Near Infrared Spectroscopy Predicts Crude Protein Concentration in Hemp Grain

Ryan V. Crawford

Jamie L. Crawford

Julie L. Hansen

Lawrence B. Smart

Virginia M. Moore

2025-07-11

Abstract

This study evaluated near-infrared spectroscopy (NIRS) for non-destructive crude protein (CP) prediction in hemp (Cannabis sativa L.) grain and validated the biological basis of spectral predictions. One hundred and forty-nine whole grain samples from 38 cultivars were collected from New York trials (2017-2021) and validated for CP by combustion. Seven preprocessing methods were tested using 100 training/testing splits, with standard normal variate transformation following Savitzky-Golay filtering selected as optimal. Comparing algorithms showed that partial least squares regression (PLSR) significantly outperformed support vector machines and random forest. The best preprocessing method and algorithm was applied to 1,000 additional splits. Optimal models contained 12 components with mean performance of RMSE = 9.94, R² = 0.84, RPD = 2.5, and RPIQ = 3.94. More than 99% of the models had, at minimum, the ability to distinguish between high and low values, with 93.2% capable of quantitative prediction.To validate biological relevance, a protein-focused model was developed using three known protein absorption bands (1200-1250, 1500-1550, and 2040-2090 nm). These models had substantially reduced performance with 86% of models capable of distinguishing between high and low values but only 14% of models capable of quantitative prediction. However, this targeted approach offers evidence that NIRS predictions are biologically grounded in protein-specific spectral features rather than spurious correlations. This research demonstrates the promise and biological validity of NIRS for hemp grain CP assessment, supporting applications in breeding programs, although applications demanding more accurate prediction will require better models.

Table of contents

CP, crude protein; MAE,mean absolute error; NIR, near-infrared; NIRS, near-infrared spectroscopy; NY, New York; PLSR, partial least squares regression; RF, random forest; RPD, relative predicted deviation, RPIQ, ratio of performance to interquartile distance; SG, Savitzky-Golay; SNV, standard normal variate, SNV-SG, standard normal variate following Savitzky-Golay; SVM, support vector machines; VIP, variable importance in projection

## 1 INTRODUCTION

Hemp (*Cannabis sativa* L.) is an annual crop with potential uses as a source of food or feed, derived from the grain, and fiber (bast or hurd), derived from the stalk. Hemp cultivars are commonly grown for one or both purposes, and a cultivar may be called a grain, fiber, or dual-purpose type (Tang et al., 2016). Because of its nutritional importance, the protein concentration of a grain crop is a prime consideration for researchers, producers, and consumers. Whole hemp grain typically contains approximately 200-300 g kg−1 crude protein (CP) (Bárta et al., 2024; Callaway, 2004; Ely & Fike, 2022; Liu et al., 2023). Crude protein is a proxy for the direct measurement of protein concentration and consists of the multiplication of nitrogen concentration by a conversion factor, often 6.25 (Hayes, 2020).

Near-infrared (NIR) spectroscopy (NIRS) technology is rapid, non-destructive, and inexpensive. It consists of the measurement of NIR radiation reflected and absorbed from a sample (the spectra). These spectra are then related to primary analytical values, typically obtained using wet chemistry assays, for components such as moisture, CP, fat, or fiber (Roberts et al., 2004). NIRS technology has been used since the 1970’s to assess forage CP concentration (Reeves, 2012; P. C. Williams, 1975). A NIRS calibration set often consists of samples from diverse genotypes of one species grown in many environments encompassing the range of expected values from the analyte or analytes (Chadalavada et al., 2022).

Partial least squares regression (PLSR), a typical method used in the agricultural and food sciences, calculates components that maximize covariance between predictor and response variables to relate spectra to analyte (Roberts et al., 2004). It uses some number of components, often selected via cross-validation in order to avoid overfitting, to fit the regression model (Wold et al., 2001). It is commonly used in spectroscopy because it tends to work well with highly correlated, noisy spectral data . Variable importance in projection (VIP) scores show the amount response variable’s variance explained by the predictor variables, with VIP scores of greater than one often serving as a criterion for variable selection (Farrés et al., 2015).

While PLSR is commonly used in agricultural applications, other algorithms may offer comparable or improved prediction accuracy. Support vector machines (SVM) and random forest (RF) algorithms have been utilized in NIRS applications, particularly where non-linear relationships exist between spectral features and analyte concentration (Chinilin et al., 2023). Support vector machines apply non-linear kernel functions to capture relationships between spectral data and analyte concentrations, with studies demonstrating competitive performance with PLSR for protein prediction in stored wheat grains Kamboj et al. (2022). Random forest (RF) algorithms use ensemble learning to combine multiple decision trees and has been used to predict protein content in grains (Chadalavada et al., 2022; Haruna et al., 2022). Comparing multiple algorithms helps ensure that the best approach is chosen.

