Journal of Experimental Botany, Vol. 72, No. 12 pp. 4373–4383, 2021
doi:10.1093/jxb/erab118 Advance Access Publication 18 March 2021

RESEARCH PAPER

Predicting photosynthetic capacity in tobacco using shortwave infrared spectral reflectance

Thomas Sexton[1], Sindhuja Sankaran[2] and Asaph B. Cousins[1,*]

[1] School of Biological Sciences, Washington State University, PO Box 644236, Pullman, WA 99164-4236, USA
[2] Department of Biological Systems Engineering, Washington State University, PO Box 646120, Pullman, WA 99164-6120, USA

* Correspondence: acousins@wsu.edu

Received 11 September 2020; Editorial decision 8 March 2021; Accepted 12 March 2021

Editor: Tracy Lawson, University of Essex, UK

Abstract

Plateauing yield and stressful environmental conditions necessitate selecting crops for superior physiological traits with untapped potential to enhance crop performance. Plant productivity is often limited by carbon fixation rates that could be improved by increasing maximum photosynthetic carboxylation capacity (V_{cmax}). However, V_{cmax} measurements using gas exchange and biochemical assays are slow and laborious, prohibiting selection in breeding programs. Rapid hyperspectral reflectance measurements show potential for predicting V_{cmax} using regression models. While several hyperspectral models have been developed, contributions from different spectral regions to predictions of V_{cmax} have not been clearly identified or linked to biochemical variation contributing to V_{cmax}. In this study, hyperspectral reflectance data from 350–2500 nm were used to build partial least squares regression models predicting *in vivo* and *in vitro* V_{cmax}. Wild-type and transgenic tobacco plants with antisense reductions in Rubisco content were used to alter V_{cmax} independent from chlorophyll, carbon, and nitrogen content. Different spectral regions were used to independently build partial least squares regression models and identify key regions linked to V_{cmax} and other leaf traits. The greatest V_{cmax} prediction accuracy used a portion of the shortwave infrared region from 2070 nm to 2470 nm, where the inclusion of fewer spectral regions resulted in more accurate models.

Keywords: Chlorophyll, hyperspectral reflectance, nitrogen, photosynthetic capacity, partial least squares regression, shortwave infrared, V_{cmax}.

Introduction

High-yielding crops will require improved or optimized physiological performance as environmental conditions become increasingly stressful. Improving the rate at which crop varieties can be accurately screened for physiological variation is necessary in order to make selection for a broader array of traits possible in modern breeding programs. One of the largest limitations to greater rates of carbon fixation by plants is maximum photosynthetic carboxylation capacity (V_{cmax}), which determines the rate-limiting step of photosynthetic carbon fixation. Crop productivity can often be limited by V_{cmax}, particularly under conditions constrained by water or nitrogen availability where sink strength does not limit plant growth. However, because measurements of V_{cmax} are too tedious and time consuming for breeding and crop improvement

programs, rapid high-throughput estimations have been proposed using hyperspectral reflectance data combined with machine learning approaches to develop regression models for prediction of V_{cmax} (Serbin *et al.*, 2012; Yendrick *et al.*, 2017; Silvia-Perez *et al.*, 2018). While many regression and machine learning approaches have been examined (Heckman *et al.*, 2018; Fu *et al.*, 2019), partial least squares regression (PSLR) has been widely used for physiological traits (Serbin *et al.*, 2012, 2014; Barnes *et al.*, 2017; Yendrick *et al.*, 2017; Silvia-Perez *et al.*, 2018; Meacham-Hensold *et al.*, 2019) because of its ability to avoid overfitting when faced with highly collinear data and greater numbers of unknown variables than observed values, both of which are common with hyperspectral reflectance measurements (Wold *et al.*, 2001).

Models predicting V_{cmax} and maximum electron transport capacity of several species generally include all wavelengths spanning 350–2500 nm across the visible (350–700 nm), near infrared (NIR; 700–1400 nm), and shortwave infrared (SWIR; 1400–2500 nm) regions, and the importance of each wavelength in the model varies (Serbin *et al.*, 2012; Ainsworth *et al.*, 2014; Yendrick *et al.*, 2017; Silvia-Perez *et al.*, 2018). Currently it is not well understood which physiological signals are being detected from hyperspectral reflectance to predict V_{cmax} (Meacham-Hensold *et al.*, 2019). Although models have been built where predictions of V_{cmax} are independent of chlorophyll content (Barner *et al.*, 2017; Meacham-Hensold *et al.*, 2019), many models predicting physiological traits consistently observe peaks of predicted model importance in the visible (400–700 nm) and red-edge (700–740 nm) regions where reflectance is often associated with pigment content (Barnes *et al.*, 2017; Yendrick *et al.*, 2017; Ely *et al* 2019; Meacham-Hensold *et al.*, 2019). It remains unclear what is driving the changes in reflectance that link to V_{cmax}, but the strong dependence on visible and red-edge reflectance (Barnes *et al.*, 2017; Ely *et al.*, 2019; Meacham-Hensold *et al.*, 2019) indicates that these predictions are detecting a link between reflectance and plant health because a change in red-edge reflectance is one of the first indications of plant stress (Dobrowski *et al.*, 2005). This implies that changes in leaf stress are coordinated with variation in leaf chemical content, structure, and photosynthetic rate, and may be partly responsible for reflectance-based predictions of physiological traits such as V_{cmax}.

These accurate predictions show great promise that V_{cmax} can be detected from reflectance data, although differences in the loading weights across the spectra between models indicate that different reflectance signatures are being detected under varying conditions. Isolating narrower signals related to V_{cmax} may increase the transferability of models, particularly if signals are independent of other traits that often co-vary with V_{cmax}. This is a needed improvement as reflectance-based predictions of V_{cmax} have not been accurate when models are applied to different datasets across species and conditions (Fu *et al.*, 2019). This raises the question of what differences in leaf composition (e.g. nitrogen content) or structure are present and responsible for the variation in spectral signals that are being extracted by regression models for prediction of V_{cmax}. Robust models that are accurate across different species and growth conditions could be developed once clearer links between reflectance and V_{cmax} are established in order to avoid the detection of confounding traits that may correlate with V_{cmax} but are actually independent.

To address the uncertainty of which spectral signals are uniquely linked to V_{cmax} and other physiological traits, we used tobacco plants with genetic modifications that altered V_{cmax} in order to avoid confounding physiological differences such as enzyme kinetics or conductance to carbon dioxide that often drive variation in V_{cmax} across species or conditions. Measurements of V_{cmax} were determined both *in vivo* and *in vitro* alongside measurements of chlorophyll, carbon, and nitrogen content and fresh leaf mass area (LMA). The separation of V_{cmax} from these other leaf traits did not appear to reduce the accuracy of modeling using hyperspectral reflectance to predict V_{cmax} or other leaf traits. Predictions of V_{cmax} with the greatest accuracy were made by building models using only reflectance from a portion of the SWIR region (1400–2500 nm).

Materials and methods

Plant material and growth conditions

Previously generated transgenic *Nicotiana tabacum* (tobacco) lines that were heterozygous (SSUX1) and homozygous (SSUX2) for the antisense gene had reduced Rubisco content and were used in this study (Hudson *et al.*, 1992). Plants were grown in TPBC-19 environmental growth chambers (BioChambers, Winnipeg, Manitoba, Canada) under high CO_2 at 3000 μmol mol^{-1} and 75% relative humidity to minimize growth differences between plant genotypes. All plants were grown in 4 liter pots in commercial potting mix (Sunshine LC-1, Sun Gro Horticulture Inc., Bellevue, WA, USA) and were fertilized weekly with Peters 20-20-20 (J.R. Peters, Allentown, PA, USA). Tobacco plants were grown under a 12 h diurnal cycle (26 °C day/22 °C night) with a daytime light intensity of 1000 μmol photons m^{-2} s^{-1} of photosynthetically active radiation (PAR). All measurements described below were conducted 30–45 d after planting.

