Received: 20 December 2023 | Accepted: 4 April 2024

DOI: 10.1002/ppj2.20106

The Plant Phenome Journal OPEN ACCESS

ORIGINAL ARTICLE

Data driven discovery and quantification of hyperspectral leaf reflectance phenotypes across a maize diversity panel

Michael C. Tross[1,2,3] | **Marcin W. Grzybowski**[1,2,3,4] | **Talukder Z. Jubery**[5,6] | **Ryleigh J. Grove**[1,2,3] | **Aime V. Nishimwe**[1,2,3] | **J. Vladimir Torres-Rodriguez**[1,2,3] | **Guangchao Sun**[1,2,3,8] | **Baskar Ganapathysubramanian**[5,6] | **Yufeng Ge**[2,7] | **James C. Schnable**[1,2,3]

[1]Quantitative Life Sciences Initiative, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[2]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[3]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[4]Department of Plant Molecular Ecophysiology, Institute of Plant Experimental Biology and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland

[5]Department of Mechanical Engineering, Iowa State University, Ames, Iowa, USA

[6]Translational AI Research and Education Center, Iowa State University, Ames, Iowa, USA

[7]Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[8]Advanced Diagnostic Laboratory, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA

Correspondence
James C. Schnable, Quantitative Life Sciences Initiative, University of Nebraska-Lincoln, Lincoln, NE, USA.
Email: schnable@unl.edu

Assigned to Associate Editor Keshav Singh.

Funding information
United States Department of Agriculture—National Institute of Food and Agriculture, Grant/Award Numbers: 2020-68013-32371, 2021-67021-35329; National Science Foundation, Grant/Award Number: OIA-1826781; U.S. Department of Energy, Grant/Award Number: DE-SC0020355; Foundation for Food and Agriculture Research, Grant/Award Number: 602757

Abstract

Estimates of plant traits derived from hyperspectral reflectance data have the potential to efficiently substitute for traits, which are time or labor intensive to manually score. Typical workflows for estimating plant traits from hyperspectral reflectance data employ supervised classification models that can require substantial ground truth datasets for training. We explore the potential of an unsupervised approach, autoencoders, to extract meaningful traits from plant hyperspectral reflectance data using measurements of the reflectance of 2151 individual wavelengths of light from the leaves of maize (*Zea mays*) plants harvested from 1658 field plots in a replicated field trial. A subset of autoencoder-derived variables exhibited significant repeatability, indicating that a substantial proportion of the total variance in these variables was explained by difference between maize genotypes, while other autoencoder variables appear to capture variation resulting from changes in leaf reflectance between different batches of data collection. Several of the repeatable latent variables were significantly correlated with other traits scored from the same maize field experiment, including one autoencoder-derived latent variable (LV8) that predicted plant chlorophyll content modestly better than a supervised model trained on the same data. In at least one case, genome-wide association study hits for variation in autoencoder-derived variables were proximal to genes with known or plausible links to leaf phenotypes expected to alter hyperspectral reflectance. In aggregate, these results suggest that an unsupervised, autoencoder-based approach can identify meaningful and genetically controlled variation in high-dimensional, high-throughput phenotyping data and link identified variables back to known plant traits of interest.

1 | INTRODUCTION

Mendel's laws of genetics (law of segregation; law of independent assortment; law of dominance) were discovered through the analysis of qualitative traits (Biffen, 1905; Weldon, 1902). The principles first discovered via study of whether peas were yellow or green, among other traits, formed the foundation for the modern field of genetics, which, in turn, lead to dramatic advances in crop productivity and stress tolerance via genetics-informed breeding over the last century. However, while the initial traits studied by the founders of the field were qualitative in nature, many traits of interest to both scientists and stakeholders are quantitative traits that can vary over a continuous range (e.g., flowering time or yield), and quantitative genetic approaches have been developed to both identify the genes responsible for controlling variation in continuous traits and build models to predict trait values from genetic data. These quantitative genetic approaches use measurements of traits of interest across large populations, combined with genetic marker data from the same population, to identify individual genes or genomic segments associated with variation in target traits. Collecting trait data from large plant field experiments are labor intensive and frequently represent the most expensive portion of quantitative genetics experiments of plant breeding efforts (Tibbs Cortes et al., 2021). Advances in high-throughput phenotyping technologies have the potential to reduce the cost of identifying genomic loci controlling traits of interest by either decreasing the resources required to score each plant/plot or reducing the marginal cost of collecting additional traits in parallel from a single field experiment (Fahlgren et al., 2015). Automated phenotyping strategies require solving two separate problems: collection of sensor data (e.g., RGB images, light detection and ranging [LIDAR] point clouds, and hyperspectral reflectance patterns) from plants of interest and using the collected sensor data to generate quantitative or qualitative estimates of specific traits (Furbank & Tester, 2011; Yang et al., 2020).

Among many sensor modalities commonly used for high-throughput plant phenotyping, spectrometers that collect leaf-level hyperspectral data in the visible, near-infrared, and shortwave-infrared regions are increasingly used to estimate a wide range of plant chemical and physiological traits. Numerous studies have shown that VIS-NIR-SWIR can estimate leaf pigments, nitrogen content, water content, photosynthesis parameters, various metabolites, and nutrient contents (M. Grzybowski et al., 2021). Many of these spectrometers are portable, enabling data collection from field-grown plants, a clear advantage compared to many lab-only instruments. Raw measurements of many individual spectral reflectance values are typically processed to generate predicted values for different traits of interest. Widely employed processing techniques for estimating known plant traits from raw spectral intensity values include narrow-band spectral indices, including the chlorophyll index (Wu et al., 2008) and the anthocyanin index (Steele et al., 2009), partial least squares regression (PLSR) (Burnett et al., 2021), and more recently, approaches based on machine-learning and deep learning methodologies (Furbank et al., 2021). Both PLSR and the set of machine-learning methods described in Furbank et al. (2021) are classified as "supervised" approaches, indicating that the models are trained using a population of data points where both the spectral reflectance values and the true values for the trait of interest (labels) are already known.

Unsupervised methods can discern variation in sensor data gathered from various populations without the need to construct models targeting specific a priori traits or relying on labeled training data. This approach is particularly useful when dealing with datasets that exhibit high dimensionality, often with only a few degrees of variability. To boost predictive accuracy, implementing low-dimensional representations is advantageous, as these highlight the fundamental characteristics of the data while filtering out unnecessary details. Techniques such as principal component analysis (PCA) and neural networks (NNs) with auto-encoding are frequently utilized for this type of dimensionality reduction. While these methods might be abstract and sometimes challenging to interpret biologically, they are effective in uncovering hidden patterns in phenotypic and genetic variations (Ubbens et al., 2020). PCA focuses on linear transformations to extract latent features, whereas NNs use a mix of linear and nonlinear transformations. Studies by Wang et al. (2016) and Fournier and Aloise (2019) have empirically demonstrated the greater effectiveness of NNs over PCA in dimensionality reduction. In the realm of plant science, the efficiency of NNs has been confirmed by several researchers. Classifications and ordering of shape categories in strawberries using 2D images achieved heritabilities comparable to direct human measurement (Feldmann et al., 2020). Gage and coworkers demonstrated that quantitative latent phenotypes extracted from LIDAR point clouds in maize fields can exhibit heritabilities similar to hand-measured traits (Gage et al., 2019). Autoencoder NNs are an unsupervised approach that reduce high-dimensional data to a smaller set of latent variables that will, ideally, represent patterns of variation present in the original higher dimensional dataset (Baldi, 2012; Rumelhart et al., 1985; Wang et al., 2016). Autoencoders comprise two NNs, an encoder and a decoder. The encoder takes as input the values of all dimensions from a single sample in the dataset and reduces the dimensionality down to a configured amount of latent variables. The decoder then takes the latent variables as input and tries to reconstruct the sample to the original dimensionality. The reconstructed data from the decoder is then compared to the input of the encoder and the reconstruction loss is calculated. The loss is then backpropagated into the encoder and decoder networks to improve the parameters in the direction of better reconstructions. Through many iterations using numerous samples, the encoder is able to produce latent phenotypes that are representative of the original data. Here, we extract latent phenotypes from hyperspectral leaf reflectance sensor data in a maize (*Zea mays*) diversity panel. We demonstrate how these latent phenotypes can be annotated and used as a proxy for traits with limited to no ground truth data and that the model we trained to extract these traits from hyperspectral leaf reflectance data exhibits transferability between maize and a related crop species, sorghum (*Sorghum bicolor*).