The biological validity of NIRS predictions can be enhanced through targeted analysis of wavelength regions corresponding to known molecular absorption bands. Models developed using biologically-relevant wavelengths are evidence that predictions are based on CP-spectral relationships rather than spurious correlations with other quality components. Three spectral regions with established CP-related vibrational assignments are: 1200-1250 nm (C-H stretch 2nd overtone from amino acid side chains), 1500-1550 nm (N-H stretch 1st overtone from peptide bonds), 2040-2090 nm (N-H + C-N combination bands)(Köllmann et al., 2023; Shi et al., 2022). Selecting spectral regions prior to model fitting avoids the bias that may result when data are used to simultaneously select a model and estimate parameters (Yates et al., 2023). Models developed using portions of the NIRS spectra be more readily transferable between instruments because they discard noisy, irrelevant portions of the spectra (Y. Wang et al., 2025).

A NIRS-scanned sample of whole grain may be used for other purposes besides the scan, including planting as a seed or consumed, benefits attractive to plant breeders, farmers, and the food industry (Caporaso et al., 2018). In wheat and corn, grain CP content has been shown to be heritable (Geyer et al., 2022; Giancaspro et al., 2019). This suggests that NIRS technology could serve as a resource to rapidly identify high concentration CP hemp germplasm, enabling the screening of germplasm as seed, before planting to the field, and facilitating the efficient development of high concentration CP hemp populations.

For this study, researchers hypothesized that a benchtop NIR spectrometer could be used to develop a model to accurately predict CP concentration based on a data set of hemp grain representing multiple years, locations, and cultivars from grain and dual-purpose hemp types using PLSR, RF, or SVM.

## 2 MATERIALS AND METHODS

### 2.1 Hemp Grain Sample Background

Spectral data were collected from whole (unground) hemp grain samples, harvested at maturity, collected from from 18 experimental trials in New York (NY) between 2017 and 2021 (149 samples). Ninety percent of the samples were collected from Ithaca, Geneva, and Freeville NY. Grain samples were collected through hand sampling or mechanical harvest and were subsequently cleaned of chaff and dried at a temperature of 30 C for six days in a forced-air dryer. All CP values were expressed as concentration dry matter. In total, 149 samples from 38 cultivars were represented in the data set. Cultivars were either grain or dual-purpose types and included both commercially available and experimental material. Seventy-eight samples were scanned and chemically assayed in 2017, 19 in 2018, 24 in 2019, and 28 in 2021. All cultivars and locations were represented in 2017, but only a selected subset of cultivar-location combinations were represented in 2018-2021 because not all cultivars were planted everywhere and only a portion of these cultivar-location combinations were sampled, scanned, and assayed due to logistical constraints. More information about hemp cultivars and locations is available in Supplemental Table S1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Sample distribution by location and type  Tally of hemp cultivars and locations. Private cultivars are labeled ‘Cultivar1’, ‘Cultivar2’, etc., while experimental cultivars are labeled ‘Experimental1’, ‘Experimental2’, etc.   | cultivar2 | chazy | freeville | geneva | ithaca | willsboro | Total | | --- | --- | --- | --- | --- | --- | --- | | altair | NA | NA | NA | 1 | NA | 1 | | anka | NA | 1 | 3 | 5 | 2 | 11 | | bialobrzeskie | NA | 1 | 3 | 4 | 1 | 9 | | canda | NA | 1 | 1 | 1 | NA | 3 | | cfx-1 | NA | 1 | 2 | 5 | NA | 8 | | cfx-2 | NA | 1 | 2 | 4 | NA | 7 | | crs-1 | 1 | 1 | 2 | 5 | NA | 9 | | cultivar1 | NA | 1 | NA | NA | NA | 1 | | cultivar2 | NA | NA | NA | 1 | NA | 1 | | cultivar3 | NA | NA | NA | 1 | NA | 1 | | cultivar4 | NA | NA | NA | 1 | NA | 1 | | earlina 8 | NA | NA | 1 | NA | NA | 1 | | experimental1 | NA | NA | NA | 1 | NA | 1 | | experimental2 | NA | NA | NA | 1 | NA | 1 | | felina 32 | NA | 1 | 2 | 3 | NA | 6 | | futura 75 | NA | 1 | 3 | 4 | NA | 8 | | grandi | NA | 3 | 3 | 4 | NA | 10 | | h-51 | NA | NA | 1 | 2 | NA | 3 | | han-fn-h | NA | NA | NA | 1 | NA | 1 | | han-nw | NA | NA | NA | 1 | NA | 1 | | helena | NA | 1 | NA | NA | NA | 1 | | henola | NA | NA | NA | 2 | NA | 2 | | hlesia | NA | NA | NA | 3 | NA | 3 | | hliana | NA | NA | 1 | 1 | NA | 2 | | joey | NA | 1 | 1 | 1 | NA | 3 | | katani | NA | 2 | 3 | 4 | NA | 9 | | nebraska (feral) | 1 | NA | NA | 1 | NA | 2 | | pewter river | NA | 1 | NA | NA | NA | 1 | | picolo | NA | 1 | 2 | 5 | NA | 8 | | portugal | NA | NA | 1 | NA | NA | 1 | | rocky hemp | NA | NA | 1 | NA | NA | 1 | | sterling gold | NA | NA | 1 | NA | NA | 1 | | swift | 1 | 1 | NA | 1 | NA | 3 | | tygra | NA | 1 | 3 | 4 | NA | 8 | | uso-31 | 2 | 1 | 2 | 4 | NA | 9 | | wojko | NA | 1 | 3 | 4 | NA | 8 | | x-59 | NA | 2 | NA | 1 | NA | 3 | | Total | 5 | 24 | 41 | 76 | 3 | 149 | |

