Received: 16 August 2021 | Accepted: 18 April 2022

DOI: 10.1002/ppj2.20047

The Plant Phenome Journal OPEN ACCESS

ORIGINAL RESEARCH

Fourier-transform infrared spectroscopy (FTIR) as a high-throughput phenotyping tool for quantifying protein quality in pulse crops

Amod Madurapperumage[1] | **Nathan Johnson**[1] | **Pushparajah Thavarajah**[1] | **Leung Tang**[2] | **Dil Thavarajah**[1]

[1]Dep. of Plant and Environmental Sciences, Clemson Univ., 113 Biosystems Research Complex, Clemson, SC 29634, USA

[2]Agilent Technologies, Harwell Campus, Becquerel Ave., Didcot OX11 0RA, UK

Correspondence
Dil Thavarajah, Dep. of Plant and Environmental Sciences, Clemson Univ., 113 Biosystems Research Complex, Clemson, SC 29634, USA.
Email: dthavar@clemson.edu

Assigned to Associate Editor Seth Murray.

Abstract

Fourier-transform mid-infrared (FT-MIR) spectroscopy is a high-throughput, cost-effective method to quantify nutritional traits, such as total protein and sulfur-containing amino acid (SAA) concentrations, in plant matter. This study used the spectroscopic technique FT-MIR coupled with attenuated total internal reflectance sampling interface to develop multivariate models for total protein concentration in chickpea (*Cicer arietinum* L.), dry pea (*Pisum sativum* L.), and lentil (*Lens culinaris* Medik.), in addition to SAA concentration in lentil. Total nitrogen data from combustion analysis and SAA data from high-performance liquid chromatography analysis following acid hydrolysis were used for model calibration and validation. Models for the total protein concentration of chickpea (calibration root mean square error [RMSE] = 0.093, R^2 = 0.948, prediction RMSE = 0.10), dry pea (calibration RMSE = 0.096, R^2 = 0.845, prediction RMSE = 0.093), and lentil (calibration RMSE = 0.13, R^2 = 0.845, prediction RMSE = 0.11) utilized infrared regions associated with protein structures, namely amide bands A, I, and II. In sulfur-related models for lentil total SAA (calibration RMSE = 0.014, R^2 = 0.827, prediction RMSE = 0.022) and methionine (calibration RMSE = 0.0075, R^2 = 0.815, prediction RMSE = 0.014) models utilized the C-S and S-CH_3 stretching and bending bands. Study findings support the conclusion that FT-MIR spectroscopy is a promising high-throughput and cost-effective phenotyping technique that will allow quantifying protein traits quickly and easily in pulse crops.

Abbreviations: AA, amino acids; ATR, attenuated total reflectance; FIR, far-infrared; FT, Fourier-transform; FTIR, Fourier-transform infrared spectroscopy; FT-MIR, Fourier-transform mid-infrared; HPLC, high-performance liquid chromatography; IR, infrared; MAS, marker-assisted selection; MIR, mid-infrared; NIR, near-infared; PLS, partial least squares; QTL, quantitative trait loci; RMSE, root means square error; SAA, sulfur-containing amino acids; ZFF, zero-fill factor.

1 | INTRODUCTION

Pulse crops, such as chickpea (*Cicer arietinum* L.), dry pea (*Pisum sativum* L.), and lentil (*Lens culinaris* Medik.), are an essential part of the global food system to provide plant-based protein, low digestible carbohydrates, and a range of micronutrients (Foyer et al., 2016; Johnson et al., 2020). These staple crops are increasing in popularity as plant-based protein sources—a trend expected to continue based on many factors such as health benefits and climate change (Graça et al., 2019; Kim et al., 2019; Pimentel & Pimentel, 2003). Pulses tend to be low in sulfur-containing amino acids (SAA) (Boye et al., 2012), so varieties high in methionine and cystine are a vital breeding objective to increase the protein quality in plant-based diets. However, measuring the concentration of amino acids (AA), particularly SAA, is challenging, as they are susceptible to acid degradation and thus require an additional protective oxidation step. A typical method takes two to three days for sample digestion before AA quantification. Instruments to measure AA concentrations, such as high-performance liquid chromatography (HPLC), are generally low-throughput, expensive, time-consuming, and require highly skilled operators. Quantitative Fourier-transform mid-infrared (FT-MIR) spectroscopy methods offer a promising alternative to conventional methods for analyzing protein and SAA. Samples can be analyzed in seconds without the chemicals and consumables required by traditional techniques.

Infrared (IR) is a low-energy region in the electromagnetic spectrum extending from 12,800 to 10 cm^{-1} (Skoog et al., 2016) and consists of the near-infrared (NIR; 12,800–4,000 cm^{-1}), mid-infrared (MIR; 4,000–200 cm^{-1}), and far-infrared (FIR; 200–10 cm^{-1}) spectrums (Skoog et al., 2016). Infrared spectroscopy using interferometers coupled with Fourier-transform (FT) algorithms are termed Fourier-transform infrared spectroscopy (FTIR) instruments and have several advantages over previous dispersive spectroscopy instruments, including (a) greater energy intensity due to the lack of slits and fewer optics to attenuate the source radiation (mechanically simpler), known as Jacquinot's (throughput) advantage; (b) simultaneous collection of multiple wavelengths (without the need for scanning), resulting in a shorter collection time and consequent increases in the signal-to-noise ratio, known as Fellgett's (multiplex) advantage; and (c) increased wavenumber accuracy inherent to the internal laser calibration and interferometer, enabling multiple scans to be collected *and* averaged, known as Connes' advantage (Perkins, 1987; Skoog et al., 2016). Fourier-transform instruments in the near, mid, and far regions probe high-frequency oscillations (vibrational overtones), fundamental vibrational modes, and low energy vibrations (Berthomieu & Hienerwadel, 2009; Capuano & van Ruth, 2015; El Khoury & Hellwig, 2017). However, the fundamental oscillations in MIR spectroscopy provide quantitative data from unique functional group oscillations (Leong et al., 2018). The overtones arising in the NIR range lack a robust quantitative background due to the complexity of unresolved bands (Capuano & van Ruth, 2015). Thus, chemometric models underlying NIR spectroscopy may not produce consistent quantitative results across diverse samples, such as grain flours from different regions or years, despite success in training sets. NIR spectroscopy was first reported for the evaluation of protein in pulses in Williams et al. (1978), yet the method has been little reported since, with even less work reported using MIR spectroscopy. The stronger absorption bands of MIR spectra provide a superior platform for consistent chemometrics with greater selectivity and sensitivity, which will not change with crop genotype, growing location, or year. Therefore, FT-MIR can be used to simultaneously identify and quantify molecules (i.e., proteins, carbohydrates, etc.) based on their distinct functional groups without further sample preparation.

