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

2025-07-03

Abstract

The protein concentration of hemp (*Cannabis sativa* L.) grain is of interest to researchers, producers, and consumers. This study was conducted to determine whether hemp grain can be non-destructively assayed for crude protein (CP) concentration using spectra obtained from near-infrared spectroscopy (NIRS) to build a prediction model for crude protein concentration using partial least squares regression (PLSR). One hundred and fourty-nine whole hemp grain samples were obtained from 18 cultivar trials in New York (NY) from 2017-2021. The samples’ NIRS spectra were collected and the samples were ground and assayed by combustion. Seven potential preprocessing methods, as well as untransformed spectra, were tested using 100 training and testing set splits of the data and the best method was selected. That method was applied to 1000 additional splits of the data set. Model fit was evaluated using RMSE, R2, relative predicted deviation (RPD), and ratio of performance to interquartile distance (RPIQ). Once a preprocessing method was selected, the optimal number of model components and prediction performance on the testing sets were examined. A preprocessing method consisting of the standard normal variate transformation following a Savitzky-Golay filter had the lowest RMSE and the highest R2, RPD and RPIQ, with RPD and RPIQ 2.1%, and 2.4% higher than a Savitzky-Golay filter by itself (significant at <0.05). All preprocessing methods outperformed untransformed spectra. Optimal final models typically consisted of 12 components. Seventy-four percent of the 1000 final models had, at minimum, the ability to distinguish between high and low values of CP concentration, with 49% of the models capable of approximating quantitative prediction. The models tested to overestimate CP concentration by 0.5% in the lowest tertile of samples and underestimate CP concentration by 0.4% in the highest tertile of samples. The worst-predicted samples tended to come from Geneva, NY, possibly as a result of the models’ class imbalance (half of the samples were from Ithaca, NY while 28% were from Geneva). The research shows the promise that NIRS offers in the non-destructive assay of CP concentration in hemp grain.

Table of contents

CP, crude protein; 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

## 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. 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 protein (Bárta et al., 2024; Callaway, 2004; Ely & Fike, 2022; Liu et al., 2023). Crude protein is often used as 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) and the relation of the spectra to primary analytical values, typically obtained using wet chemistry assays, for components such as moisture, protein, fat, or fiber (Roberts et al., 2004). NIRS technology has been used since the 1970’s to assess forage CP concentration (Reeves, 2012; 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) is a typical method used in the agricultural and food sciences to relate spectra to analyte (Roberts et al., 2004). Partial least squares regression calculates components that maximize covariance between predictor and response variables. Partial least squares regression uses some number of components, often selected via cross-validation, to fit the regression model and is commonly used in spectroscopy because it tends to work well with highly correlated, noisy spectral data (Wold et al., 2001).

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 spectroscopic applications, particularly where non-linear relationships exist between spectral features and analyte concentrations (**chinilin2023?**). 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 (**kamboj2022?**). Random forest (RF) algorithms use ensemble learning to combine multiple decision trees and has been used to predict protein content in grains (**chadalavada2022?**; **haruna2022?**). Comparing multiple algorithms helps ensure that the best approach is chosen.

The biological validity of spectroscopic 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 true protein-spectral relationships rather than spurious correlations with other grain components. Three spectral regions with established vibrational assignments are: 1180-1230 nm (C-H stretch 2nd overtone from amino acid side chains), 1480-1530 nm (N-H stretch 1st overtone from peptide bonds), 2040-2070 nm (N-H + C-N combination bands) (Köllmann et al., 2023; **shi2022?**). These models may also be more readily transferable between instruments because they discard noisy, irrelevant portions of the spectra (Wang et al., 2025).

Non-destructive sampling of grain is attractive for plant breeders. A NIRS-scanned sample of whole grain may be used for other purposes besides the scan, including planting as a seed. In wheat and corn, grain protein 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, a benchtop NIR spectrometer was used to develop a model to 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.

## 2 MATERIALS AND METHODS

### 2.1 Hemp Grain Sample Background

Spectral data were obtained from whole (unground) hemp grain samples, harvested at maturity, collected from 2017–2021 from 18 cultivar trials in New York (NY) (149 samples). Grain samples were obtained by hand sampling or mechanical harvest and were cleaned of chaff and dried at 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 grain or dual-purpose types and included both commercially available and experimental material. Seventy-eight samples were scanned and assayed in 2017, 19 in 2018, 24 in 2019, and 28 in 2021. 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 cultivar 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 spectra were averaged at each 2 nm increment. All remaining samples from 2018-2021 were collected on a per-plot basis. 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.