Gas exchange measurements

Measurements of gas exchange were made using a LI-6800 (LI-COR Biosciences, Lincoln, NE, USA) with a 6 cm^2 fluorescence head (6800-01A). The youngest fully expanded leaf of each plant was chosen for gas exchange, where assimilation rate versus intercellular CO_2 concentration (A–C_i curves) were measured, beginning at 100 μmol mol^{-1} CO_2, increased by 100 until reaching 1000 μmol mol^{-1} CO_2, and then measured at 1300, 1600, and 2000 μmol mol^{-1} CO_2 totaling 13 discrete CO_2 concentrations. During A–C_i curves, leaves were illuminated with 1500 PAR at 25 °C and were subject to fluorescence flashes at each CO_2 concentration. Curves were performed on 65 biological replicates in total, including 26 wild-type (WT), 11 SSUX1, and 28 SSUX2 plants. Dark-adapted fluorescence measurements of all plants were made immediately prior to dawn, when the leaf was allowed to acclimate to the measuring beam in the chamber for 2 min and then maximum fluorescence was measured. Maximum carboxylation capacity (V_{cmax}) was modeled from the initial slope of A–C_i curves following Sharkey *et al.* (2007) where the V_{cmax} value that minimized the difference between measured and modeled data points is solved for. Kinetic constants used in the equation

solving for V_{cmax} were taken from von Caemmerer, (2000). The maximum photosynthetic rate (A_{max}) for each plant was determined as the photosynthetic rate at 2000 μmol mol^{-1} CO_2.

Hyperspectral reflectance acquisition and pre-processing

Hyperspectral reflectance data were measured using a HR-1024i spectroradiometer (Spectra Vista Corporation, Poughkeepsie, NY, USA) using a fiber optic light guide attachment connected to a leaf clip. Two reflectance measurements per plant were averaged, with a measurement made within 10 s prior to and following gas exchange on the same area of the leaf that was measured in the gas exchange chamber. The spectroradiometer was calibrated before measuring each plant using a Spectralon 99% diffuse reflectance standard (Labsphere, North Sutton, NH, USA). The range of the spectroradiometer was from 350 nm to 2500 nm, with discrete reflectance values measured every 1.5 nm from 350 mm to 1000 nm, 3.8 nm from 1000 nm to 1885 nm, and 2.5 nm from 1885 nm to 2500 nm. Since the spectroradiometer integrated three detectors, overlapping regions were auto-corrected using the software provided by the manufacturers. The spectra comprised 992 unique spectral reflectance data points from each measurement. These raw data were subject to several pre-processing steps. Firstly, Euclidean normalization was performed, then binning by averaging reflectance data at 5 nm intervals. Following these pre-processing steps, the final number of spectral reflectance data points for each measurement was 432 with reflectance data at every 5 nm. Normalization and binning were performed to smooth noise where detectors overlap and at the ends of the measured range (Schmidt and Skidmore, 2004; Wang and Sousa, 2009). These corrections had subtle effects on final modeled results.

Statistical analysis and model development

All statistical analysis was performed in RStudio (Version 0.98.1103). Normality was tested for all variables using the Shaprio–Wilks test, while physiological parameters were assessed using a one-way ANOVA. Significant differences in ANOVA were followed by Fisher's least significant difference test. Differences in the leaf responses to CO_2 concentrations measured with gas exchange were evaluated using a two-way ANOVA.

Physiological traits were first compared with commonly used spectral vegetation indices to identify correlations. These indices included the normalized difference vegetation index (NDVI; Sims and Gamon, 2002), the water index (WI; Peñuelas *et al.*, 1993), the chlorophyll/carotenoid index (CCI; Gamon *et al.*, 2016), the physiological reflectance index (PRI; Gamon, 1992), the plant senescence reflectance index (PSRI; Merzlyak *et al.*, 1999), the lignin cellulose absorption index (LCAI; Daughtry *et al.*, 2005), the normalized difference nitrogen index (NDNI; Serrano *et al.*, 2002), the Vogelmann1, (Vogelmann *et al.*, 1993), the red-edge NDVI (RENDVI; Sharma *et al.*, 2015), the normalized difference red-edge and normalized ratio indices using 660 nm and 1550 nm (NDRE and NRI; Herrmann *et al.*, 2010), the normalized dry matter index, and the normalized difference index for LMA (NDMI and ND_{LMA}, respectively, Cheng *et al.*, 2014).

Partial least squares regression (PLSR) models were then developed using the 'pls' package in R. Models for eight leaf traits were first developed with leave-one-out cross-validation (LOOCV) using from three to 20 components. Model performance using LOOCV was evaluated by R^2 values and the root mean squared error (RMSE) of predicted values. In LOOCV modeling, performance is assessed by systematic elimination of a single data point from the dataset (n times as sample size) to estimate prediction accuracy. Additionally, another model evaluation technique is to divide the dataset into a training dataset used for model development and a testing dataset for independent assessment of the developed model. Following LOOCV, more rigorous model evaluation was done using training and testing for each trait by 100 random divisions (iterations) of the dataset into an 80% training subset used for building a PLSR model followed by using the remaining 20% of the data for (independent) testing of model prediction accuracy. Pre-processed reflectance data across the entire spectra were used for initial model development followed later by the systematic removal of reflectance values from specific spectral regions in order to evaluate the impact on model performance. Models were developed using reflectance data from only the visible, then the NIR, and then the SWIR, followed by removal of each of these regions, with other regions remaining. Additional combinations of included wavelength regions were evaluated depending on model performance for each physiological variable. For each model, loading and coefficient data were extracted in addition to the calculation of variable importance in projection (VIP) scores using the provided supplemental code to the pls R package as described by Chong and Jun (2005).

Leaf biochemistry and chemical content

Following the 65 gas exchange measurements, leaf punches 0.79 cm^2 in area were taken of the same area of the leaf measured for gas exchange and reflectance, and were immediately frozen in liquid nitrogen and transferred to –80 °C storage until further measurement. Chl *a* and *b* were measured with an Evolution 300 UV-VIS spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) following the protocol of Ritchie (2006), where punches were placed in 95% ethanol for 48 h in the dark at 4 °C. Leaf carbon and nitrogen contents was determined using an elemental analyzer (ECS 4010, Costech Analytical, Valencia, CA, USA) following combustion. Maximum carboxylation capacity of the leaf punches was estimated from the rate of NADH consumption of a crude leaf extract (Ruuska *et al.*, 1998). Fresh LMA was determined by dividing the fresh weight of each leaf punch by the area.

Results

Transgenic plant characterization and gas exchange

Plants with reduced Rubisco content typically are small and chlorotic in greenhouse conditions, but when grown in chambers with elevated CO_2, WT and transgenic plants showed no differences in growth (Fig. 1). However, there were significant (P<0.01) differences in the *in vitro* V_{cmax} values for WT, SSUX1, and SSUX2 plants of 122.9±5.1, 46.2±2.7, and 29.1±2.5 μmol m^2 s^{-1}, respectively (P<0.01). The SSUX1 seed was generated from a heterozygous plant (Ruuska *et al.*, 2000) and segregated

Fig. 1. Representative tobacco plants photographed following gas exchange, reflectance measurements, and the collection of leaf tissue. Wild-type, heterozygous (SSUX1), and homozygous (SSUX2) transgenic plants were grown in ecological growth chambers at 3000 ppm CO_2.

with an ~1:2:1 ratio of *in vitro* V_{cmax} values similar to the WT, in between the WT and SSUX2, and similar to SSUX2.