Core Ideas

- Autoencoder latent variables show stronger correlations with chlorophyll content than principal components.
- Autoencoder-derived latent variable exhibits modestly better performance than partial least squares regression supervised model.
- Latent variables derived from autoencoders are significantly associated with genetic markers.
- Latent variables capture variance in traits that are transferrable across species and years.
- Significant proportions of total variance in individual latent variables are attributable to genetics.

2 | MATERIALS AND METHODS

2.1 | Field experiment and data collection

A previously described field experiment consisting of 1680 plots subdivided into two complete replicates of a population of 752 maize inbred genotypes comprising a subset of the Wisconsin Diversity panel (Mazaheri et al., 2019), and a single repeated check genotype was planted on May 6, 2020 at the Havelock Farm Research Facility at the University of Nebraska-Lincoln (40.852 N, 96.616 W) (Sun et al., 2022). Briefly, each plot consisted of two rows with approximately 20 plants per row. Rows were 7.5-ft long, with 30-in. row spacing and 2.5-ft alleyways between sequential plots. Published plant-level phenotypes for the same experiment were taken from Mural et al. (2022). To minimize variation introduced by differences in environmental conditions or developmental stage, we sought to collect hyperspectral reflectance from the entire field experiment in the shortest interval possible. Given constraints on labor and field access in the summer of 2020 as the result of the coronavirus pandemic and associated lockdown procedures, collecting data from the entire field required 9 days of work, spread over a 13-day interval from July 8 to July 20, 2020. Hyperspectral reflectance was collected from a single fully expanded leaf from a representative plant per plot avoiding edge plants when possible using a spectroradiometer (FieldSpec4; Malvern Panalytical Ltd., formerly Analytical Spectral Devices) with a contact probe, following a previously described protocol (Ge et al., 2019). Leaf spectral was collected as described previously (Wijewardane et al., 2023). Briefly, a single plot level value was generated for each of 2151 reflectance values. Each value represented the proportion of light reflected in a 1 nm increment of light wavelengths between 350 and 2500 nm. Data from 1665 plots were initially collected. Seven plots with abnormal spectra, estimated as leaf reflectance <0 or >1, were removed from the analysis, resulting in a final dataset of 1658 plot-level reflectance spectra.

Molecular leaf traits were collected from subsets of between 243 and 318 of the leaves employed above. Chlorophyll concentration (CHL), equivalent water thickness (EWT), leaf water content (LWC, %), and specific leaf area (SLA, m^2/kg) were collected from these leaves using the methods adopted by Ge et al. (2019) and Li et al. (2023). Briefly, CHL was measured using a handheld chlorophyll meter (MC-100; Apogee Instruments, Inc.); EWT was calculated using the following formula: (fresh weight of leaves - dry weight of leaves)/leaf area; LWC was calculated using the following formula: (dry weight of leaves/fresh weight of leaves) × 100%; and SLA was calculated using the following formula: leaf area/dry weight of leaves. Phosphorus (P), nitrogen (N), potassium (K), magnesium (Mg), calcium (Ca), sulfur (S), iron (Fe), manganese (Mn), boron (B), copper (Cu), and zinc (Zn) were quantified from dried leaf samples by a commercial provider (Midwest Laboratories, Inc.).

The sorghum leaf hyperspectral and molecular trait dataset used to assess model transferability was collected from two experiments, one in the field under two nitrogen treatments and one under greenhouse conditions. The field experiment was conducted at the University of Nebraska-Lincoln's Havelock Farm facility (N 40.861, W 96.598) in 2020, where one row plots of sorghum genotypes from the sorghum association panel were grown with either 0 pounds per acre (low nitrogen) or 80 pounds per acre (high nitrogen) of supplemental nitrogen (M. W. Grzybowski et al., 2022). Hyperspectral reflectance was collected from the second leaf, counting downward from the last fully extended leaf, of a single plant per plot. Ground truth measurements of molecular leaf traits were collected for 266 samples (130 from the high nitrogen treatment and 136 from the low nitrogen treatments). The greenhouse experiment was conducted at the University of Nebraska-Lincoln's automated phenotyping facility at the Nebraska Innovation Campus. Data were collected from 321 plants, representing 236 unique sorghum genotypes, which were grown in a single common experiment in the greenhouse (Tross et al., 2021). Ground truth and hyperspectral measurements were collected as described above for maize.

2.2 | Dimensional reduction

The dimensionality of the 1658 plot-level hyperspectral reflectance values was reduced using both PCA and a trained autoencoder NN. For PCA, the values were reduced to 10 principal components (PCs) using the scikit-learn package (Pedregosa et al., 2011). These 10 components were sufficient to summarize 99% of variance in the dataset. An autoencoder architecture was implemented in Keras (v2.8.0) (Bank et al., 2020; Chollet, 2015). The empirically determined network architecture consisted of an encoder with five dense layers with 2151, 2200, 3000, 2024, and 10 neurons, respectively, and a decoder with five dense layers with 1024, 1536, 2500, 2500, and 2151 neurons (Figure S1). A scaled exponential linear unit activation function was employed for all dense layers, with the exception of the final dense layer of the decoder network, which employed a tanh activation function. Both the encoder and decoder were trained using a mean absolute error loss function, the standard gradient descent optimizer, and a learning rate of 0.1. The raw set of 1658 plot-level hyperspectral reflectance values was split 5:1 into training and validation data. Autoencoders were trained for up to 1000 epochs, or until 100 epochs passed without further improvement, whichever came first. The final autoencoder described in this manuscript was trained for 413 epochs before stopping based on a lack of further improvement.

To enable apples-to-apples comparisons of autoencoder latent variables to current state-of-the art methods for chlorophyll estimation, a PLSR model (Helland, 1990) was implemented using scikit-learn and trained using ground truth chlorophyll content measurements collected from 318 plots. The performance of the model was evaluated using fivefold cross-validation, a widely adopted approach for interpreting hyperspectral reflectance data (Chen et al., 2020, 2021; Juola et al., 2023; Manna et al., 2018; Shi et al., 2022). Comparisons to latent variable data were conducted after subsetting latent variables to only data from the same 318 plots to allow for a direct comparison between methods.

A random forest model was implemented in scikit-learn where separate instances of the model were trained to predict the value for each latent variable generated by the autoencoder given data on the values of 30 conventionally measured plant phenotypes (Mural et al., 2022). After training, the importance of each conventionally measured trait in predicting the value for a given latent variable was determined using scikit-learn's built-in feature importance function for random forest models (Breiman, 2001; Ho, 1995). Here, "feature importance" refers to mean decrease in impurity, a metric for how much each feature contributes to organizing the data into more homogeneous or pure groups at each split in the decision tree nodes (Louppe et al., 2013).

