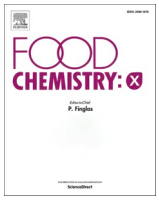
Food Chemistry: X 23 (2024) 101583

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



# High-throughput seed quality analysis in faba bean: leveraging Near-InfraRed spectroscopy (NIRS) data and statistical methods

Antonio Lippolis [a], Pamela Vega Polo [a], Guilherme de Sousa [a], Annemarie Dechesne [a], Laurice Pouvreau [b], Luisa M. Trindade [a,*]

[a] Plant Breeding, Wageningen University & Research, Droevendaalsesteeg 1, 6708PB, Wageningen, the Netherlands
[b] Wageningen Food & Biobased Research, Wageningen University & Research, Bornse Weilanden 9, 6708WG, Wageningen, the Netherlands

ARTICLE INFO

*Keywords:*
faba bean
Legumes
High-throughput phenotyping
Chemometrics
NIR spectroscopy
Machine learning
Bayesian statistics

ABSTRACT

Near-infrared spectroscopy (NIRS) provides a high-throughput phenotyping technique to assist breeding for improved faba bean seed quality. We combined chemical analysis of protein, oil content (and composition) with NIRS through chemometrics, employing Partial Least Squares (PLS), Elastic Net (EN), Memory-based Learning (MBL), and Bayes B (BB) as prediction models. Protein was the most reliably predicted trait ($R^2$ = 0.96–0.98) across field trials, followed by oil ($R^2$ = 0.82–0.86) and oleic acid ($R^2$ = 0.31–0.68). Samples for training the models were selected using K-means clustering. The optimal statistical approach for prediction was compound-specific: PLS for protein (Root Mean Squared Error - RMSE = 0.46), BB for oil (RMSE = 0.067), and EN for oleic acid content (RMSE = 2.83). Reduced training set simulations revealed different effects on prediction accuracy depending on the model and compound. Several NIR regions were pinpointed as highly informative for the compounds, using the shrinkage and variable selection capabilities of EN and BB.

## 1. Introduction

Faba bean (*Vicia faba L.*) is a legume crop that stands out for its high protein, yield potential, and nitrogen-fixing efficiency (Adhikari et al., 2021). Recently, the increasing plant-based proteins consumption has driven a growing research interest and investments in improving the seed quality of this crop in Europe. Breeding for improved quality aims to enhance protein content, reduce the level of anti-nutritional compounds (e.g., tannins, vicine and convicine), and to reduce off-flavours (e.g., lipid-derived compounds) (Lippolis et al., 2023). Breeding for quality requires phenotyping methods that are usually time-consuming and cost-inefficient, involving chemical analysis on a vast number of samples. Analysis of anti-nutritional compounds and off-flavours in faba bean seeds is typically performed using high liquid chromatography (HPLC) (Tacke et al., 2022), which is not a high-throughput method. For this reason, high-throughput phenotyping systems, including Near-Infrared Spectroscopy (NIRS), are becoming increasingly attractive to plant breeders (Gonçalves et al., 2021).

Samples scanned under NIR light produce a unique spectral response representing mainly hydrogen-containing functional groups, such as OH, CH, or NH (Ozaki & Morisawa, 2021). NIR spectra are correlated to the samples chemical compositions. Chemometrics models are pivotal to estimate compound quantities from the many overlapping peaks ('multicollinearity') present in NIRS data (Manley, 2014). Partial least squares (PLS) regression is the most used linear model in chemometrics to analyse multivariate, collinear, and noisy NIRS data. Other linear methods like Elastic Net (EN) are not as much explored in NIRS literature, despite their ability to deal with multicollinearity through variables shrinkage and selection (regularization) (Zou & Hastie, 2005). Regularization methods can also be implemented within a Bayesian statistics framework. Bayes B is one Bayesian model often used by plant breeders in Genomic Selection, a methodology used to predict plant performance from DNA (Meuwissen et al. 2021). Bayes B has been applied to NIRS data in plants in a very limited number of studies (Gonçalves et al., 2021), thus more research is needed to verify its predictive ability with spectroscopic data. Local models represent another family of methods that have been tested on NIRS data. These models capture relationships within specific data subsets by creating different models for each subset. Memory-based learning (MBL) is a powerful local modelling technique widely adopted in soil spectroscopy (Ramirez-Lopez et al., 2014), yet its application in agriculture and plants remains largely unexplored. MBL effectively resolves nonlinear

* Corresponding author.
*E-mail addresses:* antonio.lippolis@wur.nl (A. Lippolis), laurice.pouvreau@wur.nl (L. Pouvreau), luisa.trindade@wur.nl (L.M. Trindade).

https://doi.org/10.1016/j.fochx.2024.101583
Received 5 January 2024; Received in revised form 13 June 2024; Accepted 18 June 2024
Available online 26 June 2024

relationships that may be present in NIRS data by breaking down a global model into a series of simpler local models. Recently, the growing research on NIRS has led to more complex predictive methods in various fields of application such as food, agriculture, and medicine. These methods include support vector machines (SVMs), which use kernel functions, and neural networks, including Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) (Anderson et al., 2021; Mishra et al., 2022). Although they are capable of capturing complex patterns, including data non-linearity, they require extensive datasets for training, laborious and time-consuming hyperparameters tuning, and significant computational power. Thus, their use in research is justified only with large datasets. It must be said that there is no best method among the different NIRS modelling approaches, as each has its advantages and disadvantages (Lucà et al., 2017), and their specific performance depends on several experimental factors as well as on the chemical compound of interest.

A NIRS prediction pipeline starts with the development of calibration models using samples of known chemical and spectral composition (training or calibration set data). These models are then used to rapidly screening unknown samples based on their spectral data alone (Næs et al., 2002). In plant breeding, models tailored to specific years, genetic material, or growing conditions may not generalize well to new contexts (lack of robustness). A robust model typically requires a carefully selected training set that is representative of the chemical, physical (e.g., colour, seed size, etc.) and spectral variability present in a specific crop (Anderson et al., 2021; Nicolai et al., 2007). Training set samples are often randomly selected. This selection may not ensure that the training set is the best possible representation of the entire sample population, especially if the sample size is small. Sampling algorithms can be used to select samples based on their spectral characteristics to optimize the design of the training set. Currently, there are no established methods for optimizing training set design in the faba bean literature. However, more in-depth investigations have been carried out in soil science. Ramirez-Lopez et al. (2014) compared different sampling algorithms to assess their effectiveness in selecting samples representative of the spectral features of the entire population. In addition, the training set size is also a pivotal factor in building robust models. Once calibration models have been established, it is advisable to validate their accuracy with external samples (independent validation sets) (Anderson et al., 2021; Nicolai et al., 2007). This involves testing models that have been developed using data from specific field trials to predict the chemical compositions of seeds harvested in different locations, in different years, or from different genetic material. However, calibration models that report excellent performance often lack proper external validation (independent set validation), which is necessary to test the robustness of the model.

Attempts to use NIRS to predict seed quality attributes are well documented in the faba bean literature, including predictions for protein, moisture, starch, oil, total polyphenols, tannins, vicine, and convicine content (El-Sherbeeny & Robertson, 1992; Puspitasari et al., 2022; Wang et al., 2014). Lately, Johnson et al. (2023) investigated the prediction of bioactive compounds such as antioxidants (iron-reducing antioxidant power) and phenolics in faba bean flour. However, most studies lacked comprehensive methodological documentation and focused on seed quality rather than the development of high-throughput strategies. Additional efforts are needed to guide faba bean breeders and researchers in adopting a clear NIRS predictive approach, from calibration design to validation. To the best of our knowledge, no studies have yet detailed the optimal application and comparison of sampling algorithms for designing efficient training sets based on spectral features. Furthermore, it is rare to find studies that examine predictability across different growing conditions, such as field trials in various locations. Across-trials prediction has been investigated solely by Johnson et al. (2023) in the faba bean NIRS research. Moreover, PLS regression (or some modified versions) dominates the literature, and no alternative methods have been investigated for *Vicia faba* so far.

In this research, we aimed to establish clear guidelines for the efficient use of NIRS by developing a reproducible, license-free pipeline that has been validated in real plant breeding scenarios. We first investigated three different sampling algorithms to effectively design robust training sets. We provided a rapid method for comparing these algorithms, defining the best algorithm as the one that allows the selection of a training set that best represents the variability of the entire population. Secondly, we questioned whether Elastic Net (EN), Memory-Based Learning (MBL), and Bayes B (BB) could improve prediction accuracy compared to the widely used Partial Least Squares (PLS) regression, particularly in the context of predictions across breeding trials. Notably, EN and BB are not commonly used in NIRS prediction, but they are well-suited for handling high-dimensional data, such as spectroscopic datasets. Another objective was to identify cost-effective NIRS analysis strategies by investigating how different models and traits (chemical compounds) respond to different training set sizes. In summary, this work aimed to enhance the role of NIRS in faba bean phenotyping, focusing on protein, oil, and oleic acid content as key compounds. This study provides a replicable strategy for other traits and crops, contributing to informed decision making and increased efficiency in breeding and research for improved quality.

## 2. Material and methods

### 2.1. Plant material and NIR spectra acquisition

Seed samples were harvested from two different field trials in the Netherlands in September 2021 (trial 1) and in September 2022 (trial 2). In trial 1, 409 plots were harvested and in trial 2, 532 plots were harvested. In total, the experiment included 250 different genotypes. The seeds were ground to $<0.5$ mm using a Pulverisette-14 rotor mill (Fritsch, Germany) and analysed with a Vis-NIR spectrometer (FOSS, DS2500 Analyzer). The samples were scanned after each harvest to obtain spectra in the form of absorbance (log 1/reflectance) with a 2 nm resolution, and in the range from 1100 to 2500 nm. After spectra collection, the samples were stored at $-18$ °C prior to chemical analysis.