Gas exchange measurements also produced clear in net rate of CO_2 assimilation between WT plants and the two transgenic lines (Fig. 2). WT plants had greater net assimilation rates (A_{net}) in response to intercellular CO_2 concentrations (C_i) than both transgenic lines ($P<0.01$) and SSUX1 had greater A_{net} than SSUX2 ($P<0.01$) (Fig. 2). The *in vivo* V_{cmax} values determined from fitting A–C_i response curves were well correlated with *in vitro* V_{cmax} ($R^2=0.86$, $P<0.01$, Fig. 3). The WT had the greatest *in vivo* V_{cmax} values that were significantly greater than those of transgenic genotypes, where SSUX1 had significantly greater *in vivo* V_{cmax} than SSUX2 ($P<0.01$, Fig. 4).

Leaf characteristics

Total leaf chlorophyll and nitrogen content showed minor but significant ($P<0.05$) increases in WT plants relative to SSUX2, while SSUX1 was not significantly different from either genotype. Carbon content was not significantly ($P<0.05$) different between any genotypes, and fresh LMA was significantly greater in SSUX1 plants relative to the WT and SSUX2. While nitrogen content was significantly correlated with carbon content ($R^2=0.41$, $P<0.05$, not shown) and V_{cmax} ($R^2=0.32$, $P<0.05$ for both *vivo* and *vitro*, Fig. 5A), no other significant relationships were present between V_{cmax} (determined either *in vivo* or *in vitro*) and other traits including LMA, chlorophyll, or carbon content (Fig. 5).

Hyperspectral reflectance and predictive modeling

Reflectance spectra of different genotypes appeared similar, with the exception of lower reflectance of SSU1X at ~1700 nm (Fig. 6). The spectral vegetation indices evaluated showed that chlorophyll content and the NDVI were well correlated ($R^2=0.68$, $P<0.01$), while fresh LMA was significantly correlated with the WI and LCAI ($R^2=0.54$ and 0.62, respectively; $P<0.01$). No other traits including *in vivo* and *in vitro* V_{cmax} correlated with any other of the commonly used vegetation indices evaluated here (see the Materials and methods).

Predictive PLSR modeling using entire hyperspectral reflectance data with LOOCV could accurately model maximum photosynthetic rate (A_{max}), chlorophyll content, LMA, and *in vivo* and *in vitro* V_{cmax}, with $R^2>0.75$, using eight components (Table 1). The importance of each wavelength across the spectra for impact on model performance was evaluated by

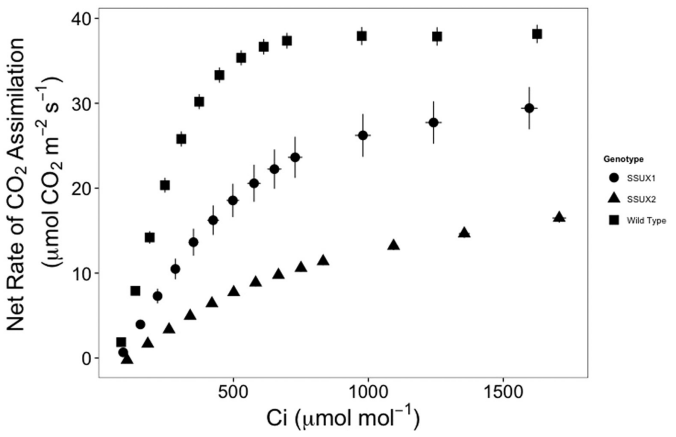

Fig. 2. The net rate of CO_2 assimilation (A_{net}) to increasing intercellular CO_2 concentration (C_i). Averages are shown for WT, heterozygous (SSUX1), and homozygous (SSUX2) transgenic tobacco plants, where each genotype include 26, 11, and 28 biological replicates respectively.

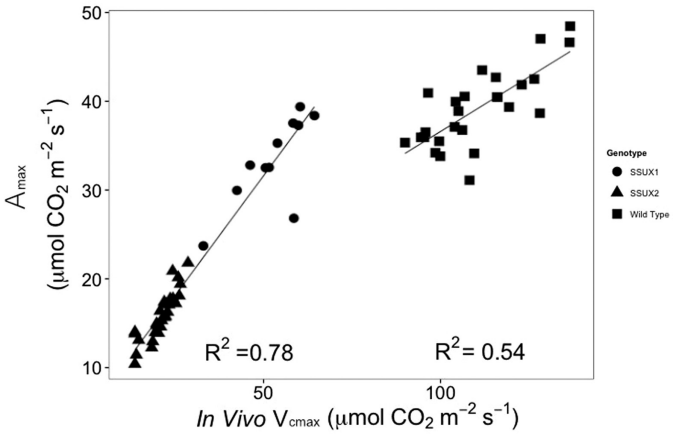

Fig. 3. The correlation between maximum photosynthetic rate determined from A–C_i curves and *in vivo* V_{cmax}. Regression lines are shown for transgenic (SSUX1 and SSUX2) and WT tobacco plants separately. Each point represents one biological replicate.

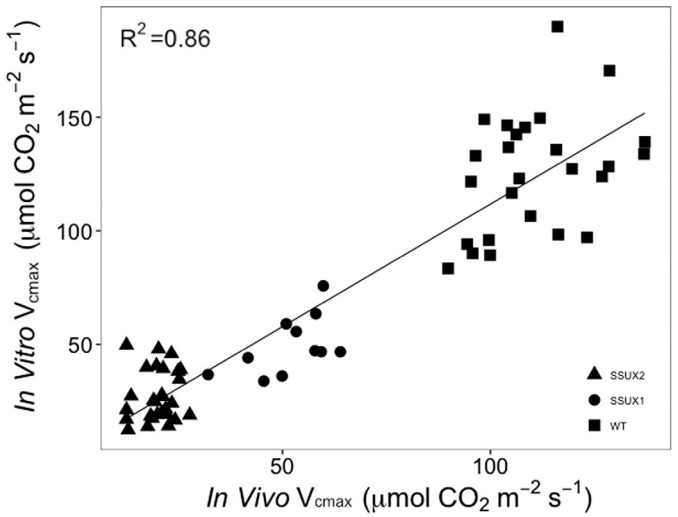

Fig. 4. The correlation observed between *in vivo* and *in vitro* V_{cmax}. A significant ($P<0.01$) correlation was present between *in vivo* V_{cmax} derived from A–C_i curves made at 1500 PAR and 25 °C, and *in vitro* V_{cmax} determined from spectrophotometric assays at 25 °C on crude extracts of leaf punches measured under saturating CO_2 where Rubisco was fully activated.

