Food Chemistry 391 (2022) 133264

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

A novel high-throughput hyperspectral scanner and analytical methods for predicting maize kernel composition and physical traits

Jose I. Varela [a], Nathan D. Miller [b], Valentina Infante [a,c], Shawn M. Kaeppler [a,d], Natalia de Leon [a], Edgar P. Spalding [b,*]

[a] Department of Agronomy, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA
[b] Department of Botany, University of Wisconsin-Madison, 430 Lincoln Drive, Madison, WI 53706, USA
[c] Department of Bacteriology, University of Wisconsin-Madison, 1550 Linden Drive, Madison, WI 53706, USA
[d] Wisconsin Crop Innovation Center, University of Wisconsin – Madison, 8520 University Green, Middleton, WI 53562, USA

ARTICLE INFO

Keywords:
Hyperspectral imaging
Maize composition
Vitreousness
NIR
PLSR
PLS-DA

ABSTRACT

Large-scale investigations of maize kernel traits important to researchers, breeders, and processors require high throughput methods, which are presently lacking. To address this bottleneck, we developed a novel flatbed platform that automatically acquires and analyzes multiwavelength near-infrared (NIR hyperspectral) images of maize kernels precisely enough to support robust predictions of protein content, density, and endosperm vitreousness. The upward facing-camera design and the automated ability to analyze the embryo or abgerminal sides of each individual kernel in a sample with the appropriate side-specific model helped to produce a superior combination of throughput and prediction accuracy compared to other single-kernel platforms. Protein was predicted to within 0.85% (root mean square error of prediction), density to within 0.038 g/cm^3, and endosperm vitreousness percentage to within 6.3%. Kernel length and width were also accurately measured so that each kernel in a rapidly scanned sample was comprehensively characterized.

1. Introduction

The chemical composition and structural characteristics of a maize (*Zea mays*) kernel determine how well suited it is for its various industrial uses. For example, wet millers prefer kernels with softer endosperm because they require less steeping time and allow better starch-protein separation while dry millers desire hard endosperms (Wu & Bergquist, 1991). The differences between hard and soft endosperms are mostly due to differences in how densely the starch granules are embedded within a complex protein matrix (Gustin et al., 2013) and to differences in the physicochemical properties of the starch itself (Xu et al., 2019). Hard endosperms have high vitreousness, referring to their glass-like optical properties while soft endosperms generally scatter light more and therefore appear opaque. Supplemental Fig. 1 displays the visible differences between hard (vitreous) and soft (opaque) endosperms.

In the case of livestock feed, the bioavailability of starch in the endosperm is highly dependent on endosperm hardness (Dias Junior et al., 2016; Philippeau & Michalet-Doreau, 1997) because the protein matrix may affect how microorganisms in the rumen gain access to the starch granules (McAllister, Phillippe, Rode, & Cheng, 1994). Furthermore, endosperm vitreousness has been shown to affect resilience during harvest, storage, resistance to insects and fungi, and other practical characteristics (Holding & Larkins, 2006). Unfortunately, this important trait is difficult to measure directly.

Measuring vitreousness percentage typically requires manually removing the pericarp and embryo and then laboriously dissecting the floury soft endosperm from the vitreous hard portion to calculate a mass ratio. Alternative methods range from visually ranking light transmission of samples on a light box to quantifying resistance to grinding as a proxy for this trait (Gustafson & de Leon, 2010). Endosperm vitreousness is highly correlated with total kernel density (Correa, Shaver, Pereira, Lauer, & Kohn, 2002), which can be measured by determining the volume of gas or liquid a known mass of kernels can displace, or by determining the percentage of kernels that float on a salt solution having a known specific gravity (Bergquist & Thompson, 1992). Both methods are low throughput, which limits the feasibility of evaluating the number of samples needed for large-scale studies. Reliable, automated, non-invasive measurement of vitreousness percentage would enable larger-

* Corresponding author.
E-mail address: spalding@wisc.edu (E.P. Spalding).

https://doi.org/10.1016/j.foodchem.2022.133264
Received 6 December 2021; Received in revised form 16 May 2022; Accepted 17 May 2022
Available online 20 May 2022

Table 1
Composition statistics of maize kernels used for calibration and validation.

Cross-validation								External validation					
Trait	Set	Genotypes (n)	Kernels (n)	Mean	SD	CV	Range	Genotypes (n)	Kernels (n)	Mean	SD	CV	Range
Vitreousness (%)	1	99	949	66.97	12.09	0.18	35.08-95.8	50	479	68.36	9.48	0.14	41.33-89.87
Protein (%)	2	107	320	12.51	2.13	0.17	8.17-19.45	53	159	12.32	2.1	0.17	8.02-18.2
Density (g/cm^3)	3	104	302	1.18	0.06	0.05	1.0-1.35	52	146	1.19	0.05	0.04	1.07-1.34
Volume (cm^3)	3	104	302	0.21	0.05	0.23	0.07-0.39	52	146	0.22	0.04	0.18	0.13-0.33
Weight (g)	3	103	299	0.25	0.06	0.23	0.08-0.42	53	149	0.27	0.05	0.18	0.14-0.41

SD = standard deviation, CV = coefficient of variation.

scale investigations of endosperm quality.

Near-infrared (NIR) spectroscopy has been widely used to infer chemical composition and physical properties of maize kernels. The methodology relies on collecting transmitted or reflected light at wavelengths between 780 nm and 2500 nm and applying chemometric methods that exploit the inherent property of the C—H, O—H, N—H and S—H organic bonds to absorb NIR light through overtone vibrations (Siesler, Ozaki, Kawata, & Heise, 2008). NIR spectra from biological material have multiple overlapping absorbance patterns due to the complex mixture of organic compounds therefore multivariate statistical approaches are required to build the equations that can predict the trait given the spectrum (Hacisalihoglu et al., 2016; Spielbauer et al., 2009)). To study cereal grains, spectra are typically measured from fine ground powders or bulk whole grain samples (Orman & Schumann, 1991). NIR spectroscopy and properly calibrated equations can produce useful predictions of major seed constituents like starch, oil and proteins quickly and inexpensively (Fox & Manley, 2014). Ngonyamo-Majee et al. (2008) used NIR spectroscopy to develop vitreousness prediction equations for maize kernels by scanning powder from ground kernels. Although accurate predictions were attained, sample preparation is time-consuming and destructive.

Commercially available NIR analyzers acquire spectra from bulk grain samples. These units capture a spectrum representing an average of many intact kernels in unknown orientations within the sample. The spectra contain features that can predict kernel composition traits. At least one study of intact kernels identified two spectral features that correlated well with endosperm hardness (Robutti, 1995). Alternatively, custom-built NIR reflectance spectrometers can capture a spectrum as a seed tumbles down a tube, displaying different positions to the sensor (Hacisalihoglu et al., 2016; Spielbauer et al., 2009). These units cannot produce information about how composition varies between grain tissues such as the starchy endosperm and oil-rich embryo, which is relatively large in maize. Studies have shown that the germinal or abgerminal side of the kernel reflect different NIR spectra (Orman & Schumann, 1992; Weinstock, Janni, Hagen, & Wright, 2006).