### 2.2. Design of the training (calibration) and validation set

#### 2.2.1. Training set

From trial 1, 125 samples were selected for further chemical analysis (training set). The training set design was based on the spectral variability and included the following steps:

1. Principal component analysis (PCA)

The data were pre-processed using Standard Normal Variate (SNV) coupled with an 11-point 2nd-order Savitzky–Golay and 2nd derivative filter, implemented in the prospectr R package (Stevens & Ramirez-Lopez, 2022). Pareto scaling was additionally applied to prevent variables with small variance and a low signal-to-noise ratio from becoming overly influential. Six principal components were selected to account for approximately 90% of the spectral variance. The presence of outliers was investigated by calculating orthogonal (Q) and score ($T^2$) distances and using standard critical limits in the mdatools R package (Kucheryavskiy, 2020).

2. Design of three different training sets

Three different potential training sets were created using three sampling algorithms implemented in the prospectr R package: 1) the Kennard-Stone algorithm (KS), using Mahalanobis distance to calculate dissimilarities on PC scores; 2) the K-means (KM) algorithm, selecting the samples nearest to each of the 125-clusters centre ($k = 125$); 3) the Puchwein algorithm (PW), using an initial distance of 0.2. These methods are described and referenced in Stevens and Ramirez-Lopez (2022).

3. Comparison of training sets

Four probability density functions (pdfs) were calculated on the PC scores from the PCA: one representing the whole population and one for each of the three training sets. The representativeness of a specific training set was assessed by comparing its pdf with that of the entire population (Ramirez-Lopez et al., 2014) using the Kullback-Leibler divergence (KL) metric. The KL distance metric is used to quantify the difference between pdfs. The closer the pdf of a training set is to the pdf of the whole population, the more representative the training set is of the whole population. Larger KL values indicates that the training set is unbalanced in terms of spectral representativeness. The algorithm producing the lowest KL value was chosen as the best option, as lower values indicate closer similarity between pdfs and thus higher representativeness.

#### 2.2.2. Validation set

From trial 2, 67 samples were selected as the validation set, using the same pre-processing as the training set for consistency. Spectral diversity between the trials was assessed using PCA with six PCs. The Kennard-Stone algorithm was forced to initialize the sampling from the training set samples. If samples from trial 1 were in the validation set, they were replaced by the most comparable samples from trial 2, based on a pairwise similarity matrix calculated using the Mahalanobis distance in the resemble R package (Ramirez-Lopez et al., 2016).

#### 2.2.3. Reduced training set size

After establishing the main training set, three additional subsets were generated. These subsets contained 70%, 40%, and 20% of the samples from the original training set, respectively. The KS algorithm was used to ensure that the samples in the smaller subsets were consistently drawn from the larger subsets.

### 2.3. Reference chemical analysis

The protein, oil, and oleic acid contents were measured for the samples included in the training and validation sets.

#### 2.3.1. Total oil

Oil was extracted from 3 g of ground seeds using 30 ml hexane. The mixture was shaken at 40 °C and 600 rpm (revolutions per minute) for 30 min, and then centrifuged at 4200 rpm for 5 min. The supernatant was evaporated in a vacuum evaporator. The extraction process was repeated three times, with each sample tested in triplicate. Oil content was determined by the difference in weight between the initial empty tubes and the same tubes containing oil after the evaporation of hexane. Data were adjusted for dry matter, which was calculated by drying the seed powder at 103 °C for 36 h.

#### 2.3.2. Total protein

Protein content was measured using the Dumas method. Approximately 250 mg of seed powder was combusted in the Rapid N Exceed analyzer (Elementar, Germany). Nitrogen oxides were quantified to determine total nitrogen (%N). A conversion factor (CF) of 6.25 was used to determine the crude protein content (%N x CF). Samples were analysed in duplicate, and data was corrected for dry matter.

#### 2.3.3. Oleic acid

Gas chromatography (GC) with a Flame Ionization Detector (FID) was used to measure the oleic acid content as a Fatty Acid Methyl Ester (FAME). Nonadecanoic acid (C19:0) triacylglycerol was used as an internal standard to check if the hydrolyses was complete. A stock solution of 1 mg $ml^{-1}$ was prepared. The previously extracted oil was mixed with the standard (0.1 ml), hexane (10 mg oil 0.9 ml $hexane^{-1}$), and 4.25 M potassium hydroxide in methanol (KOH/MeOH) solution (60 µl), and then incubated at 60 °C for 10 min. After centrifugation, the supernatant was analysed with the GC-FID (Agilent model 7890B) using a 30 m × 320 µm × 0.25 µm Agilent column (product DB-23). Hydrogen ($H_2$) was used as the carrier gas. A 1 µl sample was injected into the inlet, which was heated to 260 °C. The inlet pressure was 4.2566 psi (pound-force per square inch) and the split ratio was 1/50. The oven was held at 140 °C for 1 min before the temperature was increased at a rate of 4 °C $min^{-1}$ to 220 °C and held for 5 min. The chromatographic data were processed using MS ChemStation (Agilent Technologies, USA). The retention time of the FAMEs was compared with commercial standards for the identification of oleic acid. Oleic acid was quantified based on peak area ratios. Samples were analysed in duplicate.

### 2.4. Spectral pre-processing

The spectral pre-processing was first optimized for the PLS model on the validation set using the Q statistic (spectral reconstruction error) as suggested by Summerauer et al. (2021). Various Savitzky–Golay (SG) filter combinations were tested, including different derivatives and polynomial approximations, coupled with standard normal variate and detrend techniques. This initial investigation revealed that Savitzky–Golay filters with a first derivative minimized Q with the fewest latent variables. This pre-processing was therefore selected for a trial-and-error approach in which different window sizes were tested. The final choice on the best pre-processing was based on the Root Mean Squared Error of cross-validation ($RMSE_{cv}$) of all three compounds simultaneously (Supplementary Table 1). The same data pre-processing method was subsequently applied to EN, MBL, and BB.

### 2.5. Statistical approaches for NIRS models development

#### 2.5.1. Partial least squares (PLS) regression

Partial Least Squares models the linear relationship between spectra (X) and chemicals (Y) by projecting them into a lower dimensional space. It extracts orthogonal factors, or latent variables (LVs), from the spectra using information from X and Y simultaneously.

The PLSR model can be expressed as given in eqs. (1) and (2):

$$X = TPT' + E \quad (1)$$

$$Y = UQT' + F \quad (2)$$

where X is the spectrum matrix and Y is the response matrix. T and U are score matrices corresponding to X and Y respectively, while P and Q are their respective loading matrices. E and F symbolize the residual matrices for X and Y respectively. To establish a regression between X and Y, eq. (3) is used:

$$U = T\beta \quad (3)$$

where β is the vector of regression coefficients in the linear model. Substituting this relationship into the original model, we can derive the predictions as in eq. (4):

$Y = UQT' + T\beta QT'$ (4).

PLSR was fitted using the caret R package (Kuhn, 2008). The optimal number of LVs was determined by 10-fold cross-validation.

#### 2.5.2. Memory-based learning (MBL)

Memory-based learning is a machine learning approach that predicts new observations based on the most similar samples from the training set (called nearest neighbours), without constructing a general (global) model from the data (Ramirez-Lopez et al., 2014). MBL first creates a p-dimensional space from the spectra, where 'p' is the number of principal components. For each sample to be predicted, the algorithm identifies the nearest neighbours (closed samples) from the training set and determines their optimal number (k). A local model is then fitted to each sample using only the spectra of its nearest neighbours.

The optimal k values were determined by a leave-nearest-neighbour-out cross-validation (NNv) procedure, ranging from 8 to 125 in increments of 5, and using the Spectral Angle Mapper (SAM) as the dissimilarity metric. The pairwise dissimilarity matrix between all k-neighbours and the predictor variables were used jointly as a source of predictors.

Weighted Average Partial Least Squares regression (WAPLS) was used for the local models. Specifically, a modified version of the PLS algorithm was used. WAPLS fitted multiple models including 5 to 7 LVs, with results presented as weighted averages of all predicted values. The weight for each component was calculated as $W_j = \frac{1}{S_{1:j} \times g_j}$ ; where $S_{1:j}$ is the root mean square of the spectral reconstruction error of the unknown observations when using $j$ pls components, and $g_j$ is the root mean square of the squared regression coefficients corresponding to the $j$th PLS component.

MBL was fitted using the resemble R package (Ramirez-Lopez et al., 2016). The package includes all the references for the methods mentioned above (e.g, WAPLS).

#### 2.5.3. Elastic net (EN) regression

Elastic Net is a statistical method that regularizes or constrains the coefficient estimates of $p$ predictors by shrinking them towards zero or even exactly to zero (Zou & Hastie, 2005). The Elastic Net penalty is a convex combination of two penalties: the Ridge Regression (RR) penalty, which imposes a constraint on the sum of the squares of the regression coefficients, and the Least Absolute Shrinkage and Selection Operator (LASSO) penalty, which imposes a constraint on the sum of the absolute values of the regression coefficients. The mathematical representation of EN is given in eq. (5):

$$\min_{\beta} \sum_{i=1}^{n} \left( yi - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{J=1}^{p} \beta_j^2 \quad (5)$$

Here, eq. (5) consists of three primary components:

1. The left part of the equation represents the usual least squares criterion, which minimizes the sum of squared residuals.
2. The second term, $\lambda_1 \sum_{j=1}^{p} |\beta_j|$ , is the penalised sum of the absolute values of the regression coefficients, reflecting the LASSO penalty.
3. The third term, $\lambda_2 \sum_{J=1}^{p} \beta_j^2$ , is the penalised sum of the squared values of the regression coefficients, reflecting the RR penalty.