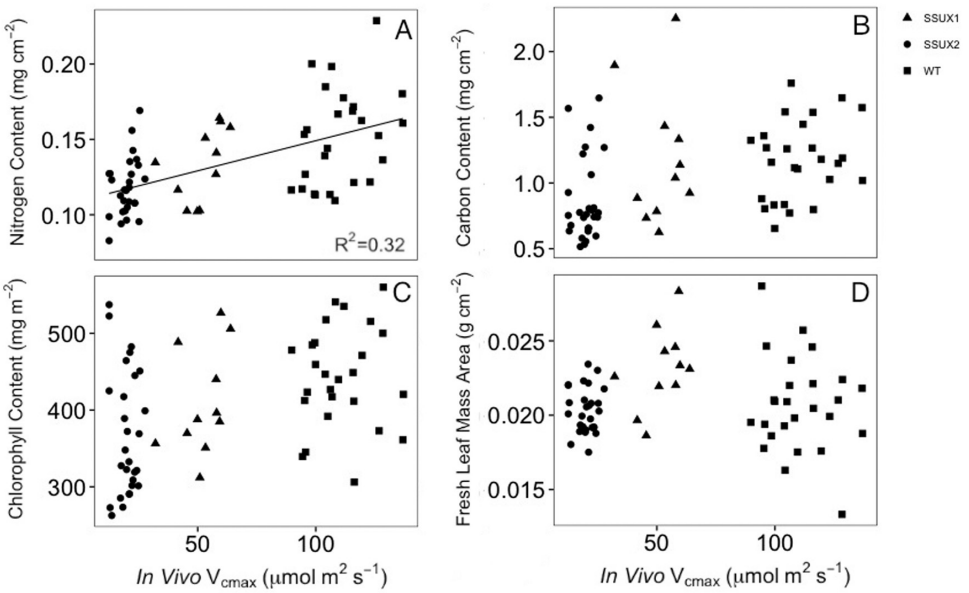

Fig. 5. Leaf traits and estimated Rubisco carboxylation capacity. Relationship between *in vivo* V_{cmax} and nitrogen content (A), carbon content (B), chlorophyll content (C), and fresh leaf mass (D) in WT, SSUX1, and SSUX2 plants. A significant correlation (P<0.01) was present between *in vivo* V_{cmax} and nitrogen content. No other variables had significant correlations.

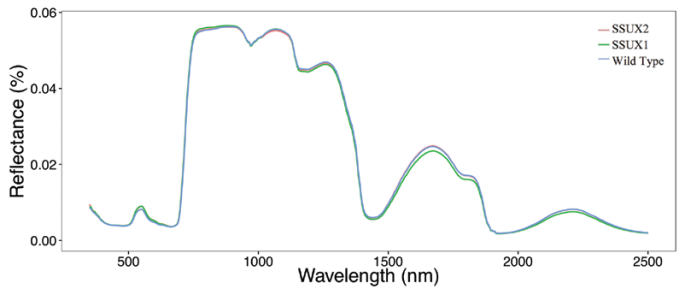

Fig. 6. Hyperspectral leaf reflectance spectra measured on WT and transgenic (SSUX1 and SSUX2) tobacco genotypes. Measurements were made across the visible, near infrared, and shortwave infrared wavelength spectra, while plants were in the growth chamber. Colored lines show the mean reflectance of each genotype where 432 individual points are smoothed and represented as a continuous line. Standard error of reflectance is not shown, and was only observable at this scale near the reflectance peak at 1700 nm.

examining coefficient and loading values as well as VIP scores (Fig. 7). Consistently,VIP scores showed that models were most dependent on reflectance in the red-edge region for all traits except fresh LMA. Modeled predictions of fresh LMA were more dependent on reflectance in the SWIR, a region typically associated with leaf thickness and water content. Similarly, loading and coefficient values also had peaks in the red edge for all traits, although these were not the only areas with apparent contributions to model performance (Fig. 7). Coefficient plots showed that models for chlorophyll and nitrogen content and V_{cmax} were most dependent on reflectance in the visible region, while loading plots indicated that the SWIR regions were of equal importance (Fig. 7). The accuracy of PLSR modeling to predict *in vivo* V_{cmax} (R^2=0.81), chlorophyll content (R^2=0.80), and LMA (R^2=0.82) was stronger than for nitrogen content R^2=0.59 (Fig. 8).

More rigorous evaluation of performance for models built with entire hyperspectral reflectance data using 80% training and 20% testing subsets showed decreased performance for all traits. Average performance of repeated training and testing for models built to predict chlorophyll content and *in vivo* V_{cmax} had an R^2 value of 0.55 for both traits, while models of *in vitro* V_{cmax}, nitrogen content, and LMA had lower average prediction abilities of R^2 of 0.49, 0.30, and 0.51, respectively (Table 1). However, selective inclusion of reflectance data from distinct spectral regions reduced the number of variables present during model development and increased model performance for both LOOCV and 80/20 training and testing for all traits other than A_{max}, stomatal conductance (g_s), and fresh LMA (Table 1). For example, prediction accuracy from 80/20 training and testing of *in vivo* V_{cmax} increased from an R^2 of 0.55 to 0.78 when reflectance data from the visible and NIR regions were excluded from the model (Table 1).

Predictions of gas exchange parameters including A_{max} and steady-state g_s were also modeled using hyperspectral reflectance data. Predictions using LOOCV models with entire hyperspectral data of A_{max} and g_s had high accuracy, with R^2 values of 0.86 and 0.71, respectively, greater than models made using any single region (Tables 1, 2). Using the entire spectral data, models with eight components evaluated using 80% training and 20% testing data yielded lower prediction accuracies, with R^2 of 0.36 for g_s and 0.70 for A_{max}. These more robust estimates of model performance were slightly improved (R^2=0.73 and 0.40 for A_{max} and g_s,

Table 1. Regression coefficient (R^2) values from partial least square regression (PLSR) modeling with eight components used for each model except leave-one-out cross-validated (LOOCV) models built with whole spectrum data

	Whole spectrum (350–2500 nm)		Selected regions for greatest model performance		
	LOOCV	80/20	Wavelengths present	LOOCV	80/20
Chlorophyll concentration (mg m^{-2})	0.77	0.55±0.18	500–1400 nm	0.77	0.64±0.18
Nitrogen content (%)	0.59	0.30±0.15	2000–2500 nm	0.76	0.38±0.19
Carbon content (%)	0.74	0.45±0.19	2000–2500 nm	0.81	0.55±0.20
Leaf mass area (g m^{-2})	0.81	0.51±0.20	2000–2500 nm	0.82	0.42±0.18
In vivo V_{cmax} (μmol CO_2 m^2 s^{-1})	0.81	0.55±0.17	2070–2470 nm	0.90	0.75±0.14
In vitro V_{cmax} (μmol CO_2 m^2 s^{-1})	0.76	0.49±0.19	2070–2470 nm	0.83	0.58±0.16
A_{max} (μmol CO_2 m^2 s^{-1})	0.86	0.70±0.18	350–1400 nm	0.85	0.73±0.14
g_s (μmol CO_2 m^2 s^{-1})	0.71	0.36±0.20	350–1400 nm	0.69	0.40±0.19

LOOCV and 80% training and 20% testing (80/20) models display the differences in predictive ability when validating model performance using different methods. Errors on 80/20 modeling represent the SD of 100 models developed and tested from randomized division of training and testing data. The left side of the table shows modeled results when the entire hyperspectral reflectance data were included in modeling, while the right shows when reflectance from only selected spectral regions was present during model development. The number of physiological measurements and corresponding reflectance spectra for each model was 65 for estimations of V_{cmax}, A_{max}, and g_s, and 68 for all other traits