Hyperspectral imaging combines NIR spectroscopy with pixel-based imaging (Feng et al., 2019). Hyperspectral imaging has been used to study maize and wheat kernels (Caporaso, Whitworth, & Fisk, 2018; Williams & Kucheryavskiy, 2016; Zhao et al., 2018) and to associate spectral signatures with specific tissues (Miao et al., 2020). The hyperspectral imaging work presented here is based on previous work demonstrating that maize kernels with categorically distinct degrees of endosperm hardness could be separated by analyzing NIR hyperspectral images (McGoverin & Manley, 2012; Williams, Geladi, Fox, & Manley, 2009). The novel flatbed imaging platform we describe acquired multiwavelength NIR images of large numbers of maize kernels that properly trained algorithms could process to predict endosperm vitreousness, kernel density, and kernel protein content while simultaneously measuring morphometric features such as kernel length and width.

2. Materials and methods

2.1. Seed population and genetic relationships

WiDiv-942 is a panel of 942 maize inbred lines that produce physiologically mature kernels in the Midwest region of the United States (Mazaheri et al., 2019). Using a subset of 501 lines from the full WiDiv population, Renk et al. (2021) demonstrated that this population contains a significant amount of variation in kernel composition phenotypes such as carbohydrates, oil, and protein. Thus, this population is appropriate for building models that can predict composition traits from NIR spectral information.

All kernels were produced in 2018 in a field at the University of Wisconsin, West Madison Agricultural Research Station (WMARS) using one-row plots of 3.7 m length and 0.76 m spacing and arranged according to a randomized complete block design. Plants were self-pollinated to avoid any potential xenia effect of foreign pollen. The uppermost ear was harvested from five plants per row after plants reached physiological maturity, the ears were then dried with forced air until approximately 12% moisture and stored indoors. The genetic relationships between inbred members of a population can be visualized by performing principal components analysis (PCA) on sets of DNA sequence features called single nucleotide polymorphisms (SNPs). To show the relationships between members of WiDiv-942 used here, PCA was performed on a set of 5000 SNPs evenly sampled from the 899,784 SNPs Mazaheri et al. (2019) originally identified. The principal components scores were computed using the package 'PCAtools' version 1.2.0 (Blighe, 2019) in R version 3.6.3 (R Core Team (2020), 2020). Supplemental Fig. 2 shows the distribution of the samples selected to build each prediction equation within the full WiDiv-942 space.

Three sets of genotypes were used to build and test prediction models for endosperm vitreousness, kernel protein, and kernel density. The three sets were formed with members of the WiDiv-942 population because of the diversity it displays for the traits under study and because of future intentions to use it in genetic studies of the compositional phenotypes. A preliminary investigation explored variation in vitreousness levels enough to indicate which genotypes may represent an even sampling of the range of values across the population. This information was used to create Set 1, comprising 1428 kernels from 149 genotypes used to build a model that predicts vitreousness. To select lines for building the protein prediction equations, the full WiDiV-942 population was scanned as powder using a commercial NIR analyzer (NIRS DS 2500, FOSS, Hilleroed, Denmark) and a model provided by the manufacturer. These results were used to select Set 2, comprising 479 kernels from 150 lines used to create a protein prediction model. The results of the preliminary vitreousness survey were used to guide genotype selection for density because density correlates highly with vitreousness. Set 3 refers to the 448 kernels extracted from 150 lines (Table 1) to create a density prediction model. Set 3 was also used to build weight and volume prediction models as those traits were directly measured for density calculations. Ten kernels per genotype were used for vitreousness and three kernels for protein and density. The total number of kernels shown in Table 1 is not exactly the product of genotypes and kernels per genotype because a few samples were discarded