The parameters $\lambda_1$ and $\lambda_2$ are the shrinkage factors for these terms, controlling the degree of penalization. Their optimal values were determined by a 10-fold cross-validation procedure. The hyperparameter $\alpha$, which control the balance between the RR and LASSO penalties, was set to 0.5.

EN was fitted using the glmnet R package (Friedman et al., 2010). Variable shrinkage and selection were visualized by plotting the absolute normalized coefficients. Normalization was carried out using the most important coefficient as the reference.

#### 2.5.4. Bayes B (BB)

Bayes B is a regression method implemented within a Bayesian framework, where model parameters are considered as random variables with prior distributions that are updated after data collection via Bayes' theorem (Meuwissen et al., 2021). This model was initially implemented for genomic prediction. BB assumes that only a fraction of the wavelengths affect the trait (chemical compounds), and that the effect of each wavelength has a different variance.

BB assigns prior distributions to a collection of unknown parameters $\theta$, including the intercept $\beta_0$, regression coefficients $\beta_j$, hyperparameters $\Omega$ and the residual variance $\sigma_e^2$. The prior density formula was as in eq. (6):

$$p(\theta) = N(\beta_0|0, 1 \times 10^5)\chi^{-2}(\sigma_e^2|df_e, S\varepsilon)\left\{\Pi_{j=1}^{n} p(\beta_j|\Omega)\right\}p(\Omega) \quad (6)$$

Here, the intercept is assigned a normal prior with a very large variance, essentially treating the intercept as a "fixed" effect. The residual variance is assigned a scaled-inverse chi-squared density ($\chi^{-2}$) with degree of freedom $df_e$ and scale parameters $S\varepsilon$. Wavelength effects have assigned priors, $p(\beta_j|\Omega)$, indexed by a set of hyperparameters $\Omega$ that are also treated as random with prior distribution $p(\Omega)$. In Bayes B, $p(\beta_j|\Omega)$ is a mixture of a point of mass at zero and a scaled $t$-density, or: $p(\beta_j|\Omega) \sim \pi \text{ x } t\left(\beta_j|df_\beta, S_\beta\right) + (1-\pi) \text{ x } 1 \ (\beta_j = 0)$. Therefore, a priori, with probability $\pi$, $\beta_j$ has an effect drawn from the $t$-density and with probability $(1-\pi)$ $\beta_j = 0$, having no effect. Using the BGLR R package (Pérez & de Los Campos, 2014), $df_\beta$ was set to 5 and the other hyperparameters were treated as random. Other parameters were: *nIter* = 200,000 (number of iterations of the sampler), *burnIn* = 15,000 (the number of initial samples discarded), and *thin* = 5 (thinning used to compute posterior means). The importance of the wavelengths for the prediction was visualized as above for EN.

### 2.6. Training and evaluation of NIRS prediction models

The model parameters optimization (training phase) of PLS, EN, and BB was conducted using 10-fold cross-validation (CV), or leave-nearest-neighbour-out cross-validation (NNv) in the case of MBL. The optimal parameters for each model were defined as those minimizing the RMSE of these cross-validations ($RMSE_{CV}$/$RMSE_{NNv}$). An example of PLS for oil prediction is shown in Supplementary Fig. 1. The performance of the final model was assessed in an independent validation set using Root Mean Squared Error (RMSE) as the primary metric, supplemented by the Residual Prediction Deviation (RPD), and the Coefficient of Determination ($R^2$). NIRS models were developed using the full size of the training set, but also using the reduced training sets (see 2.2.3). The resulting models were always evaluated on the same independent set (see 2.2.2). The full size set for oleic acid content comprised a total of 101 samples.

## 3. Results and discussion

### 3.1. K-means clustering designed a representative training set

The spectral variability of the samples harvested in trial 1 was investigated by PCA (Supplementary Fig. 2). None of the samples were outliers based on the squared orthogonal Euclidean distance (q) and Hotelling T2 distance (h). However, 27 of the 409 total samples were extreme (Supplementary Fig. 3), showing a deviating behaviour expected due to sample heterogeneity, and chemical or physical variations.

Three different training sets were selected using the Kennard-Stone, K-means, and Puchwein sampling algorithms. The Kullback & Leibler divergence (KL) metric values were 2.84, 3.04, and 3.39 for K-means, Kennard-Stone, and Puchwein, respectively. The lowest value of 2.84 suggested that K-means is the most effective algorithm for selecting 125 samples that best represent the entire spectral population. In fact, lower KL values indicate a greater similarity of the initial population by the training set in terms of score distributions (Fig. 1 A-C). Although we highlighted the K-means algorithm as the best method, it remains unclear whether this finding can be generalized or is dataset specific. Therefore, further research should compare different sampling algorithms based on this proposed rapid comparison. A more precise comparison of sampling algorithms would require the chemical analysis of all the different training sets and the development of NIRS models using each set, which would drastically increase the cost.

It is worth noting that K-means clustering ensured a broad representation of genetic diversity in the training set. Positively, 109 of the

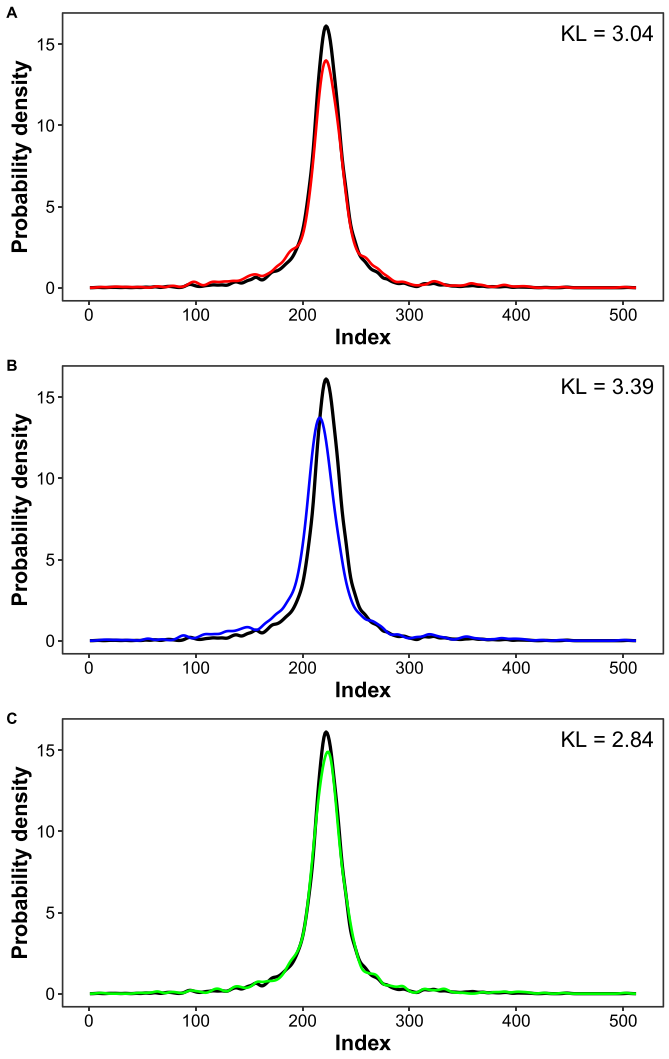



**Fig. 1. A-C.** Probability density functions of the principal component scores for the three selected calibration sets (coloured lines) compared to the density of the total population (black lines). A) Kennard stone sampling in red (KL = 3.04); B) Puchwein sampling in blue (KL = 3.39); C) K-means sampling in green (KL = 2.84). K-means produced the lowest KL value, thus the calibration set that best represented the variability of the entire dataset (overlap between the distributions). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

125 selected samples had unique genotypes, and only 16 samples represented replicates in the field of one of these 109 genotypes (i.e., replicate 1 and replicate 2 of the same genotype). The NIRS literature often fails to detail the differences in sample genetics present in the training set. However, a large number of samples with identical genetics could pose an issue during the cross-validation procedure. Cross-validation (CV) assumes that samples in the training set are independent (Rabinowicz & Rosset, 2022). A recent comprehensive study involving 60 samples from 10 Australian faba bean varieties showed that the inclusion of field replicates in the training set is indeed a common practice (Johnson et al., 2023). These samples may not be considered independent. In our study, it is positive that only ~10% of the calibration samples were field replicates; moreover, they were sampled from different spectral clusters, reducing the aforementioned dependency problem in the CV.

### 3.2. Training and validation sets differed in chemical and genetic variability

From trial 2, 67 samples were selected as the validation set to test the robustness of the models across breeding trials. The validation set differed in chemical variability from the training set (Fig. 2 A-C). This diversity is a desired feature in this specific context, as the intention was to design an independent and unbiased set for the validation and comparison of the models.

The protein content of the training set ranged from 18.39% to 34.39%, with an average of 25% and a coefficient of variation (CV) of ~11%. The validation set showed a higher average (30%), ranging from 21% to 35%, with a CV of ~10%. These values covered the full range of protein variation expected in faba bean breeding material. For example, an evaluation of 840 inbred breeding lines with a wide genetic background showed variation in protein content ranging from 18% to 30% (El-Sherbeeny & Robertson, 1992). The large variation was expected as our samples covered different botanical groups and were collected from around the world. Despite having a smaller sample size compared to Wang et al. (2014), the calibration set we designed showed a greater variability. The inclusion of extreme values (low or high protein) helps to prevent bias towards under- or over-prediction of protein content.

