creating models for tissue and cell type prediction.

Arabidopsis gene expression as a model for other flowering plants may not be the most suitable approach

Lower prediction rates are expected when a model is tested on different data than its training set (Table 2). However, the lower precision and recall scores attained when a model trained on *Arabidopsis* is tested on gene expression samples across the flowering plants undermines the foundational argument for using model species: that data from *Arabidopsis* would be predictive for plants in general. This is not to say that there is not substantial conservation of tissue-specific gene expression patterns. Our own work (Palande et al., 2023) and that of others (Julca et al., 2021) strongly supports conserved tissue-specific gene expression patterns across flowering plants, as is true of animals as well (Fukushima and Pollock, 2020). Rather, the ability to leverage and predict tissue identity from conserved gene expression profiles is diminished when building a model from a single, arbitrary species.

Details of the performance of our model hint at underlying biological considerations when using model species data. Not all tissue types are equally predictable, and the prediction of belowground tissue outperforms other tissue types (Figure 3). We hypothesized that the ability to predict

tissue identity from *Arabidopsis* may be inversely correlated with the phylogenetic distance of a sample from Brassicaceae, but we found no evidence to support this idea (Figure 4). Additionally, the precision values for predicting tissue type of flowering plant data from *Arabidopsis* are much higher than the recall values (Table 2). This may indicate that models are relatively better at calling samples with conserved tissue specificity with *Arabidopsis* (a true positive) over those without (a false negative). These results may also be a product of our classification scheme. For example, samples in the aboveground and whole plant tissue categories are often more similar to each other than to belowground tissue because they are missing roots and might more easily be misclassified with each other. The category other is composed of diverse tissues that may not have clear predictive features. These factors should be considered when evaluating the classification results (Figure 3). Nonetheless, if we run a kNN model using the original 23 tissue type labels that were aggregated to create the four labels that we focused on, similar prediction results are achieved; this is true both for exceptional results when testing and training on only *Arabidopsis* data and for poor performance when predicting on flowering plant data trained using *Arabidopsis* data. While it is important to eventually explore more defined labels describing specific tissue types (or even single cell data) across developmental stages, our results indicate that high prediction rates using only *Arabidopsis* data and lower rates predicting flowering plant data from a single model species would likely remain the case.

Our results potentially arise not only from genes with evolutionary differences in tissue-specific expression compared to *Arabidopsis*, but also from genes that may have conserved expression but differ in the ways we have culturally constructed our developmental descriptions of plant species. Such a circumstance might arise when the cell type-specific expression of a gene is truly conserved, but evolved differences in functional morphology between species lead us to apply different tissue descriptors (e.g., between a herbaceous annual and a woody perennial, or a CAM succulent compared to a weedy C_3 plant). The misalignment of tissue labels extends to more quantitative descriptors and to the molecular level, including GO terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms (Kanehisa and Goto, 2000) that ultimately become biased toward plants with sequenced genomes (Provart et al., 2016). For example, in our analysis of genes corresponding to the most positive and most negative PC1 loading values, there was a noticeable enrichment of genes associated with the glucosinolate biosynthetic and metabolic pathways in *Arabidopsis* samples (Appendix S1). However, this enrichment was absent in broader angiosperm samples, as these compounds are found almost exclusively in Brassicaceae. Glucosinolates are a diverse group of secondary metabolites that play a critical role in plant defense against herbivores and pathogens. Beyond their defensive role, they seem to be involved in growth, development, microbiota interactions, and phosphate

nutrition (Kopriva, 2021). Focusing on a single organism or on a small group of model species to predict attributes of all plants is a flawed approach from both biological (arising from evolutionary novelty) and philosophical (due to semantic, ontological, and cultural differences in how we socially construct plants) perspectives.

