# Journal Pre-proof


ADVANCES IN THE INDIVIDUAL AUTHENTICATION OF COCOA BEANS: vis/NIR SPECTROSCOPY AS A TOOL TO DISTINGUISH FERMENTED FROM UNFERMENTED BEANS AND CLASSIFY GENOTYPES IN THE EASTERN AMAZONIA

Anne Pinto, Antoine Deryck, Giulia Victória Lima, Ana Caroline de Oliveira, Fabio Gomes Moura, Douglas Fernandes Barbin, Juan Antonio Fernández Pierna, Vincent Baeten, Hervé Rogez

PII: S0956-7135(24)00276-7

DOI: https://doi.org/10.1016/j.foodcont.2024.110559

Reference: JFCO 110559

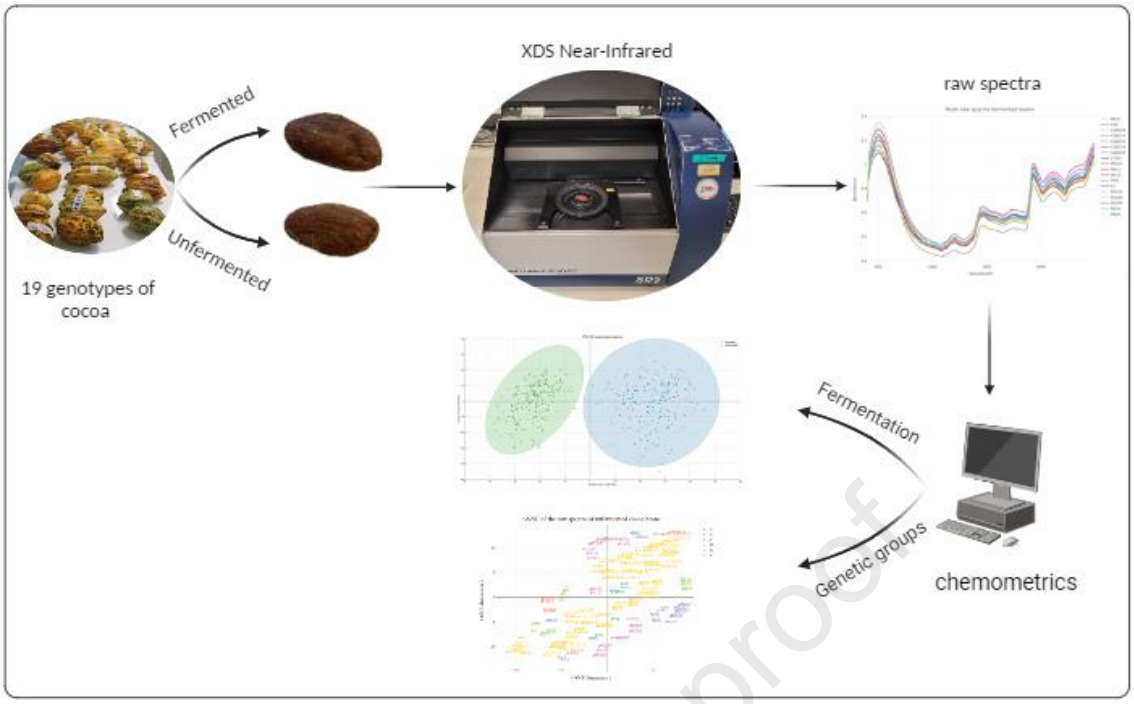
To appear in: *Food Control*

Received Date: 31 January 2024

Revised Date: 12 April 2024

Accepted Date: 6 May 2024

Please cite this article as: Pinto A., Deryck A., Lima G.V., de Oliveira A.C., Moura F.G., Barbin D.F., Fernández Pierna J.A., Baeten V. & Rogez H., ADVANCES IN THE INDIVIDUAL AUTHENTICATION OF COCOA BEANS: vis/NIR SPECTROSCOPY AS A TOOL TO DISTINGUISH FERMENTED FROM UNFERMENTED BEANS AND CLASSIFY GENOTYPES IN THE EASTERN AMAZONIA, *Food Control*, https://doi.org/10.1016/j.foodcont.2024.110559.



# ADVANCES IN THE INDIVIDUAL AUTHENTICATION OF COCOA BEANS: vis/NIR SPECTROSCOPY AS A TOOL TO DISTINGUISH FERMENTED FROM UNFERMENTED BEANS AND CLASSIFY GENOTYPES IN THE EASTERN AMAZONIA

Anne Pinto[a]; Antoine Deryck[b]; Giulia Victória Lima[a]; Ana Caroline de Oliveira[b], Fabio Gomes Moura[a]; Douglas Fernandes Barbin[c], Juan Antonio Fernández Pierna[b]; Vincent Baeten[b]; Hervé Rogez[a*]

[a]Center for Valorization of Amazonian Bioactive Compounds (CVACBA), Federal University of Pará (UFPA), Belém, PA, Brazil.
[b]Walloon Agricultural Research Center (CRA-W), Knowledge and Valorization of Agricultural Products Department, Quality and Authentication of Products Unit, Gembloux, Belgium.
[c]Department of Food Engineering, University of Campinas (UNICAMP), Campinas, SP, Brazil.
*Correspondent Author: herverogez@gmail.com

**Abstract**

Reliable analytical methods to authenticate high-quality and economically valuable cocoa beans are highly desirable, and NIR spectroscopy stands out as a rapid and non-destructive alternative. This study employs NIR for the authentication and differentiation of 19 whole Forastero cocoa beans from Eastern Amazon (Pará, Brazil), based on their fermentation status and genetic profiles. Partial Least Squares Discriminant Analysis (PLS-DA) models in wavelength ranges of 400-700 nm, 1400-1600 nm, and 1900-2500 nm, demonstrated high sensitivity and specificity in discriminating fermented from unfermented beans, regardless of genotype. Various compounds, including proteins, lipids, carbohydrates, anthocyanins, and theobromine, provide crucial insights into the spectral regions essential for distinction. The variable importance in projection (VIP) score value greater than 1 was used to select relevant variables, and Linear Discriminant Analysis (LDA) was performed in both the visible range (472 nm and 636 nm) and the infrared range (2096 nm and 2278 nm), demonstrating that absorbances at two specific wavelengths are sufficient for discrimination. The t-distributed stochastic neighbor embedding (t-SNE) indicated a segregation trend of the genotypes based on classification

by major genetic groups in unfermented beans, suggesting that the biochemical characteristics shared by them are more prominent before fermentation. The PLS-DA models based on complete vis/NIR spectra showed comparable results in discriminating the 19 cocoa genotypes in both fermented (0.14% prediction error) and unfermented beans (0.16% prediction error). The model's classification errors can be attributed to shared genetic ancestry among the samples, primarily in unfermented beans. This research corroborates the effectiveness of vis/NIR spectroscopy as a straightforward tool for whole cocoa bean authentication, providing rapid insights into genetic diversity regardless of their fermentation state.

Keywords: *Theobroma cacao; Near-Infrared spectroscopy; chemometrics; fermentation; cocoa genotypes, authentication.*

## 1. Introduction

Cocoa (Theobroma cacao L.) and its products, such as chocolate, are widely consumed globally and are valued for their flavor and health benefits. Presently, Brazil contributes with 273,873 tons of cocoa beans annually, ranking sixth in global production. Over half of the Brazilian production comes from the Amazon region, mainly from state of Pará, which is known for its competitive advantage over other regions, with high productivity (955 kg/ha in dried cocoa beans), low production cost (US$ 750.00 per ton produced) (Mapa, 2023), and potential for producing fine cocoa (Collin et al., 2023).

The cocoa from the Brazilian Amazon is traditionally classified as Forastero. In 2008, a new subclassification for this variety was proposed by Motamayor et al. (2008), including Marañon (PA), Curaray (AGU), Iquitos (IMC), Nanay (NA), Contamana (SCA), Amelonado (BE), Purús (CAB), Nacional (MO), and Guiana (CJ), but the authors reported difficulties in accessing Brazilian germplasm. Beyond its inherent genetic diversity, Pará's success in cocoa production is attributed to the planting of 20 cocoa genotypes with high yield and disease resistance, developed and selected in the 1970s by the Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC). Today, approximately 15 million seeds of these genotypes are distributed annually to local producers (MAPA, 2023).

The quality characteristics of cocoa beans are associated with both the cocoa genotype and post-harvest processing stages, particularly fermentation and drying (Santander Muñoz et al., 2020).