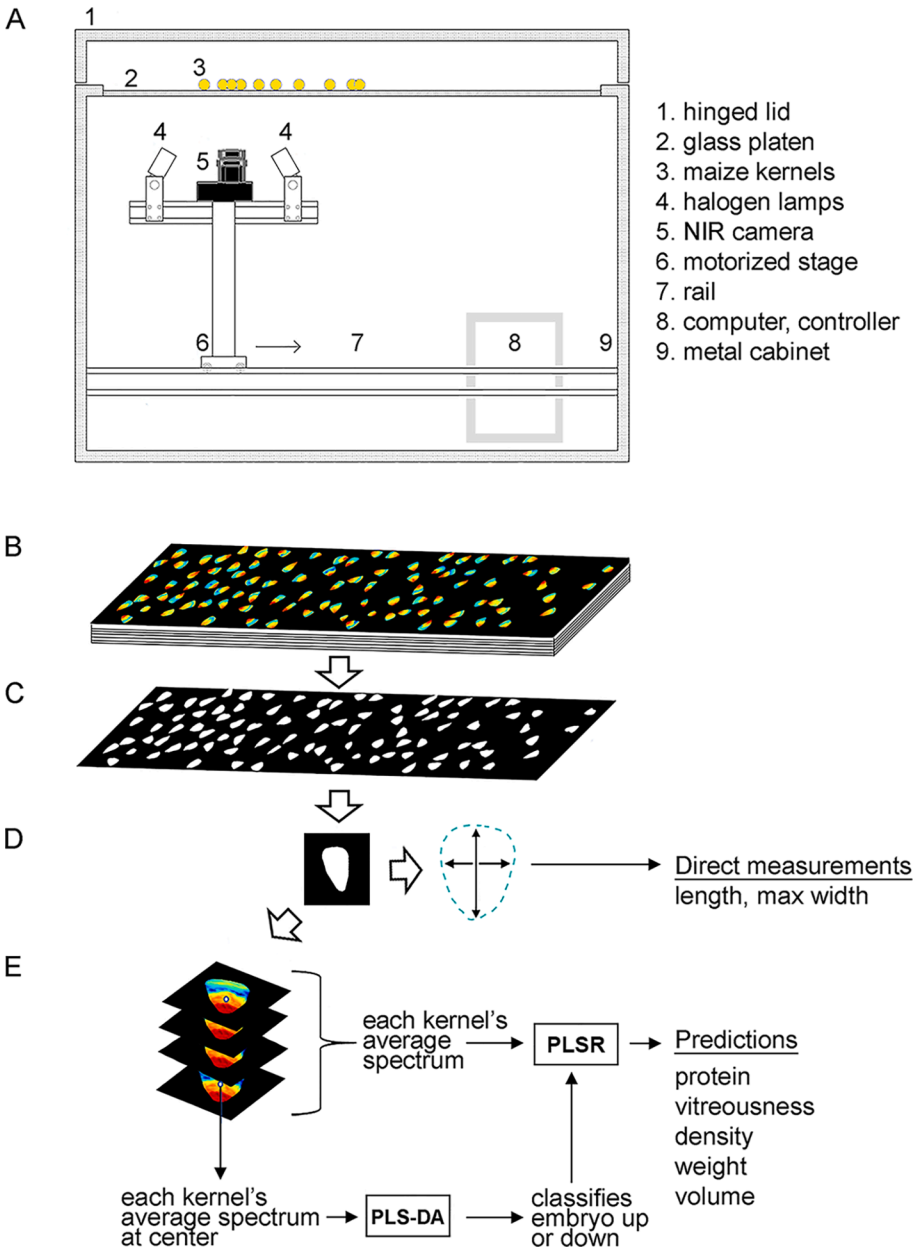

Fig. 1. Acquisition and processing of hyperspectral images of maize kernels. A. Diagram of the flatbed hyperspectral scanner. B. The scanner produces a stack of 224 images, the pixels in each one registering the amount of light in a different narrow band of wavelengths between 950 nm and 1700 nm. C. Binarization produces a mask with kernel pixels set to 1 and background pixels set to 0. D. The length and maximum width of each separate kernel is directly measured from the binarized image. E. The spectrum of each pixel in each kernel object is obtained from the stack depicted in B. The spectra from each pixel in the kernel object are averaged. To determine if the averaged spectra are from the germinal side of the kernel, or its abgerminal side, the average spectrum of pixels at the center of the kernel is used as an input to a PLS-DA classifier that is trained to distinguish between the two sides. The up or down label and the kernel's average spectrum are the inputs to a PLSR model that is trained to predict the indicated kernel traits.

during laboratory analysis.

2.2. Ground truth measurements of traits

2.2.1. Endosperm vitreousness

The 10 kernels from each of the lines in Set 1 that were scanned and then dissected for ground truth measurement of vitreousness following the methods described in Correa et al. (2002) and Ngonyamo-Majee et al. (2008) were not selected randomly. Instead, for each line, 100 kernels were randomly selected and sorted into groups of 10 kernels that were visually most like each other. One kernel was randomly chosen from each of the 10 groups to make the final sample. This process produced a representative sample set that reduced bias toward a particular size or shape. After scanning the germinal and abgerminal sides in the grid configuration, each kernel was soaked in distilled water for three minutes, then the pericarp and embryo were removed with a scalpel. The complete endosperm thus isolated was weighed on an electronic analytical balance (Ohaus AX224/E, Parsippany, New Jersey, USA). The floury endosperm component was manually removed using an electric rotary tool equipped with a 1/16 in. round engraving accessory with the aid of magnifying glasses. After all the floury endosperm was carefully removed, the weight of the remaining vitreous endosperm was recorded to calculate vitreousness as a percentage of the total endosperm weight.

2.2.2. Density

The density of 3 kernels randomly selected from each of the lines in Set 3 was determined by measuring buoyant force with an analytical balance according to the Archimedes principle. Supplemental Fig. 3 shows the apparatus used. A 50 mL beaker containing 30 mL of distilled water at 22 °C was placed on a microbalance (Ohaus AX224/E, Parsippany, NJ, USA). A kernel attached to a needle was submerged using a drill press stand to control the motion. The balance recorded an increase in mass after the kernel and a marked section of the needle were submerged. The known volume of the submerged section of the needle was subtracted from the total displaced volume. The measured weight of the fluid that the kernel displaces was converted to a volume (specific density of water is 0.998 g mL^{-1} at 22 °C). Dividing the mass of the kernel, measured separately, by the measured volume gives kernel

density. The accuracy of the method was determined to be greater than 99% by measuring kernel-sized pieces of pure minerals of known densities (pyrite, quartz, fluorite, aventurite and hematite).

2.2.3. Protein

Three kernels from each of the lines in Set 2 were randomly selected for total C and N analysis, from which total protein was calculated. Each kernel was ground using a mortar and pestle and transferred to a 2 mL microcentrifuge tube. The powder was dried at 60 °C for 48 h and stored in borosilicate glass desiccators. A microbalance (Mettler-Toledo XP6, Columbus, Ohio, USA) was used to prepare tin foil capsules containing 10 mg of powder, which were combusted in an elemental analyzer (CE Elantech EA1112, Lakewood, New Jersey, USA). BBOT (2,5-Bis (5-*tert*-butyl-benzoxazol-2-yl) thiophene) and Atropine were used as calibration standards as recommended by the manufacturer. Total protein was calculated as N × 6.25 expressed on a dry weight basis.

2.2.4. Dimensions

The length and width of 480 kernels in Set 2 were measured with a precision digital caliper (Mitutoyo 500, Aurora, Illinois) to determine the accuracy of the automated image processing-based measurements. Kernel length was measured from the kernel tip to the center of the cap. Kernel width was defined as the largest distance perpendicular to kernel length axis.

2.3. Hyperspectral imaging device

The data in this study were acquired with a hyperspectral imaging device developed in collaboration with Middleton Spectral Vision (Middleton, Wisconsin, USA). The device uses a 12-bit NIR line-scan camera (Specim model FX17e, Oulu, Finland) to collect images of samples placed above it on a horizontal 16 × 107 cm glass plate. A bank of eight broadband quartz halogen lamps provided full spectrum illumination (Fig. 1A). When the camera is fitted with a 33-mm focal length lens, its line of 640 pixels covers 102 mm, resulting in a spatial resolution of 0.16 mm. The camera and lamps are mounted beneath the rectangular glass sample bed, facing upward. A motor translates the upward-facing camera and lamps along the long axis of the sample bed at 16.5 mm s^{-1} as the camera acquires lines (frames) at a rate of 100 s^{-1}. Each pixel in the line registers the energy in 224 wavelength bands between 950 nm and 1700 nm, corresponding to a spectral resolution of 3.3 nm.

The scanner and software automatically scanned a dark and white reference image to normalize each pixel value when processing the data. A piece of porous white polytetrafluoroethylene was added to the beginning of the scanning region to collect a "white reference". A baseline "dark" reference was acquired by closing the camera shutter for 0.6 s. The spectra at each pixel in subsequent scans of biological material were corrected according to Equation (1).

$$\frac{I_0 - I_d}{I = I_w - I_d} \tag{1}$$

where I is the corrected image, I_0 is the raw image, I_d is the dark reference image, and I_w is the white reference image.

The maize kernels to be scanned were either scattered randomly on the glass sample bed or placed in a 5 × 24 grid fixture with one kernel per cell (Supplemental Fig. 4). In both sample configurations, individual kernels were the analyzed unit. The grid arrangement allows a single indexed kernel to be retrieved for a posterior use such as destructive ground-truth measurement. The scattered arrangement is much faster, but it does not allow a particular kernel to be chosen for a future use. When the grid was used, each kernel was scanned twice - once with the embryo (germinal side) facing the camera and then the kernel was flipped to capture the abgerminal side. A model was trained to distinguish between the two sides of a kernel (section 2.6). When kernels were scattered on the device, this model was used to determine which side of each kernel faced the camera. The two types of sample presentation served different purposes. The grid was used to associate image data with the same kernel used for destructive ground-truth measurements, which is necessary for building predictive models. The scattered kernel method was used in a high-throughput manner to make model-based inferences from many genotypes.