### 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 common metric used to optimize the number of samples for chemical analysis (**nogales-bueno2021?**). Prior to selection, spectra were preprocessed using SNV (standard normal variate)-detrend with settings 1,4,4,1 for the derivative, gap, smooth, and smooth-two settings respectively.

### 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 R software and packages used

### 2.5 Model Development

Training and testing sets were created by dividing samples by their laboratory CP concentration values into tertiles to ensure that a representative range of values was present in both training and testing sets. 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 the caret package using PLSR. In fitting the model, the number of components was optimized over a grid search from 1-20. Model performance was evaluated with 25 iterations of bootstrapping and minimized RMSE in selecting the number of components in the final model.

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, 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. The answer to the question of exactly which values constitute a “good” model varies depending upon the reference consulted, but for simplicity’s sake the standard established for an acceptable model was R2 > 0.80, an RPD greater than 2.5 and ideally greater than 3 (“good” to “excellent” quantitative prediction), and an RPIQ greater than 2.3 but ideally greater than 4.1 prediction on the testing set (Chadalavada et al., 2022; Luce et al., 2017; Rawal et al., 2024).

Analyses of variance were performed for each of these metrics in order to compare preprocessing methods. For each ANOVA, each data set was considered as a subject and different variances were allowed for each preprocessing method. Once the most promising preprocessing method was identified, it was used in all subsequent analyses.

To optimze the 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.

All algorithms were evaluated using identical training and testing set splits and the optimal preprocessing method identified in the initial screening. Each algorithm was permitted optimal hyperparameter tuning using R’s caret package to ensure fair comparison. One hundred random splits into training sets and testing sets were performed for each algorithm, with model performance evaluated using the same metrics as the preprocessing comparison (RMSE, R², RPD, and RPIQ). 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. Performance on the testing sets was summarized with RMSE, R2, RPD, and RPIQ. 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 wavelengths. 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 2](#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 2: 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

All preprocessing methods outperformed untransformed spectral data, as shown in [Table 3](#tbl-preproc). The SNV-SG method had the lowest RMSE and the highest R2, RPD and RPIQ averaging over all iterations. The superiority of SNV-SG by these metrics made it the best choice for the final model.

### 3.3 Preprocessing methods comparison

All preprocessing methods outperformed untransformed spectral data, as shown in [Table 3](#tbl-preproc). The SNV-SG method had the lowest RMSE and the highest R2, RPD and RPIQ averaging over all iterations. The superiority of SNV-SG by these metrics made it the best choice for the final model.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: 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 | | First derivative | 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 | | Savitzky-Golay smoothing | 10.8 ± 1.5 | 0.817 ± 0.063 | 2.36 ± 0.40 | 3.84 ± 0.88 | | 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.

## 4 Algorithm comparison

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: Performance comparison of machine learning algorithms for hemp grain protein prediction   | Algorithm | RMSE (±SD) | R² (±SD) | RPD (±SD) | RPIQ (±SD) | Quantitative Capable (%) | Total Acceptable (%) | | --- | --- | --- | --- | --- | --- | --- | | Partial Least Squares | 10.447 (±1.322) | 0.819 (±0.065) | 2.41 (±0.39) | 3.84 (±0.92) | 86% | 99% | | Random Forest | 13.685 (±1.57) | 0.697 (±0.095) | 1.83 (±0.25) | 2.91 (±0.61) | 24% | 92% | | Support Vector Machine | 12.501 (±1.555) | 0.743 (±0.08) | 2.01 (±0.3) | 3.19 (±0.71) | 45% | 97% |   Comparing algorithms |

The algorithm comparison revealed clear performance differences among the three approaches, with PLSR achieving superior performance across all evaluation metrics ([Table 4](#tbl-algorithm-comparison)). This suggests that the spectral features captured by PLSR are well-suited for hemp grain protein prediction, while the additional complexity of non-linear algorithms did not provide predictive advantage. The superior performance of PLSR 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.