All experimental trials were planted in randomized complete block design with each cultivar replicated four times. The 2017 data were comprised of samples from the same 13 cultivars sampled from six NY locations. For those trials, grain was harvested from each plot individually and aggregated by cultivar within each trial. Four subsamples were drawn from each aggregated sample and scanned separately. These 2017 spectra were averaged at each 2 nm increment. All remaining samples from 2018-2021 were collected on a per-plot basis.

### 2.2 Spectral Data Collection and Preprocessing

A benchtop NIR spectrometer (FOSS/ NIR FOSS/ NIR Systems model 5000) was used to obtain the spectra (FOSS North America, Eden Prairie, MN, USA). Spectra were collected every 2 nm from 1100-2498 nm and the logarithm of reciprocal reflectance was recorded. A 1/4 rectangular sample cup (5.7 cm × 4.6 cm) was used to scan the samples.

WINISI software version 1.02A (Infrasoft International, Port Matilda, PA, USA) was used to calculate the mean spectra in 2017 and to select samples for laboratory assay in all years. Samples were selected according to their spectral distance from their nearest neighbor within the data set with a cutoff of a distance of 0.6 H, where H is approximately equal to the squared Mahalanobis distance divided by the number of principal components used in the calculation (Garrido-Varo et al., 2019). The spectral cutoff threshold of 0.6 H is a metric used to optimize the number of samples for chemical analysis Shenk & Westerhaus (1991). Prior to selection, spectra were preprocessed using Standard Normal Variate (SNV)-detrend with settings 1,4,4,1 for the derivative, gap, smooth, and smooth-two settings respectively. These settings are standard WINISI software parameters and were applied because hemp grain samples were subject to light scatter and noise due to particle size variation, with the first derivative applied to show component absorption (Barnes et al., 1989; Towett et al., 2013).

### 2.3 Laboratory Validation

Laboratory assays were performed by Dairy One Forage Laboratory (Ithaca, NY). For those assays, 1 mm ground samples were analyzed by combustion using a CN628 or CN928 Carbon/Nitrogen Determinator. Samples from 2017 were aggregated as described above, but the remaining samples were not aggregated.

### 2.4 Software and Packages Used

All analyses were conducted using R version 4.4.1 (R Core Team, 2024) with workflow management via targets v. 1.8.0 (Landau, 2021) and package control via renv v. 1.1.4 (Ushey, 2024). Data processing utilized data.table v. 1.16.0 (Barrett et al., 2024) and tidyverse v. 2.0.0 (Wickham et al., 2019), while machine learning was implemented through caret v. 6.0.90 (Kuhn, 2021) with pls v. 2.8.0, randomForest v. 4.7-1.1, and kernlab v. 0.9-33 for PLSR, random forest, and SVM algorithms, respectively.

### 2.5 Model Development

For all models, training and testing sets were created by dividing samples by their laboratory CP concentration values into three equal parts (tertiles) to ensure that a representative range of values was present in both training and testing sets and so that the model could predict across the entire range of data. Within each tertile, 75% of the samples were randomly assigned to the training set and the remaining 25% were assigned to the testing set. For each training set, models were developed in R’s caret package using PLSR, RF, and SVM models. Model performance was evaluated with caret’s default 25 iterations of bootstrapping and minimized RMSE in selecting the number of components in the final model. unless otherwise indicated, for PSLR models, the number of components was optimized over a grid search from one to 20. For SVM models, cost and sigma parameters were optimized using a grid search over 10 values. For RF models, eight variables were considered at each split and the number of trees was set to 500. For all models, performance was summarized using RMSE, R², RPD, and RPIQ.

Initially a number of common spectral preprocessing methods were tested by creating 100 training and testing sets, as described above. Spectral data were transformed by each of the following methods: 1) first derivative; 2) Savitzky-Golay (SG) using the first derivative, third order polynomial, and a window of size five; 3) gap-segment derivative using the first derivative, a gap of 11, and a segment size of five; 4) SNV; 5) standard normal variate following Savitzky-Golay (SNV-SG) using the same SG parameters as above; 6) SNV-detrend with second order polynomial; and 7) multiplicative scatter correction. For comparison, models were also developed using untransformed spectra.