Moving forward and embracing plant and cultural diversity

Arabidopsis was selected as a model species unilaterally, over raised objections, decades ago, on the basis of primarily genetic and molecular biology considerations (Meyerowitz, 1987; Clough and Bent, 1998; Arabidopsis Genome Initiative, 2000; Bennett et al., 2003; Bevan and Walsh, 2005). Arguments in favor of selecting agricultural or ecological models or models that would better represent plant diversity were ignored. These past decisions have led to continued focus on *Arabidopsis*, and there is continuing advocacy for funding research using *Arabidopsis* as a model species at the expense of plant diversity to the current day (Provart et al., 2016; Parry et al., 2020). Since then, data science and computational approaches have gained increasing importance. After decades of acquiring sequencing data from across the flowering plants, we are able to ask objectively if focusing on a single plant allows us to predict the biology of other flowering plants better than if we had studied all plants equally from the start; the answer is no (Table 2). Using a data science approach and building machine learning models using *Arabidopsis* gene expression data to predict the tissue identity of gene expression samples from across flowering plants, as we have done here, does not preclude the consideration of other, more important qualitative arguments against the model species concept that continues to limit the potential of the plant science community. Furthermore, beyond *Arabidopsis*, there is an additional focus on agriculturally important species at the expense of all plants (Marks et al., 2023). More insidiously, the social construct of plants and their diversity arises from colonialism, as evidenced not only by the plants that we in the Global North have chosen to research and document and how we do so, but also by which plant genomes have been sequenced and by whom (Marks et al., 2021), usually through extinguishing and stealing the cultural knowledge of Indigenous peoples (Dwyer et al., 2022). The collective gene expression data of the plant science community is highly biased towards *Arabidopsis* (Marks et al., 2023). We speculate that if the global plant science community were to operate more equitably and include different cultural perspectives focusing on diverse species, the data we collect would be more varied and could be included in models that better encompass all plants.

Useful discoveries and insights have arisen from the *Arabidopsis* genome initiative that have served as a blueprint for and inspired similar genomic initiatives in numerous other plant species (Arabidopsis Genome

Initiative, 2000). Decades later, using machine learning approaches, we are now able to reevaluate our past efforts and plan a new course forward. The methodology that we present here can be used for much more than evaluating whether a particular plant species is an optimal model organism choice; rather, these methods can be used to create and evaluate models leveraging all species data in predictive frameworks. Rather than advocating for continued focus and funding for a single model species (Provart et al., 2016; Parry et al., 2020), we are long past due in addressing the historical inequities that have led to our current construction of the plant sciences and in embracing the biological and cultural diversity of the plant world, which will result in a sounder and more predictive science.

AUTHOR CONTRIBUTIONS

So.P., B.L.M., A.Y.H., A.K., Sa.P., E.M., R.V., D.H.C., and A.R.-C. acquired funding and conceived the research; J.A., P.B.-L., A.B., B.N.I.B., S.F.B., N.A.C., S.D.A., K.C.D., F.X.G.-C., M.F.G.-C., S.H., H.H.-O., A.T.H., P.I., M.J., N.A.J., W.L., A.L.-H., A.L.-D., C.M.-A., E.J.M.-G., D.C.P., J.M.S., H.S., S.C.S., A.K.S., J.E.T., D.V.-A., R.W., B.M.W., M.W., and S.Z.-A. curated data, designed the experiments, and visualized the data; So.P. led the final data analysis; R.V., D.H.C., and A.R.-C. led the class as instructors in which the research was performed; So.P., J.A., P.B.-L., A.B., B.N.I.B., S.F.B., N.A.C., S.D.A., K.C.D., F.X.G.-C., M.F.G.-C., S.H., H.H.-O., A.T.H., P.I., M.J., N.A.J., W.L., A.L.-H., A.L.-D., C.M.-A., E.J.M.-G., D.C.P., J.M.S., H.S., S.C.S., A.K.S., J.E.T., D.V.-A., R.W., B.M.W., M.W., S.Z.-A., B.L.M., A.Y.H., A.K., Sa.P., E.M., R.V., D.H.C., and A.R.-C. participated in writing and editing the manuscript. All authors approved the final version of the manuscript.