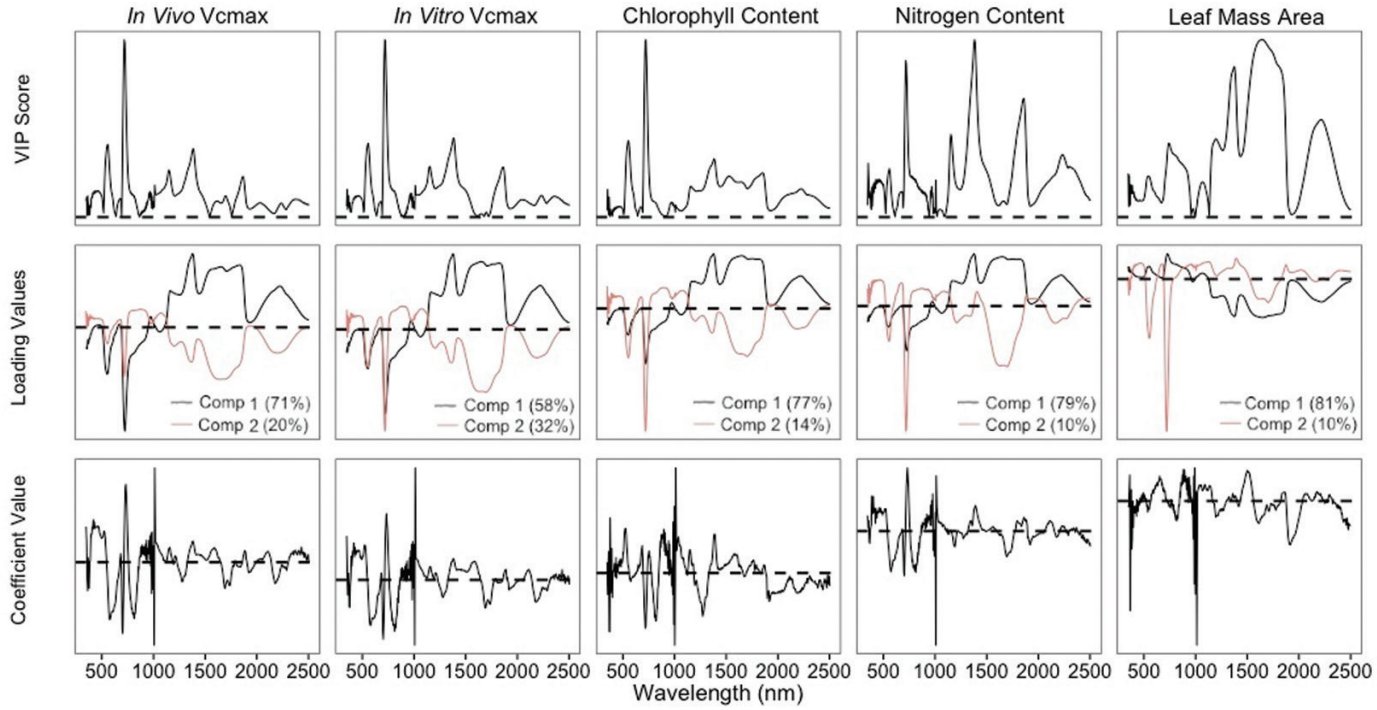

Fig. 7. Predicted importance of each variable on model performance for full spectrum leave-one-out cross-validated models. Models were generated for each of the five traits where plots show calculated variable importance projection (VIP) scores (top row), extracted loading values for the first two components (second row), and extracted coefficient values (bottom row). *y*-axis scales are variable for each plot, with axis values not shown. Differences in the scale of coefficient and loading values are due to differences in the magnitude of trait values being predicted. Zero is indicated in each graph with a dashed line, where plots are interpreted by identifying wavelengths that deviate farthest from zero, as reflectances at those wavelengths are expected to have the largest impact on model prediction.

respectively) when using reflectance in the visible and NIR regions only (Table 1).

Discussion

Leaf chemical composition can be predicted using hyperspectral reflectance data. For example, traits such as chlorophyll, nitrogen, and water content have inherent physical properties and interact with light to cause changes in reflectance at discrete wavelengths (Curran, 1989; Gamon *et al.*, 1992, 2016; Peñuelas *et al.*, 1993; Filella and Peñuelas, 1994; Gitelson *et al.*, 1996; Ceccato *et al.*, 2001). More recent work has shown that complex traits such as maximum carboxylation capacity (V_{cmax}), electron transport (Serbin *et al.*, 2012; Barnes *et al.*, 2017; Yendrick *et al.*, 2017), and

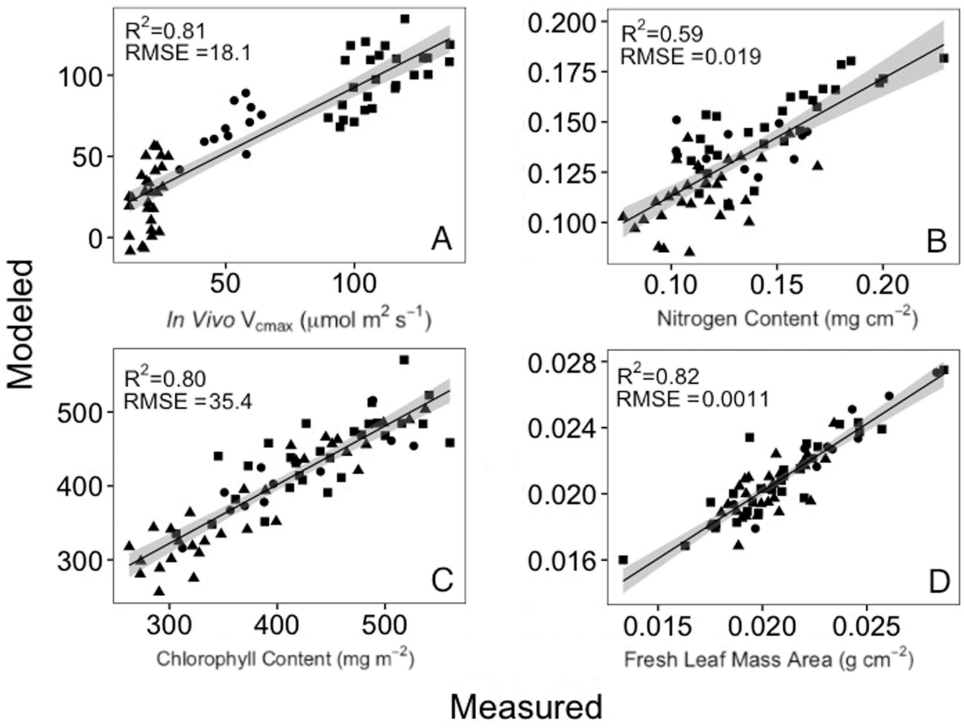

Fig. 8. Measured versus modeled values for *in vivo* V_{cmax} (A), nitrogen content (B), chlorophyll content (C), and fresh leaf mass area (D). Partial least square regression models for each trait were made using leave-one-out cross-validation with entire hyperspectral reflectance data and eight components. Linear regression lines show the 95% confidence interval.

Table 2. Regression coefficient (R^2) values from partial least square regression (PLSR) modeling of selected spectral regions using leave-one-out cross-validation (LOOCV) and eight components for all models shown

	LOOCV model performance for selected spectral regions		
	Visible (350–700 nm)	**NIR (700–1400 nm)**	**SWIR (1400–2500 nm)**
Chlorophyll concentration (mg m^{-2})	0.75	0.76	0.64
Nitrogen content (%)	0.35	0.50	0.59
Carbon content (%)	0.56	0.71	0.78
Leaf mass area (g m^{-2})	0.50	0.77	0.79
In vivo V_{cmax} (μmol CO_2 m^2 s^{-1})	0.60	0.62	0.73
In vitro V_{cmax} (μmol CO_2 m^2 s^{-1})	0.55	0.60	0.69
A_{max} (μmol CO_2 m^2 s^{-1})	0.60	0.74	0.70
g_s (μmol CO_2 m^2 s^{-1})	0.53	0.59	0.60

The number of physiological measurements and corresponding reflectance spectra for each model was 65 for estimations of V_{cmax}, A_{max}, and g_s, and 68 for all other traits

even yield can be predicted from reflectance data using regression approaches (el-Hendawy *et al.*, 2019). However, it is not clear which structural, compositional, or stress status differences are being used to predict traits via spectral reflectance analyses. Here we used reflectance to accurately predict a wide range of V_{cmax} data determined both *in vivo* and *in vitro* in a genetically altered population in order to identify spectral reflectance signatures uniquely related to V_{cmax} while controlling for nitrogen and chlorophyll content, as these traits are often closely associated with V_{cmax}. Reflectance models were improved by reducing the number of wavelengths used, indicating that models for physiological traits should be made using only relevant spectral information. Models with the highest V_{cmax} prediction accuracy used only SWIR reflectance, indicating links between V_{cmax} and internal leaf structure and composition.