2.3 | Quantitative genetic analyses

The repeatability of plant phenotypes in this study was calculated using the equation $\rho = \sigma^2{}_G/(\sigma^2{}_G + \sigma^2{}_e/2)$, where $\sigma^2{}_G$ is the total amount of variance explained by genetics and $\sigma^2{}_e$ is the total amount of residual variance. Variance components were derived by fitting a linear model with the formula $yi = \mu + ti + ei$, where yi is the mean value of the genotype, μ is the overall mean, ti is the effect of genotype i, and ei is the residual error of genotype i. Linear models were fit to each dataset using software package lme4 (v1.1-23) (Bates et al., 2015) implemented in the R programming language (V4.0.4) (R Core Team, 2020).

Genome-wide association studies were conducted using the mixed linear model approach implemented within the GEMMA software package (v0.98.1) (Zhou & Stephens, 2012) with a set of 16.6 million segregating genetic markers, a subset from the set of genetic markers published in M. W. Grzybowski et al. (2023), and filtered to include only those with a minor allele frequency ≥0.05 and a proportion of heterozygous calls ≤0.05. A total of three PCs of variation in the genetic marker data were calculated using PLINK (v1.90b4) software package (Purcell et al., 2007) and incorporated as covariates into the model employed for genome-wide associations. The threshold for statistical significance in this study, 2.20×10^{-8}, was determined by applying the Bonferroni correction to the estimated 2,269,711 effective number of independent statistical tests represented by the

16.6 million markers employed in this study. The effective number of independent markers was calculated by pruning those initial 16.6 million markers using a sliding window of 500 bp, step size of 100, and removing all single-nucleotide polymorphisms (SNPs) above a linkage disequilibrium of 0.2 using PLINK. For the spatially corrected analysis of latent variable 3, best linear unbiased predictors (Robinson, 1991) were calculated using spatial modeling R package SpATS (Velazco et al., 2017) with day of collection included as an additional fixed effect. Genome-wide association studies were then conducted as described above.

An expression quantitative trait loci (eQTL) mapping analysis as described in Torres-Rodriguez et al. (2023) was conducted using the rMVP (V1.0.6) (Yin et al., 2021) implementation of the mixed linear model. Briefly, gene expression for the Zm00001eb29707 gene model across genotypes was transformed using the Box-Cox method (Osborne, 2010), and the genetic marker dataset was processed as previously described. A kinship matrix generated using the VanRaden method (VanRaden, 2008) and three PCs of variation in the genetic marker dataset were used as covariates in the analysis.

3 | RESULTS

3.1 | Hyperspectral leaf reflectance values

The proportion of light reflected by the adaxial surfaces of maize leaves was measured for leaves harvested from a replicated field study including more than 700 genotypes. This reflectance varied between 0.2% and 52% for individual 1 nm wide bands of light between 350 and 2500 nm (3,566,358 observations, 2151 individual wavelengths × 1658 leaves). Pairwise correlations between the intensity with which different 1 nm wide wavelengths of light were reflected by different maize leaves in the dataset ranged from −0.1 to 0.99 (Spearman's rho) with blocks of wavelengths exhibiting high correlations (Figure 1A). The proportion of the variance in the reflectance of individual 1 nm wide wavelengths, which could be explained by differences between maize genotypes ranged from 0 to 0.45 (Figure 1B).

The substantial correlations observed among the reflectance values of many individual wavelengths of light suggested the potential to summarize leaf reflectance using a smaller number of variables. Variation in reflectance across all 2151 individual wavelengths was summarized using autoencoders trained to summarize individual leaf reflectance spectra between 1 and 20 latent variables, and then reconstruct the original 2151 variable data from the smaller number of variables. The architecture of these autoencoders differed only in the number of variables passed from the encoder to the decoder (see Methods). The resulting models were assessed in two ways: first, by the reconstruction loss observed on validation data not used to train the autoencoders, and second, by the correlation of the autoencoder variables with a set of 15 molecular leaf traits we quantified from maize plants in the same field experiment. The five lowest minimum reconstruction losses on the validation data of 20 trained models were 0.0122, 0.0132, 0.0128, 0.0129, and 0.0125 for models having 5, 7, 10, 11, and 17 latent variables, respectively (Figure S2). The maximum correlation of each latent variable with any of the molecular traits was determined, resulting in one maximum correlation value for each latent variable. The maximum value of these per-latent variables across all latent variables (the maximum of maximums) was employed as a metric for evaluating the relationship between modeled variables and observed plant properties. The five highest maximum correlation values were observed for models employing 3, 8, 9, 10, and 13 latent variables. The specific maximum correlation values produced by these models were 0.65, 0.66, 0.61, 0.66, and 0.63, respectively. Based on a combination of minimizing reconstruction loss and maximizing correlation with molecular traits measured from the same field experiment, the 10 latent variables model was selected for the analyses presented below.

The final trained encoder was used to summarize leaf reflectance data from each of 1658 plots as 10 total latent variables. The repeatability of four of these 10 latent variables exceeded that of any individually measured wavelength. The highest observed repeatability of a latent variable was 0.64, while the highest observed repeatability of an individual wavelength was 0.45 (Figure 2). A control was employed using PCA to summarize the same dataset to 10 PCs. Among the first 10 PCs, one had a repeatability of at least 0.5 (two less than the autoencoder approach) and a maximum repeatability of 0.59 (0.64 for the autoencoder model) (Figure S3). In several cases, latent variables and a few PCs with low repeatability appeared to represent variation between leaves analyzed on different days (Figures S4 and S5). The latent variable most correlated with chlorophyll content exhibited a correlation of ($R^2 = 0.59$) (Figure 2F), which was much greater than the highest correlation observed between any of the first 10 PCs and chlorophyll ($R^2 = 0.31$) (Figure S6). It matched, and in fact modestly exceeded, the predictive accuracy of supervised models (partial least squares) individually trained on different 80% subsets of the same 318 ground truth chlorophyll measurements used to evaluate both models ($R^2 = 0.58$) (Figure 2E).

The transferability of the autoencoder-derived variables with plant traits was assessed using two sorghum leaf reflectance datasets and associated ground truth data (M. W. Grzybowski et al., 2022; Wijewardane et al., 2023). The pre-trained encoder described above was used to summarize variation from hyperspectral leaf reflectance collected from 321 sorghum plants grown under greenhouse conditions in