The functional groups of proteins (N-H and C = O) and SAA (C-S and C-H of S-CH_3) have permanent dipole moments, and such groups can be readily probed with FT-MIR spectroscopy (Barth, 2007; Berthomieu & Hienerwadel, 2009). Total protein and SAA offer a helpful picture of protein quality in pulses since pulses are high in protein but limited by SAA (Bhatty, 1988; Sarwar & Peace, 1986). Standard laboratory approaches for measuring protein and SAA include the Dumas method (nitrogen analysis through combustion), Kjeldahl method, UV-visible spectroscopy (Chang & Yan, 2019), and various chromatography techniques, such as HPLC with diode array detection. Most of the above approaches are destructive to the sample, require extensive analysis time, chemicals, and skills and are thus expensive. Amino acid analysis, for example, costs over $100 per sample. Total protein analysis is less expensive at ~$6 but remains a constraint when analyzing thousands of samples. Consequently, these methods do not qualify as high-throughput workflows desired in nutritional breeding programs. In contrast, FT-MIR

Core Ideas

- Fourier-transform infrared spectroscopy (FTIR) spectroscopy is used to measure pulse total protein and S containing amino acids.
- FTIR is a unique tool to measure functional groups of a nutrient trait with low concentrations.
- These Fourier-transform mid-infrared prediction models utilized the C-S and S-CH3 stretching and bending bands.

TABLE 1 HPLC gradient method and conditions (max pressure: 400 bar; column temp: 40 ˚C)

Time	A	B	Flow rate
min	——————% MP——————		mL/min
0.00	100.0	0.0	0.25
3.00	100.0	0.0	0.25
10.40	81.5	18.5	0.62
23.00	43.0	57.0	0.62
23.10	0.0	100.0	0.62
27.00	0.0	100.0	0.62
27.10	100.0	0.0	0.62
27.90	100.0	0.0	0.62
28.00	100.0	0.0	0.25
33.00	100.0	0.0	0.25

Note. MP, mobile phase.

spectroscopy is a nondestructive, high-throughput approach requiring little operating costs or training. Therefore, the objectives of this paper are two-fold: (a) demonstrate FT-MIR as a potential high-throughput, nondestructive, and cost-effective phenotyping technique for pulse nutritional traits, and (b) present multivariate models for the quantification of protein and SAA in pulse crops based on FT-MIR spectra.

2 | MATERIALS AND METHODS

2.1 | FTIR instrumentation and data analysis software

A Cary 630 FTIR spectrometer with a diamond attenuated total reflectance (ATR) module (Agilent Technologies) was used to acquire all MIR spectroscopic data. The data acquisition was performed within a spectral range of 650–4,000 cm^{-1} under Happ-Genzel apodization. The instrument acquisition parameters were optimized for each trait to enable the collection of spectral data with sufficient selectivity and sensitivity for quantitative analysis (Table 2). The data were analyzed with MicroLab Expert software (version 1.1) developed by Agilent Technologies for multivariate statistical modeling (chemometric modeling). Scatter plots were generated, and pooled *t*-tests were performed in JMP Pro (14.0.0).

2.2 | Chickpea, dry pea, and lentil seed samples

All pulse seed samples were collected from U.S. breeding programs, specifically the USDA-ARS chickpea breeding program at Washington State University and the organic pulse nutritional breeding program at Clemson University. For chickpea and dry pea, a total of 100–150 dry seeds were selected from each breeding line and ground to a maximum particle size of 0.5 mm, using a cyclone sample grinder (UDY Corporation). Likewise, 10–50 seeds were selected from each lentil line and ground using a blade coffee grinder (KitchenAid) and sieved to a maximum particle size of 0.5 mm. The powdered subsamples were stored before analysis in a cold room maintained at 10 ˚C with a humidity level of ~50%.

2.3 | Total nitrogen analysis

The total nitrogen content of all pulse flours was analyzed on a combustion nitrogen analyzer at the Clemson Agricultural Service Laboratory (Clemson, SC). The final protein concentration was determined by multiplying total nitrogen by a factor of 6.25 (Salo-väänänen & Koivistoinen, 1996).

2.4 | Sulfur-containing AA analysis

Lentil SAA concentrations were determined using an acid hydrolysis method with a pre-oxidation step, followed by HPLC analysis. The hydrolysis method was adapted from Gehrke et al. (1985) and Manneberg et al. (1995). In brief, 40 mg of lentil flour was weighed into glass culture tubes (16 × 125 mm, polytetrafluoroethylene [PTFE] lined cap). A lentil lab reference standard was included in each batch to monitor batch-to-batch variation. Five mL of chilled performic acid (9:1 ratio of formic acid and hydrogen peroxide) was added to each tube to convert the SAA to stable derivatives, methionine sulfone, and cysteic acid. The tubes were gently swirled on a vortex mixer and refrigerated in an ice bath overnight (16 h). Caps were removed, and PTFE boiling rods (1/8 in. × tube length) were added. Samples were evaporated to dryness in an oil bath under vacuum (~70–80 ˚C, ~610 mmHg; 3 gal. resin trap; BACOENG). The tube rack was elevated with a stir bar underneath to improve consistent evaporation across the batch. The pressure was slowly lowered to prevent bumping. Tubes were removed, and residual oil was wiped off. Caps were removed, and 4.9 mL of 6 M HCl (hydrochloric acid) was added, along with 0.1 mL internal standard mix (25 mM norvaline and sarcosine each). Tubes were tightly capped and gently swirled. Proteins were hydrolyzed in an oven at 110 ˚C for 24 hr. Tubes were then allowed to cool to room temperature and vortex mixed. Samples were filtered (0.22 µm polypropylene syringe filter), and 1 mL was added to a clean glass tube to be evaporated to dryness as before. Samples were reconstituted with 1 mL mobile phase A and loaded into HPLC vials for analysis.