AFFILIATIONS

¹Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan, USA

²Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA

³Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH), Universidad Nacional Autónoma de México, Juriquilla, Querétaro, Mexico

⁴Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA

⁵Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA

⁶Kellogg Biological Station, Michigan State University, East Lansing, Michigan, USA

⁷Department of Horticulture, Michigan State University, East Lansing, Michigan, USA

⁸Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

⁹Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

¹⁰Unidad de Investigación Médica en Inmunología e Infectología, Instituto Mexicano del Seguro Social, Ciudad de México, Mexico

¹¹Programa de Posgrado en Ciencias Biológicas, Facultad de Medicina, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

¹²Laboratory of Agrigenomic Sciences, Escuela Nacional de Estudios Superiores Unidad León, Universidad Nacional Autónoma de México, León, Guanajuato, Mexico

¹³Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

¹⁴Laboratorio de Ecología Evolutiva y Conservación de Anfibios y Reptiles, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

¹⁵Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, Michigan, USA

¹⁶Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

¹⁷Molecular Plant Sciences Program, Michigan State University, East Lansing, Michigan, USA

¹⁸Genetics and Genome Sciences, Michigan State University, East Lansing, Michigan, USA

¹⁹Computational Population Genetics Group, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

²⁰Colección Nacional de Aves, Posgrado en Ciencias Biológicas, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

²¹Departamento de Botánica, Posgrado en Ciencias Biológicas, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

²²Cell and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

²³Plant Breeding, Genetics, and Biotechnology, Michigan State University, East Lansing, Michigan, USA

²⁴Crop and Soil Sciences Program, Michigan State University, East Lansing, Michigan, USA

²⁵MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, Michigan, USA

²⁶Programa de Posgrado en Ciencias Biológicas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

²⁷Programa de Posgrado en Ciencias Biológicas, Escuela Nacional de Estudios Superiores (ENES), Unidad Morelia, Universidad Nacional Autónoma de México, Morelia, Michoacán, Mexico

²⁸Department of Biology, Grinnell College, Grinnell, Iowa, USA

²⁹Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

³⁰Department of Biomedical Informatics, Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

³¹Department of Mathematics, Michigan State University, East Lansing, Michigan, USA

³²Plant Resilience Institute, Michigan State University, East Lansing, Michigan, USA

³³Plantec National Laboratory, ENES-León, León, Guanajuato, Mexico

ACKNOWLEDGMENTS

This work was funded primarily by a National Science Foundation Research Traineeship (NSF-NRT) grant (NSF 1828149), which established the Integrated training Model in Plant And Compu-Tational Sciences (IMPACTS) program at Michigan State University. This grant funded fellows within this program as well as the project-based curriculum for the Plants and Python Course that formed the backbone of this paper. This work is also supported by the NSF Plant Genome Research Program (awards IOS-2310355, IOS-2310356, and IOS-2310357) and the NSF Plant, Fungal and Microbial Developmental Mechanisms award (IOS-2039489). This project was supported by the United States Department of Agriculture (USDA) National

Institute of Food and Agriculture and by Michigan State University AgBioResearch.

DATA AVAILABILITY STATEMENT

The code to reproduce the results in this manuscript is available at <https://github.com/PlantsAndPython/arabidopsis-gene-expression>; the data are available at <https://datadryad.org/stash/dataset/doi:10.5061/dryad.4b8gthtn7> (Chitwood and Palande, 2024).