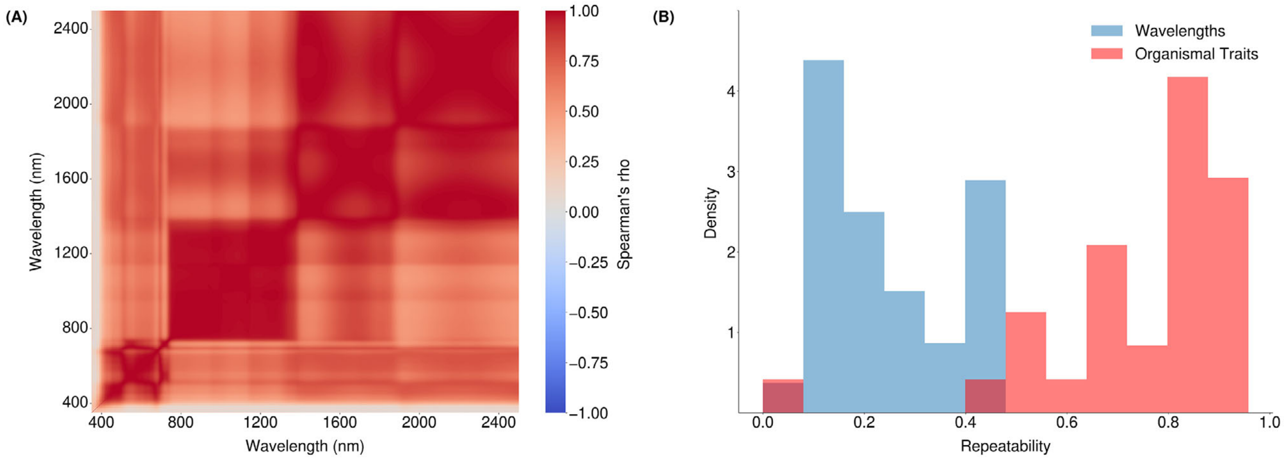

FIGURE 1 Correlations among the reflectance values of individual wavelengths and the proportion of variance attributable to differences between maize genotypes. (A) Spearman's rho values across 1658 maize leaves measured in a 2020 field experiment for all possible pairwise combinations of the proportion of light reflected for individual 1 nm wide wavelengths between 350 and 2500 nm. (B) Comparison of the distribution of repeatability values, defined here as the proportion of variance attributed to differences between replicates of the same genotype in a single environment, for the reflectance of 2151 individual 1 nm wide wavelengths and for a set of 30 hand-measured traits scored from maize plants in the same field experiment (Mural et al., 2022).

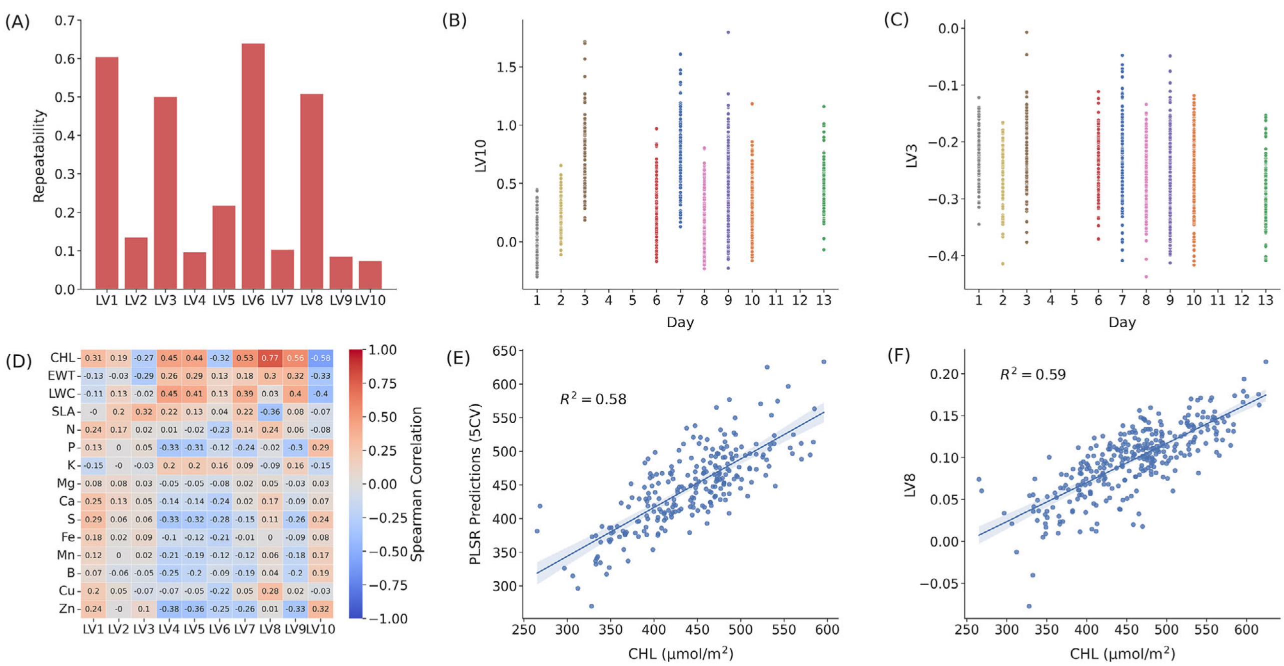

FIGURE 2 Association of latent variables with genetic, environmental, and plant phenotype factors. (A) Comparison of the repeatability of each of the 10 latent variables derived from hyperspectral leaf reflectance measured in this study. (B) Latent variable 10 value of the leaf reflectance compared with the day of collection of leaf reflectance for each plot. (C) Latent variable 3 value of the leaf reflectance compared with the day of collection of the leaf reflectance for each plot. (D) Associations measured by Spearman's rho between individual latent variables and ground truth measurements for 15 traits each scored in subsets of between 243 and 318 maize plots from which leaf reflectance data were also collected in 2020. (E) Association between observed chlorophyll content and fivefold cross-validation predictions from partial least squares regression model for all experimental plots. (F) Association between observed chlorophyll content and latent variable 8 of the leaf reflectance values for all experimental plots. CHL, chlorophyll concentration.

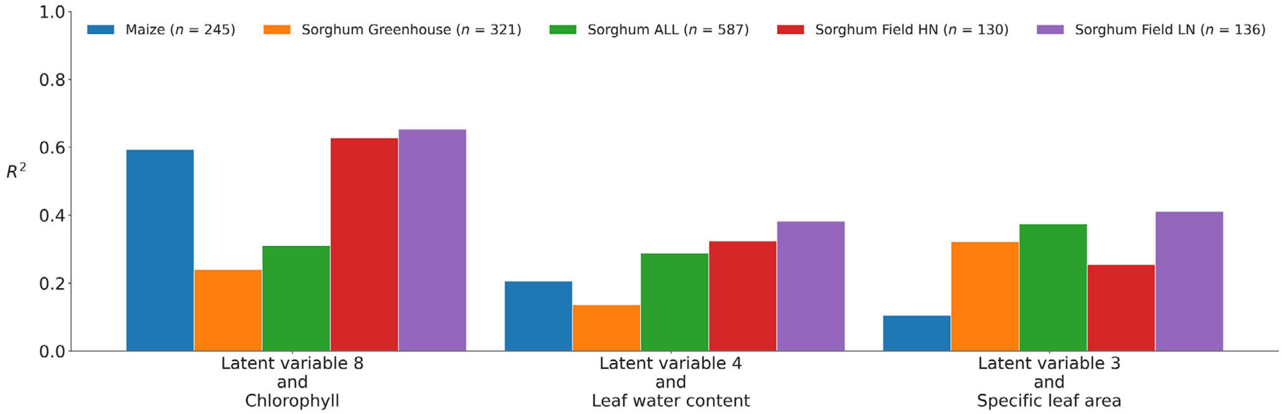

FIGURE 3 Associations between latent variables of leaf hyperspectral reflectance of sorghum populations with the same traits across species and years. All latent variables are derived from the same autoencoder model trained on hypersepctral reflectance data from a maize species. Latent variables derived from maize field reflectance data (2020), sorghum leaf reflectance field data (2020) (Grzybowski et al., 2022) and greenhouse data (2018) (Wijewardane et al., 2023), predicted from a model trained on leaf reflectance in maize (2020) were all associated with the same molecular traits. ALL, all the samples including sorgham greenhouse, sorgum low nitrogen (LN), and sorghum high nitrogen (HN).