Amino acid concentrations were measured using an HPLC method adapted from Agilent application notes (Agilent

TABLE 2 Instrument acquisition and model parameters

Model name	Instrument scans # (background/sample)	Resolution cm^{-1}	Zero-fill factor	Preprocessing	Calibration breeding lines #	Validation breeding lines #	Calibration spectra #	Validation spectra #
Chickpea total protein	36/64[a]	4	None	D+S	55	22	154	84
Dry pea Total Protein	36/64[a]	4	None	N, D+S	40	22	135	59
Lentil total protein	200/100[b]	2	2	N	32	18	57	25
Lentil SAA	200/100[b]	2	2	N	37	24	53	34
Lentil methionine	200/100[b]	2	2	N	26	22	39	31

Note. D+S, Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21), N, Normalization (0 to 1).
[a]64 scans ≈ 30 s at 4 cm^{-1} resolution.
[b]100 scans ≈ 75 s at 2 cm^{-1} resolution.

Application Note, 2010; Long, 2015). An Agilent 1100 series system (Agilent Technologies) was used for analysis. A diode array detector (DAD) collected spectra at 338 nm, 10 nm bandwidth (reference 390 nm, 20 nm bandwidth) and 262 nm, 10 nm bandwidth (reference 390 nm, 20 nm bandwidth). Mobile phase A consisted of 10 mM Na_2HPO_4 (sodium phosphate), 10 mM $Na_2B_4O_7 \bullet 10H_2O$ (sodium tetraborate decahydrate), and 5 mM NaN_3 (sodium azide) and was adjusted to pH 8.2 with concentrated HCl and subsequently filtered through 0.2 µm regenerated cellulose membrane. Solution B consisted of acetonitrile/methanol/water (45:45:10, v/v/v). Separation was achieved on an Agilent Poroshell HPH-C18 3 × 100 mm analytical column (Part Number 695975-502; Agilent Technologies) with the corresponding Poroshell HPH-C18 3 × 5 mm guard column (Part Number 823750–928). The G1329A autosampler derivatized AAs with OPA (*o*-phthalaldehyde) and FMOC (9-fluorenylmethyl chloroformate). Vials of borate buffer (Part Number 5061-3339), H_2O (water) needle wash, and injection diluent (100 mL solution A, 0.4 mL H_3PO_4 conc.) were also required. The injection method was as follows (default speed and offset were used except where noted): (a) draw 2.5 µL from borate buffer, (b) draw 0.5 µL from a sample, (c) mix 3 µL from the air for five times, (d) wait 0.2 min, (e) draw 0 µL from needle wash, (f) draw 0.5 µL from OPA (vial insert) using 2 mm offset, (g) mix 3.5 µL from the air for six times, (h) draw 0 µL from needle wash, (i) draw 0.4 µL from FMOC (vial insert) using 2 mm offset, (j) mix 3.9 µL from the air for 10 times, (k) draw 32 µL from injection diluent, (l) mix 20 µL from the air for eight times, and (m) inject. See Table 1 for instrument method and conditions. Dilution series were made for calibration standard curves from 9 to 900 pmol/µL with norvaline (primary AA) and sarcosine (secondary AA) as internal standards at 500 pmol/µl. Calibration curves were generated for each AA from the ratio of AA/internal standards. Standards included cysteic acid, aspartic acid, glutamic acid, asparagine, serine, glutamine, histidine, glycine, threonine, methionine sulfone, arginine, alanine, tyrosine, cystine, valine, methionine, tryptophan, phenylalanine, isoleucine, leucine, lysine, hydroxyproline, and proline.

2.5 | Chickpea total nitrogen model

The diamond ATR surface was cleaned with HPLC grade methanol (Fisher Scientific) before spectra of the ground chickpea samples (fully homogenized by mixing) were collected. Instrument and model parameters are available in Table 2. The instrument acquisition parameters were set to absorbance mode with 64 scans (~30 s) per spectrum (Table S1), 4 cm^{-1} resolution, and no zero-fill factor (ZFF). Each breeding line was analyzed seven times. The most stable spectra with constant intensity were selected without averaging for calibration. Background corrections (36 scans) were performed between each spectral collection. Protein is a macronutrient with easily resolved IR bands, requiring less stringent acquisition parameters than SAA, as discussed below. The calibration set included 55 breeding lines (154 spectra) from the 2018 chickpea population, and the validation set included 22 breeding lines (84 spectra) from the 2020 chickpea population for the partial least squares (PLS-1) model (Tobias, 1995). The Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21) was applied to all spectra. The model was calibrated with nitrogen values obtained from a nitrogen analyzer. The PLS-1 model was developed based on the regions sensitive to the total protein concentration (3,682.61–3,006.98 cm^{-1}, N-H stretch; 1,718.30–1,487.21 cm^{-1}, amide bands I and II), and eight PLS model factors were included in the model. The model was run with full cross-validation.

2.6 | Dry pea total nitrogen model

The same background correction and data acquisition steps as for chickpea were followed (Table 2). However, the calibration set included 40 breeding lines (135 spectra) from the 2019 dry pea population, and the validation set included 22 breeding lines (59 spectra) from the 2020 dry pea population. The spectra were initially normalized to a scale of 0 to 1, and the Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21) was applied. The model was calibrated with total nitrogen values, as done for the chickpea model. The PLS-1 model was developed based on the same spectral ranges as the total nitrogen model above; however, 11 PLS model factors were included in the model. The model was run with full cross-validation.