ORCID

Daniel H. Chitwood  <http://orcid.org/0000-0003-4875-1447>

REFERENCES

- Angiosperm Phylogeny Group, M. W. Chase, M. J. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, et al. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181(1): 1–20.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814): 796–815.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics* 25(1): 25–29.
- Azodi, C. B., J. Pardo, R. VanBuren, G. de Los Campos, and S. H. Shiu. 2020. Transcriptome-based prediction of complex traits in maize. *The Plant Cell* 32(1): 139–151.
- Bennett, M. D., I. J. Leitch, H. J. Price, and J. S. Johnston. 2003. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* genome initiative estimate of approximately 125 Mb. *Annals of Botany* 91(5): 547–557.
- Bergstra, J., D. Yamins, and D. D. Cox. 2013. Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*, Vol. 13.
- Bevan, M., and S. Walsh. 2005. The *Arabidopsis* genome: A foundation for plant research. *Genome Research* 15(12): 1632–1642.
- Chitwood, D., and S. Palande. 2024. Data from: Expression-based machine learning models for predicting plant tissue identity. Dryad Dataset. <https://datadryad.org/stash/dataset/doi:10.5061/dryad.4b8gthtn7> [accessed September 2024].
- Clough, S. J., and A. F. Bent. 1998. Floral dip: A simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal* 16(6): 735–743.
- Coppens, F., N. Wuyts, D. Inzé, and S. Dhondt. 2017. Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology* 4: 58–63.
- Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20: 273–297.
- Cover, T., and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1): 21–27.
- Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Cerón-Rojas, X. Zhang, et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112(1): 48–60.
- Dwyer, W., C. N. Ibe, and S. Y. Rhee. 2022. Renaming Indigenous crops and addressing colonial bias in scientific language. *Trends in Plant Science* 27: 1189–1192.
- Fukushima, K., and D. D. Pollock. 2020. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nature Communications* 11(1): e4459.
- Garnier, S., N. Ross, R. Rudis, P. A. Camargo, M. Sciaini, and C. Scherer. 2024. viridis(Lite): Colorblind-friendly color maps for R, version 0.6.5. Website: <https://sjmgarnier.github.io/viridis/> [accessed 17 September 2024].
- Haykin, S. 1998. *Neural networks: A comprehensive foundation*. Prentice Hall, Hoboken, New Jersey, USA.
- Ho, T. K. 1995. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, 14–16 August 1995, 278–282. IEEE, New York, New York, USA.
- Hogeweg, P. 2011. The roots of bioinformatics in theoretical biology. *PLoS Computational Biology* 7(3): e1002021.
- Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 4(1): 44–57.
- Ij, H. 2018. Statistics versus machine learning. *Nature Methods* 15(4): e233.
- Julca, I., C. Ferrari, M. Flores-Tornero, S. Proost, A. C. Lindner, D. Hackenberg, L. Steinbachová, et al. 2021. Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. *Nature Plants* 7(8): 1143–1159.
- Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873): 583–589.
- Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1): 27–30.
- Knapp, S., L. Bohs, M. Nee, and D. M. Spooner. 2004. Solanaceae—A model for linking genomics with biodiversity. *Comparative and Functional Genomics* 5(3): 285–291.
- Kopriva, S. 2021. Glucosinolates revisited—A follow-up of ABR volume 80: Glucosinolates, 249–274. In J. P. Jacquot [ed.], *Advances in Botanical Research*, Vol. 100. Academic Press, London, United Kingdom.
- Li, F. W., and A. Harkess. 2018. A guide to sequence your favorite plant genomes. *Applications in Plant Sciences* 6(3): e1030.
- Lim, P. K., X. Zheng, J. C. Goh, and M. Mutwil. 2022. Exploiting plant transcriptomic databases: Resources, tools, and approaches. *Plant Communications* 3(4): 100323.
- Marks, R. A., S. Hotaling, P. B. Frandsen, and R. VanBuren. 2021. Representation and participation across 20 years of plant genome sequencing. *Nature Plants* 7(12): 1571–1578.
- Marks, R. A., E. J. Amézquita, S. Percival, A. Rougon-Cardoso, C. Chibici-Revneanu, S. M. Tebele, J. M. Farrant, et al. 2023. A critical analysis of plant science literature reveals ongoing inequities. *Proceedings of the National Academy of Sciences, USA* 120(10): e2217564120.
- Mason, L., J. Baxter, P. Bartlett, and M. Fread. 1999. *Boosting algorithms as gradient descent*. In *Advances in Neural Information Processing Systems*, vol. 12. MIT Press, Cambridge, Massachusetts, USA.
- Mazzocchi, F. 2015. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports* 16(10): 1250–1255.
- Meyerowitz, E. M. 1987. *Arabidopsis thaliana*. *Annual Review of Genetics* 21(1): 93–111.
- Meyerowitz, E. M. 2001. Prehistory and history of *Arabidopsis* research. *Plant Physiology* 125(1): 15–19.
- Michael, T. P., and S. Jackson. 2013. The first 50 plant genomes. *The Plant Genome* 6(2). <https://doi.org/10.3835/plantgenome2013.03.0001in>
- Mitchell, C. E., A. A. Agrawal, J. D. Bever, G. S. Gilbert, R. A. Huffbauer, J. N. Klironomos, J. L. Maron, et al. 2006. Biotic interactions and plant invasions. *Ecology Letters* 9(6): 726–740.
- Palande, S., J. A. Kaste, M. D. Roberts, K. S. Aba, C. Claucherty, J. Dacon, R. Doko, et al. 2023. The topological shape of gene expression across the evolution of flowering plants. *PLoS Biology* 21(12): e3002397.
- Parry, G., N. J. Provart, S. M. Brady, B. Uzilday, Multinational *Arabidopsis* Steering Committee. 2020. Current status of the multinational *Arabidopsis* community. *Plant Direct* 4(7): e00248.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.
- Proost, S., and M. Mutwil. 2018. CoNekT: An open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic Acids Research* 46(W1): W133–W140.
- Provart, N. J., J. Alonso, S. M. Assmann, D. Bergmann, S. M. Brady, J. Brkljacic, J. Browse, et al. 2016. 50 years of *Arabidopsis* research: Highlights and future directions. *New Phytologist* 209(3): 921–944.