Genetic diversity influences the composition in the beans, such as proteins, lipids, carbohydrates, and phenolic compounds, affecting the microbial profile of the pulp and the biochemical changes that occur in the beans during fermentation, responsible for the development of color and formation of the flavor of commercial cocoa beans (Santander Muñoz et al., 2020). After fermentation, the beans are dried and supplied to traders. A common adulteration practice involves mixing fermented beans with unfermented beans due to high demand (Aikpokpodion & Dongo, 2010).

The co-plantation of genotypes and the blending of beans with different post-harvest processing conditions complicate the identification of high-economic-value genotypes and the assurance of the quality of derived products, like chocolate.

In light of these challenges, near-infrared spectroscopy (NIR) has been employed as a rapid method to predict biochemical quality parameters, offering qualitative and quantitative methods for the characterization, classification, and authentication of cocoa and chocolate samples, as reviewed by Teye et al. (2020).

The application of NIR for differentiation between fermented and unfermented beans has been the subject of intensive studies (Sunoj, Igathinathane & Visvanathan, 2016; Hashimoto et al, 2018; Hernandez et al., 2022), with recent research highlighting the potential of NIR for the classification and differentiation of intact cocoa bean genotypes (Barbin et al., 2018, Cruz-Tirado et al., 2020), offering benefits such as process speed and waste reduction. However, both studies were limited to analyzing only five genotypes.

This study aims to evaluate the efficacy of vis/NIR spectroscopy in distinguishing Forastero cocoa genotypes from the Brazilian Amazon and reducing the complexity and cost of analysis in differentiating between fermented and unfermented beans. A comprehensive sample set was used, along with physicochemical and genetic data, to enrich the interpretation of NIR spectra. This approach provided a more robust understanding of the unique characteristics of cocoa beans and contributed to their authentication.

## 2. Materials and Methods

### *2.1. Sample collection*

Nineteen Forastero cocoa genotypes from eastern Amazonia were selected based on their importance to the cocoa industry and are presented in (Table 1). Around 70 fruits of each genotype were kindly collected in July 2020 by the Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC) in Medicilândia and Tucumã, Pará, Brazil. Since the genotype determines the basic chemical composition of the beans and fermentation induces additional chemical changes, the beans were removed from the fruits and approximately 1 kg of each genotype was fermented within the same fermentation box (with genotypes isolated in nylon bags) for 6 days under the same temperature and relative humidity conditions. Three genotypes were randomly chosen to be fermented in duplicate (P7, CCN51, and CAB270). Unfermented and fermented beans were sun-dried for 5 days until their moisture content reached <8 %, and stored under refrigeration until analyses. No additional processing, such as grinding or peeling, was performed.

*2.2 Genetic classification of the cocoa genotypes*

The genetic diversity of the 19 cocoa genotypes in the Eastern Amazonia was analyzed by De Oliveira et al. (Unpublished results) based on DNA polymorphisms using 15 standard cocoa microsatellite markers. Genetic data were analyzed using the STRUCTURE v2.3.4 software (Pritchard et al., 2000), which employs a Bayesian approach to model the probability of a sample belonging to one of K groups (K=10, representing the groups proposed by Motamayor et al. (2008): Marañón, Curaray, Criollo, Iquitos, Nanay, Contamana, Amelonado, Purús, Nacional, and Guiana). For each genotype, STRUCTURE provided a set of membership coefficients (Q values) representing the estimated proportion of its ancestry.

The coefficient of membership (Q) for an individual in each specific group indicates the proportion of their ancestry attributed to that group, ranging from 0 to 1, where 0 denotes no ancestry in the group, and 1 indicates complete contribution. The two highest membership coefficients (Q1 and Q2) represent the predominant genetic composition and are presented in Table 1.

**Table 1**. Cocoa genotypes from Eastern Amazonia, origin, and classification of the majority genetic groups based on the coefficient of membership (Q).

| Genotype | Origin | Q1 | Genetic group | Q2 | Genetic group |
|---|---|---|---|---|---|
| CA6 | Medicilândia | 0.3239 | Iquitos | 0.2041 | Nanay |
| PA169 | Tucumã | 0.4665 | Marañón | 0.2535 | Amenolado |
| PA121 | Medicilândia | 0.9240 | Marañón | 0.924 | Marañón |

| | | | | | |
|---|---|---|---|---|---|
| PA195 | Tucumã | 0.6744 | Marañón | 0.6744 | Marañón |
| BE10 | Medicilândia | 0.3022 | Nanay | 0.2576 | Manañón |
| CAB499 | Tucumã | 0.5731 | Purús | 0.5731 | Purús |
| CCN51 | Medicilândia | 0.4608 | Criollo | 0.2930 | Iquitos |
| IMC67 | Medicilândia | 0.6602 | Iquitos | 0.6602 | Iquitos |
| CAB324 | Tucumã | 0.462 | Purús | 0.4598 | Nanay |
| CAB214 | Medicilândia | 0.5386 | Purús | 0.4155 | Contamana |
| MA11 | Tucumã | 0.4192 | Purús | 0.2835 | Amelonado |
| P7 | Medicilândia | 0.5295 | Nanay | 0.4257 | Contamana |
| RB36 | Tucumã | 0.9511 | Purús | 0.9511 | Purús |
| RB40 | Medicilândia | 0.8646 | Purús | 0.8646 | Purús |
| CAB270 | Medicilândia | 0.3238 | Purús | 0.2792 | Guiana |
| MO1 | Medicilândia | 0.3729 | Amelonado | 0.265 | Purús |
| CAB208 | Medicilândia | 0.7415 | Purús | 0.7415 | Purús |
| MA15 | Medicilândia | 0.8389 | Purús | 0.8389 | Purús |
| CAB 314 | Tucumã | 0.4860 | Purús | 0.3489 | Nanay |

*2.3 Physicochemical analysis*

The physicochemical characteristics of cocoa beans (Table 2) were obtained using standard analytical methods according to the AOAC (2023): moisture (931.04), lipid (963.15), total soluble solids (932.12), and protein content (970.22). The pH was measured according to the protocol of Senanayake et al. (1997). The fermentation index (FI) was determined using the spectrophotometric method described by Gourieva & Tserrevitinov (1979), based on the degradation of anthocyanins during fermentation and calculated by the ratio of the absorbance at 460 nm and 530 nm. The cut test correlates visual characteristics with chemical composition: unfermented beans have a predominant violet color while the brown color is characteristic of properly fermented cocoa. A longitudinal section was carried out on 30 randomly selected cocoa beans of each genotype to evaluate the degree of fermentation and the results were expressed as a percentage of violet, partially brown, and brown beans (ISO 2451, 2017). The color was measured directly on the inner surface (cotyledons) of the beans after the cut test using a Minolta colorimeter and the yellowness parameter (b*) was evaluated (Barbin et al., 2018).

137 **Table 2.** Values obtained for the physicochemical analyses of fermented and unfermented cocoa beans.