2.7 | Lentil total nitrogen and SAA models

The diamond ATR window was cleaned with HPLC grade methanol and allowed to dry before each spectrum was collected. The background was collected every 30 min or less for convenience. Fourier-transform mid-infrared spectra were collected for 50 lentil breeding lines, and six spectra were collected per breeding line. Acquisition parameters included 200 scans per background and 100 scans (~75 s) per spectrum at a resolution of 2 cm^{-1} and a ZFF of 2 (Table 2). All spectra were normalized to a scale of 0 to 1. Unlike the previous models, the spectra were not derivatized by the Savitzky-Golay algorithm because the spectra were highly structured and informative at a resolution of 2 cm^{-1} and with a ZFF of 2. The increased scan number and resolution generated detailed spectra and allowed for the quantification of SAA, which are at low concentrations in lentil. For ease, the same spectra were used for the protein model. Additionally, this allows for the models to be combined into a single method for generating protein and SAA data simultaneously.

A PLS-1 model for total nitrogen in lentil flour was developed using Agilent MicroLab Expert software. The most stable spectra were applied in calibration without averaging. The calibration set included 32 breeding lines (57 spectra), and the validation set included 18 breeding lines (25 spectra). The model utilized the same spectral regions as in chickpea and dry pea and included five PLS model factors. PLS-1 models for total SAA and methionine were similarly attempted. In the model for total SAA, the calibration set included 37 breeding lines (53 spectra), and the validation set included 24 breeding lines (34 spectra). The model utilized 721.24–867.07, 1,231.88–1,469.96, 1,904.20–2,241.99 and 2,825.78–2,994.91 cm^{-1} spectral regions and included eight PLS model factors. Furthermore, the methionine model included 26 breeding lines (39 spectra) and 22 breeding lines (31 spectra) for calibration and validation, respectively. The model utilized 674.65–808.37, 1,182.03–1,484.41, 1,975.49–2,158.59, and 2,658.52–2,991.19 cm^{-1} spectral regions with eight PLS model factors. All lentil models were run with full cross-validatation.

3 | RESULTS AND DISCUSSION

This study successfully demonstrated that FT-MIR is a robust, nondestructive tool for measuring protein and SAA in pulse crops. Proteins and SAA have polar functional groups sensitive to MIR energy. The functional groups of proteins (N-H and C = O) in chickpea, dry pea, and lentil flour were analyzed through FT-MIR spectroscopy. Associated IR bands were identified at ~1,550 cm^{-1} (amide II bands), ~1,650 cm^{-1} (amide I band), and between 3,310 and 3,270 cm^{-1} (amide A band) (Tiwari & Singh, 2012). Multivariate models (PLS-1) were developed associating these regions with total nitrogen content. In chickpea, predicted protein concentrations of the validation set ranged from 18.3 to 23.9%, with a mean of 20.9% (Table 3). The chickpea total nitrogen model achieved an R^2 of 0.948, a calibration root means square error (RMSE) of 0.093, and a prediction RMSE of 0.10 (Figure 1b and Table 4). For dry pea, the predicted total protein concentration of the validation set ranged from 18.1 to 23.1%, with a mean of 21.2%. The dry pea total nitrogen model achieved a calibration RMSE of 0.096, an R^2 of 0.845, and a prediction RMSE of 0.093 (Figure S1b). For lentil, predicted protein concentrations ranged from 25.4 to 33.3%, with a mean of 28.3%. The lentil total nitrogen model achieved an R^2 of 0.845, a calibration RMSE of 0.13, and a prediction RMSE of 0.11 (Figure S2b). These models predicted mean protein concentrations in chickpea, dry pea, and lentil within the cited ranges in the literature (chickpea: 15.6–22.4%, dry pea: 20–25%, and lentil: 20.6–31.4%), demonstrating the applicability of the method in the field (Jarpa-Parra, 2018; Khan et al., 2016; Upadhyaya et al., 2016). Furthermore, pooled two-tailed *t*-tests performed on each crop (chickpea: $P > |t| = 0.93$; dry pea: $P > |t| = 0.97$; lentil: $P > |t| = 0.82$) targeting the means of actual and predicted protein concentrations of validation data showed no significant difference.

The functional groups of SAA (C-S and C-H of S-CH_3) in lentil flour were similarly analyzed. SAA is a valuable nutritional breeding trait because lentil (and other pulse crops) is nutritionally limited by SAA, methionine, and cysteine, despite being high in total protein. These low concentrations present a challenge for IR band resolution and consequent quantification. However, this study successfully identified bands in the lentil MIR spectrum (~751–685,~2,493–2,157, and ~2,977–2,861 cm^{-1}) associated with C-H stretching of methyl mercaptan (S-CH_3) and C-S stretching in pure

TABLE 3 Actual vs. model predicted data

Model name	Actual calibration set range)	Actual calibration set true mean	Actual validation set range	Actual validation set true mean	Predicted validation set range	Predicted validation set true mean	*t*-test
	% protein						
Chickpea total protein	15.4–24.6	20.0	18.1–24.6	20.3	18.3–23.9	20.9	NS
Dry pea total protein	18.3–23.9	21.1	18.4–23.6	21.0	18.1–23.1	21.2	NS
Lentil total protein	25.7–33.7	29.7	24.7–31.1	29.6	25.4–33.3	28.3	NS
Lentil SAA	0.211–0.348	0.279	0.197–0.321	0.265	0.207–0.326	0.258	NS
Lentil methionine	0.185–0.264	0.224	0.2007–0.251	0.221	0.194–0.294	0.222	NS

Note. NS, actual and predicted means of validation data were not significant at $P < .05$; SAA, sulfur-containing amino acid.