For each of these preprocessing methods, PLSR models were fit and predictions were made on the corresponding testing set. Since there were seven preprocessing methods as well as untransformed spectra, eight separate models were fit for each of the 100 sets. The relationship between the predicted and actual values of the test set were calculated via RMSE, R2, relative predicted deviation (RPD), and ratio of performance to interquartile distance (RPIQ), four common model assessment metrics. Larger R2, RPD and RPIQ values and smaller RMSE values are best. There is a long history of using RPD to evaluate chemometric models although the statistic has been criticized as inadequately reflecting the distribution of skewed populations, a situation which RPIQ was designed to address (Bellon-Maurel et al., 2010). The answer to the question of exactly which values constitute a “good” model varies depending upon the reference consulted, but a simple rubric for model quality following Chadalavada et al. (2022) is set forth in [Table 2](#tbl-model-rubric)There are more conservative definitions of model quality for which these categories would not suffice. For example, P. C. Williams & Sobering (1993) suggests a minimal RPD of 2.5 for screening in breeding programs with values of 5-10 for quality control (P. Williams, 2013).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Model assessment rubric  Model Evaluation Criteria for NIRS Prediction Models   | Category | RPD | Interpretation | | --- | --- | --- | | Excellent | > 3.0 | Excellent quantitative prediction | | Good | 2.0 - 3.0 | Good quantitative prediction | | Fair | 1.5 - 2.0 | Moderate prediction (quantitative) | | Poor | < 1.5 | Poor prediction | |

Analyses of variance were performed for each of these metrics in order to compare preprocessing methods. For ANOVAs involving preprocessing methods, each data set was considered as a subject and different variances were allowed for each preprocessing method .Post hoc comparisons were performed with Tukey method to compare estimates set to = 0.05. Once the most promising preprocessing method was identified, it was used in all subsequent analyses.

To select an algorithm to fit a final model for these data, researchers compared three machine learning algorithms: PLSR, SVM (with radial basis function kernel), and random forest (RF). This comparison was conducted to validate the choice of PLSR through empirical evaluation against alternative algorithms and to assess whether more complex modeling approaches might improve predictive accuracy. One hundred random splits into training sets and testing sets were performed and models were constructed using each of the three algorithms. Statistical significance of performance differences between algorithms was assessed using analysis of variance.

Using the optimal preprocessing method identified above, one thousand training and testing sets were generated and models were developed. The larger number of iterations was feasible once the preprocessing method was selected, allowing for more robust performance estimates. The pattern of errors, expressed as the difference between the actual and predicted values for a given sample, was examined.

To validate the biological basis of spectroscopic predictions and ensure that models were responding to protein-related spectral features rather than chance correlations, a protein-focused model was developed using the same PLSR methodology as the full-spectrum model but restricted to three protein-relevant bands and a maximum of 15 components. Performance was compared to the full-spectrum model to assess whether targeted wavelength selection could maintain prediction accuracy while providing greater biological interpretability.

## 3 RESULTS AND DISCUSSION

### 3.1 Laboratory Assay CP Values

Laboratory assay CP concentration values are summarized in [Table 3](#tbl-protein-summary). These are similar to the range of values observed in the literature, indicating an reasonable basis for a chemometric model. The values were left-skewed (skewness of -0.29) and two thirds of the samples contained more than 250 g kg -1 CP.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Summary of laboratory-assayed CP values  Summary of Laboratory Assayed CP Values (g kg-1)   | Mean | SD | Minimum | First Quartile | Median | Third Quartile | Maximum | | --- | --- | --- | --- | --- | --- | --- | | 261 | 25 | 208 | 239 | 264 | 282 | 308 | |

### 3.2 Preprocessing Methods Comparison

Among the eight methods compared via ANOVA, the SNV-SG and SG preprocessing methods performed the best and were in the best highest-performing post-hoc comparison group. The SNV-SG method had the lowest RMSE and the highest R2, RPD and RPIQ averaging over all iterations ([Table 4](#tbl-preproc)). The superiority of SNV-SG by these metrics made it the best choice for the final model.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Evaluation of Preprocessing Methods by Metric ± Standard Error  Preprocessing Method Comparison (Mean ± SD)   | Preprocessing Method | RMSE (g/kg) | R² | RPD | RPIQ | | --- | --- | --- | --- | --- | | SNV + Savitzky-Golay | 10.3 ± 1.3 | 0.829 ± 0.062 | 2.48 ± 0.40 | 4.03 ± 0.87 | | Savitzky-Golay smoothing | 10.4 ± 1.2 | 0.826 ± 0.056 | 2.45 ± 0.39 | 3.98 ± 0.86 | | Raw spectra | 10.8 ± 1.5 | 0.816 ± 0.063 | 2.35 ± 0.41 | 3.83 ± 0.87 | | First derivative | 10.9 ± 1.4 | 0.806 ± 0.066 | 2.32 ± 0.37 | 3.77 ± 0.78 | | Standard normal variate (SNV) | 11.1 ± 1.7 | 0.804 ± 0.068 | 2.30 ± 0.43 | 3.75 ± 0.88 | | SNV + detrending | 11.1 ± 1.5 | 0.807 ± 0.062 | 2.29 ± 0.39 | 3.72 ± 0.81 | | Gap-segment derivative | 11.2 ± 1.7 | 0.805 ± 0.063 | 2.28 ± 0.40 | 3.70 ± 0.83 | | Multiplicative scatter correction | 11.3 ± 1.6 | 0.801 ± 0.067 | 2.26 ± 0.39 | 3.68 ± 0.86 | |