Predicting chlorophyll and nitrogen content from reflectance

To assess the ability of reflectance data to correctly predict traits with well-established reflectance characteristics, and to ensure that models of other leaf traits were independent of

V_{cmax}, chlorophyll and nitrogen content were first modeled using the reflectance data (Fig. 7). This also allowed for the evaluation of the effect of different model validation methods and inclusion of different spectral regions on model performance of traits with known spectral properties.

Although leaf chlorophyll and nitrogen content showed a narrow range of measured values, these were still accurately predicted with LOOCV using the entire spectrum of the hyperspectral reflectance data (R^2=0.8 and 0.59, respectively, Fig. 8). These relationships appeared to be primarily dependent on light reflected from the visible and red-edge regions for both traits (Fig. 7). Determining which regions of the spectra have the greatest impact on model performance is often done by examining loading values as well as regression coefficients across all wavelengths to identify the relative contribution of each wavelength to the final predicted value (Barnes *et al.*, 2017; Silvia-Perez *et al.*, 2018; Meachan-Hensold *et al.*, 2019). Additionally, VIP scores are also often calculated using the weight and variance of each predictor in the model to provide an estimate of the relative importance of each wavelength in overall model performance (Serbin *et al.*, 2014; Ely *et al.*, 2019). Alternatively, removing regions of the spectra from the data provided to the model allows the impact on performance to be evaluated for different spectral regions with less error in interpretation. Here we developed PLSR models with different regions of the hyperspectral spectral data included in order to identify which regions contained the most valuable input for models to make robust and accurate predictions.

Chlorophyll content has been shown to be mechanistically linked to spectral reflectance in the green and red-edge regions as a result of light absorption (Gitelson *et al.*, 1996, 2003), and this was predicted by high VIP scores as well as coefficient and loading values in the reflectance model developed here to predict chlorophyll content. In agreement with these loading values, VIP scores, and coefficient values, the red-edge region was important for model performance, and removal of this region resulted in reduced model accuracy for chlorophyll content. Contrastingly, the NIR region contained low VIP scores and loading values, but removal of this region decreased prediction ability of chlorophyll. In the SWIR region, high VIP scores and loading values were present, but predictions of chlorophyll were more accurate after removing this region. The variability in predicted versus observed importance for each region was also apparent for nitrogen where, despite large predicted importance, removing the visible and NIR regions improved model performance (Table 1). Additionally, the SWIR region (1400–2500 nm) had low predicted model importance by VIP scores and coefficient values, but using reflectance only from this region resulted in the model with the highest prediction ability. This agrees with other work that has observed nitrogen having the greatest effect on reflectance in the SWIR (Curran *et al.*, 1989; Smith, 2003). By examining model performance through inclusion or removal of distinct spectral regions, it was made clear which regions contained information allowing for accurate prediction of each trait. These regions were known from previous work that has established links between reflectance and the chemical properties of chlorophyll and nitrogen (Curan *et al.*, 1989; Filella and Peñuelas, 1994; Merzlyak *et al.*, 1999; Gitelson *et al.*, 2003; Smith, 2003) but could not be reliably predicted by VIP scores, loading values, or coefficient values.

Evaluating the performance of PLSR models is commonly done using LOOCV where a single estimate of model performance is made by systematically removing a single value from the dataset to evaluate prediction accuracy. Such approaches are implemented for smaller datasets, or when variability within a dataset is high. Alternatively, a more robust and unbiased method of evaluating model prediction ability, but requiring a greater number of observations, is to divide the dataset into training and testing subsets (Barnes *et al.*, 2017). The majority of observations are used for training where a model is developed, while the remaining observations, usually 10–25%, are used for testing (independent validation). The predicted values from the testing subset are compared with known values withheld from the model. While slight reductions in performance are expected with this approach due to the reduction in sample size used for testing, large differences in performance parameters between these two approaches also indicate that models generated using LOOCV may be overfit. In this study, models using entire hyperspectral data for prediction of chlorophyll and nitrogen content showed high prediction accuracy using LOOCV, but had lower accuracy when evaluated with 80/20 training and testing. Reducing the model complexity by only using specific wavelength regions for regression and reducing the number of components in the model did not change accuracy for chlorophyll using LOOCV, but did increase 80/20 training and testing performance for both chlorophyll and nitrogen content (Table 1). The improvements in accuracy of models excluding extraneous spectral information probably resulted from reducing model complexity. This highlights the importance of identifying discrete spectral regions that are related to the physiological signal of interest, demonstrating the importance of both limiting the number of wavelengths for model development and using evaluation methods more rigorous than LOOCV. We expect these models to better maintain accuracy across diverse and fluctuating conditions such as those in the field.

Predicting V_{cmax} from reflectance

With few exceptions (Serbin *et al.*, 2012), models generated to predict complex metabolic traits such as V_{cmax} generally use reflectance data across the entire measured spectra from 350 nm to 2500 nm (Ainsworth *et al.*, 2014; Serbin *et al.*, 2014; Barnes *et al.*, 2017; Ely *et al.*, 2019; Meacham-Hensold *et al*, 2019). Similarly, modeling of *in vivo* and *in vitro* V_{cmax} using the entire hyperspectral region in this study produced models with high accuracy where R^2 values were 0.81 and 0.76 using

LOOCV, respectively. When we used the model developed by Serbin *et al.* (2012) for V_{cmax} estimation using data acquired in this study, the predicted values were not correlated with measured data. Furthermore, within this study, more rigorous evaluation of model performance determined by dividing data into 80% training and 20% testing subsets resulted in decreased model performance for *in vivo* and *in vitro* V_{cmax} to 0.55 and 0.49, respectively. This further questions the ability of models constrained by LOOCV evaluation to detect variation from datasets collected under even slightly different conditions.

Model performance was improved by building a PLSR model using only selected spectral regions (Table 1); specifically only including reflectance in a portion of the SWIR from 2070 nm to 2470 nm. This was unpredicted because examination of loading, coefficient, and VIP values from full spectrum data indicated that the red-edge and NIR regions were most important to model performance (Fig. 7), and removal of these regions is expected to reduce prediction accuracy. Here the model using entire spectra was more dependent on reflectance in the visible and NIR regions, where stronger correlations with V_{cmax} (Fig. 9) but less variation (Fig. 7) were present. However, forcing the model to use reflectance only in the SWIR region where more subtle correlations, but greater variation, were present allowed for greater prediction accuracy. The entire hyperspectral reflectance model's dependence on visible and NIR reflectance agrees with most other models using reflectance to predict V_{cmax} (Barnes *et al.*, 2017;Yendrick *et al.*, 2017; Meacham-Hensold *et al.*, 2019; Ely *et al.*, 2019), while a smaller number of V_{cmax} models show large peaks of predicted model dependence in the SWIR region that support

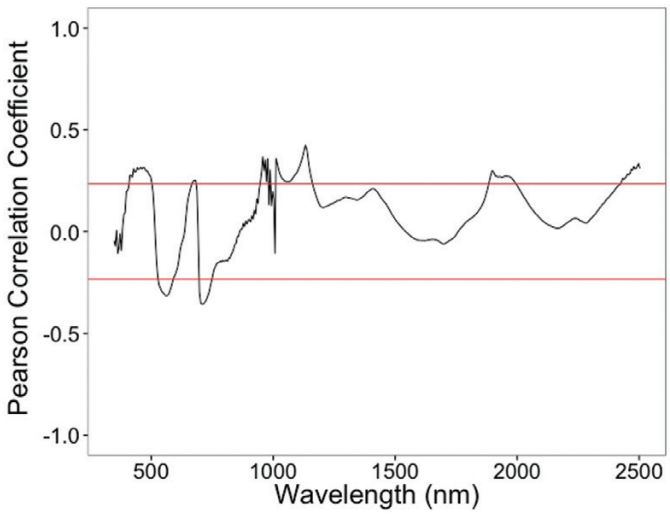

Fig. 9. Correlation between *in vivo* V_{cmax} and spectral reflectance at each wavelength across all genotypes (n=65). A total of 432 correlations were examined at discrete wavelengths from 350 nm to 2500 nm, shown here as a continuous line. Red lines represent the threshold for significant correlations (P<0.05), where values >0.23 and < –0.23 indicate a significant correlation between V_{cmax} and reflectance values across all measured plants.

the observations here where reflectance in the SWIR region was informative of V_{cmax} (Ainsworth *et al.*, 2014; Silvia-Perez *et al.*, 2018).