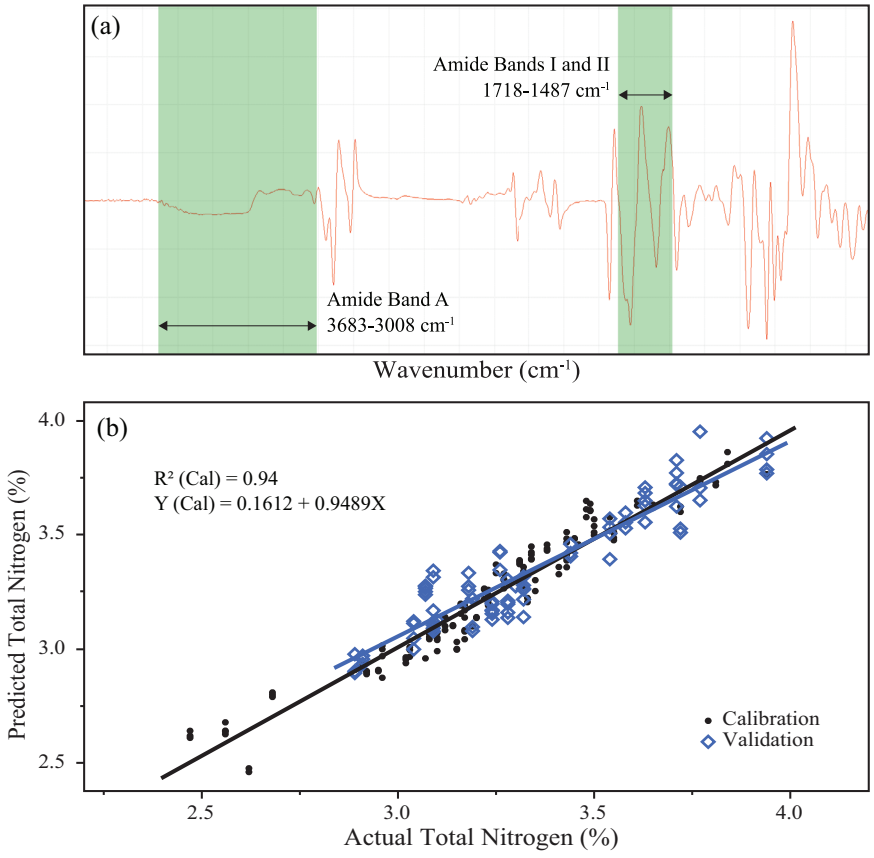

FIGURE 1 (a) Average chickpea mid-infrared first-derivative absorbance spectrum. Regions in green were selected for the total nitrogen model in chickpea. (b) Scatter plot of actual vs. predicted total nitrogen (%) of calibration and validation data with lines of best fit

TABLE 4 Chemometric model statistics

Model name	R^2	RMSEC	RMSECV	RMSEP	SEP	Bias
Chickpea total protein	0.948	0.093	0.093	0.10	0.10	−0.0057
Dry pea total protein	0.845	0.096	0.096	0.093	0.091	0.0039
Lentil total protein	0.845	0.13	0.13	0.11	0.11	0.016
Lentil SAA	0.827	0.014	0.014	0.022	0.021	−0.0066
Lentil methionine	0.815	0.0075	0.0075	0.014	0.014	0.0011

Note. RMSEC, root mean square error of calibration; RMSECV, root mean square error of cross validation; RMSEP, root mean square error of prediction; SEP, standard error of prediction.

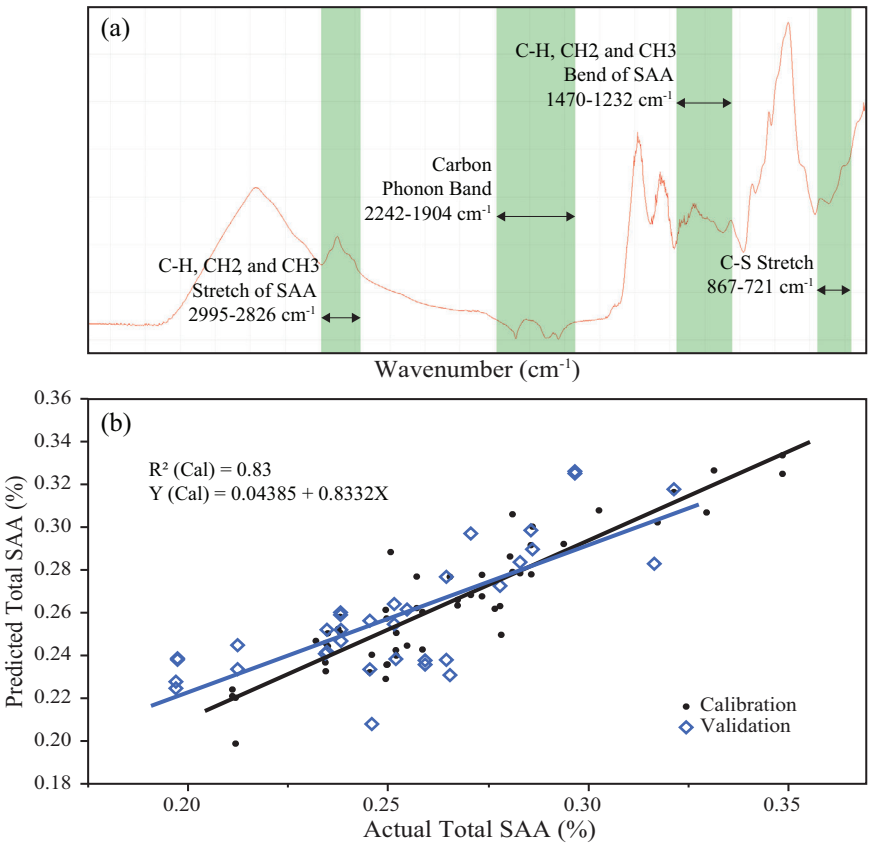

FIGURE 2 (a) Average lentil MIR absorbance spectrum. Regions in green were selected for the total sulfur-containing amino acid (SAA) model in lentil. (b) Scatter plot of actual vs. predicted total SAA (%) of calibration and validation data with lines of best fit