The average oil content of the training set was 1.69%, higher than the average content of the validation set (1.44%). The range of variation was from 1.35% to 2.17% for the calibration set and from 1.17% to 1.92% for the validation set, with CVs of ~9% and 10%, respectively. Wang et al. (2014) recorded an oil range from 0.48% to 1.99% in faba bean. Therefore, material with very low oil (<1%) was not included in this study. The slight differences in oil content between studies could be due to the different solvents and extraction methods used. However, if the very low oil content is due to genetic factors, the inclusion of this genetics would extend the calibration curves and facilitate the breeding of ultra-low oil varieties. For oleic acid content, the training set consisted of 101 samples, ranging from 13.72% to 30.10%, with a mean of 18.65% and a CV of ~13%. The validation set of 57 samples ranged from 13.72% to 26.76%, with a mean of 18.65% and a CV of ~15%. Oleic acid's variability was consistent with previous chemical analyses in faba bean (Welch and Wynne Griffiths, 1984).

Moreover, the validation set differed in genetic background, with 68% of the samples being genetically distinct from the calibration set. The KS algorithm identified new spectral variation introduced into the dataset, suggesting a correlation between spectral variation and genetic background. When applied to a combined dataset of trial 1 and trial 2, the algorithm (using the calibration set samples as forced initializing points) selected only 8 out of 67 samples from trial 1. This approach effectively increased the chemical, spectral, and genetic variability of the population used to train and validate the NIRS models. The main benefits were an unbiased validation set and the potential to expand the predictive range of the model by later incorporating these validation samples into future training sets.

### 3.3. Savitzky–Golay filter removed noise from raw NIRS data

Easily visualized peaks in the raw spectra were located at 1192–1204 nm, 1452–1502 nm, 1738–1782 nm, 1932–1944 nm, 2094–2140 nm, and 2306–2348 nm (Fig. 2 D). The raw data showed sloping baselines due to light scattering effects and other instrumental variations. It is well known that multivariate analysis performed on the spectra can be affected by additive and multiplicative scattering effects (Rinnan et al., 2009). A Savitzky–Golay (SG) filter with a first derivative and a first order polynomial applied to a window size of 7 points was used to minimize such effects. This pre-processing removed the baseline shift and highlighted informative peaks (Fig. 2 E). The results of the trial-and-error approach used to select the best pre-processing are reported in the Supplementary Table 1.

### 3.4. Prediction accuracy decreased from protein to oil and oleic acid

Partial Least Squares (PLS), Elastic Net (EN), Memory-based Learning (MBL), and Bayes B (BB), were employed to estimate protein, oil, and oleic acid content in faba bean seeds. In terms of overall predictive ability, protein content was the most reliably predictable trait ($R^2$ = 0.96–0.98; RPD = 4.05–6.6), followed by oil content ($R^2$ = 0.82–0.86; RPD = 1.53–2.34), and oleic acid content ($R^2$ = 0.31–0.68, RPD = 0.68–1.04). The difference in predictive ability is illustrated by the $R^2$ in Fig. 3 A.

Wang et al. (2014) previously found oil less predictable than protein in faba bean, aligning with other findings in pea (Hacisalihoglu et al., 2020). Notably, faba beans and peas often store ten times more protein than oil. The amount of specific chemical compounds influences the prediction accuracy, as it affects the detection threshold (Johnson et al., 2020). This may partly explain why oil content was less predictable than protein. However, differences in the molecular composition of these two components also play an important role (Cem & Kahriman, 2012). To date, faba bean literature lacks information on predicting oil composition, mainly because of the lack of interest in fatty acids in non-olinageous crops. Most NIRS validations have focused on crops with medium to high oil content, including soybean (14–24%) (Leite et al., 2020) and sesame (40.7–58.4%) (Tsegay et al., 2023). A potential limitation of our study regarding oleic acid prediction arises from using the relative peak areas from the chromatogram as response variables. This widely adopted approach does not provide absolute quantification. Absolute quantification could further improve the prediction, as prediction accuracy is highly dependent on the accuracy and precision of the reference method (Manley, 2014).

Protein content has been effectively predicted in previous studies on faba bean. Wang et al. (2014) predicted proteins with an RMSE of 0.34, an $R^2$ of 0.94, and an RPD of 4.05. Their results align closely with our PLS model, which exhibited even higher $R^2$ (0.98) and RPD (6.66). In our study, the models' robustness was assessed on independent samples, which were collected from a different field trial and featured new genetic backgrounds. Such comprehensive validation is rare in the faba

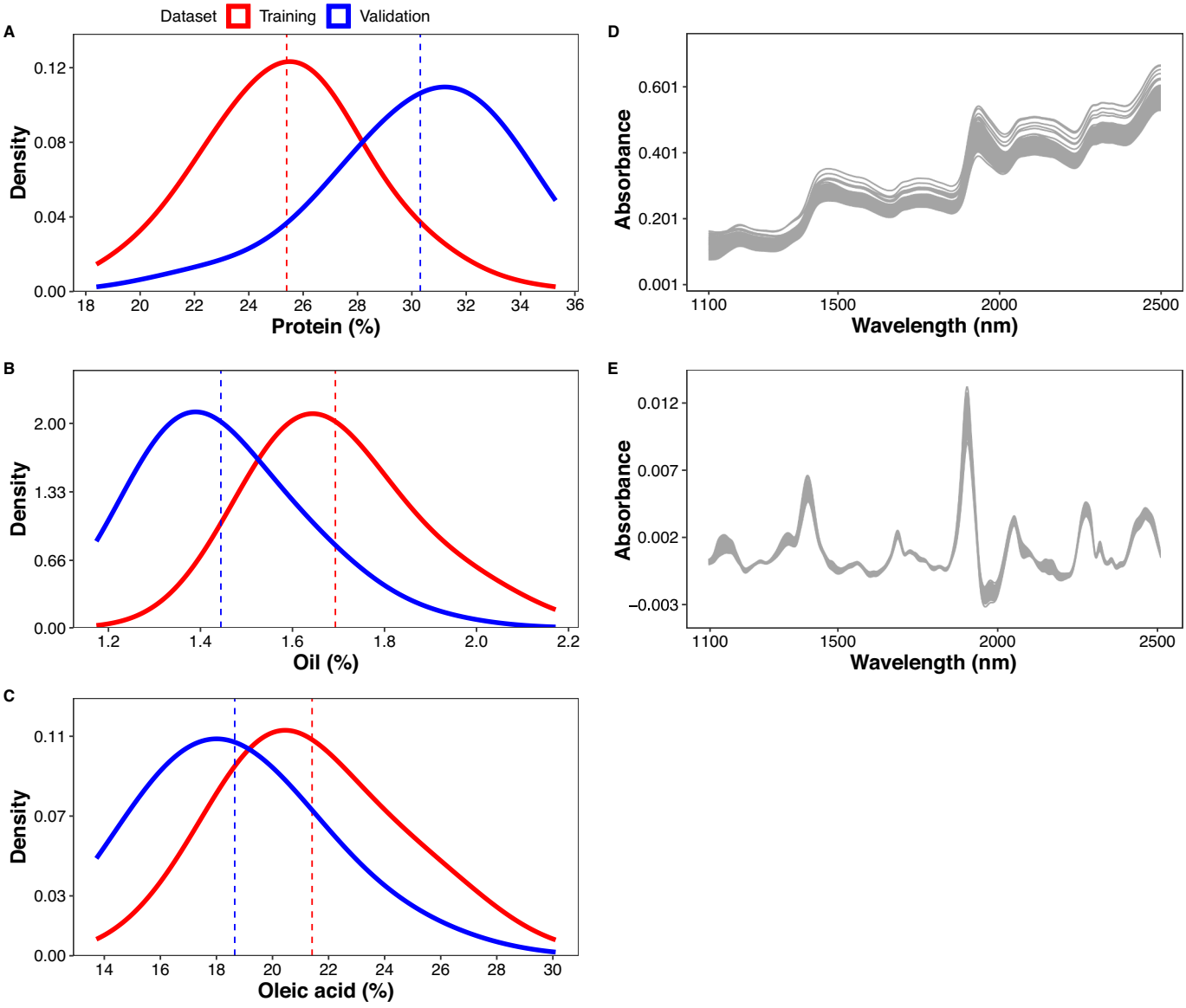



**Fig. 2. A-C.** Variability in protein (A), oil (B), and oleic acid (C) of the training set (red) and validation set (blue). The dashed lines indicate the respective mean values. **D-E.** The figures depict raw (D) and pre-processed (E) NIR spectra. D shows differences in absorbance values (y-axis) at the starting wavelength (x-axis), indicating a baseline shift due to various factors including light scattering. E shows the NIR spectra after pre-processing, which removed the baseline shift and associated noise. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