In addition to V_{cmax}, the accuracy of predicting nitrogen and carbon content increased when using reflectance from the SWIR region only. However, these models did not appear synonymous as V_{cmax} data were not correlated with carbon content and only mildly correlated with nitrogen content. Although some variation in modeled predictions of V_{cmax} may be partly linked with nitrogen content, especially considering the impact of nitrogen on reflectance in the SWIR region, the majority of variation in V_{cmax} could not be explained by nitrogen. This work is in agreement with results published by Meacham-Hensold *et al.*, (2019), where efforts were made to ensure that predictions of V_{cmax} from reflectance were independent of nitrogen content.

We propose that reflectance in the SWIR region is a key component of successfully modeling variation in V_{cmax} primarily as a result of the link between internal leaf structure and reflectance. A portion of the relationship between reflectance and V_{cmax} is also likely to be a result of dry matter content due to the links between reflectance in the SWIR region and carbon and nitrogen content, as well as the small portion of V_{cmax} variance correlating with nitrogen and correctly predicted by these models. Reflectance in SWIR is generally thought to be most highly dependent on absorbance by water content, since strong water absorbance peaks occur at 1450, 1940, and 2500 nm. However, differences in leaf structure and composition have been associated with reflectance in the SWIR (Jacquemoud *et al.*, 1996; Ceccato *et al.*, 2001; Ollinger, 2011). For example, changes in SWIR reflectance have been linked to variation in internal structure and dry mass content (Ollinger, 2011), and it is likely that accurate predictions of V_{cmax} from SWIR reflectance were due to changes in composition, arrangement, and internal leaf structure that changed in response to different demands for carbon dioxide.

Additional insights on the association between carboxylation capacity and reflectance came from the difference in prediction ability between reflectance models of *in vivo* and *in vitro* determinations of V_{cmax}. Even though *in vivo* and *in vitro* V_{cmax} data were well correlated (R^2=0.87, P<0.01), predictions using 80/20 training and testing from reflectance data were higher at R^2 =0.78 for *in vivo* measurements compared with R^2=0.58 for *in vitro* measurements. The greater accuracy of the *in vivo* data indicates that reflectance predictions of carboxylation capacity are at least partly linked to physiological status. This is because *in vitro* determinations of V_{cmax} are separated from the conditions experienced in the leaf as measurements occur after Rubisco has been fully activated at saturating CO_2 and constant pH, which probably do not represent the same environment within the leaf and did not vary between plants as may have been likely. Thus, measured reflectance signals appear be more informative of physiological status, as opposed to the true maximum

carboxylation capacity of Rubisco. Measurements of reflectance are probably capturing optical variation that is coordinated with changes in physiological activity such as enzyme activation and conductance to CO_2. On longer time scales, these can include pigment content and leaf thickness, but here we speculate that shorter scale responses also resulted in reflectance changes in the SWIR. We propose that detected changes here may also include accumulation of water or solutes that play roles in shifting cell and plastid shape and size, as well as movement of organelles within cells. As mentioned previously, these coordinated changes with carboxylation capacity and internal structure may have been partly responsible for the accurate predictions of V_{cmax} using SWIR reflectance.

The challenges presented in trying to discriminate reflectance signals between related physiological parameters are well illustrated with A_{max} and *in vivo* V_{cmax}. These two traits have similar values and were well correlated across WT and transgenic plants (Fig. 3), despite having a different physiological basis. This caused a notable shift in the relationship between A_{max} and V_{cmax}, as A_{max} increased linearly with Rubisco content in transgenic plants, while in WT plants A_{max} appeared limited by the maximum rate of electron transport (Fig. 1). Prediction accuracy of both V_{cmax} and A_{max} was high, and PLSR models were dependent on entirely different wavelengths (Table 1). Predictions of V_{cmax} were greatest using only reflectance at the far end of the SWIR region, while the best predictions of A_{max} used reflectance only in the visible and NIR regions of the spectra. Predictions of A_{max} were heavily influenced by visible reflectance and probably linked to pigment content. This specifically may have increased prediction ability for WT plants, where limitations to A_{max} result from electron transport capacity that is largely determined by the content of various pigments. When modeling A_{max} and V_{cmax} values using whole-spectrum data, wavelengths are primarily selected in the visible region, resulting in the creation of nearly synonymous models. However, completely independent models that improve accuracy of each trait can be developed, but only when using stringent criteria to manually restrict reflectance wavelengths available for regression.

Conclusions

This work demonstrated that variation occurring in V_{cmax} independent of other leaf traits could be detected using spectral reflectance data. While prediction accuracy was high using entire hyperspectral data, the most accurate predictions using reflectance were derived solely from a portion of the SWIR region, 2070–2470 nm. Improved prediction accuracy using reflectance in the SWIR region could not be predicted based on loading, coefficient, or VIP values from modeling performed on the entire spectra. Despite many studies successfully building regression models to predict many complex leaf-level traits such as V_{cmax} using reflectance, an understanding of which reflectance wavelengths contribute to observed physiological variation has been convoluted. Here, we demonstrated that limiting the amount of spectral data included in a model allowed for accurate predictions that performed better under rigorous validation and were able to more clearly link physiological variation to hyperspectral signals. Future work isolating spectral regions that provide the most accurate predictions of V_{cmax} in other species and conditions are needed in order to develop a widely transferable model.

Acknowledgements

The authors are grateful to Benjamin Harrow for the assistance in processing and measuring leaf samples for nitrogen and carbon content.

Author contributions

TS: performing experiments, data collection, and analysis. TS, SS, AC: writing and approval.

Conflict of interest

The authors declare no conflicts of interest.

Data availability

The data supporting the findings of this study are available on Zenodo: https://zenodo.org/record/4480331#.YBSNFSMrJdh

References

Ainsworth EA, Serbin SP, Skoneczka JA, Townsend PA. 2014. Using leaf optical properties to detect ozone effects on foliar biochemistry. Photosynthesis Research **119**, 65–76.

Barnes ML, Breshears DD, Law DJ, Van Leeuwen WJD, Monson RK, Fojtik AC, Barron-Gafford GA, Moore DJP .2017. Beyond greenness: detecting temporal changes in photosynthetic capacity with hyperspectral reflectance data. PLoS One **12**, e0189539.

Ceccato P, Flasse S, Tarantola S, Jacquemoud S, Grégoire J-M. 2001. Detecting vegetation leaf water content using reflectance in the optical domain. Remote Sensing of Environment **77**, 22–33.

Cheng T, Rivard B, Sánchez-Azofeifa AG, Féret J-B Jacquemoud S, Ustin SL. 2014. Deriving leaf mass per area (LMA) from foliar reflectance across a variety of plant species using continuous wavelet analysis. ISPRS Journal of Photogrammetry and Remote Sensing **87**, 28–38.