methionine were recognized (Figures 2a, S3a–S4). The bands apparent at ~2,991–2,659 cm^{-1} and 1,470–1,232 cm^{-1} represent the total C-H, C-CH_2, and C-CH_3 oscillations in lentil flour. The region between ~2,159-1,975 cm^{-1} (the phonon band arising due to the oscillations of the carbon lattice of ATR- diamond) strengthened the prediction of the multivariate regression models for total SAA and methionine. The lentil SAA model achieved an R^2 of 0.827, and the predicted validation data ranged from 0.207 to 0.326%, with a mean of 0.258%. In this model, the calibration RMSE was 0.014, and the prediction RMSE was 0.022 (Figure 2b). Further, the methionine model achieved an R^2 of 0.815 and predicted the validation results between 0.194–0.294%, with a mean of 0.222%. The methionine model had the calibration and prediction RMSEs at 0.0075 and 0.014, respectively (Figure S3b). The lines of best fit for the validation data (Figures 1b – 2b, S1b–S3b; blue lines) have deviated slightly from that of the calibration data (Figures 1b - 2b, S1b–S3b; black lines). The *t*-tests performed for total SAA and methionine ($P > |t| = 0.35$ and $P > |t| = 0.76$, respectively) returned no significant differences between actual and predicted means. The predicted lentil methionine mean, 0.22%, agrees well with the literature (0.22%, USDA ARS, 2019). Total SAA makes up ~2% of the total protein content of lentils, whereas SAA comprise ~4% of beef and chicken protein and ~8% of chicken egg protein (USDA ARS, 2019). Lentil and other pulse crops are not a good source of SAA; however, genetic selection and breeding may help increase their SAA concentrations. Developing lentil varieties with high SAA concentrations could help improve the dietary intake of better-quality protein and develop food products, such as protein powder, that contain high-quality protein without adding another high-SAA source.

Chemometric models with well-recognized and consistent underlying bands will aid in the development of analytical methods and accurate, consistent modeling regardless of differing sample origins. While the prediction RMSEs indicate these models have high predictive ability for each sample, the *t*-tests indicate they also accurately predict the population means. The calibration data were not used in model validation, and the purpose of calibration data was to build the model, whereas validation data was to test the model. Thus, these total protein, total SAA, and methionine chemometric models have consistent applicability over these pulse crops regardless of sample origin. Accordingly, FT-MIR spectroscopy provides added advantages for stable and straightforward chemometric modeling compared with methods associated with the NIR range, which lacks a strong quantitative foundation (Guo et al., 2016).

Traditional univariate statistical regression modeling based on Beer-Lambert was unsuitable for complex sample systems like lentil and chickpea. Partial least squares regression (a multivariate statistical regression algorithm) was applied with chemometric modeling throughout this study, where the

best predictive use of spectral variables can be enhanced. The use of PLS regression reduced the dimensionality of the multivariate space in a supervised manner, maintaining a good correlation between dependents (absorbance values) and independent (analyte concentrations) variables (Saikat et al., 2008). Therefore, PLS-1 proved to be an excellent choice for correlating nutrient data with the spectral regions associated with protein functional groups. Fourier-transform mid-infrared spectroscopic data were utilized with minimal mathematical pre-processing (averaging, normalization, and the Savitzky-Golay derivative and smoothing algorithm). In FT-MIR spectroscopy, the spectra are always associated with functional groups and molecular skeletal structures (Yadav, 2005). Fully resolved functional group bands act as fingerprints for traits (analytes). In proteins, the A, I, and II amide bands (Figures 1a and S1a–S2a) were significantly associated with protein content in our models. The C-H stretching bands of methyl mercaptans and C-S stretching bands in methionine are mainly associated with our total SAA and methionine models. Other spectral regions common to both lentil flour and the standard compounds (Figure S4) were also selected to enhance the regression in the chemometric models. Notably, different spectral acquisition parameters were followed in the lentil models than the chickpea and dry pea models during spectral sampling. This was to ensure sufficiently high resolution and scan number in the lentil spectra to observe the minor bands associated with methionine at low concentrations in the sample matrix. Once highly resolved spectra were employed, the number of spectra required for a consistent model in the lentil models was lower than for the chickpea and dry pea models, which employed lower resolution parameters and had fewer spectral details (data points) in each spectrum. However, high resolution is not required for a bulk trait such as total protein because the associated amide bands are distinct and quickly resolved. The use of first derivatives in the chickpea and dry pea spectra further strengthened the predictive ability of the two respective chemometric models related to total proteins.

Breeding programs require the generation of large amounts of phenotypic data. Nutritional traits are no exception, yet higher costs are associated with collecting these data than traditional agronomic traits such as yield. With the great promise of molecular-based breeding approaches, such as marker-assisted backcrossing and genomic selection along with genome-wide association studies, large datasets are needed to discover quantitative trait loci (QTL) and elucidate underlying gene pathways associated with traits (Liu et al., 2020; Roorkiwal et al., 2016; Sab et al., 2020; Upadhyaya et al., 2016). The application of conventional protocols in quantifying nutrients (nutritional phenotyping) is not suitable for the large volume of samples from the field. Significant challenges with traditional quantitative analysis techniques include long analysis times, highly trained workers, chemical costs, chemical disposal, and instrument maintenance. Fourier-transform infrared spectroscopy analysis time is short (i.e., less than a minute), and the method does not require a skilled operator (Capuano & van Ruth, 2015). It also requires minimal sample preparation, minimizing the risks of hazardous chemical usage and chemical cost. Compared with the complex compartmentalization typical of liquid and gas chromatography systems, the compact instrumentation occupies little space and is relatively simple in construction. Maintenance costs are also considerably lower than other analytical instruments (Minali & Rein, 2015). Therefore, FT-MIR spectroscopy can support a high-throughput and efficient workflow for the quantitative analysis of nutritional traits.