From the literature, these results are readily explained. Standard normal variate and SNV-detrend both correct light scatter, which is often a function of differences in particle size and sample packing density, although SNV-detrend is often used for densely-packed, powdered samples (Barnes et al., 1989). SG is a smoothing filter that regresses on the signal over a series of windows, removing noise while preserving the signal’s shape and features (Li et al., 2020; Luo et al., 2005). Derivatives, here including SG, gap-segment, and first derivatives pretreatments may remove additive and multiplicative effects, but not necessarily light scatter; as well, derivatives may increase spectral noise (Rinnan et al., 2009). Here, hemp grain was neither powdered nor densely packed but samples were subject to light scatter and noise due to differences in particle size in the hemp grain.

### 3.3 Algorithm Comparison

The algorithm comparison revealed clear performance differences among the three approaches, with PLSR achieving significantly superior performance across all evaluation metrics ([Table 5](#tbl-algorithm-comparison)). This suggests that the spectral features captured by PLSR are well-suited for hemp grain protein prediction, at least for screening, while the additional complexity of non-linear algorithms did not provide predictive advantage. This contrasts with researchers who obtained their best predictions of grains or seeds using machine learning methods other than PLSR (SVM and convolutional neural networks,respectively) , although in the present case the training sets were less than half the size of those used by the other researchers (Chadalavada et al., 2022; Huang et al., 2013).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5: Algorithm performance using RPD-primary unified evaluation criteria. Primary classification by RPD: Excellent (>3.0), Good (2.0-3.0), Fair (1.5-2.0), Poor (<1.5). RPIQ and R² provide supporting evidence.   | Algorithm | RMSE (±SD) | R² (±SD) | RPD (±SD) | RPIQ (±SD) | Excellent (%) | Good (%) | Fair (%) | Poor (%) | Quantitative (%) | Qualitative (%) | Total Usable (%) | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Partial Least Squares | 10.447 (±1.322) | 0.819 (±0.065) | 2.41 (±0.39) | 3.84 (±0.92) | 5% | 81% | 13% | 1% | 86% | 13% | 99% | | Random Forest | 13.685 (±1.57) | 0.697 (±0.095) | 1.83 (±0.25) | 2.91 (±0.61) | 0% | 24% | 68% | 8% | 24% | 68% | 92% | | Support Vector Machine | 12.501 (±1.555) | 0.743 (±0.08) | 2.01 (±0.3) | 3.19 (±0.71) | 0% | 45% | 52% | 3% | 45% | 52% | 97% |   Comparing algorithms |

The superior performance of PLSR in this relatively small data set validates its selection as the primary modeling algorithm. Statistical analysis confirmed significant differences between algorithms across all four performance metrics (p < 0.05), indicating robust performance differences despite the relatively modest sample size.

### 3.4 PLSR Full Spectrum Model Development

The model improved most rapidly as the number of components increased from one to seven, with the inclusion of each additional component being associated with a decrease in RMSE of 6%-11%. From eight to 12 components, model performance continued to improve, although gains were more modest. With 13 or more components, performance gains were minimal and the relative ranks of the models tended to be stable ([Figure 1](#fig-model-calibration)). The visible plateau in the plot at 12 components showed the point where the inclusion of additional components failed to improve model performance on the cross validation holdout set and indicated that the model was optimized and not overfit.

|  |
| --- |
| Figure 1: Decreasing RMSE with increasing number of components for 1000 training sets |

The performance of the final models on the testing sets is summarized in [Figure 2](#fig-performance-boxplot). The means of the final models were: RMSE = 9.94, R2 = 0.84, RPD = 2.5, and RPIQ = 3.94. More than 99% of the models had, at minimum, the ability to distinguish between high and low values, with 93.2% having some capacity for quantitative prediction (RPD > 2.0). Despite generally reasonable model performance, a subset of poor models can be seen in the boxplots.

|  |
| --- |
| Figure 2: Final model testing set performance over 1000 iterations |

This is a more lax standard for RPD than that set forth by P. Williams (2013), who recommended a minimum acceptable RPD value of 2.5 as suitable rough screening, such as in plant breeding programs. Utilizing those higher benchmarks, 59% of the models were sufficient for rough screening and 6% were suitable for screening, although none sufficed for quality or process control. However, models with lower RPD values can still be of use to researchers (Bellon-Maurel et al., 2010).

The pattern of test set errors was examined on a per-sample basis by calculating the difference between the actual and predicted values for the samples in the test sets ([Figure 3](#fig-validation-errors)). The models showed systematic bias, with overestimation in the lowest tertile and underestimation in the highest tertile samples. It is known that PLSR predictions are biased towards the mean values in the training set (Bellon-Maurel et al., 2010). The behavior of these predictions follows that pattern. The variance of the errors did not increase appreciably as CP concentration increased.

|  |
| --- |
| Figure 3: Testing set prediction errors on a per-sample basis. Actual sample value set to zero and samples ranked from least to greatest actual CP concentration value |

The 15 (10%) best and 15 worst predicted samples as measured by the mean absolute error of prediction were identified and their backgrounds examined. Overall, half of the samples in the data set came from Ithaca, while 28% were collected from Geneva, NY. However, of the 15 worst-predicted samples, ten were from Geneva, while four of the 15 best-predicted samples were from Geneva (by contrast, five of the best-predicted and five of the worst-predicted samples came from Ithaca). Overall, samples from Geneva had the highest mean absolute error (MAE) of prediction among locations, 63% greater than samples from Ithaca and 153% greater than samples from Freeville, NY (the only locations where more than 20 samples were assayed).