- Reiser, L., E. Bakker, S. Subramaniam, X. Chen, S. Sawant, K. Khosa, T. Prithvi, and T. Z. Berardini. 2024. The Arabidopsis Information Resource in 2024. *Genetics* 227: iyae027.
- Strable, J., and M. J. Scanlon. 2009. Maize (*Zea mays*): A model organism for basic and applied research in plant biology. *Cold Spring Harbor Protocols* 10: pdb-emo132.
- Sultan, S. E. 2000. Phenotypic plasticity for plant development, function and life history. *Trends in Plant Science* 5(12): 537–542.
- Yu, Y., H. Zhang, Y. Long, Y. Shu, and J. Zhai. 2022. Plant public RNA-seq database: A comprehensive online database for expression analysis of ~45 000 plant public RNA-seq libraries. *Plant Biotechnology Journal* 20(5): 806–808.
- Zhang, N., M. Wang, and N. Wang. 2002. Precision agriculture—A worldwide overview. *Computers and Electronics in Agriculture* 36(2–3): 113–132.
- Zhang, H., F. Zhang, Y. Yu, L. I. Feng, J. Jia, B. O. Liu, B. Li, et al. 2020. A comprehensive online database for exploring ~20,000 public Arabidopsis RNA-seq libraries. *Molecular Plant* 13(9): 1231–1233.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Enriched gene ontology (GO) terms associated with principal component (PC) loading values.

How to cite this article: Palande, S., J. Arsenaault, P. Basurto-Lozada, A. Bleich, B. N. I. Brown, S. F. Buysse, N. A. Connors, et al. 2024. Expression-based machine learning models for predicting plant tissue identity. *Applications in Plant Sciences* 12: e11621. <https://doi.org/10.1002/aps3.11621>