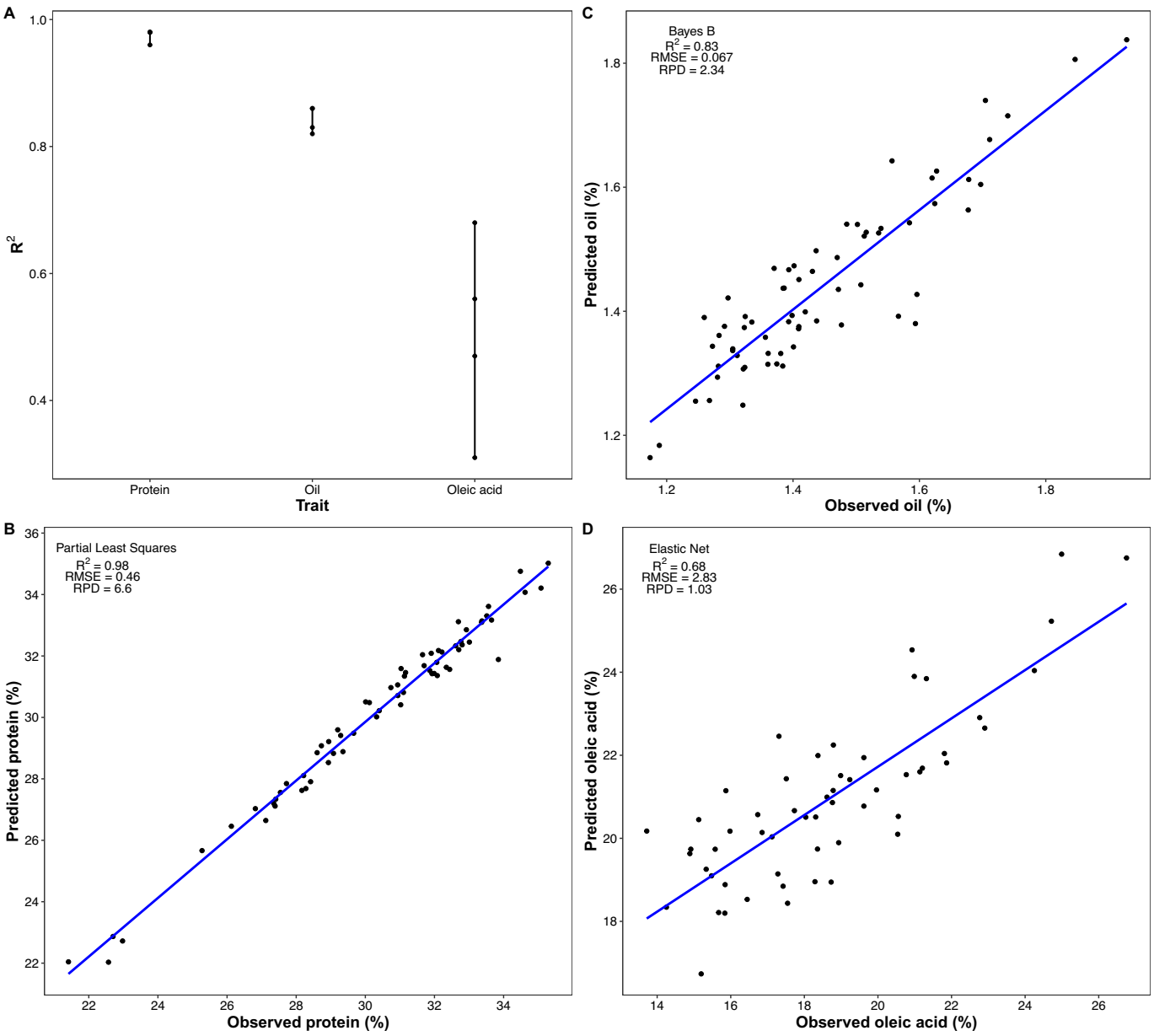



**Fig. 3. A.** Coefficient of determination ($R^2$) for the prediction of protein, oil, and oleic acid. The intervals defined by multiple black dots represent the output from the four different predictive models. Where fewer than four points are shown, this indicates that some models gave identical $R^2$. The data suggest that the lower the predictability of a compound by NIRS (e.g., oleic acid), the greater the variability of $R^2$ between models. **B—D.** Plots of observed versus predicted values for protein (B), oil (C), and oleic acid (D). The $R^2$, RMSE, and RPD values are displayed, and they refers to Partial Least Squares for protein, Bayes B for oil, and Elastic Net for oleic acid.

bean literature. Similar to our approach, only Johnson et al. (2023) validated protein predictions across different field trials. However, their study was more limited in scope, relied on a smaller sample size of only 10 varieties, and included a narrowed protein variation range of 26.5–30.2%. The Bayes B model predicted oil content with an RMSE of 0.067, an $R^2$ of 0.83, and an RPD of 2.34, outperforming the only model available in literature ($R^2$ 0.68, RMSE 0.16, RPD 1.79) (Wang et al., 2014). Furthermore, Elastic Net was the best model for predicting oleic acid with an $R^2$ of 0.68. Ideally, NIRS-based models should have an $R^2 >$ 0.90 for excellent prediction (Saeys et al., 2005). However, it can be argued that at the selection stage in breeding, a lower $R^2$ is sufficient and that the definition of 'acceptable model' depends on the specific application of the model. When breeders are tasked with selecting materials from thousands of options, the correlation between actual and predicted values bears importance. A model with an $R^2$ of 0.68 has a good correlation index that can effectively aid this decision making process. An $R^2$ of 0.68 would correspond to a Pearson correlation coefficient ($r$) of about 0.84 between predicted and observed chemical values, bearing in mind that $r$ can be approximated by the square root of $R^2$. Overall, the NIRS models developped in this study are robust and valuable tools for faba bean breeding, validated across trials including a wide range of chemical and genetic diversity, and outperform previously reported models.

### *3.5. The best statistical approach for prediction was compound-specific: Partial least squares for protein, Bayes B for oil, elastic net for oleic acid content*

Summary statistics for the 10-fold cross-validation and external validation are shown in Table 1:. The PLS model included 13 latent

**Table 1**
Prediction accuracy for protein, oil and oleic content by using Partial Least Squares (PLS), Elastic Net (EN), Memory-based Learning (MBL), and Bayes B (BB). For the calibration (10-fold cross-validation), the Root Mean Square Error of cross-validation ($RMSE_{cv}$), and the Coefficient of Determination of cross-validation ($R^2_{cv}$) are reported. For the validation, Root Mean Square Error (RMSE), Coefficient of Determination ($R^2$), and Ratio Residual Deviation (RPD) are reported.

| Breeding trait | Modelling method | Cross-validation[d] | | | External validation[d] | | |
|---|---|---|---|---|---|---|---|
| | | Tuning parameter | $RMSE_{cv}$ | $R^2_{cv}$ | RMSE | $R^2$ | RPD |
| **Protein** | **PLS** | LV[a]=13 | 0.35 | 0.98 | 0.46 | 0.98 | 6.66 |
| | EN | $\lambda^b = 0.0056$ | 0.37 | 0.98 | 0.54 | 0.98 | 5.71 |
| | MBL | $K^c = 28$ | 1.17 | 0.93 | 0.63 | 0.96 | 4.86 |
| | BB | | 0.41 | 0.98 | 0.75 | 0.98 | 4.05 |
| **Oil** | PLS | LV = 13 | 0.056 | 0.89 | 0.1 | 0.82 | 1.53 |
| | EN | λ = 0.0025 | 0.05 | 0.92 | 0.09 | 0.86 | 1.66 |
| | MBL | K = 98 | 0.067 | 0.80 | 0.09 | 0.86 | 1.84 |
| | **BB** | | 0.056 | 0.88 | 0.067 | 0.83 | 2.34 |
| **Oleic acid** | PLS | LV = 13 | 1.89 | 0.65 | 4.3 | 0.56 | 0.68 |
| | **EN** | λ = 0.034 | 1.22 | 0.84 | 2.83 | 0.68 | 1.03 |
| | MBL | K = 83 | 1.80 | 0.41 | 2.83 | 0.47 | 1.03 |
| | BB | | 2.29 | 0.47 | 2.8 | 0.31 | 1.04 |

[a] LV indicates the number of latent variables (also known as 'factors' or 'components'). [b] λ is a hyperparameter that balances the contribution of the penalty terms. [c] K indicates the number of neighbours used. [d] Cross-validation refers to 10-fold cross-validation for PLS, EN, BB, and nearest neighbour validation (NNv) for MBL, while external validation refers to independent samples.

variables (LV) for all three compounds. The best λ for EN was 0.0056 for protein, 0.0025 for oil and 0.034 for oleic acid. The best number of k in MBL was 28, 98, and 83 for protein, oil, and oleic acid, respectively.

The optimal predictive approach was compound-specific (Fig. 3 B-D). For protein content, the standard PLS regression was the most accurate. Specifically, PLS yielded an RMSE value of 0.46, outperforming EN, MBL, and BB values of 0.54, 0.63, and 0.75, respectively. As RMSE is a performance measure expressed in the same unit as the compound being analysed, these values suggest that PLS can predict the protein content in unknown and independent samples with an accuracy margin of ± RMSE%. The RPD values for PLS, EN, MBL, and BB were 6.66, 5.71, 4.86, and 4.5 respectively, further supporting PLS as the model with the lowest mean prediction error. Variable shrinkage and selection carried out by EN and BB highlighted the most informative wavelengths for predicting protein content (Table 2). As expected, EN showed greater stringency in variable selection (51 variables selected) than BB, which tends to assign very small effects to certain variables while keeping them in the equations. The two models prioritised different wavelengths (Fig. 4). These differences reflect the distinctive approaches to variable selection, but could also be influenced by the compound's complex molecular architecture and by multicollinearity. Variable selection did not result in superior performance compared to PLS.