2.4. Computational methods and feature extraction pipeline

The raw images the device produces are multichannel images, which means they are $m \times n \times z$ matrices where m and n are the width and length of the image in pixels and z is the number of wavelength bins (224 in this work) at which the photon fluence has been measured. The following image processing steps were coded in the MATLAB (version R2019b) computer language to extract information from these hyperspectral images of kernels in either the gridded or scattered configurations.

i. Raw images that were spectrally corrected with white and dark reference scans using Eq.1 were converted to absorbance values using Eq. (2).

$$Absorbance = \log_{10}\left(\frac{1}{reflectance}\right) \tag{2}$$

ii. Sweeping a line-scan camera does not necessarily produce square pixels. Absorbance images were resized using a bicubic interpolation method (*imresize* in MATLAB) and the inner distance of the first cell of the grid as a reference.
iii. Each pixel having an absorbance value at 1090 nm less than 1.4 was set to zero. This threshold value, determined by inspection, accurately segmented the kernels from the background to create a binary mask.
iv. Small objects in the binary image, those containing fewer than 700 pixels, were dust or debris and therefore removed from the binary mask.
v. The average absorbance of a 3 × 3 pixels square region centered at the kernel's center of mass was calculated for each wavelength in the absorbance array. These data were used to determine if a kernel in a scattered sample was germinal-facing or abgerminal-facing as explained in section 2.6.
vi. For each object in a size-filtered binary mask (i.e., kernel), the absorbance at each wavelength (absorption spectra) were averaged across pixels to create an average absorption spectrum. This average spectrum was associated with a unique center-of-mass coordinate pair.
vii. Kernel length, width and area were extracted based on the size-filtered binary mask using the methods Miller et al. (2017) previously developed. Briefly, analysis of contour curvature initiates a process that identifies the tip of the kernel, which marks one end of the major axis (kernel length) and the longest orthogonal segment is kernel width.
viii. For each scanned image containing many kernels, the averaged absorbances were exported to the R programing environment were the orientation and PLSR predicted traits (vitreousness, protein, density, weight and volume) were calculated with methods described in 2.5.

2.5. Prediction model construction, calibration, and validation.

Models to predict vitreousness, protein, density, weight, and volume, were built and tested using the 'pls' package version 2.8–0 (Liland, Mevik, & Wehrens, 2021) in R. Separate models were created for germinal-facing images and abgerminal facing images. The dependent variables were spectral data extracted from hyperspectral images

Table 2
Performance of the PLS model for five kernel traits and two orientations using the optimal spectral pretreatment.

Kernel trait	Set	Kernel Orientation	Spectra pretreatment	PLS factors	Cross-validation			External validation	
					RMSE	PRESS	R^2	RMSE	R^2
Vitreousness (%)	1	Abgerminal	SG.D1W11	13	7.42	52361.7	0.62	6.3	0.56
		Germinal	SG	14	7.1	48611.9	6.5	6.6	0.51
Protein (%)	2	Abgerminal	SNV1D	11	1.06	360.9	0.75	0.85	0.84
		Germinal	SNVSG	17	1.06	358.3	0.76	0.92	0.79
Density (g/cm^3)	3	Abgerminal	SG	7	0.036	0.41	0.67	0.038	0.43
		Germinal	SNV	8	0.042	0.53	0.52	0.04	0.52
Volume (cm^3)	3	Abgerminal	Raw	17	0.029	0.26	0.63	0.027	0.54
		Germinal	Raw	12	0.029	0.24	0.67	0.022	0.69
Weight (g)	3	Abgerminal	Raw	17	0.032	0.31	0.68	0.032	0.57
		Germinal	SG.D1W11	9	0.033	0.34	0.66	0.027	0.69

RMSE, Root mean square error of prediction; R^2, Coefficient of determination; PRESS, Predicted residual sum of squares.
Spectral pretreatments: SG.D1W11, Savitzky-Golay + 1st derivative using windows size of 11; SG, Savitzky-Golay; SNV1D, Standard Normal Variate + 1st derivative; SNVSG, Standard Normal Variate + Savitzky-Golay; SNV, Standard Normal Variate.

obtained in the grid configuration and the independent variables were trait measurements (section 2.2) made on the same kernels. For each trait, two-thirds of the kernels (n = 320 for protein, n = 302 for density, and n = 949 for vitreousness) were used for building the partial least squares regression (PLSR) model and one third (n = 159 protein, n = 146 for density, and n = 479 for vitreousness) was held out to validate the model. The Kennard-Stone algorithm (Kennard & Stone, 1969) was used to partition the raw spectral data into these training and validation sets using the 'prospectr' package version 0.2.1 (Stevens & Ramirez-Lopez, 2020) in R. Samples from the same genotype were assigned to one group or the other but not both to reduce associations between the calibration and validation data sets that could inflate prediction accuracy.

After partitioning the raw spectral data into training and validation sets, each spectrum was subjected to one of 13 filters included in the 'waves' package (Hershberger & Gore, 2020). This step determines the spectral filtering and smoothing pretreatment that enables a PLSR model to produce the most accurate predictions. To guard against overfitting, cross-validation with the "one-sigma heuristic" strategy included in the 'selectNcomp' function of the 'pls' package was used to determine the lowest number of latent factors the model could use before the error of prediction (RMSEP) increased more than one standard error from the best result achievable using any number of factors. The selected model was used to perform an external validation using the remaining one third of the kernels. Performance statistics were calculated for the cross-validation of the calibration set and the external validation set (Table 1). For each trait-kernel orientation (germinal/abgerminal) combination, the pretreatment method with the lowest Root Mean Square Error of Prediction (RMSEP) was chosen.

2.6. Kernel side classification

Because composition differs between grain tissues, separate models were created for predicting traits from germinal-side and abgerminal-side images. To apply the correct model to each kernel in a scattered kernel image, the orientation of each kernel must be analytically determined. Partial least squares discriminant analysis (PLS-DA), a type of supervised learning (Barker & Rayens, 2003; Lee, Liong, & Jemain, 2018; Ruiz-Perez, Guan, Madhivanan, Mathee, & Narasimhan, 2020), was used to classify each kernel in an image as germinal or abgerminal based on spectral information obtained from the 3 × 3 pixel square section at the center of mass. Images from kernel Set 3 (protein) were used to train and test the PLS-DA model using the function 'plsda' from the package 'mdatools' version 0.11.3 (Kucheryavskiy, 2020) using the "SIMPLS" algorithm in R. Four hundred and eighty (480) individualized kernels were scanned using the grid. The kernels were scanned in both orientations generating 8 images with 120 kernels each were the spectral data was extracted from each kernel. From a total of 960 manually classified kernels (480 germinal side and 480 abgerminal side), half were randomly selected to train the algorithm. The optimal number of latent variables was determined using the 'leave one out cross-validation' (LOOCV) scheme. The lowest RMSECV value determined the optimal number of latent variables (LV). Finally, the computed model was used to predict samples in the held-out validation set. To test the effect of spectral pretreatments on model performance, the same 13 spectral pretreatments evaluated during construction of the PLSR models were applied and subsequently submitted to PLS-DA calibration and cross-validation. The pretreatment with the highest classification accuracy defined as the number of correct predictions divided by the total number of predictions was selected and used. Classification accuracy with the validation set was calculated to generate a confusion matrix, along with specificity and sensitivity rates (Supplemental Table 1).

2.7. Complete pipeline

The processes, measurements, and analyses described in Sections 2.3-2.6 were combined to produce a pipeline shown in Fig. 1B-E. All of the code created to execute the analyses is available in this repository, https://github.com/jivarelao/Hyperspectral_Scanner.