Chong IG, Jun C-H. 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems **78**, 103–112.

Curran PJ. 1989. Remote sensing of foliar chemistry. Remote Sensing of Environment **30**, 271–278.

Daughtry C, Hunt E Jr, Doraiswamy P, McMurtrey J III. 2005. Remote sensing the spatial distribution of crop residues. Agronomy Journal **97**, 864–871.

Dobrowski SZ, Pushnik JC, Zarco-Tejada PJ, Ustin SL. 2005. Simple reflectance indices track heat and water stress-induced changes in steady-state chlorophyll fluorescence at the canopy scale. Remote Sensing of Environment **97**, 403–414.

El-Hendawy SE, Al-Suhaibani NA, Elsayed S, Hassan WM, Dewir YHa Refay Y, Abdella KA. 2019. Potential of the existing and novel spectral reflectance indices for estimating the leaf water status and grain yield of spring wheat exposed to different irrigation rates. Agricultural Water Management **217**, 356–373.

Ely KS, Burnett AC, Lieberman-Cribbin W, Serbin SP, Rogers A. 2019. Spectroscopy can predict key leaf traits associated with source–sink balance and carbon–nitrogen status. Journal of Experimental Botany **70**, 1789–1799.

Filella I, Peñuelas, J. 1994. The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. International Journal of Remote Sensing **15**, 1459–1470.

Fu P, Meacham-Hensold K, Guan K, Bernacchi CJ. 2019. Hyperspectral leaf reflectance as proxy for photosynthetic capacities: an ensemble approach based on multiple machine learning algorithms. Frontiers in Plant Science **10**, 730.

Gamon JA, Huemmrich KF, Wong CYS, Ensminger I, Garrity S, Hollinger DY, Noormets A, Peñuelas J. 2016. A remotely sensed pigment index reveals photosynthetic phenology in evergreen conifers. Proceedings of the National Academy of Sciences, USA **113**, 13087–13092.

Gamon JA, Peñuelas J, Field CB. 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sensing of Environment **41**, 35–44.

Gitelson AA, Gritz Y, Merzlyak MN. 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of Plant Physiology **160**, 271–282.

Gitelson AA, Merzlyak MN, Lichtenthaler HK. 1996. Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. Journal of Plant Physiology **148**, 501–508.

Heckmann D, Schlüter U, Weber APM. 2017. Machine learning techniques for predicting crop photosynthetic capacity from leaf reflectance spectra. Molecular Plant **10**, 878–890.

Herrmann I, Karnieli A, Bonfil DJ, Cohen Y, Alchanatis V. 2010. SWIR-based spectral indices for assessing nitrogen content in potato fields. International Journal of Remote Sensing **31**, 5127–5143.

Hudson GS, Evans JR, von Caemmerer S, Arvidsson YBC, Andrews TJ. 1992. Reduction of ribulose-1,5-bisphosphate carboxylase/oxygenase content by antisense RNA reduces photosynthesis in transgenic tobacco plants. Plant Physiology **98**, 294–302.

Jacquemoud S, Ustin SL, Verdebout J, Schmuck G, Andreoli G, Hosgood B. 1996. Estimating leaf biochemistry using the PROSPECT leaf optical properties model. Remote Sensing of Environment **56**, 194–202.

Meacham-Hensold K, Montes CM, Wu J, *et al.* 2019. High-throughput field phenotyping using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity. Remote Sensing of Environment **231**, 111176.

Merzlyak MN, Gitelson AA, Chivkunova OB, Rakitin VY. 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiologia Plantarum **106**, 135–141

Ollinger SV. 2011. Sources of variability in canopy reflectance and the convergent properties of plants. New Phytologist **189**, 375–394.

Peñuelas J, Filella I, Biel C, Serrano L, Save R. 1993. The reflectance at the 950–970 nm region as an indicator of plant water status. International Journal of Remote Sensing **14**, 1887–1905.

Peñuelas J, Gamon JA, Fredeen AL. Merino J, Field CB. 1994. Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves. Remote Sensing of Environment **48**, 135–146.

Ritchie RJ. 2006. Consistent sets of spectrophotometric chlorophyll equations for acetone, methanol and ethanol solvents. Photosynthesis Research **89**, 27–41.

Ruuska S, Andrews T, Badger M, Hudson G, Laisk A, Price G, von Caemmerer S. 1998. The interplay between limiting processes in C3 photosynthesis studied by rapid-response gas exchange using transgenic tobacco impaired in photosynthesis. Australian Journal of Plant Physiology **25**, 859– 870.

Ruuska SA, Badger MR, Andrews TJ, von Caemmerer S. 2000. Photosynthetic electron sinks in transgenic tobacco with reduced amounts of Rubisco: little evidence for significant Mehler reaction. Journal of Experimental Botany **51**, 357–368.

Schmidt KS, Skidmore AK. 2004. Smoothing vegetation spectra with wavelets. International Journal of Remote Sensing **25**, 1167–1184.

Serbin SP, Dillaway DN, Kruger EL, Townsend PA. 2012. Leaf optical properties reflect variation in photosynthetic metabolism and its sensitivity to temperature. Journal of Experimental Botany **63**, 489–502.

Serbin SP, Singh A, McNeil BE, Kingdon CC, Townsend PA. 2014. Spectroscopic determination of leaf morphological and biochemical traits for northern temperate and boreal tree species. Ecological Applications **24**, 1651–1669.

Serrano L, Penuelas J, Ustin S. 2002. Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: decomposing biochemical from structural signals. Remote Sensing of Environment **81**, 355–364.

Sexton T, Sankaran S, Cousins A. 2021. Data from: Predicting photosynthetic capacity in tobacco using shortwave infrared spectral reflectance. Zenodo https://zenodo.org/record/4480331#.YBSNFSMrJdh

Sharkey TD, Bernacchi CJ, Farquhar GD, Singsaas EL. 2007. Fitting photosynthetic carbon dioxide response curves for C3 leaves. Plant, Cell & Environment **30**, 1035–1040.

Sharma LK, Bu H, Denton A, Franzen DW. 2015. Active-optical sensors using red NDVI compared to red edge NDVI for prediction of corn grain yield in North Dakota, U.S.A. Sensors (Basel, Switzerland) **15**, 27832–27853.

Silva-Perez V, Molero G, Serbin SP, Condon AG, Reynolds MP, Furbank RT, Evans JR. 2018. Hyperspectral reflectance as a tool to measure biochemical and physiological traits in wheat. Journal of Experimental Botany **69**, 483–496.

Sims DA, Gamon JA. 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sensing of Environment **81**, 337–354.

Smith M-L, Martin ME, Plourde L, Ollinger SV. 2003. Analysis of hyperspectral data for estimation of temperate forest canopy nitrogen concentration: comparison between an airborne (AVIRIS) and a spaceborne (Hyperion) sensor. IEEE Transactions on Geoscience and Remote Sensing **41**, 1332–1337.

Vogelmann JE, Rock BN, Moss DM. 1993. Red edge spectral measurements from sugar maple leaves. International Journal of Remote Sensing 14, 1563–1575.

von Caemmerer S. 2000. Biochemical models of leaf photosynthesis. Collingwood: CSIRO Publishing.

Wang LE, Sousa WP. 2009. Distinguishing mangrove species with laboratory measurements of hyperspectral leaf reflectance. International Journal of Remote Sensing **30**, 1267–1281.

Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems **58**, 109–130.

Yendrek CR, Tomaz T, Montes CM, Cao Y, Morse AM, Brown PJ, McIntyre LM, Leakey ADB, Ainsworth EA. 2017. High-throughput phenotyping of maize leaf physiological and biochemical traits using hyperspectral reflectance. Plant Physiology **173**, 614–626.