2018, 130 sorghum plants grown in the field in 2020 under sufficient nitrogen conditions, and 136 sorghum plants grown in the field under nitrogen-deficient conditions in 2020. The same latent variable continued to exhibit correlations with ground truth chlorophyll measurements in greenhouse grown sorghum ($R^2 = 0.24$), field-grown sorghum with optimal nitrogen conditions ($R^2 = 0.63$), and field-grown sorghum under nitrogen-deficient conditions ($R^2 = 0.65$) (Figure 3).

3.2 | Latent variables capture information on variation in organismal traits

Random forest models were trained to predict latent variables from a suite of 30 traits measured from maize plants in the same field experiment. This was done to determine if latent variables were associated with variation in other traits of interest in the same maize population. Feature importance (mean decrease in gini impurity) was assessed for each organismal trait in random forest models trained to predict each latent variable. This produced an assessment of which plant traits contained significant information about the value of each latent variable. Leaf width exhibited the highest mean decrease in impurity across five folds (0.12) for latent variable 3, with the second largest decrease exhibited by the trait number of branches per tassel (0.08) (Figure 4). The severity of southern rust lesions, a leaf pathogen observed later in the growing season in the same field, emerged as the most influential organismal trait in predicting latent variable 8 (Figure S7). On the other hand, flowering time (measured as the number of days to pollen and the number of days to silking) played a more significant role in predicting the two latent variables with the highest repeatabilities (Figure 2A) (latent variables 1 and 6) (Figure S7).

3.3 | Linking latent variables to causal genes via genome-wide association

Many latent variables were not strongly associated with molecular or organismal traits. Genome-wide association studies were conducted on all latent variables to better understand what types of mechanisms might underlie the variation captured by each variable. Most latent variables exhibited statistically significant association with at least one genetic marker in the maize genome (Figure S8). Latent variable 3 had a significant marker that was located 15,811 base pairs downstream from Zm00001eb134990. The Arabidopsis ortholog of this gene, CYCD5;1 (AT4G37630), is believed to play a role in controlling endoreduplication during leaf development, a process associated with trichomes and other specialized protruding cells from the leaf surface (Sterken et al., 2012). A genetic marker 7713 base pairs upstream of Zm00001eb297070 was significantly associated with variation in latent variable 5 (Figure 5G). Notably, this signal was also only 988 base pairs away from the peak SNP of an eQTL-associated variation in the expression of that same gene (Zm00001eb297070) (Torres-Rodriguez et al., 2023) in mature leaf tissue (Figure 5H). A significant hit for latent variable 6 was located within the annotated gene model for Zm00001eb434330 (Figure 5F), a gene expressed primarily in developing leaves in maize (Hoopes et al., 2019; Stelpflug et al., 2016).

4 | DISCUSSION

The cost and throughput of collecting accurate measurements of plant traits across large field experiments is increasingly the rate-limiting step in both plant quantitative genetics research

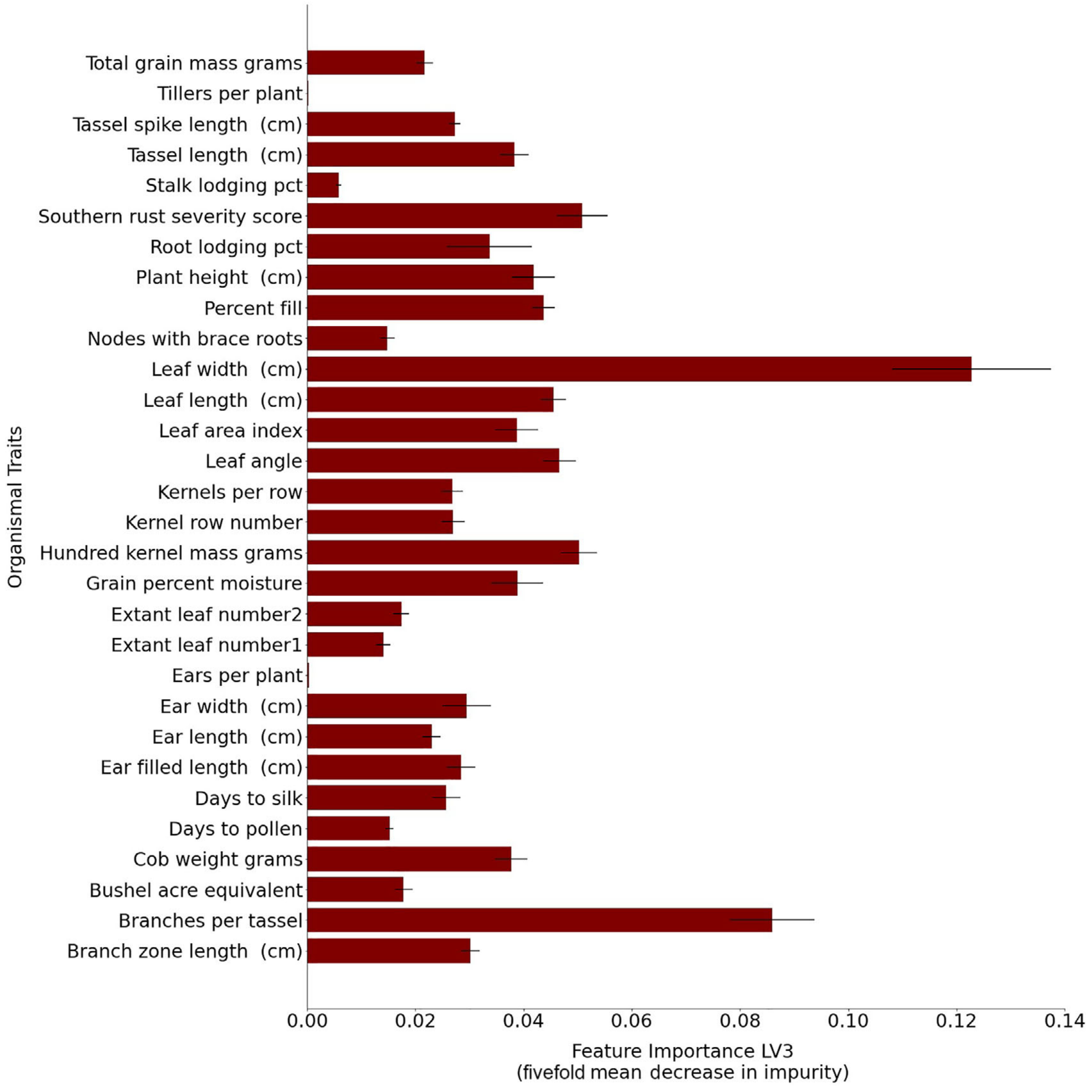

FIGURE 4 The relative importance of 30 hand-measured traits in predicting latent variable 3 of the hyperspectral leaf reflectance values. This importance is derived from the mean decrease in impurity of each node in the decision trees within a trained random forest model, attributed to each feature (trait). *X*-axis indicates the fivefold mean decrease in impurity calculated for each trait. Error bars represent standard error across the five folds for each trait. *Y*-axis indicates the 30 hand-measured traits traits used to predict the latent variable 3.

and plant breeding. Approaches that substitute sensor data and prediction models for direct human measurements of traits have been adopted for some applications and show promise in others. However, common approaches to training models to predict traits for sensor data using supervised models require large and expensive datasets to train, making them inaccessible to many researchers working on specialty crops, genetic models, or previously poorly studied traits. Here, we aimed to quantify traits using a high-dimensional hyperspectral leaf reflectance dataset combined with data-driven approaches, which have the potential to mitigate some of the logistical challenges of supervised training models.