| Genotypes | Total soluble solids (°Brix) | | | | pH | | | | Lipids (g/100g DW) | | Proteins (g/100g DW) | | Fermented index | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | External | | Internal | | External | | Internal | | | | | | | |
| | Fermented | unfermented | Fermente | unferment | Fermented | unfermented | Fermented | unferment | Fermented | unfermented | Fermented | unfermented | Fermented | unferment |
| CAB499 | 2.96±0.19 | 12.57±0.35 | 110±0.00 | 10.00±0.00 | 5.98±0.00 | 4.49±0.00 | 5.11±0.21 | 6.71±0.01 | 32.85±1.61 | 32.10±1.95 | 18.59±0.45 | 17.87±0.30 | 0.97±0.00 | 0.64±0.02 |
| MA 15 | 3.08±0.00 | 12.80±1.20 | 13±0.00 | 11.00±0.00 | 5.81±0.01 | 4.81±0.00 | 4.99±0.01 | 6.68±0.00 | 30.27±0.06 | 30.93±0.37 | 18.97±0.60 | 17.93±0.00 | 1.14±0.00 | 0.64±0.03 |
| IMC 67 | 3.37±0.02 | 12.54±0.71 | 14±0.00 | 10.50±0.71 | 6.15±0.01 | 4.47±0.01 | 4.91±0.01 | 6.71±0.00 | 31.30±1.36 | 29.46±2.47 | 17.05±2.11 | 16.30±0.14 | 1.13±0.02 | 0.64±0.02 |
| MA 11 | 4.96±0.06 | 7.53±0.81 | 11.5±0.07 | 11.50±0.71 | 5.75±0.07 | 4.74±0.01 | 4.82±0.01 | 6.65±0.01 | 35.40±0.44 | 31.59±0.68 | 19.15±0.01 | 18.12±0.30 | 1.11±0.19 | 0.41±0.00 |
| RB 36 | 4.35±0.23 | 16.63±0.54 | 12±0.00 | 7.50±0.71 | 5.91±0.01 | 5.42±0.00 | 4.87±0.00 | 6.77±0.01 | 30.05±1.39 | 32.93±1.31 | 18.44±0.15 | 17.72±0.60 | 1.06±0.01 | 0.58±0.02 |
| RB40 | 5.09±0.07 | 8.41±0.01 | 11±0.00 | 10.50±0.71 | 6.04±0.01 | 5.72±0.00 | 5.32±0.04 | 6.73±0.00 | 30.37±0.93 | 32.61±1.84 | 20.29±0.74 | 19.16±0.60 | 1.35±0.03 | 0.75±0.02 |
| BE10 | 3.72±0.18 | 9.01±1.04 | 13.5±0.07 | 12.50±0.71 | 5.79±0.00 | 4.56±0.01 | 4.85±0.00 | 6.59±0.01 | 31.67±2.09 | 30.49±0.62 | 16.71±0.75 | 16.62±0.00 | 1.04±0.02 | 0.66±0.05 |
| PA 169 | 3.98±0.39 | 10.40±0.00 | 13±0.14 | 11.50±0.71 | 5.83±0.00 | 5.33±0.01 | 4.92±0.01 | 6.67±0.01 | 36.52±0.90 | 38.48±2.00 | 17.89±0.31 | 18.32±0.01 | 1.11±0.04 | 0.69±0.06 |
| PA121 | 3.66±0.10 | 9.3±0.58 | 14±0.00 | 13.50±0.71 | 5.76±0.01 | 4.87±0.00 | 4.94±0.01 | 6.53±0.00 | 34.78±0.31 | 31.55±1.29 | 18.09±0.30 | 16.81±0.30 | 1.13±0.03 | 0.77±0.01 |
| MO1 | 3.78±0.17 | 8.16±1.30 | 12±0.14 | 11.50±0.71 | 5.91±0.00 | 4.59±0.01 | 5.01±0.01 | 6.38±0.02 | 28.34±0.90 | 32.93±0.34 | 16.74±0.46 | 16.52±0.45 | 1.13±0.10 | 0.76±0.02 |
| CA6 | 3.20±0.05 | 8.43±0.37 | 12±0.00 | 9.00±0.00 | 6.08±0.00 | 4.90±0.01 | 5.21±0.01 | 6.65±0.01 | 29.74±2.87 | 29.76±1.18 | 18.57±0.15 | 18.59±0.15 | 1.53±0.02 | 0.57±0.01 |
| CAB324 | 3.01±0.07 | 8.06±1.14 | 12±0.00 | 12.50±0.71 | 6.02±0.00 | 4.47±0.02 | 5.15±0.01 | 6.51±0.02 | 30.50±2.14 | 30.62±0.63 | 16.40±0.30 | 17.26±0.60 | 1.34±0.05 | 0.80±0.08 |
| CAB208 | 4.45±0.26 | 11.07±0.23 | 12.5±0.07 | 10.50±0.71 | 6.33±0.04 | 5.30±0.00 | 5.17±0.04 | 6.67±0.02 | 26.31±0.51 | 27.48±2.54 | 17.70±0.90 | 16.32±0.15 | 1.54±0.04 | 0.66±0.01 |
| CAB314 | (nd) | 11.05±0.70 | (nd) | 10.87±0.50 | (nd) | 4.27±0.00 | (nd) | 6.07±0.00 | (nd) | 31.89±1,25 | (nd) | 17.98±0.45 | (nd) | 0.66±0.00 |
| PA 195 | 5.27±0.23 | 12.31±1.37 | 11±0.00 | 12.00±0.00 | 5.65±0.21 | 5.01±0.01 | 5.08±0.03 | 6.63±0.01 | 30.64±1.30 | 33.76±1.97 | 18.43±0.45 | 19.33±0.30 | 0.96±0.03 | 0.96±0.02 |
| CAB214 | 4.39±0.16 | 5.05±0.01 | 9.5±0.07 | 9.50±0.71 | 5.93±0.00 | 5.80±0.01 | 5.12±0.00 | 6.71±0.02 | 29.51±0.67 | 30.32±0.42 | 18.94±0.31 | 18.65±0.01 | 0.97±0.02 | 0.97±0.00 |
| P7 | 2.81±0.06 | 10.25±0.87 | 10±0.05 | 9.50±0.71 | 5.79±0.06 | 4.41±0.01 | 5.08±0.05 | 6.66±0.01 | 33.86±2.68 | 33.93±2.42 | 18.92±0.96 | 17.75±0.16 | 1.01±0.04 | 0.52±0.00 |
| CAB270 | 3.00±0.26 | 10.43±0.32 | 15±0.00 | 12.00±0.00 | 5.99±0.15 | 4.67±0.01 | 4.99±0.19 | 6.66±0.01 | 30.23±1.76 | 30.09±1.56 | 19.82±1.38 | 15.75±0.90 | 1.08±0.01 | 0.72±0.02 |
| CCN51 | 3.61±0.24 | 10.40±0.01 | 12.5±0.10 | 11.00±0.00 | 5.84±0.04 | 4.36±0.01 | 5.08±0.10 | 6.41±0.02 | 30.74±1.39 | 31.77±1.03 | 17.96±1.57 | 15.76±0.01 | 1.16±0.11 | 0.68±0.02 |
| Mean | 3.72 | 10.5 | 12.28 | 10.95 | 5.91 | 4.90 | 5.03 | 6.62 | 31.34 | 31.66 | 18.35 | 17.49 | 1.14 | 0.66 |
| Range | 2.77-5.27 | 7.53-16.63 | 9.5-15 | 9.5—13.5 | 5.75-6.33 | 4.36-5.72 | 4.82-5.32 | 6.37-6.77 | 26.31-36.64 | 27.48-38.48 | 16.4-20.85 | 15.75-19.33 | 0.95-1.54 | 0.41-0.79 |

138 DW: Dry Weight; (nd)- not determined: genotype was not evaluated for fermented beans

### *2.4 Spectral acquisition*

Spectral data from unfermented and fermented cocoa beans were obtained in reflectance mode and recorded as absorbance (log 1/R) using a XDS Near-Infrared-Rapid Content Analyzer (Foss NIRSystems, Denmark). The wavelength range spanned from 400 to 2500 nm, with a resolution of 2 nm. Both the visible and near-infrared (vis/NIR) ranges were included. For each of the 19 genotypes, spectra were obtained from 10 whole beans randomly selected from a set of approximately 1 kg, except for the duplicate fermented samples P7, CCN51, and CAB270, represented by 20 beans. Cocoa beans were scanned using a ring sample cup, and for each cocoa bean, the spectra acquisition was carried out on both sides in stationary mode with 32 scans taken at a single spot. The spectra were preprocessed by auto-linearization and the average spectrum was kept. All the beans were measured randomly (it means that the 10 or 20 beans of each genotype were not measured consecutively). The mean spectra per genotype were calculated and are presented in Fig.S.1(Supplementary material)

### *2.5 Data processing*

The data analysis were performed on R version 4.2.2 (RStudio Team, 2020) with the caret (Kuhn, 2022), rchemo (Lesnoff, 2022), and mdatools (Kucheryavskiy, 2020) packages. The Mahalanobis distance and the Z-score method were used to check possible outliers, but no samples were discarded (Pierna et al., 2002; Aggarwal et al. 2019). In this research, 4 different pre-processing combinations were tested: Standard Normal Variable (SNV), Savitzky-Golay (SG), SNV followed by SG and SG followed by SNV. The PCA of the raw and pre-processed data were observed, and the pre-processing that presented the best separation between the two groups was selected The preprocessing applied was Savitzky-Golay (SG) with a window size of 21 points (width = 21), a derivative of the first order (dorder = 1), and a polynomial degree of the second order (porder = 2).

#### *2.5.1 Exploratory data analysis*

For exploratory analysis of NIR spectra, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were applied to evaluate possible separation patterns of cocoa beans (Sentellas & Saurina, 2023; Oña et al., 2020). To assess the genetic diversity of the 19 genotypes, the spectra were analyzed separately in two datasets: fermented and unfermented beans.

### *2.5.2 Discriminant Analysis*

PLS-DA models were chosen for the bean discriminations as it has been frequently used to classify cocoa samples (Teye et al., 2020; Sentellas & Saurina, 2023). The data was split into calibration and validation sets.