Accordingly, the chemometric regression (PLS-1) models for total protein and methionine could be an essential part of this high-throughput phenotyping workflow. This analytical technique could lower costs in breeding programs globally and open possibilities for developing and under-resourced countries to adopt the technique in their breeding programs. The methods and models presented in this study can accelerate nutritional breeding programs by reducing the time and cost of analysis and by being incorporated into QTL discovery pipelines. Rapid, low-cost data generation is advantageous for efficiently increasing sample size and power in genome-wide association studies. Once QTLs are detected, flanking markers can be used in marker-assisted selection (MAS) to verify the presence or absence of favorable alleles in progeny. MAS could be an effective technique for nutritional traits because the phenotype can be predicted without processing and analyzing the seed. Seedlings could be genetically tested and selected or discarded before flowering, allowing for same-generation hybridization, essentially cutting generation time in half.

4 | CONCLUSIONS

Fourier-transform mid-infrared spectroscopy is conveniently applicable with simple chemometric modeling to predict the concentrations of total proteins and SAA in chickpea, dry pea, and lentil. Well-recognized functional groups (bands) associated with total protein content and SAA content in the MIR range make multivariate modeling relatively simple. Therefore, the present work on FT-MIR spectroscopy creates a platform for high-throughput and nondestructive phenotyping with minimal costs and chemical hazards. Further, these techniques can reduce breeding program expenses globally and allow under-resourced countries to expand into nutritional phenotypes, such as those with improved protein content. Future studies may benefit from exploration of different modeling techniques and larger sample sizes for calibration and validation.

ACKNOWLEDGMENTS

This project was supported by the American people via the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID, award no 7200AA19LE00005/subaward no 89915-11295 awarded to DT); the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018–02799) of the United States Department of Agriculture, National Institute of Food and Agriculture, the Pulse Health Initiative (USDA-ARS awarded to DT); the Good Food Institute, and the USDA National Institute of Food and Agriculture, [Hatch] project [1022664] awarded to DT. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the USDA, USAID, or the United States Government. The authors thank Drs. George Vandermark, USDA-ARS, Washington State, and Siv Kumar, ICARDA, Morocco, providing pulse seed samples for model development.

AUTHOR CONTRIBUTIONS

Amod Madurapperumage: Conceptualization; Data curation; Formal analysis; Methodology; Software; Writing – original draft. Nathan Johnson: Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Writing – original draft; Writing – review & editing. Pushparajah Thavarajah: Conceptualization; Data curation; Formal analysis; Methodology; Validation; Writing – original draft; Writing – review & editing. Leung Tang: Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Writing – original draft; Writing – review & editing. Dil Thavarajah: Conceptualization; Funding acquisition; Investigation; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Dil Thavarajah https://orcid.org/0000-0002-4251-7476

REFERENCES

Agilent Application Note. (2010). *Separation of two sulfurated amino acids with other seventeen amino acids by HPLC with pre-column derivatization*. Agilent Technologies.

Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta - Bioenergetics*, *1767*(9), 1073–1101. https://doi.org/10.1016/j.bbabio.2007.06.004

Berthomieu, C., & Hienerwadel, R. (2009). Fourier transform infrared (FTIR) spectroscopy. *Photosynthesis Research*, *101*(2–3), 157–170. https://doi.org/10.1007/s11120-009-9439-x

Bhatty, R. S. (1988). Composition and quality of lentil (*Lens culinaris* Medik): A review. *Canadian Institute of Food Science and Technology Journal*, *21*(2), 144–160. https://doi.org/10.1016/S0315-5463(88)70770-1

Boye, J., Wijesinha-Bettoni, R., & Burlingame, B. (2012). Protein quality evaluation twenty years after the introduction of the protein digestibility corrected amino acid score method. *British Journal of Nutrition*, *108*(2, Suppl), S183–S211. https://doi.org/10.1017/S0007114512002309

Capuano, E., & van Ruth, S. M. (2015). Infrared spectroscopy: Applications. Encyclopedia of food and health (1st ed.). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-384947-2.00644-9

Chang, S. K. C., & Yan, Z. (2019). Protein analysis. In S. S. Nielsen (Ed.), *Food analysis* (5th ed., pp. 315–327). Springer. https://doi.org/10.1007/978-3-319-45776-5

El Khoury, Y., & Hellwig, P. (2017). Far infrared spectroscopy of hydrogen bonding collective motions in complex molecular systems. *Chemical Communications*, *53*(60), 8389–8399. https://doi.org/10.1039/C7CC03496B

Foyer, C. H., Lam, H.-M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H., … Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants*, *2*, 16112. https://doi.org/10.1038/nplants.2016.112

Gehrke, C. W., Wall, L. L., Sr., Absheer, J. S., Kaiser, F. E., & Zumwalt, R. W. (1985). Sample preparation for chromatography of amino acids: Acid hydrolysis of proteins. *Journal of Association of Official Analytical Chemists*, *68*(5), 811–821. https://doi.org/10.1093/jaoac/68.5.811

Graça, J., Godinho, C. A., & Truninger, M. (2019). Reducing meat consumption and following plant-based diets: Current evidence and future directions to inform integrated transitions. *Trends in Food Science and Technology*, *91*(July), 380–390. https://doi.org/10.1016/j.tifs.2019.07.046

Guo, T., Feng, W. H., Liu, X. Q., Gao, H. M., Wang, Z. M., & Gao, L. L. (2016). Fourier transform mid-infrared spectroscopy (FT-MIR) combined with chemometrics for quantitative analysis of dextrin in Danshen (Salvia miltiorrhiza) granule. *Journal of Pharmaceutical and Biomedical Analysis*, *123*, 16–23. https://doi.org/10.1016/j.jpba.2015.11.021

Jarpa-Parra, M. (2018). Lentil protein: A review of functional properties and food application. An overview of lentil protein functionality. *International Journal of Food Science and Technology*, *53*(4), 892–903. https://doi.org/10.1111/ijfs.13685