For oil content, the alternative models outperformed the standard PLS method. The RMSE values for BB, EN, and MBL were 0.067, 0.09, and 0.09, respectively, while the RMSE of PLS was 0.1. The RPD values confirmed the superior performance of BB. The RPD values for BB, MBL, EN, and PLS were 2.34, 1.84, 1.66, and 1.53, respectively. Both EN and BB detected a strong signal for oil content in the 1720 to 1734 nm wavelength range, most likely because the casual oil signal is retained in this range (Fig. 4). Specifically, BB detected a prominent peak at 1728 nm with a coefficient ten times greater than the second most prominent peak (1732 nm). Existing literature has pinpointed intervals such as 1600–1800 nm and 2100–2380 nm as relevant to oil and fatty acids (Manley, 2014; Sato et al., 1991). In particular, the 1620–1730 nm region is thought to be related to the first overtone of the C—H vibration found in chemical groups such as -CH3, -CH2, and =CH2 (Manley, 2014; Sato et al., 1991). The band position at 1732 nm is reported for the C—H bonds in aliphatic hydrocarbons (Ciurczak et al., 2021; Workman Jr & Weyer, 2012). Therefore, the vibrational frequency of the hydrocarbon chains in fatty acids, and thus in the total oil, may be the cause of the prominent peak detected.

For predicting oleic acid content, PLS was not the optimal approach. PLS produced an RMSE value of 4.3, whereas EN, MBL, and BB performed better with values of 2.83, 2.83, and 2.8, respectively. Accordingly, the RPD values for PLS, EN, MBL, and BB were 0.68, 1.03, 1.03, and 1.04 respectively. Although BB had a slightly lower RMSE, EN was selected as the superior method due to its considerably higher $R^2$. This indicates that, despite a comparable prediction error, BB did not ensure a strong linear relationship between observed and predicted values. As a general trend, the RMSEs were always higher in the external validation than in the cross-validation. An important significant discrepancy was observed for oleic acid. Using the EN method as an example, the RMSEcv was 1.22, while the RMSE in the validation increased to 2.83. This difference suggests that the model could be over-fitted to the calibration set. The reduced generalization may also be related to the smaller calibration set for oleic acid, highlighting the importance of achieving an appropriate sample size. EN removed noisy signals and improved the prediction by variable selection, retaining 89 variables. The most informative peaks were located between 1698 nm and 1788 nm. Considering that 1728 nm was the most prominent peak for oil in this study, finding very close band positions (e.g., 1716 nm, 1718 nm, 1720 nm, etc.) indicates that this region contains important chemical signals associated with lipids. The peaks at 1714 and 1699 nm are reported to correspond to the C—H overtone positions for aliphatic hydrocarbons (Ciurczak et al., 2021; Workman Jr & Weyer, 2012), which constitute the structural chains in fatty acids.

Our study is the first to explore such diverse statistical approaches in faba bean research, and to introduce the use of EN and BB in legumes to the existing literature. We introduced these metodologies as they are frequently employed in genomic prediction, a common tool in breeding to predict plant or animal performance based on molecular information (Daetwyler et al., 2013; Wang et al., 2019). The use of models that can be applied to different datasets (e.g., molecular and NIRS data) could provide an advantage in breeding pipelines by making more efficient use of various data sources. Furthermore, our findings support an initial research hypothesis that these advanced statistical models could further improve the predictive performance of faba bean seeds' quality. Specifically, BB improved oil content prediction accuracy by ~33%, while EN improved oleic acid prediction by ~34%. This suggests that for compounds with lower predictive ability, the elimination or minimization of non-contributing wavelengths by EN and BB aids to filter out noisy information. Ferragina et al. (2015) have already noted the strong performance of Bayesian models in predicting "difficult-to-predict" dairy traits, such as milk fatty acids. The BB and EN models identified the

**Table 2**
Variable importance for the prediction of protein, oil, and oleic acid using Bayes B and Elastic Net. The table shows the top 10 selected variables (Variable Importance) for each model and compound. Relevant literature information potentially linking specific wavelengths to the respective compounds is reported.

| Model | Compound | Variable Importance | Literature information[a] |
|---|---|---|---|
| Bayes B | Protein | 1330 nm<br>1406 nm<br>1426 nm<br>1430 nm<br>1432 nm<br>1440 nm<br>1880 nm<br>1918 nm<br>2048 nm<br>2226 nm | The region between 1450 and 1600 nm has been associated with protein content (Ciurczak et al., 2021; Workman Jr & Weyer, 2012). In particular, peaks around 1460 nm are linked to the N—H first overtone bond vibration. The N—H first overtone is associated with regions from 1510 to 1530 nm, and the aromatic C—H stretch of amino acids is typically found around 1570 nm ( Ciurczak et al., 2021; Workman Jr & Weyer, 2012). In addition, the 2050 to 2070 nm interval is associated with vibrations characteristic of protein structures, specifically the N—H second overtone and combinations. Notably, around 2048 nm and 2226 nm, signals associated with the N—H second overtone and C—H and C═O stretching combinations in proteins are expected. The traditional wavelength identified for protein characterization is 2185 nm (Ciurczak et al., 2021; Workman Jr & Weyer, 2012). |
| Elastic Net | Protein | 1542 nm<br>1544 nm<br>1546 nm<br>1672 nm<br>1674 nm<br>1676 nm<br>1742 nm<br>1744 nm<br>1746 nm<br>2294 nm | |
| Bayes B | Oil | 1356 nm<br>1684 nm<br>1700 nm<br>1726 nm<br>1728 nm<br>1732 nm<br>1894 nm<br>2222 nm<br>2304 nm<br>2434 nm | Intervals such as 1600–1800 nm and 2100–2380 nm have been identified as pertinent to oil and fatty acids (Manley, 2014; Sato et al., 1991). In particular, the 1620–1730 nm region is thought to be associated to the first overtone of the C—H vibration found in chemical groups such as -CH3, -CH2, and = CH2 (Manley, 2014; Sato et al., 1991). Furthermore, 2140 nm is identified as C-H/C=O from lipids.<br>In addition, vibrations of C—H and C—C bonds associated with oil are observed between 2300 nm and 2385 nm, with 2315 nm reported as a traditional wavelength for lipids/oils. The 2310–2390 nm range is also described as corresponding to C—H of lipids. Finally, the 2470–2480 nm range is reported as corresponding to C—H from lipids and aliphatic compounds (Ciurczak et al., 202; Workman Jr & Weyer, 2012). |
| Elastic Net | Oil | 1564 nm<br>1566 nm<br>1696 nm<br>1724 nm<br>1726 nm<br>1728 nm<br>1730 nm<br>1732 nm<br>2226 nm<br>2372 nm | |
| Bayes B | Oleic acid | 2274 nm<br>2276 nm<br>2280 nm<br>2312 nm<br>2314 nm<br>2316 nm<br>2318 nm<br>2320 nm<br>2428 nm<br>2488 nm | Oleic acid, a major constituent of oil, shares similar spectral characteristics to those reported above. Band positions around 1700–1788 nm correspond to important signals for oil, indicating the presence of signals associated with lipids (oil and fatty acids) in this region. Specifically, peaks at 1714 nm and 1699 nm are reported in the literature to correspond to C—H overtone positions for aliphatic hydrocarbons (Ciurczak et al., 2021; Workman Jr & Weyer, 2012). In addition, peaks around 2313 nm and 2315 nm are expected to correspond to C—H overtone bands of aliphatic hydrocarbons and lipids, respectively. Around 2323 nm, peaks are likely to represent C—H overtone from aliphatic hydrocarbons, with the region between 2385 nm and 2482 nm also indicative of these compounds (Ciurczak et al., 2021; Workman Jr & Weyer, 2012). |
| Elastic Net | Oleic acid | 1698 nm<br>1700 nm<br>1716 nm<br>1718 nm<br>1720 nm<br>1782 nm<br>1784 nm<br>1786 nm<br>1788 nm<br>2210 nm | |

[a] The table summarizes information from NIR regions of the spectra identified as highly predictive by either Bayes B or Elastic Net models, suggesting potential links to specific chemical bonds based on prior information in the literature. However, establishing causality from high-dimensional and collinear data such as NIRS using only the predictive properties of sparse models is challenging.

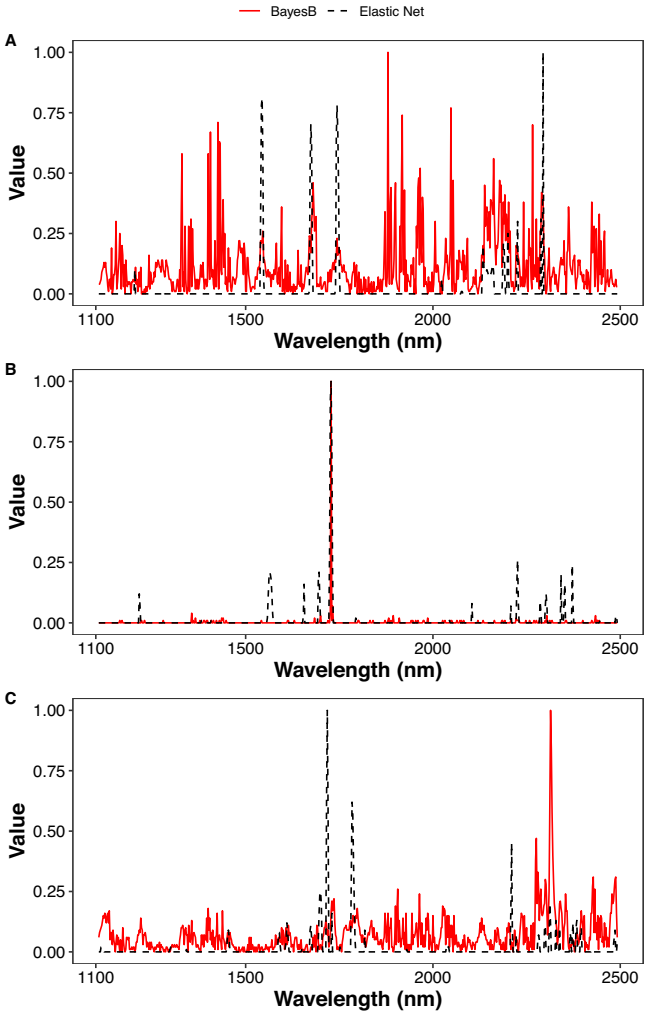



**Fig. 4.** Variable importance and selection for A) protein, B) oil, and C) oleic acid for Bayes B (red solid lines) and Elastic Net (black dashed lines). The plots display the predictor variables (wavelengths in nm) on the x-axis, and the magnitude of their importance in the prediction (y-axis). The values displayed are normalized [0, 1]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