3. Results and discussion

3.1. Variability of maize kernel traits in ground-truth sets

Directly measured traits ranged widely across the kernel samples (Table 1). Kernel volume displayed the largest range (5.6-fold). Kernel weight was second at 5-fold, followed by vitreousness (2.7-fold), protein (2.4-fold) and density (1.4-fold). The protein range of 8.02%–19.45% agreed reasonably well with a previous single-kernel study (Baye, Pearson, & Settles, 2006), and a density range of 1.0–1.35 g cm^{-3} was in accord with the single-kernel findings of Gustin et al. (2013). The ranges were also consistent with values from commercial hybrids (Correa et al., 2002). The range of kernel weight (0.08–0.42 g) and volume (0.07–0.39 cm^3) were likewise as expected for a diverse collection of inbred lines (Gustin et al., 2013). The large phenotypic diversity found in WiDiv-942 produced a large range of endosperm vitreousness (35%-95%), comparable to previous studies of this trait (Correa et al., 2002; Ngonyamo-Majee et al., 2008). This amount of variation represented in a sample of more than 1400 kernels endowed the training data with enough vitreousness variation to be generally useful in studies of maize.

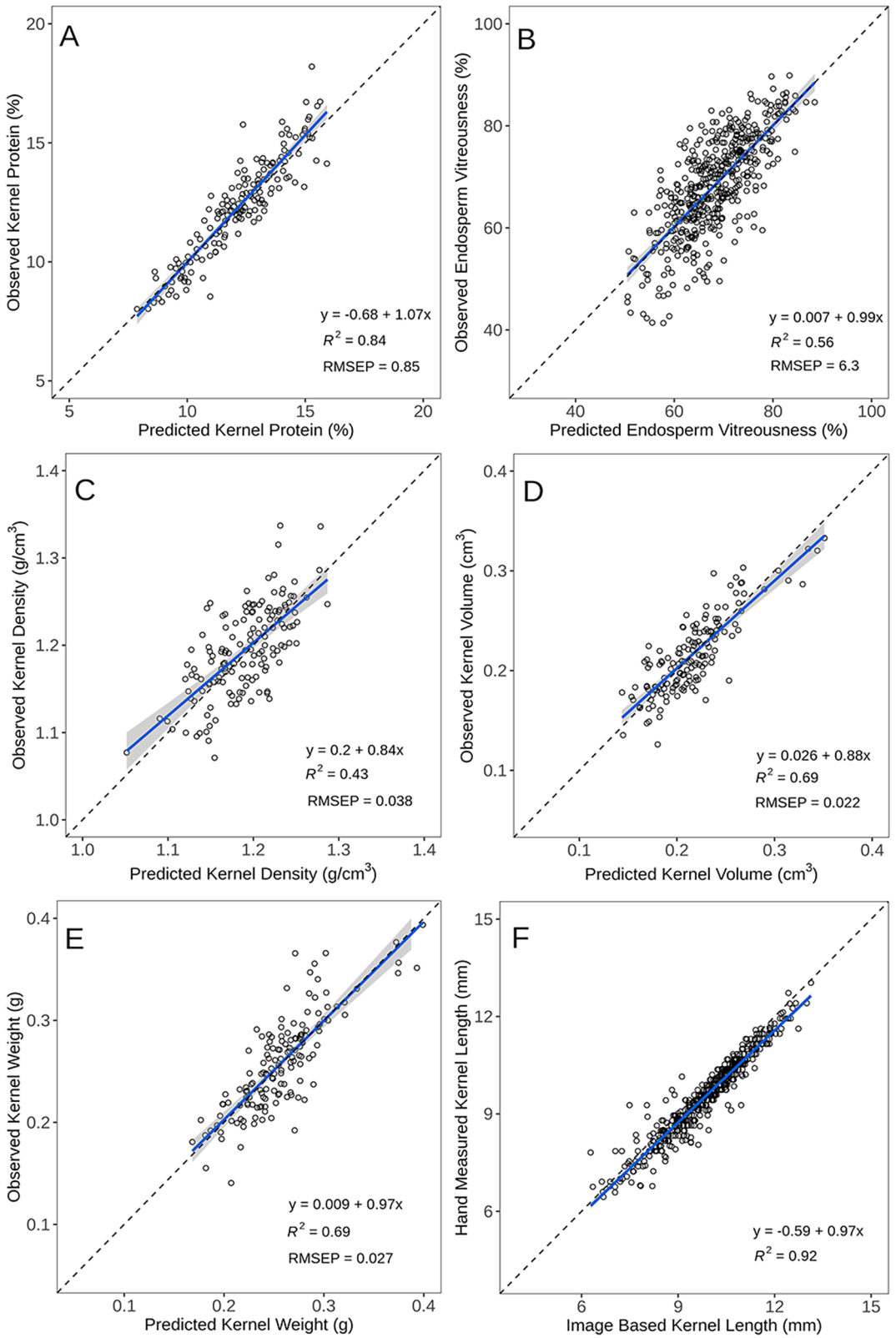

Fig. 2. Scatter plots of observed (ground truth) measurement versus spectral-based model predictions or direct measurements of dimensions. (A) kernel protein, (B) endosperm vitreousness, (C) kernel density, (D) kernel volume, (E) kernel weight. A-E, NIR-predicted and lab measured kernel and endosperm traits. (F) image-based kernel length versus hand measured kernel length. Each panel displays values for the external validation. The blue line represents the linear regression line based on the equation on the upper left. The dotted line shows the perfect agreement. R^2, Coefficient of determination; RMSEP, root mean square error of prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Single kernel predictions based on NIR absorbance.

The first derivative of the Standard Normal Variate (SNV) scatter correction method was found to be the best pretreatment for protein prediction. With it, we obtained a RMSEP of 0.85% and an r^2 of 0.84 when predicting protein from the abgerminal-side spectra. The effects of different spectral pretreatments on the model's performance are compared in Supplemental Fig. 5. The highest accuracy obtained from abgerminal side agrees with the results reported by Jiang et al., 2007 who also scanned maize kernels from both orientations to predict protein using NIR. Our protein results closely agreed with Gustin et al., 2013 who obtained an r^2 of 0.86 and SEP = 0.89 using a single kernel NIR based prediction models. The protein results reported here also performed very similar to the ones shown by Spielbauer et al., 2009 who obtained SEP = 0.81 and r^2 = 0.91 when scanning maize kernels in a custom made NIR machine.

Only moderate r^2 (0.43) but very low RMSEP (0.038 g/cm^3) was achieved when total kernel density was predicted given abgerminal-side spectral pretreated with the Savitzky-Golay smoothing filter (Table 2). Gustin et al. (2013) reported SEP for single seed NIR prediction of total kernel density of 0.067 g cm^{-3} using microcomputed tomography as the reference method. Our low r^2 value may be due to the narrow distribution of density values in the subset of the diversity panel selected. By contrast, NIR studies of maize kernels oftentimes include endosperm mutants or genotypes that have undergone selection for extreme endosperm characteristics, which may extend the range in a way that raises r^2, a statistic that is highly dependent on the range of the validation set (Davies & Fearn, 2006).