For the discrimination of fermented from unfermented samples, a hold-out validation was performed. The calibration set included 280 beans (70 % of all the beans), 147 being from fermented beans and 133 from unfermented ones. The validation set included 120 beans (30 % of all beans), 63 being from fermented beans and 57 from unfermented ones. The separation between the calibration and the validation sets was created by the R function "createDataPartition" from the caret package (Kuhn, 2022), assuring an equal repartition of fermented and unfermented beans in both datasets. The optimal number of latent variables of those models was estimated by 10-fold cross-validation based on the area under the receiver operating characteristic curve (AUC), which represents the overall ability of the model to correctly classify predictions Eq. (A.1).

A PLS-DA model was then constructed on the whole spectral range (400-2500 nm) and the variable importance in projection (VIP) method was used as a strategy to select the most important wavelengths to distinguish groups of fermented and unfermented beans (Oliveira et al., 2023). The VIP score value greater than 1 was used to select relevant variables, and new PLS-DA models were constructed in the regions 400-700 nm, 1400-1600 nm, 1900- 2500 nm, 2000-2250 nm, and 2250-2350 nm. In addition to the PLS-DA models, Linear Discriminant Analysis (LDA) models were constructed using absorbances at two wavelengths in both the visible (472 nm and 636 nm) and infrared (2096 nm and 2278 nm) ranges.

For discrimination models between genotypes in fermented cocoa beans, the calibration set was formed with 147 beans (70 % of all beans), with each of the 18 genotypes represented by 7 beans, except for the duplicate genotypes (P7, CCN51, and CAB270) represented by 14 beans. The validation set was formed with 63 beans (30 % of all samples), with each genotype represented by 3 samples, except P7, CCN51, and CAB270 represented by 6 samples.

For the unfermented beans, the same logic was employed. The calibration set was therefore formed with 133 beans (70 % of all beans), with each of the 19 genotypes represented by 7 beans. The validation set was formed with 57 beans (30 % of the total

beans), with each genotype represented by 3 beans. PLS-DA models were then constructed over the entire spectral range (400–2500 nm).

The performance of the models was evaluated using three metrics derived from the confusion matrix, which compares the class assigned to the model with the real class of the samples. These metrics include sensitivity, indicating the model's ability to detect positive cases among truly positive samples Eq. (A.2); specificity, reflecting the model's ability to identify negative cases among truly negative samples Eq. (A.3); and accuracy, representing the model's ability to correctly classify the samples Eq. (A.4) (Hossin & Sulaiman, 2015).

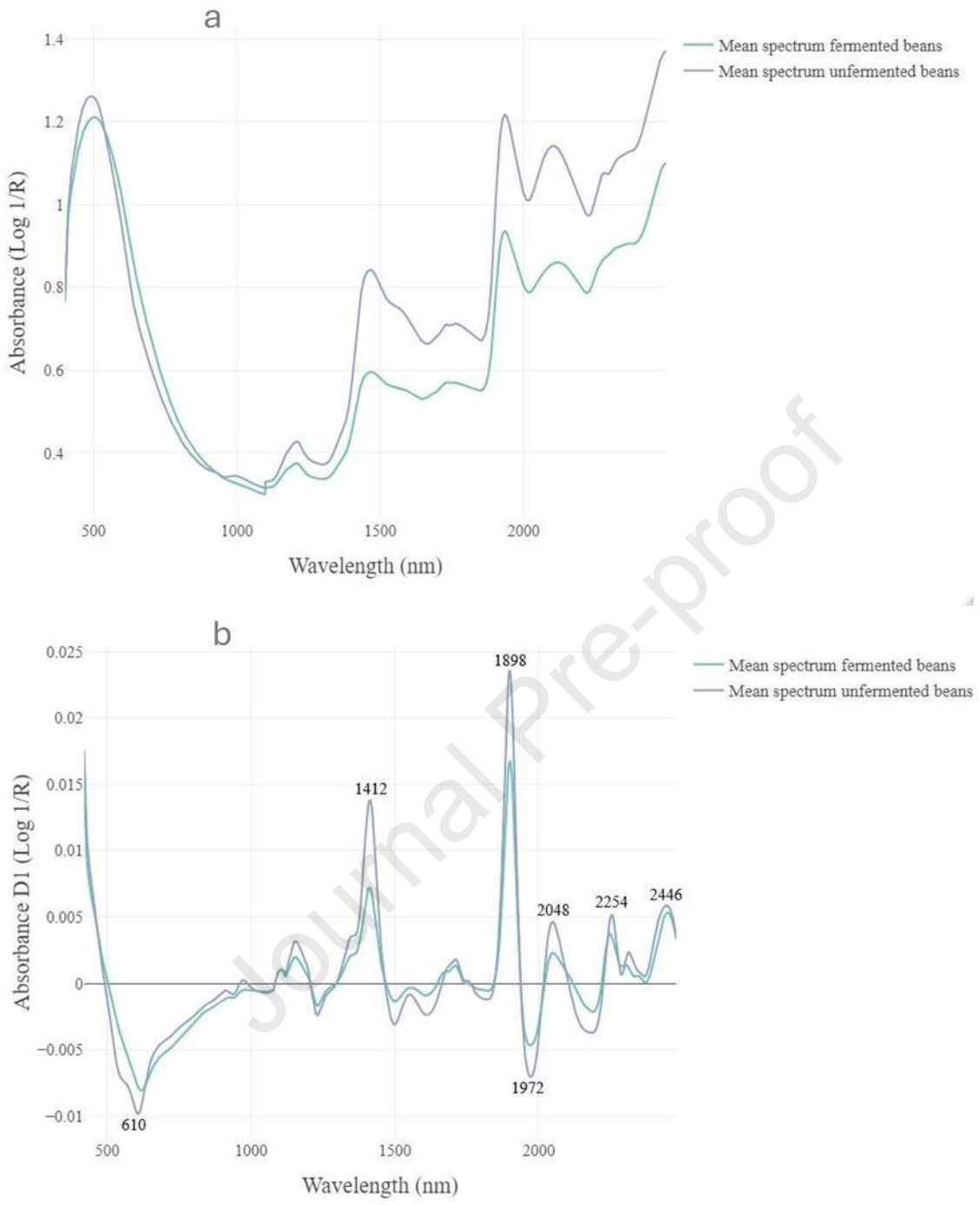
## 3. Results and Discussion

### *3.1. Discrimination between fermented and unfermented beans*

#### *3.1.1. Exploratory data analysis*

The mean raw spectra of all fermented and unfermented genotypes were obtained and presented in Fig. 1a. The similarity of spectral profiles is inherent to the species (*Theobroma cacao*) and is comparable to findings in other studies (Barbin et al., 2018; Mandrile et al., 2019; Cruz-Tirado et al, 2020; Drees et al., 2023). Both sets of samples exhibited a similar trend in absorbance but differed mainly around 500 nm and in the range between 1500 and 2500 nm, where unfermented samples showed higher absorbance than fermented ones (Fig. 1a). Differences in absorbance may be linked to the biochemical changes in the composition of the cocoa beans after fermentation (Quelal-Vásconez et al., 2019).

Spectra was preprocessed using first derivative Savitzky-Golay (SG) second polynomial order with 21 points (Fig. 1b), aiming to remove absolute variations in absorbance and unwanted scatter additive effects due to differences in the optical path length and fluctuations of the light source that commonly affect NIR spectra.




227

228 **Fig. 1**. Mean spectra of the fermented (grey line) and unfermented (green line) cocoa bean
229 samples. a. raw b. after the first derivative Savitzky-Golay preprocessing (width = 21,
230 order = 2a).

231 PCA was performed to identify possible clusters based on the pre-processed spectra of
232 the different datasets. Despite the samples coming from different origins, there was no
233 influence of location (Medicilândia and Tucumã) on the results, but in line with
234 expectations, the fermentation process caused evident clustering of the NIR spectra.

235 Confidence ellipses with a 99 % confidence level were added and showed an important
236 potential to discriminate fermented from unfermented beans (Fig. 2).


237

238 **Fig. 2.** PCA of NIR spectra of the fermented and unfermented cocoa bean samples.
239 Ellipses with confidence levels of 99 % were drawn for each group.

240 The spectrum of the MA 11 genotype in fermented beans and CA 6 in unfermented
241 beans showed atypical variations, suggesting potential errors during sample collection or
242 analysis. A remaining unfermented spectrum of the CAB 214 genotype was not assigned
243 to any group but was positioned close to the spectra of fermented samples in the PCA.
244 This might be related to the physical-chemical characteristics of the beans of this
245 genotype, which have external similarities with fermented beans (Table 2).

246 Some parameters are crucial in evaluating the quality of fermentation. For instance, to
247 achieve a high content of aromatic compounds, an internal pH of around 5 is expected for
248 fermented beans (Castro-Alayo et al., 2019). During the fermentation process, the sucrose
249 concentration tends to decrease. This reduction is indicative of microbial activity and is
250 reflected in the total soluble solids (°Brix). The beans must have a fermentation index
251 (FI) greater than 1 to be considered adequately fermented (BARIAH et al., 2014).
252 According to the data presented in Table 2, the parameters used to characterize fermented
253 beans are within the expected range. This observation may justify the successful
254 distinction of the Principal Component Analysis (PCA) of the NIR spectra.