A greater proportion of latent variables produced by autoencoder-based summaries of leaf reflectance data exhibited higher repeatabilities than PCs calculated from the same leaf reflectance dataset (Figure S3). However, repeatable traits can still be of limited utility for plant breeding and genetics if those traits are not linked to known plant properties. One autoencoder-derived latent variable captured variation in chlorophyll content with an accuracy that matched or modestly exceeded that of a supervised model trained with labeled data, while none of the top 10 PCs approached the performance of the supervised model (Figure 2 and Figure S6). For many large quantitative genetics studies, variation in phenotypic measurements due to variation in the time during which different plants or plots are scored creates additional non-genetic variance and reduces power to either identify causal genes or build trait prediction models. One potential avenue to further improve performance not investigated in this paper would be to adopt widely used methods to

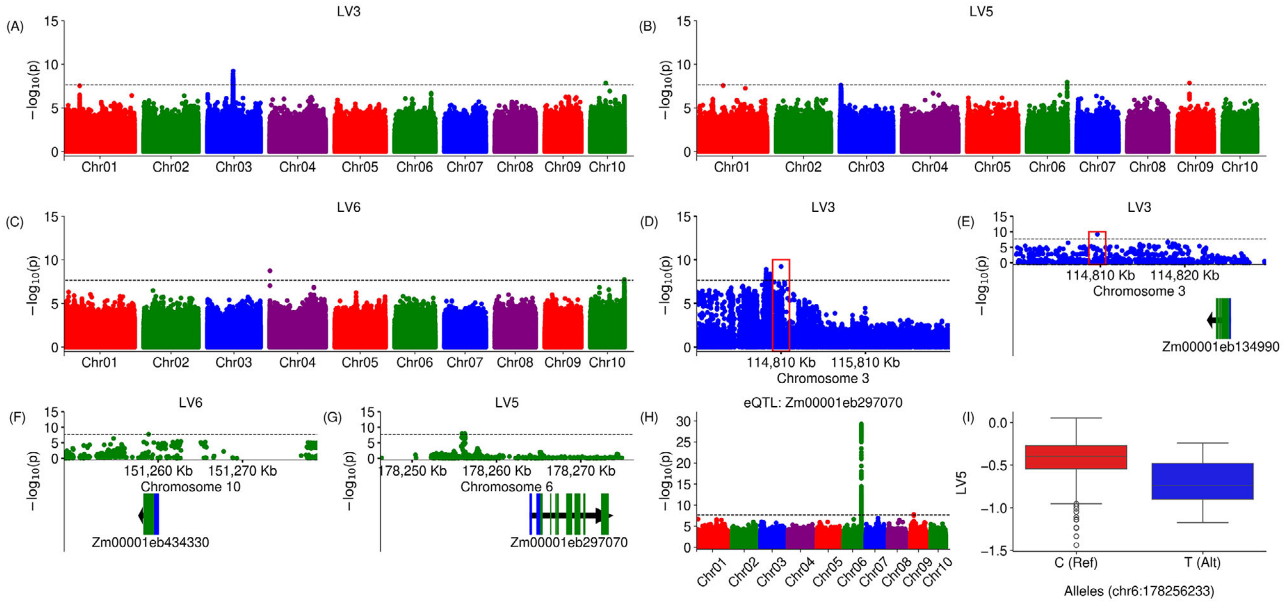

FIGURE 5 Genetic markers are significantly associated with latent variables 3, 5, and 6 of the autoencoder trained on reflectance data. Each point indicates the statistical significance of a marker (y axis) and its exact position on the genome (x axis). The dashed black lines indicate the statistical threshold cut off of 2.20×10^{-8}, which was derived from an α value of 0.05 with a Bonferonni adjustment for 2,269,711 effective genetic markers. Annotated black arrows indicate the position of the nearest gene/gene model. Blue rectangles indicate the untranslated regions while green rectangles indicate the coding sequence of the gene models. Genome-wide association studies of (A) latent variable 3; (B) latent variable 5; (C) latent variable 6; (D) zoomed in region of the significant markers on chromosome 3 of latent variable 3 starting at 113,809 – 116,809 kb. Red bounding box indicates the region of the most significant single-nucleotide polymorphism (SNP); (E) zoomed in region of the significant markers on chromosome 3 of latent variable 3 and the nearest gene/gene model starting at 114,799 – 114,829 kb. Red bounding box indicates the region of the most significant SNP; (F) zoomed in region of the of the significant markers on chromosome 10 of latent variable 6 and the nearest gene/gene model; (G) zoomed in region of the significant markers on chromosome 6 for latent variable 5 and the nearest gene/gene model; (H) eQTL analysis of the Zm00001eb297070 gene model; and (I) comparisons of the distributions of latent variable 5 for genotypes that are homozygous for the "C" reference allele versus the "T" alternate allele.

control for spatial variation across field experiments as a pre-processing step prior to feeding data to an autoencoder. In principle, by reducing the impact of non-genetic effects, this approach might result in latent variables that better represent variation in traits attributable to genetics. However, current approaches to correcting for spatial variation typically operate on individual traits. Given the amount of information captured by the relationships between the reflectance intensities of individual wavelengths within hyperspectral spectra, an optional approach to correcting for spatial variation would be spatially correct entire vectors of reflectance intensities jointly, rather than treating each individual wavelength separately. Until such approaches become feasible, the apparent partitioning of genetic and non-genetic sources of variance into separate latent variables that we observe here is encouraging (Figure 2B,C and Figure S4), and we observed largely comparable results when conducting genome-wide association study (GWAS) either using a latent variable directly or applying spatial correction to reported latent variable values prior to GWAS (Figure S9). The variance partition-like behavior of the autoencoder model employed in this study may explain, at least in part, the greater correlation of some latent variables with other, ground truth, plant phenotypes.

Beyond chlorophyll, a trait that can already be predicted with high accuracy with a number of linear models trained on large datasets, the strongest correlation of any of the latent variables calculated in this study with a panel of molecular traits was approximately $R^2 = 0.2$. Similarly, low correlations were observed with a panel of whole-plant phenotypes (Figure S10). However, linear regression may not capture non-linear relationships between traits or cases where a single latent variable reflects variation across multiple molecular or whole-plant traits. Feature importance values calculated from random forest models, which can capture both non-linear relationships and the influence of multiple traits on single latent variables, enabled the identification of whole-plant phenotypes, including leaf width, flowering time, and susceptibility to a specific foliar pathogen, associated with multiple individual latent variables (Figure S7). However, it must be noted that this approach was unsuccessful for a number of latent variables. Success depended on access to large datasets of conventionally scored traits from the same populations from which leaf reflectance data were collected, potentially reducing the logistical advantages of this approach relative to training supervised models. Another potential strategy for linking autoencoder-derived latent variables to known plant properties is via quantitative genetics (Ubbens et al., 2020). If a given latent variable is associated with multiple genes known to control a specific plant trait of interest, this would serve as significant evidence that the latent variable reflects variation in the same trait. We were successful in identifying one or more genomic intervals that were significantly associated with variation in seven out of 10 latent variables. In at least one case, a GWAS hit was associated with genes with plausible links to leaf-reflectance-related phenotypes. A genetic marker significantly associated with latent variable 3 (Figure 5A) was identified as 16 kb from a maize gene whose Arabidopsis ortholog is associated with leaf development and differential organ growth in different environments (Sterken et al., 2012). In another case, a genetic marker associated with latent variable 5 (Figures 5B,G) was also associated with cis-eQTL for a variable in the expression of an adjacent (approximately 1 kb distant) gene (Figure 5H). However, this approach was limited by both the number of GWAS hits identified per latent variable and the relatively modest number of maize genes linked with high confidence to roles in determining plant phenotypes. The former issue can potentially be addressed in the future by collecting hyperspectral leaf reflectance data from larger populations, experiments with higher levels of biological replication within a single environment, and/or across greater numbers of environments. Each of these would increase our power to identify significant associations in genome-wide association studies. The capacity of encoders trained on a single environment to continue to accurately reflect variation in the same plant phenotype across datasets collected in multiple environments and from multiple species (e.g., maize and sorghum) suggests that this approach may indeed be feasible (Figure 3). Underlying distributions of trait values can influence reported R^2 values even from identical models. An example of this is apparent in assessing the transferability of the maize autoencoder's ability to predict chlorophyll content to several different sorghum field experiments (Figure S11A). The two field studies exhibit a more dispersed distribution of ground truth chlorophyll values and thus higher R^2 with the autoencoder-derived latent variable, while the greenhouse experiment consisted largely of plants with similar chlorophyll values, so the same latent variable, produced using the same methodology, exhibits a lower R^2 (Figure S11B).