Moderately high r^2 (0.56) was achieved for endosperm vitreousness (Fig. 2) using the abgerminal side and using the Savitzky-Golay smoothing and first order derivative with a windows size of 11. An RMSEP of 6.3% shows that this system compares favorably in terms of throughput and accuracy with previous reports. Ngonyamo-Majee et al. (2008) reported a prediction accuracy of 6.04 % for this trait, but used pre-processed ground kernels as the sample. An early application of NIR hyperspectral imaging to infer endosperm quality in intact kernels (Williams et al., 2009) proved the principle by showing that nondestructive analyses could distinguish different categories of hardness. The results in Fig. 2 show how the present platform predicted continuous values of endosperm vitreousness from spectral images of intact kernels to produce results that could be used in quantitative genetics and gene mapping studies.

Vitreousness is defined as a mass ratio of two completely separable solid phases, which sometimes is difficult to achieve, at least with the currently available mechanical separation methods. Oftentimes there is as a very thin transition zone of hard floury to vitreous area rather than a completely distinguishable boundary between a floury and vitreous part. Despite this potential source of noise in the ground truth data, the RMSEP (~6%) indicates this non-destructive method that takes kernel orientation into account will be effective and broadly applicable in studies of vitreousness.

Fig. 2 shows that our models for predicting total kernel volume and weight showed moderate-high r^2 (0.69 both) and low RMSEP values (0.022 cm^3 and 0.027 g, respectively,) that were very similar to previously reported predictions based on single-kernel spectra (Spielbauer et al., 2009).

Correlation between the genotypic means of vitreousness, kernel protein and kernel density were calculated for the genotypes that are shared in the three Sets (Supplemental Fig. 6). A positive significant Pearson's correlation coefficient of 0.67 (pval < 0.001) was found between endosperm vitreousness and kernel density. This result agrees with the value reported by Correa et al., 2002 who found a positive significant correlation between these two traits (r = 0.87) and suggested that density may be a reliable tool for screening large maize data sets for vitreousness. Total protein was positively correlated (r = 0.39, pval < 0.01) with endosperm vitreousness, in close agreement with the r = 0.41 found by Ngonyamo-Majee et al. (2008).

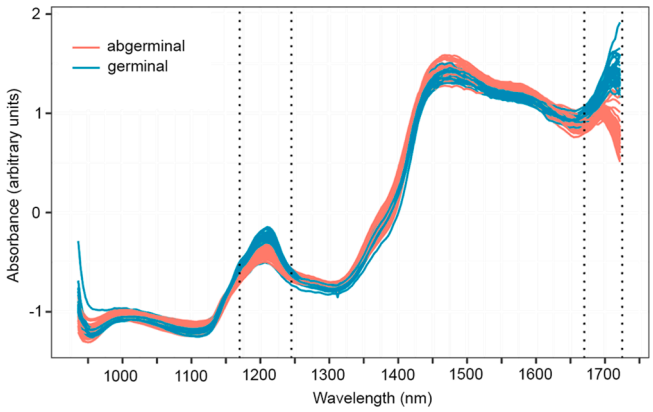

Fig. 3. Reflectance spectra in the NIR region for 80 maize kernel samples. Half of the samples were manually oriented with the germinal (embryo) side facing the camera (red lines, n = 40) and the other half with the abgerminal side facing the camera (light blue lines, n = 40). The vertical dotted lines indicate spectral regions that differed most between the two sides of the kernel. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Kernel position detection based on embryonic reflectance for high throughput scanning in bulk configuration

When scattering the kernels over the scanning region, they will randomly land with either the germinal side (embryo) or abgerminal side facing the camera. In order to use the most accurate NIR model developed in section 3.2 without manually orienting each kernel, we explored using PLS-DA to classify the orientation of each kernel in the image based on NIR absorbance information. We hypothesized that a successful orientation classifier could be based on a reliable difference in some regions of the NIR spectra between the embryo and the endosperm (Orman & Schumann, 1992). To test this, 80 kernels from Set 3 were randomly chosen and the average absorbance for both positions was plotted. The region in the vicinity of 1200 and 1720 nm has distinguishable higher absorbances for the embryo compared to endosperm as shown in Fig. 3. Osborne, Fearn, and Hindle (1993) has described that the major absorption band in oil is due to a long chain fatty acid moiety that gives rise to CH_2 second overtone at 1200 nm and the band near 1180 nm has been assigned as the second overtone of the fundamental C—H absorption of pure fatty acids containing *cis* double bonds such as oleic acid (Sato, Kawano, & Iwamoto, 1991). Cho and Iwamoto (1989) correlated the absorption bands at 1710 and 1725 nm to linoleic and oleic acids, respectively. The larger absorbances observed in the aforementioned regions in our study could, therefore, be correlated to the two dominant fatty acids presents in corn, which are mostly presented in the embryonic region (Barrera-Arellano, Badan-Ribeiro, & Serna-Saldivar, 2019).

Instead of using the average absorbance of the whole kernel exposed to the camera as input to train the classification algorithm, a 3 × 3 pixel square at the center of mass was sampled. This region invariably overlaid the embryo when it faced the camera, and the endosperm when the abgerminal side faced down. Absorbance data from this centrally-located group of pixels provided a substantially clearer signal than averaging the whole region. Absorbance data of Set 3 were used to calibrate and test the PLS-DA algorithm (section 2.6). Half of the manually classified samples were randomly selected for calibration while the other half was used for validation. Almost perfect classification accuracy was achieved with over 99.5% of correct position identification in both the calibration and validation dataset (Supplemental Table 1).