The loadings of the first principal component of this PCA were plotted to see which spectral regions had the most influence on the separation between the two groups. However, the loadings showed the importance of numerous regions and did not bring much information about specific bands.

Despite being an unsupervised method, it already showed its potential for discrimination between fermented and unfermented cocoa beans in our samples. However, it did not highlight specific discriminating bands. Therefore, a PLS-DA (a supervised method) was performed to test the classification ability of the spectra.

*3.1.2. Discriminant Analysis*

The supervised PLS-DA method was used on the whole spectral range (400-2500 nm) and the VIP scores of this model were then calculated and plotted (Fig. 3) to explore the discriminating capacity of specific variables in relation to the classes of interest. VIP is a commonly applied method for selecting relevant variables and indicates the relative importance of wavelengths in a PLS model. Higher values indicate more significant contributions to the model (Wise et al., 2006).

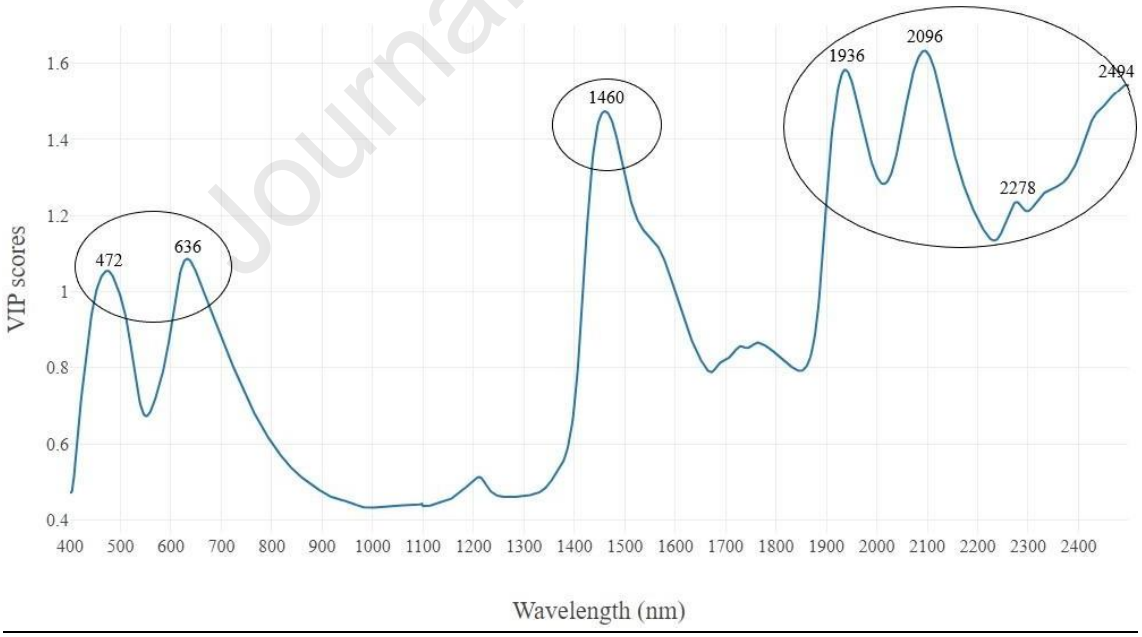



**Fig. 3.** VIP scores of the PLS-DA model (400-2500 nm) for the discrimination of fermented and unfermented cocoa beans.

The PLS-DA model constructed using only the most important wavelengths may provide better models than using the entire spectrum, in certain applications (Oliveira et

al., 2023). Acknowledging this, new models were built based on the wavelength ranges associated with VIP scores above one: the first within 400-700 nm, the second within 1400-1600 nm, and the third within 1900-2500 nm, which was subsequently subdivided into the 2000-2250 nm range and the 2250-2350 nm range (Table 3).

**Table 3**. Characteristics and performances of the PLS-DA models for discriminating fermented from unfermented cocoa beans.

| | Raw spectra | 400-700 nm | 1400-1600 nm | 1900-2500 nm | 2000-2250 nm | 2250-2350 nm |
|---|---|---|---|---|---|---|
| NLV* | 3 | 3 | 2 | 3 | 2 | 2 |
| Sensitivity | 1 | 1 | 1 | 1 | 1 | 1 |
| Specificity | 1 | 1 | 0.984 | 1 | 0.984 | 0.984 |
| Accuracy | 1 | 1 | 0.992 | 1 | 0.992 | 0.992 |

* Number of latent variables

The tested models all exhibited high accuracy in distinguishing between fermented and unfermented beans. Similarly, the specificity was consistently high. This highlights the intricate variances in chemical composition among the samples, which influence the wavelengths across all chosen spectral ranges.

Our results demonstrate that, despite the genetic variability of the cocoa genotypes analyzed, the distinctive characteristics between fermented and unfermented beans remained distinguishable through NIR spectroscopy analysis. This observation is supported by the physical-chemical differences induced by fermentation, as can be observed through PCA (Fig. S.2 see supplementary material). For example, fermented samples have lower acidity levels and total soluble solids content and are less bitter and astringent.

Teye et al. (2014) and Hernandez-Hernandez et al. (2022) have also reported that fermentation significantly modifies the NIR spectral profiles of cocoa beans. Kutsanedzie et al. (2017) note that fermentation changes the content of phenolic compounds and other metabolites in cocoa beans, impacting the chemical composition and sensory characteristics.

*3.1.2.1 PLS-DA on the spectral range of 400-700 nm*

The wavelength range of 400-700 nm corresponds to the visible part and is influenced by pigments. The bands observed at 472 nm and 636 nm (Fig. 3) are associated with the presence of anthocyanins, abundant in Forastero cocoa (Strayer, 1995; Camu et al., 2008). During fermentation, anthocyanins are hydrolyzed and cause a color change in the beans from violet to brown (Melo et al., 2021). The reduction in anthocyanin concentrations is

evident in the spectra of fermented and unfermented beans, mainly in the 610 nm band (Fig. 1b). The results obtained through the spectra align consistently with the results of traditional methods based on changes in the color of the beans.

The widely used cut test involves visual analysis, displaying the percentage of violet and brown beans (Fig. 4a). Complementary chemical methods, such as the fermentation index (FI), can also be used (Fig. 4b) (Bariah, 2014), and color analysis using the CIEL* system provides another alternative for evaluating fermentation (Fig. 4c). These results suggest that the internal biochemical transformations occurring in cocoa beans during the fermentation process can be easily observed and quickly accessed from whole beans.

The analysis in the visible range of the spectra is valuable for both cocoa producers and industries, as it avoids the cost of using more expensive NIR technologies, in addition to providing information about the fermentative quality of the beans without the drawbacks of traditional methods, such as subjectivity (cut test), long processing time, and the use of toxic substances (fermentation index) and destructiveness (color analysis). Furthermore, an LDA analysis was performed at wavelengths 472 nm and 636 nm and showed a good performance (Table S.1), reinforcing the use of visible spectral range in discrimination from whole beans, in addition to potentially reducing the transferability costs of the analysis.