Johnson, N., Johnson, C. R., Thavarajah, P., Kumar, S., & Thavarajah, D. (2020). The roles and potential of lentil prebiotic carbohydrates in human and plant health. *Plants, People, Planet*, *2*, 310–319. https://doi.org/10.1002/ppp3.10103

Khan, T. N., Meldrum, A., & Croser, J. S. (2016). Pea: Overview. In C. Wrigley, H. Corke, K. Seetharaman, & J. Faubion (Eds.), *Encyclopedia of food grains* (2nd ed.; pp. 324–333). Academic Press. https://doi.org/10.1016/B978-0-12-394437-5.00037-1

Kim, H., Caulfield, L. E., Garcia-Larsen, V., Steffen, L. M., Coresh, J., & Rebholz, C. M. (2019). Plant-based diets are associated with a lower risk of incident cardiovascular disease, cardiovascular disease mortality, and all-cause mortality in a general population of middle-aged adults. *Journal of the American Heart Association*, *8*(16), e012865. https://doi.org/10.1161/JAHA.119.012865

Leong, S. S., Ng, W. M., Lim, J. K., & Yeap, S. P. (2018). Dynamic light scattering: Effective sizing technique for characterization. of magnetic nanoparticles. In S. Sharma (Ed.), *Handbook of materials characterization* (pp. 77–111). Springer. https://doi.org/10.1007/978-3-319-92955-2_3

Liu, X., Qin, D., Piersanti, A., Zhang, Q., Miceli, C., & Wang, P. (2020). Genome-wide association study identifies candidate genes related to oleic acid content in soybean seeds. *BMC Plant Biology*, *20*(1), 1–14. https://doi.org/10.1186/s12870-020-02607-w

Long, W. (2015). *Automated amino acid analysis using an* Agilent Poroshell HPH-C18 Column. Agilent Technologies.

Manneberg, M., Lahm, H. W., & Fountoulakis, M. (1995). Quantification of cysteine residues following oxidation to cysteic acid in the presence of sodium azide. *Analytical Biochemistry*, *231*(2), 349–353. https://doi.org/10.1006/abio.1995.9988

Minali, D., & Rein, A. (2015). *The Agilent Cary 630 FTIR Spectrometer quickly identifies and qualifies pharmaceuticals*. Agilent Technologies.

Perkins, W. D. (1987). Fourier transform infrared spectroscopy: II. Advantages of FT-IR. *Journal of Chemical Education*, *64*(11), A269. https://doi.org/10.1021/ed064pa269

Pimentel, D., & Pimentel, M. (2003). Sustainability of meat-based and plant-based diets and the environment. *American Journal of Clinical Nutrition*, *78*(3, Suppl.), 660–663. https://doi.org/10.1093/ajcn/78.3.660s

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., Gaur, P. M., Chellapilla, B., Tripathi, S., Li, Y., Hickey, J. M., Lorenz, A., Sutton, T., Crossa, J., Jannink, J. L., & Varshney, R. K. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Frontiers in Plant Science*, *7*, 1666. https://doi.org/10.3389/fpls.2016.01666

Sab, S., Lokesha, R., Mannur, D. M., Somasekhar, J. K., Mallikarjuna, B. P., Laxuman, C., Yeri, S., Valluri, V., Bajaj, P., Chitikineni, A., Vemula, A., Rathore, A., Varshney, R. K., Shankergoud, I., & Thudi, M. (2020). Genome-wide SNP discovery and mapping QTLs for Seed iron and zinc concentrations in chickpea (*Cicer arietinum* L.). *Frontiers in Nutrition*, *7*, 559120. https://doi.org/10.3389/fnut.2020.559120

Saikat, M., Jun, Y., Maitra, S., & Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Casualty Actuarial Society*, 79–90.

Salo-väänänen, P. P., & Koivistoinen, P. E. (1996). Determination of protein in foods: Comparison of net protein and crude protein (N × 6.25) values. *Food Chemistry*, *57*(1), 27–31. https://doi.org/10.1016/0308-8146(96)00157-4

Sarwar, G., & Peace, R. W. (1986). Comparisons between true digestibility of total nitrogen and limiting amino acids in vegetable proteins fed to rats. *The Journal of Nutrition*, *116*(7), 1172–1184. https://doi.org/10.1093/jn/116.7.1172

Skoog, D. A., Hanlan, J., & West, D. M. (2016). *Principles of instrumental analysis* (7th ed.). Cengage.

Tiwari, B., & Singh, N. (2012). *Pulse chemistry and technology*. The Royal Society of Chemistry.

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proceedings of the 20th Annual SAS Users Group International Conference*. SAS Institute Inc.

USDA ARS. (2019). *FoodData Central* (NDB# 16069). https://fdc.nal.usda.gov/fdc-app.html#/food-details/172420/nutrients

Upadhyaya, H. D., Bajaj, D., Narnoliya, L., Das, S., Kumar, V., Gowda, C. L. L., Sharma, S., Tyagi, A. K., & Parida, S. K. (2016). Genome-wide scans for delineation of candidate genes regulating seed-protein content in chickpea. *Frontiers in Plant Science*, *7*, 302. https://doi.org/10.3389/fpls.2016.00302

Williams, P. C., Stevenson, S. G., Starkey, P. M., & Hawtin, G. C. (1978). The application of near infrared reflectance spectroscopy to protein-testing in pulse breeding programmes. *Journal of the Science of Food and Agriculture*, *29*(3), 285–292. https://doi.org/10.1002/jsfa.2740290315

Yadav, L. D. S. (2005). Organic spectroscopy. Springer. https://doi.org/10.1007/978-1-4020-2575-4

SUPPORTING INFORMATION

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

How to cite this article: Madurapperumage, A., Johnson, N., Thavarajah, P., Tang, L., & Thavarajah, D. (2022). Fourier-transform infrared spectroscopy (FTIR) as a high-throughput phenotyping tool for quantifying protein quality in pulse crops. *The Plant Phenome Journal*, *5,* e20047. https://doi.org/10.1002/ppj2.20047