wavelenght regions that were most predictive, which are directly or indirectly linked to the chemical compounds. Fig. 4 shows the absolute values of the effects of each wavelength by EN and BB. The value of each coefficient was expressed relative to the most prominent, normalizing the values within a range [0,1]. However, in reality, the models are characterised by positive and negative coefficients. As expected, EN was more stringent than BB in selecting variables, often reducing a larger number of wavelengths to zero effect for the three compounds. The extent of shrinkage and selection varies between predictors depending on their relationship with the response variable and with each other, but also varies according to the model assumption on the variance of each effect. High collinearity in NIR spectroscopic data, typical band overlap, and the presence of redundant noise can also contribute to differences in the informative wavelengths identified. We would like to emphasize that finding the most informative regions through variable selection criteria may not always be driven by genuine causal correlations. Instead, it could be the result of an empirical search for accuracy gains, typically measured by the RMSE. Despite several reported methods for extracting information from the NIR spectra, achieving unambigous results remains a challenge (Manley, 2014).

### *3.6. Reduced sample size decreased prediction accuracy depending on model type and compound*

To assess potential cost savings, different training set sizes were simulated by sampling 70%, 40%, and 20% of the full 125-sample training set. These reduced training sets retained the chemical variation of the original samples (Fig. 5 A-C; Supplementary Table 2). Summary predictive statistics of the models trained using these training sets are reported in Supplementary Tables 3, 4, and 5.

Overall, smaller sample sizes reduced prediction accuracy for the three compounds (Fig. 5 D—F), in line with the widely accepted notion in literature that larger sample sizes improve model accuracy and prevent overfitting (Schoot et al., 2020). However, the effect of reducing the sample size on model performance varied across models and compounds. Notably, protein content prediction showed greater resilience to reductions in sample size, with EN and MBL demonstrating the greatest robustness. Specifically, EN maintained an RMSE of 0.54 with only 70% of the samples. Similarly, MBL exhibited lower prediction errors with the 70% and 40% calibration sets. Although the highest accuracy for protein was achieved by PLS trained on the full training set, NIRS users may consider a small trade-off in accuracy for reduced analysis costs. In this scenario, EN and MBL trained with smaller sample sizes were cost-effective alternatives in our study. Yet, this is not a universal rule and these results should not be generalized without considering the specificities of each context.

For oil content, the RMSE consistently increased with reduced sample size, except for EN. EN retained an RMSE of 0.9, showing the same performance when trained with the full or 70% training set. Remarkably, BB also remained robust with reduced sample size, providing a better alternative to PLS, MBL, and EN on the full dataset when considering RMSE, $R^2$, and RPD simultaneously. For oleic acid, the accuracy of all models declined with decreasing sample size. This underlines the importance of a larger number of samples for compounds with lower predictive ability. Visual inspection of the trend shown in Fig. 5 D-F suggests that a larger sample size would increase the prediction accuracy for this compound.

A caveat to our research is the use of a fixed approach to data preprocessing. While previous research has demonstrated that modifying the data pre-processing strategy can improve accuracy with reduced sample sizes (Schoot et al., 2020), our study aimed to provide a straightforward pipeline for NIRS users. Constantly adjusting the data pre-processing with each update of the calibration set would be a labour-intensive task. Hence, our methodology seeks to strike a balance between efficiency and effectiveness. In addition, we advocate further research in the area of sample size reduction, and recommend that studies start with a larger sample size to determine the optimal minimum required.

## 4. Conclusion

In this study, we established a reproducible, license-free NIRS pipeline to predict protein, oil, and oleic acid content in faba bean using NIRS data. Our investigation revealed that the K-means clustering algorithm is an efficient algorithm for designing a representative training set. Interestingly, the predictive ability of different statistical models varied depending on the target compounds. It was found that for compounds with lower predictive ability (e.g., oleic acid and oil), exploring different modelling approaches can led to a higher gain in prediction accuracy compared to others that can be easily predicted using the standard PLS method (e.g., protein content). Typically, a standard PLS model developed for a specific year or condition is updated with new chemical samples to compensate any loss in prediction accuracy. We demonstrated that prediction accuracy can be maintained across breeding trials by exploring alternative statistical approaches without the need for new chemical analyses (and the associated costs), at least to some extent and for some compounds. Notably, this research introduced

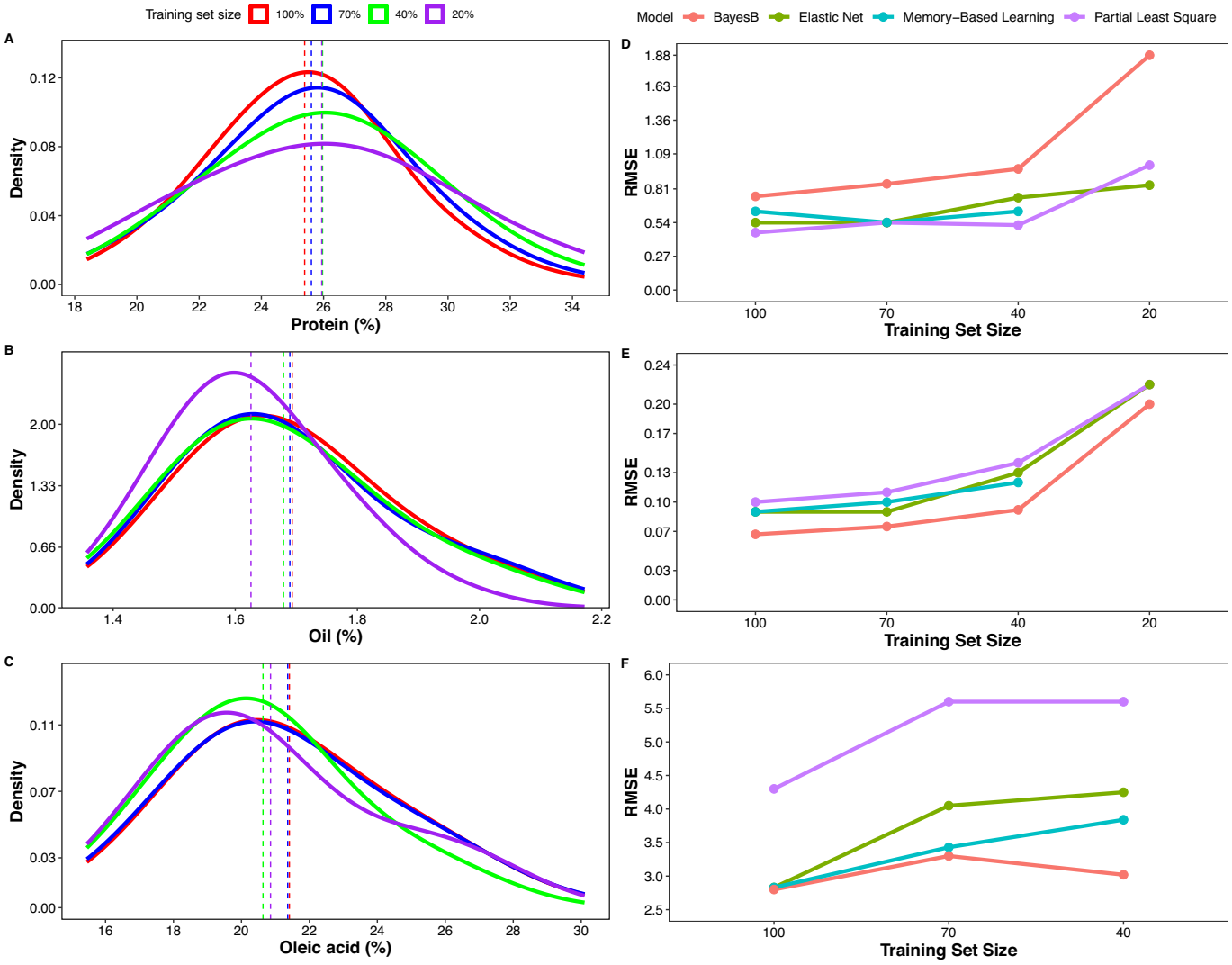



**Fig. 5. A-C.** Variability for protein (A), oil (B), and oleic acid (C) for four training sets of different size (100%, 70%, 40%, 20%) represented by different colours (legend). The dashed lines indicate the respective mean values. **D—F.** Changes in Root Mean Squared Error (y-axis) for BB, EN, MBL and PLS at different training set sizes (x-axis). D represents protein, E oil, F oleic acid. For oleic acid, the smaller training set (20%) is not shown, as the investigation stopped at 40%. For MBL, the investigation stopped at 40% for all the three compounds.

the application of both BB and EN to the NIRS analysis with a focus on legume seeds quality. These models excelled and outperformed PLS for predicting oil and oleic acid respectively, underscoring that variable selection improves predictions. In conclusion, the models we have developed are invaluable tools for the faba bean research community as well as for breeding or applications in the food industry.