Trials in Geneva and Ithaca were generally planted and harvested near the same date, and samples were assayed with the NIRS instrument within the same few weeks between field seasons. The likeliest explanation is variation in sample-handling, particularly harvesting and grain-cleaning, or variation in environmental condititions across growing seasons. However, the exact reason for these differences is likely unknowable.

### 3.5 Protein-Focused Model Validation

The biological basis of the NIRS predictions was validated by developing a protein-focused model using wavelengths within known protein absorption bands. That model was compared to full-spectrum model performance in [Table 7](#tbl-performance-summary). The protein-focused model utilized five components, aligning with the advice to use 20 samples for calibration per principal component in PLSR (P. Williams, 2013). By that rubric, a model using five principal components would require at least 100 samples to calibrate it. Here 111 samples were used in calibration, exceeding that threshold. However, the full spectrum model significantly outperformed the full-spectrum model across all performance metrics.Eighty-six percent of the protein-focused models were minimally acceptable for qualitative screening (RPD > 1.5) but only 14% of the models were capable of approximate-quantitative screening or better.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 7: Comparison of Protein-Focused and Full Spectrum Models  Model Performance Summary Statistics (n=1000 iterations)   | Analysis Type | RMSE (g/kg) | R² | RPD | RPIQ | Components | | --- | --- | --- | --- | --- | --- | | Protein-Focused | 14.3 ± 1.6 | 0.669 ± 0.082 | 1.74 ± 0.23 | 2.74 ± 0.54 | 5.0 ± 0.1 | | Full Spectrum | 10.4 ± 1.3 | 0.822 ± 0.054 | 2.40 ± 0.36 | 3.78 ± 0.80 | 12.8 ± 1.7 | |

Variable Importance in Projection (VIP) scores revealed that the protein-focused model concentrated predictive power in the expected molecular absorption regions ([Figure 4](#fig-model-comparison)). The 1500-1550 nm region (N-H stretch from peptide bonds) showed the highest importance, followed by the 1200-1250 nm region (N-H + C-N combination bands). VIP scores greater than one appeared in all 1000 model iterations at the 1246, 1248, 1546, and 1548 nm wavelengths. This pattern aligns with established protein spectroscopy literature (Lorenz-Fonfria, 2020; Lukacs et al., 2024). Thirty-one full spectrum models contained a VIP score of greater than one. These scores were concentrated in the range between 1876 and 1908 nm a region associated with protein absorption (Wu et al., 2023). The VIP scores from either model could be used in subsequent model development (Z. Wang et al., 2022).

|  |
| --- |
| Figure 4: Comparing models developed using full and protein-specific spectra |

Despite its poorer performance, the protein-focused model provided biological validation of NIRS predictions as shown by the high VIP scores around 1250 and 1550 nm. This approach improved model interpretability by explicitly connecting predictions to known CP chemistry. That could improve model robustness and transferability across NIR instruments or environments, potentially enabling the transfer of calibrations to handheld or other instruments (Chadalavada et al., 2022; Lukacs et al., 2024). It is evidence that successful predictions are based on genuine CP-spectral relationships rather than correlations with other grain components such as oil, starch, or moisture.

## 4 Conclusions

The research showed the promise of the use of NIRS in order to make predictions, at minimum, suitable for use in a breeding program, concerning CP concentration in hemp grain using PLSR. Promising preprocessing methods were identified and a model was validated. Further research could refine the model by including more samples or exploring using other mechanisms for variable selection. In the present case it would likely be particularly useful to rectify the class imbalance between Geneva and Ithaca.

This study is limited in that it represents the creation of one model based upon spectra collected from one machine. This is insufficient for use in an industrial setting, where a minimum of three machines are suggested and the prediction achieved here is lower than those needed for many industries (P. Williams, 2013). NIRS calibrations can be unique to a particular machine, even if the machines compared are of the same model (Reeves, 2012). As well, the testing and training sets are relatively small.

However, these predictions may be sufficient for screening by a breeding program. If so, the biological validation of spectroscopic predictions opens new possibilities for accelerating hemp protein improvement through plant breeding. The non-destructive nature of NIRS allows breeders to analyze individual seeds while preserving their viability for planting, enabling selection decisions based on protein content before committing land and resources to field evaluation. For genomic selection programs, the ability to phenotype large populations (thousands of individuals) cost-effectively using NIRS could substantially improve the accuracy of genomic selection for protein content, accelerating genetic gains compared to traditional phenotyping approaches.