## 5 Final model development and summary

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 5%-12%. 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)).

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

The performance of the final models on the testing sets are summarized in [Figure 2](#fig-performance-boxplot). The means of the final models were: RMSE = NA, R2 = NA, RPD = NA, and RPIQ = NA. Seventy-four percent of the models had, at minimum, the ability to distinguish between high and low values with % having the ability for quantitative prediction. Despite the generally good model performance, a subset of poor models can be seen in the boxplots.

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

Finally, 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. 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, NY (“Ithaca”), while 28% were collected from Geneva, NY (“Geneva”) Table 3. However, of the 15 worst-predicted samples, nine were from Geneva, while three of the 15 best-predicted samples were from Geneva (by contrast, seven of the best-predicted and five of the worst-predicted samples came from Ithaca). Overall, samples from Geneva had the highest mean absolute error of prediction among locations, 61% greater than samples from Ithaca and 155% 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. However, the exact reason for these differences is unknowable.

## 6 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.

Enhanced performance comparison demonstrating biological validation through protein-specific wavelength selection

| Model | Wavelengths | Components | RMSE | R\_squared | Efficiency |
| --- | --- | --- | --- | --- | --- |
| Full Spectrum | 700 | 15 | 10.02667 | 0.8411829 | Baseline |
| Protein-Focused | 68 | 5 | 12.97435 | 0.7344186 | 9.7% of variables |

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

The protein-focused model successfully maintained predictive capability while using only 5% of the available spectral data, demonstrating that hemp grain protein predictions are indeed based on biologically-relevant absorption features ([Figure 4](#fig-model-comparison), (**tbl\_mod\_compare?**)). The targeted model achieved an average R² of 0.78 compared to 0.82 for the full-spectrum model, representing only a 5% reduction in explained variance despite the dramatic reduction in input variables.

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 1480-1530 nm region (N-H stretch from peptide bonds) showed the highest importance, followed by the 2040-2070 nm region (N-H + C-N combination bands). This pattern aligns with established protein spectroscopy literature and confirms that the models are responding to genuine protein-related molecular vibrations rather than spurious correlations.

The performance of the protein-focused model provides biological validation of the spectroscopic predictions. This approach improved model interpretability by explicitly connecting predictions to known protein chemistry., suggesting that future hemp protein calibrations could focus on these specific wavelength regions. This could potentially improve model robustness and transferability across NIR instruments or environments. This approach is evidence that successful predictions are based on genuine protein-spectral relationships rather than correlations with other grain components such as oil, starch, or moisture.

This study is limited in that it represents the creation of one model based upon spectra collected from one machine. 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.

The research showed the promise of the use of NIRS in order to make predictions concerning CP concentration in hemp grain using PLS. Promising preprocessing methods were identified and a model was validated. Further research could refine the model by including more samples, particularly by rectifying the class imbalance between Geneva and Ithaca.

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 reliable protein phenotyping enabled by NIRS provides the high-quality phenotypic data necessary for training genomic prediction models. 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 limited by analytical cost and throughput. The heritability of grain protein content in related cereals (wheat, barley) ranges from 0.6-0.8, suggesting that genetic improvement in hemp protein content is achievable with adequate phenotyping. NIRS provides the phenotyping capability necessary to exploit this genetic variation, particularly in early generation selection where large numbers of low-protein genotypes can be efficiently eliminated before expensive field testing.

## 7 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.

## 8 CONFLICT OF INTEREST

The authors declare no conflict of interest.

**Generated:** 2025-07-03 10:33:27  
**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>

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>

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>

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>

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>

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

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>

Luce, M. S., Ziadi, N., Gagnon, B., & Lévesque, V. (2017). Prediction of total carbon, total nitrogen, and pH of organic materials using visible near-infrared reflectance spectroscopy. *Canadian Journal of Soil Science*, *98*(1), 175–179. <https://doi.org/10.1139/cjss-2017-0109>

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>

Rawal, A., Hartemink, A., Zhang, Y., Wang, Y., Lankau, R. A., & Ruark, M. D. (2024). Visible and near-infrared spectroscopy predicted leaf nitrogen contents of potato varieties under different growth and management conditions. *Precision Agriculture*, *25*(2), 751–770. <https://doi.org/10.1007/s11119-023-10091-z>

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.

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>

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

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>