Employing autoencoders for dimensionality reduction requires a substantially greater amount of user time and input than PCA. Data conversion, network architecture design, hyperparameter tuning, and access to the necessary types and scale of computer resources are all barriers of entry relative to current widely used methods of dimensionality reduction in plant biology applications, including PCA and current widely used supervised classification models such as PLSR. In addition, while the collection of leaf hyperspectral reflectance data for large plant populations is less labor intensive than manual scoring of large panels of plant traits from the same population, the costs of the necessary equipment are high and the labor requirements are nontrivial. However, current rapid advances in robotics and imaging technologies have the potential to address the challenge of data collection. Improved artificial intelligence and machine-learning frameworks may address the first challenges of data conversion, network architecture design, hyperparameter tuning, and scaling of computer resources. If so, the approaches described here may provide significant utility in assisting plant geneticists and plant breeding in extracting the maximum amount of useful information from these new data types.

AUTHOR CONTRIBUTIONS

Michael C. Tross: Conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing—original draft; writing—review and editing. **Marcin W. Grzybowski**: Data curation; investigation; writing—review and editing. **Talukder Z. Jubery**: Investigation; methodology; software; supervision; writing—review and editing. **Ryleigh J. Grove**: Formal analysis; writing—review and editing. **Aime V. Nishimwe**: Data curation; formal analysis. **J Vladimir Torres-Rodriguez**: Formal analysis; writing—review and editing. **Guangchao Sun**: Conceptualization; data curation; methodology. **Baskar Ganapathysubramanian**: Conceptualization; investigation; methodology; resources; software; supervision; writing—review and editing. **Yufeng Ge**: Conceptualization; data curation; funding acquisition; investigation; methodology; project administration; writing—review and editing. **James C. Schnable**: Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

The authors thank Christine Smith and Mackenzie Zwiener for supervising and designing the 2020 maize and sorghum fields described in this manuscript. The authors thank Nuwan K. Wijewardane and Abbas Atefi for both expert guidance and help in the collection of the hyperspectral reflectance data described in this study. The authors thank Nikee Shrestha for assistance in curating datasets. The authors thank Addie Thompson and Linsey Newton for packaging and shipping the seeds during the 2020 coronavirus lockdown that enabled the maize field experiment described in this study. The authors thank Nathaniel Pester, Leighton Wheeler, Sierra

Conway, Isaac Stevens, and Luke Micek for assisting in field maintenance and data collection.

CONFLICT OF INTEREST STATEMENT

James C. Schnable has equity interests in Data2Bio, Dryland Genetics, and EnGeniousAg. He is a member of the scientific advisory board of GeneSeek and currently serves as a guest editor for *The Plant Cell*. The remaining authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The code used in this study is available at https://github.com/mtross2/autoencoder_hyperspec_ref. The hyperspectral dataset, molecular leaf traits, latent variables and weights for autoencoder model generated in this study is available at https://doi.org/10.6084/m9.figshare.24808.

ORCID

Michael C. Tross https://orcid.org/0000-0002-4410-9679
Baskar Ganapathysubramanian https://orcid.org/0000-0002-8931-4852
James C. Schnable https://orcid.org/0000-0001-6739-5527

REFERENCES

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *JMLR workshop and conference proceedings* (pp. 37–49). PMLR.

Bank, D., Koenigstein, N., & Giryes, R. (2020). *Autoencoders*. *arxiv*. https://arxiv.org/abs/2003.05991

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Biffen, R. H. (1905). Mendel's laws of inheritance and wheat breeding. *The Journal of Agricultural Science*, *1*, 4–48. https://doi.org/10.1017/S0021859600000137

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Burnett, A. C., Anderson, J., Davidson, K. J., Ely, K. S., Lamour, J., Li, Q., Morrison, B. D., Yang, D., Rogers, A., & Serbin, S. P. (2021). A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. *Journal of Experimental Botany*, *72*, 6175–6189. https://doi.org/10.1093/jxb/erab295

Chen, S., Hu, T., Luo, L., He, Q., Zhang, S., Li, M., Cui, X., & Li, H. (2020). Rapid estimation of leaf nitrogen content in apple-trees based on canopy hyperspectral reflectance using multivariate methods. *Infrared Physics & Technology*, *111*, 103542.

Chen, S., Hu, T., Luo, L., He, Q., Zhang, S., & Lu, J. (2021). Prediction of nitrogen, phosphorus, and potassium contents in apple tree leaves based on in-situ canopy hyperspectral reflectance using stacked ensemble extreme learning machine model. *Journal of Soil Science and Plant Nutrition*, *22*, 1–15.

Chollet, F. (2015). *Keras*. https://keras.io

Fahlgren, N., Gehan, M. A., & Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*, *24*, 93–99. https://doi.org/10.1016/j.pbi.2015.02.006

Feldmann, M. J., Hardigan, M. A., Famula, R. A., López, C. M., Tabb, A., Cole, G. S., & Knapp, S. J. (2020). Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. *GigaScience*, *9*, giaa030. https://doi.org/10.1093/gigascience/giaa030

Fournier, Q., & Aloise, D. (2019). Empirical comparison between autoencoders and traditional dimensionality reduction methods. In *IEEE second international conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 211–214). IEEE.

Furbank, R. T., Silva-Perez, V., Evans, J. R., Condon, A. G., Estavillo, G. M., He, W., Newman, S., Poiré, R., Hall, A., & He, Z. (2021). Wheat physiology predictor: Predicting physiological traits in wheat from hyperspectral reflectance measurements using deep learning. *Plant Methods*, *17*, Article 108. https://doi.org/10.1186/s13007-021-00806-6

Furbank, R. T., & Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*, 635–644. https://doi.org/10.1016/j.tplants.2011.09.005

Gage, J. L., Richards, E., Lepak, N., Kaczmar, N., Soman, C., Chowdhary, G., Gore, M. A., & Buckler, E. S. (2019). In-field whole-plant maize architecture characterized by subcanopy rovers and latent space phenotyping. *The Plant Phenome Journal*, *2*, 1–11. https://doi.org/10.2135/tppj2019.07.0011

Ge, Y., Atefi, A., Zhang, H., Miao, C., Ramamurthy, R. K., Sigmon, B., Yang, J., & Schnable, J. C. (2019). High-throughput analysis of leaf physiological and chemical traits with VIS–NIR–SWIR spectroscopy: A case study with a maize diversity panel. *Plant Methods*, *15*, Article 66. https://doi.org/10.1186/s13007-019-0450-8

Grzybowski, M., Wijewardane, N. K., Atefi, A., Ge, Y., & Schnable, J. C. (2021). Hyperspectral reflectance-based phenotyping for quantitative genetics in crops: Progress and challenges. *Plant Communications*, *2*, 100209. https://doi.org/10.1016/j.xplc.2021.100209

Grzybowski, M. W., Mural, R. V., Xu, G., Turkus, J., Yang, J., & Schnable, J. C. (2023). A common resequencing-based genetic marker data set for global maize diversity. *The Plant Journal*, *113*, 1109–1121. https://doi.org/10.1111/tpj.16123

Grzybowski, M. W., Zwiener, M., Jin, H., Wijewardane, N. K., Atefi, A., Naldrett, M. J., Alvarez, S., Ge, Y., & Schnable, J. C. (2022). Variation in morpho-physiological and metabolic responses to low nitrogen stress across the sorghum association panel. *BMC Plant Biology*, *22*, Article 433.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, *17*, 97–114.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). IEEE.