The previously developed PLS-DA model (section 2.6) was included

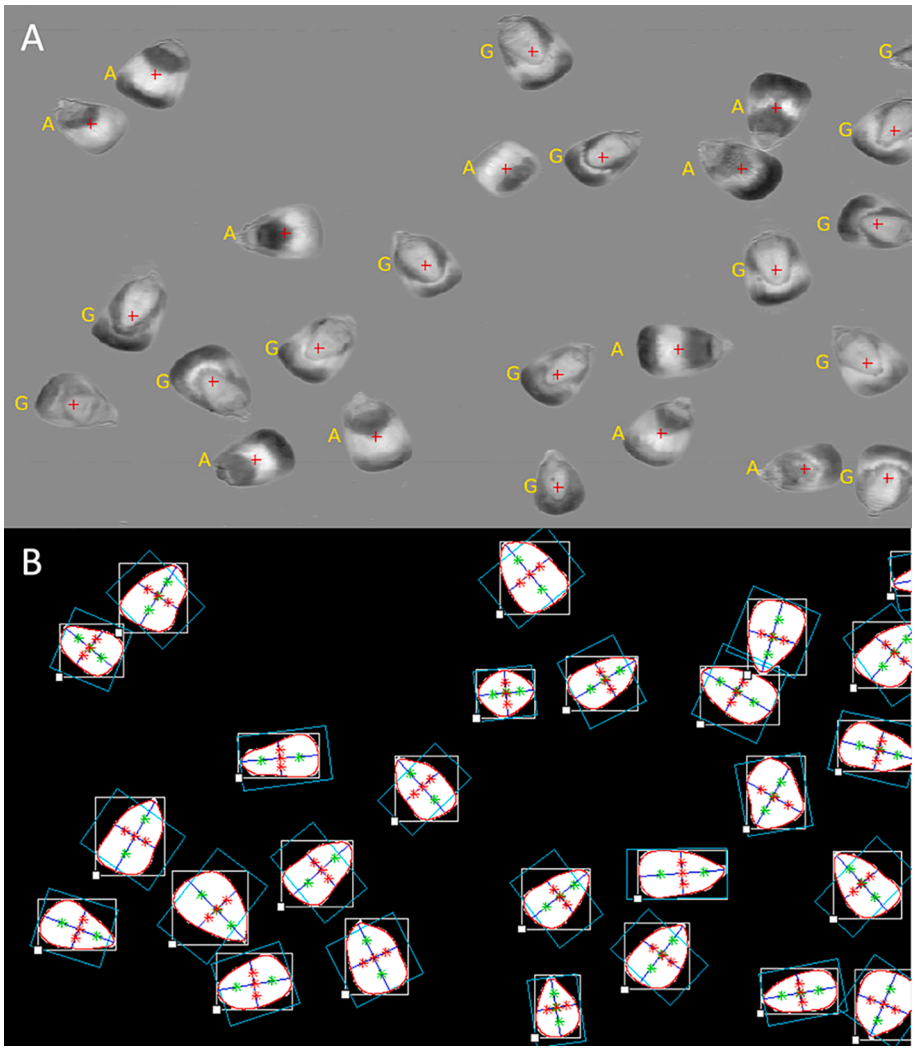

Fig. 4. Hyperspectral image of kernels and binary morphology measurements. (A) Third Principal Component score image of raw NIR absorbances showing a section of the bulk scanning system surface. The red crosses represent the center of mass of each kernel. In yellow the PLS-DA kernel orientation classification output, G = Germinal and A = Abgerminal side of the kernel facing the camera. (B) Binary mask of a section of the bulk scanning system surface generated from the NIR image collected in the flatbed scanner. Kernel contours are depicted in red, kernel major axis and minor axis are depicted as a blue line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in the high throughput NIR processing pipeline to classify each kernel (Fig. 1E) so its orientation can determine which trait-specific NIR prediction model is used (Fig. 4).

3.4. Single kernel morphometric feature extraction based on image analysis

The binary mask used to extract NIR absorbances from individual kernels was also used to measure kernel length and width (Fig. 1D). Measuring the kernel length (major axis) and width (minor axis) depends on correctly identifying the kernel tip. The algorithm developed by Miller et al. (2017) was slightly modified to measure kernel length and width from the images this novel flatbed NIR scanner produces. The correlation between hand and image-based measurements of kernel length was 0.95 (Fig. 2F) and 0.9 for kernel width. In general, this correlation may be limited by inaccuracy of hand measurements or inability of the image processing algorithm to correctly identify the tip of each kernel, particularly in kernels having imperfectly flat faces and rounded edges.

4. Conclusions

This upward-focused flatbed hyperspectral imaging scanner generated data that custom models used to predict important maize kernel traits with high throughput (75 kernels nondestructively prepared and measured in approximately 60 s). Endosperm vitreousness of a single kernel was predicted to a useful degree across a wide range of kernel types, indicating that this automated alternative to a laborious manual method would be generally useful rather than population dependent. Accuracy of kernel protein and density predictions, which are relevant to multiple grain markets, were similar to or greater than previous reports. The imaging capabilities of the platform allowed germinal or abgerminal side-specific models to be applied, which improved accuracy for some traits, and it enabled kernel size dimensions to be measured at the same time. These features could make the platform described here useful to food processors, livestock feed producers, maize researchers, and breeders. The platform may prove useful in the study of other seed crops.

CRediT authorship contribution statement

Jose I. Varela: Formal analysis, Methodology, Validation, Software, Writing – original draft. **Nathan D. Miller:** Software. **Valentina Infante:** Investigation. **Shawn M. Kaeppler:** Supervision, Writing – review & editing. **Natalia de Leon:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Edgar P. Spalding:** Conceptualization, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgements

This work is based on the research supported in part by United States Department of Agriculture (USDA) grant WIS03049 to NDL and National Science Foundation grant 1940115 to EPS. The authors wish to thank Dr. Gabor Kemeny, Chris Draves, Jack Heese and Stuart Smith from Middleton Spectral Vision for their technical support with the NIR hyperspectral system. The authors are grateful to Maggie Phillips for her technical support with the C/N analyzer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2022.133264.

References

Barker, M., & Rayens, W. (2003). Partial Least Squares for Discrimination. *Journal of Chemometrics., 17*, 166–173. https://doi.org/10.1002/cem.785

Barrera-Arellano, D., Badan-Ribeiro, A. P., & Serna-Saldivar, S. O. (2019). Corn Oil: Composition, Processing, and Utilization. In S. O. Serna-Saldivar (Ed.), *Corn: Chemistry and Technology* (pp. 539–613). Elsevier.

Baye, T. M., Pearson, T. C., & Settles, A. M. (2006). Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. *Journal of Cereal Science., 43*(2), 236–243.

Bergquist, R., & Thompson, D. (1992). Corn grain density characterized by two specific gravity techniques. *Crop Science., 32*, 1287–1290.

Blighe, K. (2019). PCAtools: Everything Principal Component Analysis. *R package version, 1*(2). https://github.com/kevinblighe/PCAtools.

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2018). Protein content prediction in single wheat kernels using hyperspectral imaging. *Food Chemistry., 240*, 32–42. https://doi.org/10.1016/j.foodchem.2017.07.048

Cho, R. K., & Iwamoto, M. (1989). The purity identification of sesame oil by near infrared reflectance spectroscopy. *Proceedings of the 2nd International NIRS conference*.

Correa, C. E. S., Shaver, R. D., Pereira, M. N., Lauer, J. G., & Kohn, K. (2002). Relationship between corn vitreousness and ruminal in situ starch degradability. *Journal of Dairy Science., 85*, 3008–3012.

Davies, A. M. C., & Fearn, T. (2006). Back to basics: calibration statistics. *SpectroscopyEurope., 18*(2), 31–32.

Dias Junior, G. S., Ferraretto, L. F., Salvati, G. G. S., de Resende, L. C., Hoffman, P. C., Pereira, M. N., & Shaver, R. D. (2016). Relationship between processing score and kernel-fraction particle size in whole-plant corn silage. *Journal of Dairy Science., 99*(4), 2719–2729. https://doi.org/10.3168/jds.2015-10411

Feng, L., Zhu, S., Liu, F., He, Y., Bao, Y., & Zhang, C. (2019). Hyperspectral imaging for seed quality and safety inspection: a review. *Plant Methods, 15*, 1–25. https://doi.org/10.1186/s13007-019-0476-y

Fox, G., & Manley, M. (2014). Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals. *Journal of the Science of Food and Agriculture, 94*(2), 174–179.