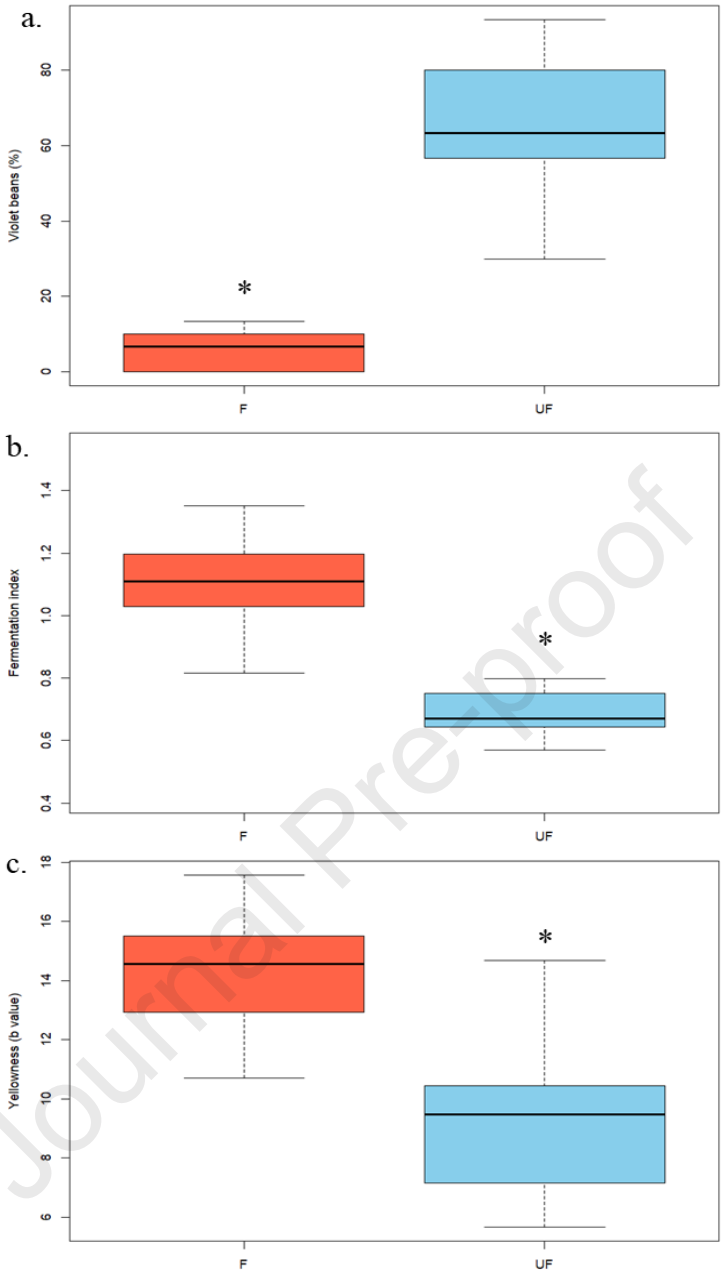



Fig. 4 Traditional methods based on color changes to classify fermented (F) and unfermented (UF) cocoa beans: a. Percentage of violet beans observed after a cut test, characteristic of unfermented beans. b. Fermentation index: FI values ≥1 indicate well-fermented beans, while FI<1 corresponds to poorly or unfermented beans. c. Yellowness: higher b* values correspond to brown pigments characteristic of fermented beans. *Student's t-test, $p < 0.05$.

*3.1.2.2 PLS-DA on the spectral range of 1400-1600 nm*

The model constructed in the wavelength range of 1400 nm to 1600 nm was also efficient in discriminating between fermented and unfermented beans. The band around 1450 nm is related to OH vibration, found in water and also in carbohydrates and polyphenols (Afoakwa et al., 2013). However, since both sample groups were dried until

reaching moisture content close to 8%, water should not be the variable justifying the performance of this model.

Proteins, with a characteristic band around 1470 nm linked to the $NH_2$ structure (Osborne, Fearn, and Hindle, 1993), might be a differentiator between the groups. Furthermore, the spectral range between 1470 nm and 1639 nm has been associated with carbohydrates (Krähmer et al., 2015), and the band at 1596 nm with starch and glucose (Mandrile et al., 2019). Unfermented beans, even when dried, retain a layer of pulp rich in carbohydrates and soluble solids, unlike fermented ones, where this layer is consumed by fermentation. This variation is reflected in the total soluble solids values of the outer part of the beans, with the average of the fermented ones (3.72 ºBrix) being lower than that of the unfermented ones (10.5 ºBrix).

*3.1.2.3 PLS-DA on the spectral range of 1900-2500 nm*

The third model was built in the spectral range of 1900-2500 nm and effectively discriminated between fermented and unfermented cocoa beans. The region near 1950 nm can be associated with the O-H combination band (Forte et al., 2022), the peak around 2100 nm corresponds mainly to starch (Cozzolino, Degner & Eglinton, 2014), cellulose is associated with the wavelength of 2199 nm (Okiyama et al., 2017; Wang et al., 2018) and the absorbance around 2057 nm can be attributed to protein (Caporaso et al., 2018).

Several other compounds may have influenced the efficiency of this model, given that the contents of cocoa bean shells include lipids, proteins, starch, theobromine, and caffeine, among others (Mandrile et al, 2019). New PLS-DA models were then developed in the 2000-2250 nm and 2250-2350 nm ranges, selected according to the VIP scores, to assess whether smaller spectral regions, and therefore spectrometers with reduced spectral ranges, could provide performances close to those obtained with the entire spectral range.

In the range of 2000-2250 nm, the bands around 2050 nm (Forte et al., 2022) and 2180 nm (Samadi, Wajizah & Zulfahrizal, 2021) are associated with the presence of proteins. Interestingly, the average protein contents of fermented whole beans (18.35 g/100g DW) and unfermented beans (17.49 g/100g DW) were very similar (Table 2).This region may be relevant due to qualitative differences in the proteins that undergo hydrolysis by the action of the aspartic endoprotease and carboxypeptidase enzymes during fermentation (Santander Muñoz et al., 2020)

The range of 2250-2350 nm is associated with lipid content. According to Veselá et al. (2007), the most important bands related to lipid variation are at 2322, 2334, and 2360 nm. This component is abundant in cocoa beans and tends to decrease during fermentation (Aremu, Agiang & Ayatse, 1995). However, similar to proteins, the average lipid values for unfermented beans (31.66 g/100g DW) and fermented beans (31.64 g/100g DW) are very close (Table 2). The model may have been influenced by differences in the nature of these components or their distribution in the cocoa bean shells, reinforcing that the differences indicated by the NIR spectra do not necessarily reflect the internal characteristics of the beans and should be investigated further.

Compared to other studies, cocoa from state of Pará displays less variability between genotypes concerning protein and lipid values (Table 2). Different cocoa genotypes from Mexico have a protein content ranging from 11.93 to 29.13 g/1 and lipid content ranging from 18.65 to 49.48 g/100 g DW (Hernández-Hernández et al., 2022). Another study conducted in Peru on 30 cocoa genotypes revealed a protein content ranging from 17.51 to 30.87 g/100g DW (Oliva-Cruz et al., 2021), and Colombian cocoa presented an average protein content of 30.82 g/100g DW for different genotypes, with a coefficient of variation of 21.81% (Chang et al., 2014).

Finally, an LDA model was built using two wavelengths (2096 nm and 2278 nm). This may be related to the concentrations of starch (2100 nm), sucrose (2088 nm), theobromine (2094 nm), and polyphenols (2150-2250 nm), which are present in the cocoa bean shells and are affected by the biochemistry of fermentation (Hernández-Hernández et al., 2022).

The LDA model was able to perfectly discriminate between fermented and unfermented cocoa beans, achieving maximum sensitivity, specificity, and accuracy parameters. This suggests that discrimination can be effectively achieved using the absorbances at two specific wavelengths, eliminating the need for a wide-range spectrometer.

### *3.2. Discrimination of Forastero cocoa genotypes from the Brazilian Amazon*

#### *3.2.1. Exploratory data analysis*

Our work, for the first time, explores the genetic diversity of 19 Forastero cocoa genotypes from the Brazilian Amazonia through NIR spectroscopy. The differences in absorbance intensities of the genotypes' NIR spectra suggest that their particular characteristics can be detected based on spectroscopic information (Fig. S.1).

Previous works already mentioned that genetic differences could be observed through NIR spectral information (Castro et al., 2022; Cruz-Tirado et al., 2020). Nevertheless, the fermentation process may lead to a homogenization of differences among genotypes, thereby posing a challenge in distinguishing fermented samples, potentially due to the maintenance of each genotype's intrinsic biochemical attributes (Hernandez-Hernandez et al., 2022). In contrast, Ferreira et al. (2022) have demonstrated that cocoa fermentation can assist in classifying samples based on their geographical origins.

To comprehensively investigate the spectral variations, we analyzed the two distinct data sets: fermented and unfermented beans. Preliminary exploratory investigations were carried out, utilizing both raw and pre-processed data in PCA and t-SNE analyses but no clustering patterns were observed among the different genotypes. However, it is worth noting that the t-SNE of the raw spectra from the unfermented beans showed a tendency towards segregation based on the majority genetic groups described in Table 1 (Fig. 5). Despite an insufficient separation for selective discrimination, this result suggests that the genetic information and inherent biochemical characteristics of the beans are more evident before fermentation.