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N. J., & Buell, C. R. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *The Plant Journal*, *97*, 1154–1167.

Juola, J., Hovi, A., & Rautiainen, M. (2023). Classification of tree species based on hyperspectral reflectance images of stem bark. *European Journal of Remote Sensing*, *56*, 2161420. https://doi.org/10.1080/22797254.2022.2161420

Li, J., Wijewardane, N. K., Ge, Y., & Shi, Y. (2023). Improved chlorophyll and water content estimations at leaf level with a hybrid radiative transfer and machine learning model. *Computers and Electronics in Agriculture*, *206*, 107669. https://doi.org/10.1016/j.compag.2023.107669

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberge (Eds.), *Advances in neural information processing systems* (Vol. 26, pp. 1–9). Curran Associates, Inc.

Manna, B., Samanta, B., Chakravarty, D., Dutta, D., Chowdhury, A., Santra, A., & Banerjee, A. (2018). Hyperspectral signature analysis using neural network for grade estimation of copper ore. *IOP Conference Series: Earth and Environmental Science*, *169*, 012108.

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C. B., Kono, T. J. Y., Kaeppler, H. F., Spalding, E. P., Hirsch, C. N., Robin Buell, C., De Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*, Article 45. https://doi.org/10.1186/s12870-019-1653-x

Mural, R. V., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Andorf, C. M., Woodhouse, M. R., Thompson, A. M., Sigmon, B., & Schnable, J. C. (2022). Association mapping across a multitude of traits collected in diverse environments in maize. *GigaScience*, *11*, giaco080. https://doi.org/10.1093/gigascience/giac080

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research and Evaluation*, *15*(12), 1–9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucker, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

R Core Team. (2020). *R: A language and environment for statistical computing*. R Core Team.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*, 15–32.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (Technical Report No. ICS-8506). Institute for Cognitive Science, California University.

Shi, S., Xu, L., Gong, W., Chen, B., Chen, B., Qu, F., Tang, X., Sun, J., & Yang, J. (2022). A convolution neural network for forest leaf chlorophyll and carotenoid estimation using hyperspectral reflectance. *International Journal of Applied Earth Observation and Geoinformation*, *108*, 102719. https://doi.org/10.1016/j.jag.2022.102719

Steele, M. R., Gitelson, A. A., Rundquist, D. C., & Merzlyak, M. N. (2009). Nondestructive estimation of anthocyanin content in grapevine leaves. *American Journal of Enology and Viticulture*, *60*, 87–92. https://doi.org/10.5344/ajev.2009.60.1.87

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., De Leon, N., & Kaeppler, S. M. (2016). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *The Plant Genome*, *9*, plantgenome2015.04.0025. https://doi.org/10.3835/plantgenome2015.04.0025

Sterken, R., Kiekens, R., Boruc, J., Zhang, F., Vercauteren, A., Vercauteren, I., De Smet, L., Dhondt, S., Inzé, D., De Veylder, L., Russinova, E., & Vuylsteke, M. (2012). Combined linkage and association mapping reveals cycd5; 1 as a quantitative trait gene for endoreduplication in arabidopsis. *Proceedings of the National Academy of Sciences*, *109*, 4678–4683. https://doi.org/10.1073/pnas.1120811109

Sun, G., Mural, R. V., Turkus, J. D., & Schnable, J. C. (2022). Quantitative resistance loci to southern rust mapped in a temperate maize diversity panel. *Phytopathology*, *112*, 579–587. https://doi.org/10.1094/PHYTO-04-21-0160-R

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*, e20077. https://doi.org/10.1002/tpg2.20077

Torres-Rodriguez, J. V., Li, D., Turkus, J., Newton, L., Davis, J., Lopez-Corona, L., Ali, W., Sun, G., Mural, R. V., Grzybowski, M. W., Thompson, A. M., & Schnable, J. C. (2023). Population level gene expression can repeatedly link genes to functions in maize. *BioRxiv*. https://doi.org/10.1101/2023.10.31.565032

Tross, M. C., Gaillard, M., Zwiener, M., Miao, C., Grove, R. J., Li, B., Benes, B., & Schnable, J. C. (2021). 3D reconstruction identifies loci linked to variation in angle of individual sorghum leaves. *PeerJ*, *9*, e12628. https://doi.org/10.7717/peerj.12628

Ubbens, J., Cieslak, M., Prusinkiewicz, P., Parkin, I., Ebersbach, J., & Stavness, I. (2020). Latent space phenotyping: Automatic image-based phenotyping for treatment studies. *Plant Phenomics*, *2020*, 1–13. https://doi.org/10.34133/2020/5801869

Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Velazco, J. G., Rodríguez-Álvarez, M. X., Boer, M. P., Jordan, D. R., Eilers, P. H. C., Malosetti, M., & Van Eeuwijk, F. A. (2017). Modelling spatial trends in sorghum breeding field trials using a two-dimensional p-spline mixed model. *Theoretical and Applied Genetics*, *130*, 1375–1392. https://doi.org/10.1007/s00122-017-2894-4

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, *184*, 232–242. https://doi.org/10.1016/j.neucom.2015.08.104

Weldon, W. F. R. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika*, *1*, 228–233. https://doi.org/10.1093/biomet/1.2.228

Wijewardane, N. K., Zhang, H., Yang, J., Schnable, J. C., Schachtman, D. P., & Ge, Y. (2023). A leaf-level spectral library to support high-throughput plant phenotyping: Predictive accuracy and model transfer. *Journal of Experimental Botany*, *74*, 4050–4062. https://doi.org/10.1093/jxb/erad129

Wu, C., Niu, Z., Tang, Q., & Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, *148*, 1230–1241. https://doi.org/10.1016/j.agrformet.2008.03.005

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., Xiong, L., & Yan, J. (2020). Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. *Molecular Plant*, *13*, 187–214. https://doi.org/10.1016/j.molp.2020.01.008

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A memory-efficient,

The Plant Phenome Journal OPEN ACCESS

visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics*, *19*, 619–628.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*, 821–824. https://doi.org/10.1038/ng.2310

SUPPORTING INFORMATION

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

How to cite this article: Tross, M. C, Grzybowski, M. W, Jubery, T. Z, Grove, R. J, Nishimwe, A. V, Torres-Rodriguez, J. V., Sun, G., Ganapathysubramanian, B., Ge, Y., & Schnable, J. C. (2024). Data driven discovery and quantification of hyperspectral leaf reflectance phenotypes across a maize diversity panel. *The Plant Phenome Journal*, *7*, e20106. https://doi.org/10.1002/ppj2.20106