## 5 ACKNOWLEDGMENTS

This work would not have been possible without the efforts of the field staff, undergraduate, and graduate students who planted, maintained, monitored and harvested these trials. Funding was provided by New York State through a grant from Empire State Development (AC477). We are grateful to those who provided seed for this project, including: Uniseeds, Verve Seeds, Winterfox Farms, International Hemp, Fiacre Seeds, and KonopiUS.

## 6 CONFLICT OF INTEREST

The authors declare no conflict of interest.

**Generated:** 2025-07-11 10:26:23  
**Framework:** Targets-based reproducible workflow

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*(5), 772–777. <https://doi.org/10.1366/0003702894202201>

Bárta, J., Roudnický, P., Jarošová, M., Zdráhal, Z., Stupková, A., Bártová, V., Krejčová, Z., Kyselka, J., Filip, V., Říha, V., Lorenc, F., Bedrníček, J., & Smetana, P. (2024). Proteomic profiles of whole seeds, hulls, and dehulled seeds of two industrial hemp (cannabis sativa l.) cultivars. *Plants*, *13*(1), 111. <https://doi.org/10.3390/plants13010111>

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, *29*(9), 1073–1081. <https://doi.org/10.1016/j.trac.2010.05.006>

Callaway, J. C. (2004). Hempseed as a nutritional resource: An overview. *Euphytica*, *140*(1), 65–72. <https://doi.org/10.1007/s10681-004-4811-6>

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2018). Near-infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews*, *53*(8), 667–687. <https://doi.org/10.1080/05704928.2018.1425214>

Chadalavada, K., Anbazhagan, K., Ndour, A., Choudhary, S., Palmer, W., Flynn, J. R., Mallayee, S., Pothu, S., Prasad, K. V. S. V., Varijakshapanikar, P., Jones, C. S., & Kholová, J. (2022). NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. *Sensors (Basel, Switzerland)*, *22*(10). <https://doi.org/10.3390/s22103710>

Chinilin, A. V., Vindeker, G. V., & Savin, I. Yu. (2023). Vis-NIR spectroscopy for soil organic carbon assessment: A meta-analysis. *Eurasian Soil Science*, *56*(11), 1605–1617. <https://doi.org/10.1134/S1064229323601841>

Ely, K., & Fike, J. (2022). Industrial hemp and hemp byproducts as sustainable feedstuffs in livestock diets. In D. C. Agrawal, R. Kumar, & M. Dhanasekaran (Eds.), *Cannabis/hemp for sustainable agriculture and materials* (pp. 145–162). Springer. <https://doi.org/10.1007/978-981-16-8778-5_6>

Farrés, M., Platikanov, S., Tsakovski, S., & Tauler, R. (2015). Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, *29*(10), 528–536. <https://doi.org/10.1002/cem.2736>

Garrido-Varo, A., Garcia-Olmo, J., & Fearn, T. (2019). A note on mahalanobis and related distance measures in WinISI and the unscrambler. *Journal of Near Infrared Spectroscopy*, *27*(4), 253–258. <https://doi.org/10.1177/0967033519848296>

Geyer, M., Mohler, V., & Hartl, L. (2022). Genetics of the inverse relationship between grain yield and grain protein content in common wheat. *Plants*, *11*(16), 2146. <https://doi.org/10.3390/plants11162146>

Giancaspro, A., Giove, S. L., Blanco, A., & Gadaleta, A. (2019). Genetic variation for protein content and yield-related traits in a durum population derived from an inter-specific cross between hexaploid and tetraploid wheat cultivars. *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.01509>

Haruna, S. A., Li, H., Wei, W., Geng, W., Yao-Say Solomon Adade, S., Zareef, M., Ivane, N. M. A., & Chen, Q. (2022). Intelligent evaluation of free amino acid and crude protein content in raw peanut seed kernels using NIR spectroscopy paired with multivariable calibration. *Analytical Methods*, *14*(31), 2999–2989. <https://doi.org/10.1039/d2ay00875k>

Hayes, M. (2020). Measuring protein content in food: An overview of methods. *Foods*, *9*(10), 1340. <https://doi.org/10.3390/foods9101340>

Huang, Z., Sha, S., Rong, Z., Chen, J., He, Q., Khan, D. M., & Zhu, S. (2013). Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed. *Industrial Crops and Products*, *43*, 654–660. <https://doi.org/10.1016/j.indcrop.2012.08.015>

Kamboj, U., Guha, P., & Mishra, S. (2022). Comparison of PLSR, MLR, SVM regression methods for determination of crude protein and carbohydrate content in stored wheat using near infrared spectroscopy. *Materials Today: Proceedings*, *48*, 576–582. <https://doi.org/10.1016/j.matpr.2021.04.540>

Köllmann, N., Schreuders, F. K. G., Mishra, P., Zhang, L., & Goot, A. J. van der. (2023). Near-infrared spectroscopy-based quantification of sunflower oil and pea protein isolate in dense mixtures for novel plant-based products. *Journal of Food Composition and Analysis*, *121*, 105414. <https://doi.org/10.1016/j.jfca.2023.105414>