## Disclosure statement

During the preparation of this work the author(s) used ChatGPT and Grammarly in order to edit/improve the flow of the text and check grammar errors. After using this tool/service, the author(s) reviewed and edited the content as needed and take (s) full responsibility for the content of the publication.

## Funding

This research is part of project 'Pulses optimized for flavour and functionality' co-financed by the Top Consortium for Knowledge and Innovation Agri & Food by the Dutch Ministry of Economic Affairs under contract number LWV19028.

## CRediT authorship contribution statement

**Antonio Lippolis:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Pamela Vega Polo:** Formal analysis. **Guilherme de Sousa:** Formal analysis. **Annemarie Dechesne:** Methodology, Formal analysis. **Laurice Pouvreau:** Project administration, Funding acquisition. **Luisa M. Trindade:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101583.

## References


Adhikari, K. N., Khazaei, H., Ghaouti, L., Maalouf, F., Vandenberg, A., Link, W., & O'Sullivan, D. M. (2021). Conventional and molecular breeding tools for accelerating genetic gain in Faba bean (Vicia Faba L.). Front. *Plant Science, 12*, Article 744259. https://doi.org/10.3389/fpls.2021.744259

Anderson, N., Walsh, K., Flynn, J., & Walsh, J. P. (2021). Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models. *Postharvest Biology and Technology, 171*, Article 111358. https://doi.org/10.1016/j.postharvbio.2020.111358

Cem, Ö., & Kahriman, E. (2012). Determination of quality parameters in maize grain by NIR reflectance spectroscopy. *Tarım Bilimleri Dergisi, 18*(1), 31–42. https://doi.org/10.1501/Tarimbil_0000001190

Ciurczak, E. W., Igne, B., Workman Jr, J., & Burns, D. A. (2021). Handbook of near-infrared analysis. CRC press. Boca Raton, Florida.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., & Hickey, J. M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics, 193*(2), 347–365. https://doi.org/10.1534/genetics.112.147983

El-Sherbeeny, M. H., & Robertson, L. D. (1992). Protein content variation in a pure line faba bean (Vicia faba) collection. Journal of the Science of Food and Agriculture, 58 (2), 193–196. https://dx.doi.org/https://doi.org/10.1002/jsfa.2740580206.

Ferragina, A., de Los Campos, G., Vazquez, A., Cecchinato, A., & Bittante, G. (2015). Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. *Journal of Dairy Science, 98*(11), 8133–8151. https://doi.org/10.3168/jds.2014-9143

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1. https://doi.org/10.18637/jss.v033.i01

Gonçalves, M. T. V., Morota, G., Costa, P. M., & d. A., Vidigal, P. M. P., Barbosa, M. H. P., & Peternelli, L. A.. (2021). Near-infrared spectroscopy outperforms genomics for predicting sugarcane feedstock quality traits. *PLoS One, 16*(3), Article e0236853. https://doi.org/10.1371/journal.pone.0236853

Hacisalihoglu, G., Freeman, J., Armstrong, P. R., Seabourn, B. W., Porter, L. D., Settles, A. M., & Gustin, J. L. (2020). Protein, weight, and oil prediction by single-seed near-infrared spectroscopy for selection of seed quality and yield traits in pea (Pisum sativum). *Journal of the Science of Food and Agriculture, 100*(8), 3488–3497. https://doi.org/10.1002/jsfa.10389

Johnson, J. B., Walsh, K., & Naiker, M. (2020). Application of infrared spectroscopy for the prediction of nutritional content and quality assessment of faba bean (Vicia faba L.). Legume. *Science, 2*(3), Article e40. https://doi.org/10.1002/leg3.40

Johnson, J. B., Walsh, K. B., & Naiker, M. (2023). Assessment of bioactive compounds in faba bean using infrared spectroscopy. Legume. *Science, e203*. https://doi.org/10.1002/leg3.203

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

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Leite, D. C., Corrêa, A. A. P., Júnior, L. C. C., de Lima, K. M. G., de Morais, C., & d. L. M., Vianna, V. F., de Almeida Teixeira, G. H., Di Mauro, A. O., & Unêda-Trevisoli, S. H.. (2020). Non-destructive genotypes classification and oil content prediction using near-infrared spectroscopy and chemometric tools in soybean breeding program. *Journal of Food Composition and Analysis, 91*, Article 103536. https://doi.org/10.1016/j.jfca.2020.103536

Lippolis, A., Roland, W. S. U., Bocova, O., Pouvreau, L., & Trindade, L. M. (2023). The challenge of breeding for reduced off-flavor in faba bean ingredients. *Frontiers in Plant Science, 14*, 1286803. https://doi.org/10.3389/fpls.2023.1286803

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma, 288*, 175–183. https://hdl.handle.net/20.500.11770/334443.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. *Chemical Society Reviews, 43*(24), 8200–8214. https://doi.org/10.1039/C4CS00062E

Mishra, P., Passos, D., Marini, F., Xu, J., Amigo, J. M., Gowen, A. A., … Rutledge, D. N. (2022). Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry, 116804*. https://doi.org/10.1016/j.trac.2022.116804

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification* (Vol. 6). Chichester, UK: NIR Publications.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology, 46*(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Ozaki, Y., & Morisawa, Y. (2021). Principles and characteristics of NIR spectroscopy. In *Near-infrared spectroscopy: Theory, spectral analysis, instrumentation, and applications, 11–35*. Singapore: Springer. https://doi.org/10.1007/978-981-15-8648-4_2

Pérez, P., & de Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics, 198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Puspitasari, W., Aleman, B., Angra, D., Appleyard, H., Ecke, W., Möllers, C., Nolte, T., Purves, R. W., Renner, C., & Robertson-Shersby-Harvie, T. (2022). NIRS for vicine and convicine content of faba bean seed allowed GWAS to prepare for marker-assisted adjustment of seed quality of German winter faba beans. Journal of Cultivated Plants, 74(01−02). http://dx.doi.org/https://doi.org/10.5073/JfK.2022.01-02.01.

Rabinowicz, A., & Rosset, S. (2022). Cross-validation for correlated data. *Journal of the American Statistical Association, 117*(538), 718–731. https://doi.org/10.1080/01621459.2020.1801451

Ramirez-Lopez, L., Schmidt, K., Behrens, T., Van Wesemael, B., Demattê, J. A., & Scholten, T. (2014). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma, 226*, 140–150. https://doi.org/10.1016/j.geoderma.2014.02.002

Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsez, C., Wadoux, A., & Breure, T. (2016). Resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. *R package version, 1*(2).

Rinnan, Å., Van Den Berg, F., & 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

Saeys, W., Mouazen, A. M., & Ramon, H. (2005). Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosystems Engineering, 91*(4), 393–402. https://doi.org/10.1016/j.biosystemseng.2005.05.001

Sato, T., Kawano, S., & Iwamoto, M. (1991). Near infrared spectral patterns of fatty acid analysis from fats and oils. *Journal of the American Oil Chemists Society, 68*, 827–833. https://doi.org/10.1007/BF02660596

Schoot, M., Kapper, C., van Kollenburg, G. H., Postma, G. J., van Kessel, G., Buydens, L. M. C., & Jansen, J. J. (2020). Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample size. *Chemometrics and Intelligent Laboratory Systems, 204*, Article 104105. https://doi.org/10.1016/j.chemolab.2020.104105

Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package. *R package version, 0*(2), 6.

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., Boeckx, P., Kearsley, E., & Van Oost, K. (2021). *The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis.*. https://doi.org/10.5194/soil-7-693-2021

Tacke, R., Ecke, W., Höfer, M., Sass, O., & Link, W. (2022). Fine-mapping of the major locus for vicine and convicine in faba bean (Vicia faba) and marker-assisted breeding of a novel, low vicine and convicine winter faba bean population. *Plant Breeding, 141*(5), 644–657. https://doi.org/10.1111/pbr.13039

Tsegay, G., Ammare, Y., & Mesfin, S. (2023). Development of non-destructive NIRS models to predict oil and major fatty acid contents of Ethiopian sesame. *Journal of Food Composition and Analysis, 115*, Article 104908. https://doi.org/10.1016/j.jfca.2022.104908

Wang, J., Liu, H., & Ren, G. (2014). Near-infrared spectroscopy (NIRS) evaluation and regional analysis of Chinese faba bean (Vicia faba L.). *The Crop Journal, 2*(1), 28–37. https://doi.org/10.1016/j.cj.2013.10.001

Wang, X., Miao, J., Chang, T., Xia, J., An, B., Li, Y., Xu, L., Zhang, L., Gao, X., & Li, J. (2019). Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. *PLoS One, 14*(2), Article e0210442. https://doi.org/10.1371/journal.pone.0210442

Welch, R. W., & Wynne Griffiths, D. (1984). Variation in the oil content and fatty acid composition of field beans (Vicia faba) and peas (Pisum spp.). *Journal of the Science of Food and Agriculture, 35*(12), 1282–1289. https://doi.org/10.1002/jsfa.2740351203

Workman, J., Jr., & Weyer, L. (2012). *Practical guide and spectral atlas for interpretive near-infrared spectroscopy*. Boca Raton, Florida: CRC press.