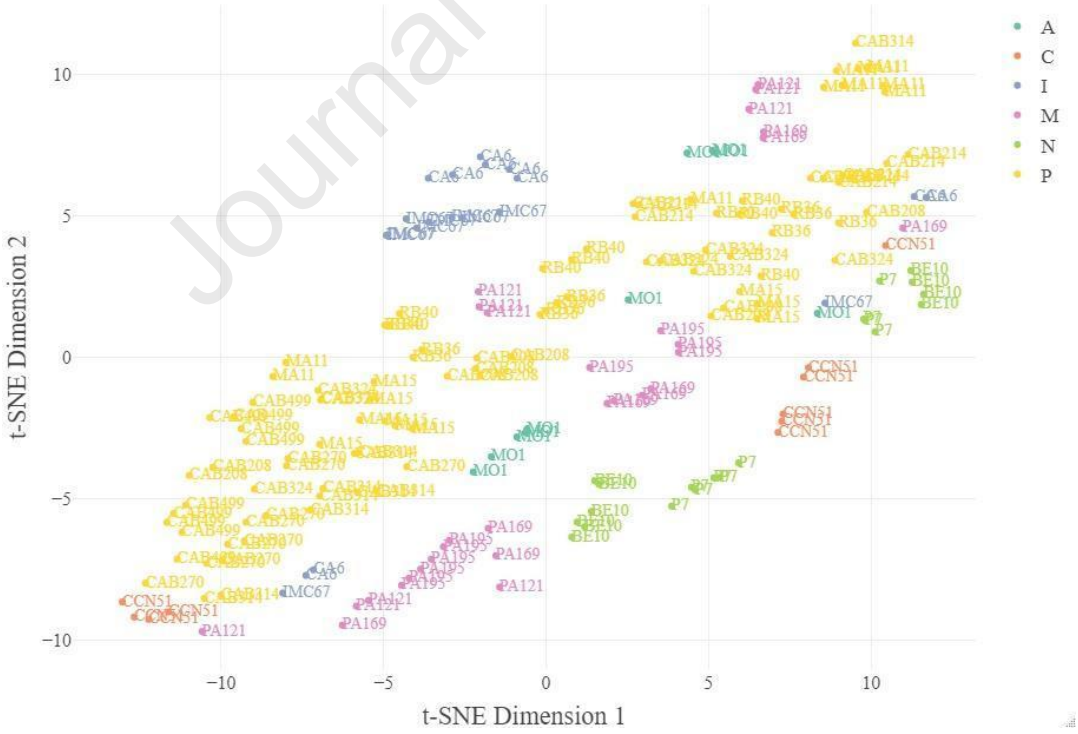



**Fig. 5.** t-SNE on the raw spectra of the unfermented cocoa beans. The legend displays the genetic groups of the samples. “A” stands for Amenolado, “C” for Contamana, “I” for Iquitos, “M” for Marañón, “N” for Nanay, and “P” for Purús.

418 However, unlike methods such as PCA or PLS, t-SNE does not explicitly provide a
419 measure of the importance of variables. In the context of t-SNE, the main focus is the
420 visualization and representation of similarity patterns, not the direct interpretation of
421 individual variables. To further investigate the genetic diversity of the samples based on
422 NIR spectra, discriminatory analyzes were performed.

423 *3.2.2. Discriminant Analysis*

424 The construction of PLS-DA models over the entire spectral range (400–2500 nm) for
425 discriminating cocoa genotypes was proposed using NIR spectra of both fermented and
426 unfermented beans. The raw and pre-processed spectra (SG) were tested and accuracy
427 was the metric used to select the best models.
428 For fermented beans, the model using data pre-processed by SG was the most
429 effective, achieving an accuracy of 0.86. Among all 63 validation samples, 9 were
430 classified incorrectly. For unfermented beans, the best model was the one constructed
431 from raw data, with an accuracy of 0.84. Out of the 57 validation samples, 9 were
432 incorrectly classified. The confusion matrices of the models are shown, respectively, in
433 Tables S.2 and S.3 (Supplementary Material).

434 Various spectral regions are crucial in differentiating cocoa genotypes, as indicated by
435 the VIP scores in the models. Although the physicochemical data of the genotypes
436 showed low variability, the composition differences in the beans might relate to
437 components not assessed in this study, like carbohydrates, phenolic compounds,
438 alkaloids, pectin, cellulose, hemicellulose, etc. Additionally, the physicochemical
439 analyses were performed on whole beans, but NIR measurements might have limitations,
440 as NIR assesses the proportion of light reflected, and deeper layers in solid samples might
441 not reflect light effectively.

442 The trend of clustering by genetic group (see 3.2.1) was used to investigate model
443 classification errors. In fermented beans, only one reference genotype presented a genetic
444 group in common with the genotype predicted by the PLS-DA model (Table 4). However,
445 in the model for unfermented beans, most misclassified genotypes shared genetic
446 ancestry, suggesting that genetic influences on biochemical similarities are more apparent
447 before fermentation (Table 4).

**Table** 4. Comparison of the reference and predicted genotypes for the misclassified samples for the PLS-DA model built on SG data for the discrimination of the genotypes of fermented and unfermented cocoa beans.

| | Reference genotype | Reference genotype groups (and associated Q value) | Predicted genotype | Predicted genotype groups (and associated Q value) |
|---|---|---|---|---|
| FERMENTED | CA6 | Iquitos (0.32) - Nanay (0.20) | RB40 | Purús (0.86) |
| | CAB208 | Purús (0.74) | BE10 | Nanay (0.30) - Marañón (0.26) |
| | CAB499 | Purús (0.57) | CCN51 | Criollo (0.46) - Iquitos (0.29) |
| | CCN51 | Criollo (0.46) - Iquitos (0.29) | CAB324 | Purús (0.46) - Nanay (0.46) |
| | CCN51 | Criollo (0.46) - Iquitos (0.29) | P7 | Nanay (0.53) - Contamana (0.43) |
| | MA15 | Purús (0.84) | P7 | Nanay (0.53) - Contamana (0.43) |
| | P7 | Nanay (0.53) - Contamana (0.43) | MO1 | Amelonado (0.37) - Purús (0.27) |
| | PA195 | Marañón (0.67) | CAB214 | Purús (0.54) - Contamana (0.42) |
| | PA195 | Marañón (0.67) | PA121 | Marañón (0.92) |
| UNFERMENTED | BE10 | Nanay (0.30) - Marañón (0.26) | CA6 | Iquitos (0.32) - Nanay (0.20) |
| | CAB270 | Purús (0.32) - Guiana (0.28) | PA169 | Marañón (0.47) - Amelonado (0.25) |
| | MO1 | Amelonado (0.37) - Purús (0.27) | CAB324 | Purús (0.46) - Nanay (0.46) |
| | MO1 | Amelonado (0.37) - Purús (0.27) | MA15 | Purús (0.84) |
| | P7 | Nanay (0.53) - Contamana (0.43) | IMC67 | Iquitos (0.66) |
| | P7 | Nanay (0.53) - Contamana (0.43) | MA11 | Purús (0.42) - Amelonado (0.28) |
| | PA195 | Marañón (0.67) | BE10 | Nanay (0.30) - Marañón (0.26) |
| | RB36 | Purús (0.95) | CAB208 | Purús (0.74) |
| | RB36 | Purús (0.95) | CAB214 | Purús (0.54) - Contamana (0.42) |

Both models exhibited low performance with the P7 genotype, and none of the erroneously predicted genotypes shared common genetic groups with this reference genotype, indicating unique compositional traits in these beans. Factors other than genetics might influence model performance. Notably, the reference genotype P7 and the predicted genotype IMC67 in the PLS-DA model for unfermented beans exhibit similar aromatic compositions (Collin et al., 2023). Furthermore, cocoa genotypes descending from P7 showed significant classification errors in a PLS-DA model using hyperspectral NIR images for hybrid classification (Cruz-Tirado et al., 2020). The same authors

reported a 4.4-34.4% prediction error using a PLS-DA model to discriminate five cocoa hybrids.

The genetic complexity of Brazilian Amazon cocoa beans is challenging even for conventional analyses, as shown by coefficients of membership that demonstrate a mix of contributions from different groups to the same genotype (Table 1). While genetic analyses provide valuable information on the authenticity of cocoa beans, they may not be the most thorough or practical approach. These analyses are expensive, time-consuming, require specialized equipment and technical know-how, making them less accessible and limiting their applicability in certain contexts. Furthermore, misidentification of cocoa genotypes occurs in about 15% to 44% of cases (Motamayor et al., 2008).

The diversity and complexity are reflected in the bean spectra. Despite this, both developed PLS-DA models showed great effectiveness in distinguishing Amazonian cocoa genotypes. They could be further enhanced by broadening the sampling plan to evaluate the NIR method under realistic conditions, including the incorporation of comprehensive information about the composition and natural interferences present in the samples.

The findings of this study are particularly valuable due to the high genetic variability of cocoa beans. These results emphasize that NIR spectroscopy, being rapid and non-destructive, is a feasible tool for authenticating cocoa genotypes in both fermented and unfermented whole beans. This understanding is crucial for the continual improvement of NIR models and for developing more effective selection and genetic improvement strategies in the cocoa sector.