Li, Y., Huang, Y., Xia, J., Xiong, Y., & Min, S. (2020). Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies. *Vibrational Spectroscopy*, *108*, 103060. <https://doi.org/10.1016/j.vibspec.2020.103060>

Liu, M., Toth, J. A., Childs, M., Smart, L. B., & Abbaspourrad, A. (2023). Composition and functional properties of hemp seed protein isolates from various hemp cultivars. *Journal of Food Science*, *88*(3), 942–951. <https://doi.org/10.1111/1750-3841.16467>

Lorenz-Fonfria, V. A. (2020). Infrared difference spectroscopy of proteins: From bands to bonds. *Chemical Reviews*, *120*(7), 3466–3576. <https://doi.org/10.1021/acs.chemrev.9b00449>

Lukacs, M., Zaukuu, J.-L. Z., Bazar, G., Pollner, B., Fodor, M., & Kovacs, Z. (2024). Comparison of multiple NIR spectrometers for detecting low-concentration nitrogen-based adulteration in protein powders. *Molecules*, *29*(4), 781. <https://doi.org/10.3390/molecules29040781>

Luo, J., Ying, K., He, P., & Bai, J. (2005). Properties of savitzky–golay digital differentiators. *Digital Signal Processing*, *15*(2), 122–136. <https://doi.org/10.1016/j.dsp.2004.09.008>

Nogales-Bueno, J., Rodríguez-Pulido, F. J., Baca-Bocanegra, B., Pérez-Marin, D., Heredia, F. J., Garrido-Varo, A., & Hernández-Hierro, J. M. (2021). Reduction of the number of samples for cost-effective hyperspectral grape quality predictive models. *Foods*, *10*(2), 233. <https://doi.org/10.3390/foods10020233>

Reeves, J. B. (2012). Potential of near- and mid-infrared spectroscopy in biofuel production. *Communications in Soil Science and Plant Analysis*, *43*(1), 478–495. <https://doi.org/10.1080/00103624.2012.641844>

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201–1222. <https://doi.org/10.1016/j.trac.2009.07.007>

Roberts, C. A., Workman, J., & Reeves, J. B. (2004). *Near-infrared spectroscopy in agriculture*. American Society of Agronomy.

Shenk, J. S., & Westerhaus, M. O. (1991). Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. *Crop Science*, *31*(2), cropsci1991.0011183X003100020049x. <https://doi.org/10.2135/cropsci1991.0011183X003100020049x>

Shi, D., Hang, J., Neufeld, J., Zhao, S., & House, J. D. (2022). Estimation of crude protein and amino acid contents in whole, ground and defatted ground soybeans by different types of near-infrared (NIR) reflectance spectroscopy. *Journal of Food Composition and Analysis*, *111*, 104601. <https://doi.org/10.1016/j.jfca.2022.104601>

Tang, K., Struik, P. C., Yin, X., Thouminot, C., Bjelková, M., Stramkale, V., & Amaducci, S. (2016). Comparing hemp (*cannabis sativa* l.) cultivars for dual-purpose production under contrasting environments. *Industrial Crops and Products*, *87*, 33–44. <https://doi.org/10.1016/j.indcrop.2016.04.026>

Towett, E. K., Alex, M., Shepherd, K. D., Polreich, S., Aynekulu, E., & Maass, B. L. (2013). Applicability of near-infrared reflectance spectroscopy (NIRS) for determination of crude protein content in cowpea (vigna unguiculata) leaves. *Food Science & Nutrition*, *1*(1), 45–53. <https://doi.org/10.1002/fsn3.7>

Wang, Y., Zhao, C., Tian, H., Xing, Z., Yue, X., Liu, S., He, Y., Bai, J., Hao, L., Zhu, M., & Dong, D. (2025). NIRS-based detection advances in agriculture: Data enhancement, characteristic wavelength selection and modelling techniques. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *343*, 126611. <https://doi.org/10.1016/j.saa.2025.126611>

Wang, Z., Wu, Q., & Kamruzzaman, M. (2022). Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour. *Food Control*, *138*, 108970. <https://doi.org/10.1016/j.foodcont.2022.108970>

Williams, P. (2013). Calibration development and evaluation methods b. Set-up and evaluation. *NIR News*, *24*(6), 20–24. <https://doi.org/10.1255/nirn.1391>

Williams, P. C. (1975). Application of near infrared reflectance spectroscopy to analysis of cereal grains and oilseeds. *Cereal Chemistry*, *52*(4), 576–561.

Williams, P. C., & Sobering, D. C. (1993). Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy*, *1*(1), 25–32. <https://doi.org/10.1255/jnirs.3>

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. <https://doi.org/10.1016/S0169-7439(01)00155-1>

Wu, X., Zeng, S., Fu, H., Wu, B., Zhou, H., & Dai, C. (2023). Determination of corn protein content using near-infrared spectroscopy combined with a-CARS-PLS. *Food Chemistry: X*, *18*, 100666. <https://doi.org/10.1016/j.fochx.2023.100666>

Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, *93*(1), e1557. <https://doi.org/10.1002/ecm.1557>