Gustafson, T. J., & de Leon, N. (2010). Genetic analysis of Maize (Zea mays L.) endosperm vitreousness and related hardness traits in the intermated B73 x Mo17 recombinant inbred line population. *Crop Science., 50*, 2318–2327.

Gustin, J. L., Jackson, S., Williams, C., Patel, A., Armstrong, P., Peter, G. F., & Settles, A. M. (2013). Analysis of maize *(Zea mays)* kernel density and volume using microcomputed tomography and single-kernel Near-Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry., 61*(10872–10880). https://doi.org/10.1021/jf403790v

Hacisalihoglu, G., Gustin, J. L., Louisma, J., Armstrong, P., Peter, G. F., Walker, A. R., & Settles, A. M. (2016). Enhanced single seed trait prediction in soybean (*Glycine max*) and robust calibration model transfer with Near-Infrared Reflectance Spectroscopy. *Journal of Agricultural and Food Chemistry., 64*, 1079–1086. https://doi.org/10.1021/acs.jafc.5b05508

Hershberger, J., & Gore, M. A. (2020). waves: Vis-NIR Spectral Analysis Wrapper. *R package version*, (1).

Holding, D. R., & Larkins, B. A. (2006). The development and importance of zein protein bodies in maize endosperm. *Maydica., 51*, 243–254.

Jiang, H. Y., Zhu, Y. J., Wei, L. M., Dai, J. R., Song, T. M., Yan, Y. L., & Chen, S. J. (2007). Analysis of protein, starch and oil content of single intact kernels by near infrared reflectance spectroscopy (NIRS) in maize (Zea mays L.). *Plant Breeding., 126*(5), 492–497. https://doi.org/10.1111/j.1439-0523.2007.01338.x

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics., 11*(1), 137–148.

Kucheryavskiy, S. (2020). Mdatools – R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems., 198*, Article 103937. https://doi.org/10.1016/j.chemolab.2020.103937

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst., 143*, 3526–3539.

Liland, K.H., Mevik, B-H., & Wehrens, R. (2021). pls: Partial Least Squares and Principal Component Regression. R package version 2.8-0. https://CRAN.R-project.org/package=pls.

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., … Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology., 2019*(19), 45. https://doi.org/10.1186/s12870-019-1653-x

McAllister, T. A., Phillippe, R. C., Rode, L. M., & Cheng, K. J. (1994). Effect of the protein matrix on the digestion of cereal grains by ruminal microorganisms. *Journal of Animal Science., 71*, 205–212.

McGoverin, C. M., & Manley, M. (2012). Classification of Maize Kernel Hardness Using near Infrared Hyperspectral Imaging. *Journal of Near Infrared Spectroscopy., 20*(5), 529–535. https://doi.org/10.1255/jnirs

Miao, C., Pages, A., Xu, Z., Rodene, E., Yang, J., & Schnable, J. C. (2020). Semantic segmentation of sorghum using hyperspectral data identifies genetic associations. *Plant Phenomics, 1*, 11. https://doi.org/10.34133/2020/4216373

Miller, N. D., Haase, N. J., Lee, J., Kaeppler, S. M., De León, N., & Spalding, E. P. (2017). A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. *The Plant Journal, 89*(169), 178. https://doi.org/10.1111/tpj.13320

Ngonyamo-Majee, D., Shaver, R. D., Coors, J. G., Sapienza, D., Correa, C. E. S., Lauer, J. G., & Berzaghi, P. (2008). Relationship between kernel vitreousness and dry matter degradability for diverse corn germplasm I. Development of near-infrared reflectance spectroscopy calibrations. *Animal Feed Science Technology., 142*, 247–258. https://doi.org/10.1016/j.anifeedsci.2007.09.023

Orman, B. A., & Schumann, R. A. (1991). Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil, and starch in maize grain. *Journal of the American Oil Chemists' Society., 69*, 1036–1038.

Orman, B. A., & Schumann, R. A. (1992). Nondestructive single-kernel oil determination of maize by near-infrared transmission spectroscopy. *Journal of the American Oil Chemists' Society., 69*, 1036–1038.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy: With applications in food and beverage analysis*. Longman Scientific and Technical.

Philippeau, C., & Michalet-Doreau, C. (1997). Influence of genotype and stage of maturity of maize on rate of ruminal starch degradation. *Animal Feed Science Technology., 68*, 25–35.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Renk, J. S., Gilbert, A. M., Hattery, T. J., O'Connor, C. H., Monnahan, P. J., Anderson, N., … Hirsch, C. N. (2021). Genetic control of kernel compositional variation in a maize diversity panel. *The Plant Genome*, e200115. https://doi.org/10.1002/tpg2.20115

Robutti, J. L. (1995). Maize kernel hardness estimation in breeding by near infrared transmission analysis. *Cereal Chemistry, 72*, 632–636.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So, you think you can PLS-DA? *BMC Bioinformatics., 21*, 2. https://doi.org/10.1186/s12859-019-3310-7

Sato, T., Kawano, S., & Iwamoto, M. (1991). Near infrared spectral patterns of fatty acid analysis from fats and oils. *Journal of the American Oil Chemistry Society., 68*, 827–833.

Siesler, H. W., Ozaki, Y., Kawata, S., & Heise, H. M. (2008). *Near-infrared spectroscopy: Principles, instruments, applications*. New York: John Wiley and Sons Inc.

Spielbauer, G., Armstrong, P., Baier, J. W., Allen, W. B., Richardson, K., Shen, B., & Settles, A. M. (2009). High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. *Cereal Chemistry., 86*(5), 556–564.

Stevens, A., & Ramirez-Lopez, L. (2020). An introduction to the 'prospectr' package. *R package Vignette R package version*, (2), 1.

Weinstock, B. A., Janni, J., Hagen, L., & Wright, S. (2006). Prediction of oil and oleic acid concentrations in individual corn (Zea mays L.) kernels using near-infrared reflectance hyperspectral imaging and multi- variate analysis. *Applied Spectroscopy., 60*, 9–16.

Williams, P., Geladi, P., Fox, G., & Manley, M. (2009). Maize kernel hardness classification by near infrared (NIR) hyper-spectral imaging and multivariate data analysis. *Analytica Chimica Acta, 653*, 121. https://doi.org/10.1016/j.aca.2009.09.005

Williams, P. J., & Kucheryavskiy, S. (2016). Classification of maize kernels using NIR hyperspectral imaging. *Food Chemistry., 209*, 131–138. https://doi.org/10.1016/j.foodchem.2016.04.044

Wu, Y. V., & Bergquist, R. (1991). Relation of corn grain density to yields of dry milling products. *Cereal Chemistry., 68*(5), 542–544.

Xu, A., Lin, L., Guo, K., Liu, T., Yin, Z., & Wei, C. (2019). Physicochemical properties of starches from vitreous and floury endosperms from the same maize kernels. *Food Chemistry., 291*, 149–156. https://doi.org/10.1016/j.foodchem.2019.04.024

Zhao, Y., Zhu, S., Zhang, C., Feng, X., Feng, L., & He, Y. (2018). Application of hyperspectral imaging and chemometrics for variety classification of maize seeds. *RSC Advances., 8*(3), 1337–1345. https://doi.org/10.1039/C7RA05954J