## 4. Conclusion

This study reaffirms the effectiveness of NIR spectroscopy in conjunction with multivariate analysis techniques in the authentication of cocoa beans, providing valuable insights into specific bands associated with crucial biochemical components. Both visible and infrared spectral regions are efficient for discriminating between fermented and unfermented whole grains, as well as an LDA with only two wavelengths (472 nm and 636 or 2096 nm and 2278 nm), suggesting the design of specific spectra sensors for smaller, cheaper, and more accurate applications. Additionally, we highlight that NIR spectroscopy can capture subtle variations in genetic characteristics. The PLS-DA models

showed good performance in discriminating cocoa genotypes in both fermented and unfermented beans, with accuracies of 0.86 and 0.84, respectively. The results have significant practical implications for the cocoa industry, offering a practical and efficient solution to address challenges associated with traditional methods of quality control and authentication. This non-invasive approach aligns with the growing industry focus on sustainability, efficiency, and the adoption of environmentally friendly methods. Furthermore, we are investigating the potential of other techniques in the discrimination and authentication of Amazonian cocoa beans, such as Raman spectroscopy and Hyperspectral Imaging.

## 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.

## CRediT authorship contribution statement

**Anne Pinto**: Investigation, Writing - original draft, Formal analysis and Visualization. **Antoine Deryck**: Software, Methodology, Writing - original draft and Formal analysis. **Giulia V. Lima**: Formal analysis, Investigation, Writing - original draft. **Ana Caroline de Oliveira**: Investigation **Fábio Gomes Moura:** Investigation **Juan Antonio Fernández Pierna**: Writing - review & editing. **Douglas Barbin** Writing - review & editing **Vincent Baeten**: Conceptualization, Data curation, Writing - review & editing, Resources, Project administration, Funding acquisition. **Hervé Rogez**: Conceptualization, Writing-review & editing, Resources, Supervision, Project administration, Funding acquisition.

## Acknowledgements

This study was financed by a cooperation project between the Wallonie-Bruxelles International and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES-WBI/2017-2019/361394).

Authors are grateful to CEPLAC (Pará, Brazil), CNPQ (Conselho Nacional de Desenvolvimento, Científico e Tecnológico) and PROPESP/UFPA for the grant and financial support.

## Reference

Afoakwa, E. O., Kongor, J. E., Takrama, J. F., & Budu, A. S. (2013). Changes in acidification, sugars and mineral composition of cocoa pulp during fermentation of pulp pre-conditioned cocoa (Theobroma cacao) beans. *International Food Research Journal* 20(3): 1215-1222.

Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & Sharma, N. (2019). Detection of spatial outlier by using improved Z-score test. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 788-790). IEEE.

Aikpokpodion, P. E., & Dongo, L. N. (2010). Effects of fermentation intensity on polyphenols and antioxidant capacity of cocoa beans. *Int. J. Sustain. Crop Prod*, v. 5, n. 4, p. 66-70.

AOAC, 2023. Official Methods of Analysis of the Association of Analytical Chemists International, 22nd ed. Gathersburg, MD, USA, Official Methods.

Aremu, C. Y., M. A. Agiang, et J. O. I. Ayatse. (1995) Nutrient and Antinutrient Profiles of Raw and Fermented Cocoa Beans. *Plant Foods for Human Nutrition* 48 (3): 217-23.

Barbin, D. F., Maciel, L. F., Bazoni, C. H. V., Ribeiro, M. D. S., Carvalho, R. D. S., Bispo, E. D. S., & Hirooka, E. Y. (2018). Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses. *Journal of food science and technology*, *55*, 2457-2466.

Bariah, K. (2014). Impact of fermentation duration on the quality of Malaysian cocoa beans using shallow box. *Asia-Pacific Journal of Science and Technology*, *19*, 74-80.

Camu, N., De Winter, T., Addo, S. K., Takrama, J. S., Bernaert, H., & De Vuyst, L. (2008). Fermentation of cocoa beans: influence of microbial activities and polyphenol concentrations on the flavour of chocolate. *Journal of the Science of Food and Agriculture*, *88*(13), 2288-2297.

Caporaso, N., Whitworth, M. B., Fowler, M. S., & Fisk, I. D. (2018). Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans. Food Chemistry, 258, 343–351. https://doi.org/10.1016/j.foodchem.2018.03.039.

Castro, W., De-la-Torre, M., Avila-George, H., Torres-Jimenez, J., Guivin, A., & Acevedo-Juárez, B. (2022). Amazonian cacao-clone nibs discrimination using NIR spectroscopy coupled to naïve Bayes classifier and a new waveband selection approach. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *270*, 120815.

Castro-Alayo, E. M., Idrogo-Vásquez, G., Siche, R., & Cardenas-Toro, F. P. (2019). Formation of aromatic compounds precursors during fermentation of Criollo and Forastero cocoa. Heliyon, 5(1).

Chang, J. F. V., Torres, C. V., Morán, D. E. P., Véliz, J. M., Remache, R. R., & Rodríguez, W. M. (2014). Atributos físicos-químicos y sensoriales de las almendras de quince clones de cacao nacional (Theobroma cacao L.) en el Ecuador. Ciencia y Tecnología, 7(2), 21-34.

Collin, S., Fisette, T., Pinto, A., Souza, J., & Rogez, H. (2023). Discriminating aroma compounds in five cocoa bean genotypes from two brazilian states: white kerosene-like catongo, red whisky-like FL89 (Bahia), Forasteros IMC67, PA121 and P7 (Pará). *Molecules*, *28*(4), 1548.

Cozzolino, D., S. Degner, et J. Eglinton. (2014). A Review on the Role of Vibrational Spectroscopy as An Analytical Method to Measure Starch Biochemical and Biophysical Properties in Cereals and Starchy Foods. *Foods* 3 (4): 605-21. https://doi.org/10.3390/foods3040605.

Cruz-Tirado, J. P., Pierna, J. A. F., Rogez, H., Barbin, D. F., & Baeten, V. (2020). Authentication of cocoa (Theobroma cacao) bean hybrids by NIR-hyperspectral imaging and chemometrics. *Food Control*, *118*, 107445.

Drees, A., Brockelt, J., Cvancar, L., & Fischer, M. (2023). Rapid determination of the shell content in cocoa products using FT-NIR spectroscopy and chemometrics. *Talanta*, *256*, 124310.

Ferreira, F. N., Albuquerque Chagas-Junior, G. C., Santana de Oliveira, M., Rodrigues Barbosa, J., Chaves Oliveira, M. E., & Santos Lopes, A. (2022). Geographical Discrimination of Ground Amazon Cocoa by Near-Infrared Spectroscopy: Influence of Sample Preparation. Journal of Food Quality, 2022.

Forte, M., Currò, S., Van de Walle, D., Dewettinck, K., Mirisola, M., Fasolato, L., & Carletti, P. (2022). Quality Evaluation of Fair-Trade Cocoa Beans from Different Origins Using Portable Near-Infrared Spectroscopy (NIRS). *Foods*, 12(1), 4.

Gourieva, K.B., & Tserrevitinov, O.B. (1979). Method of evaluating the degree of fermentation of cocoa beans. USSR Patent no.646254.

Hashimoto, J. C., Lima, J. C., Celeghini, R. M., Nogueira, A. B., Efraim, P., Poppi, R. J., & Pallone, J. A. (2018). Quality control of commercial cocoa beans (Theobroma cacao L.) by near-infrared spectroscopy. Food analytical methods, 11, 1510-1517.

Hernández-Hernández, Carolina, Víctor M. Fernández-Cabanás, Guillermo Rodríguez-Gutiérrez, África Fernández-Prior, et Ana Morales-Sillero. (2022). Rapid Screening of Unground Cocoa Beans Based on Their Content of Bioactive Compounds by NIR Spectroscopy. *Food Control* 131 (janvier): 108347. https://doi.org/10.1016/j.foodcont.2021.108347.

Hossin, Mohammad, et Sulaiman M.N. 2015. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process* 5 (mars).

ISO 2451. (2017). *Cocoa beans — Specification and quality requirements* (Issue 67.140.30 Cocoa, pp. 1–19). International Organization for Standardization. https://www.iso.org/standard/68202.html

Krähmer, A., Engel, A., Kadow, D., Ali, N., Umaharan, P., Kroh, L. W., & Schulz, H. (2015). Fast and neat–Determination of biochemical quality parameters in cocoa using near infrared spectroscopy. *Food chemistry*, *181*, 152-159.

Kucheryavskiy, S. (2020). mdatools–R package for chemometrics. Chemometrics and Intelligent Laboratory Systems, 198, 103937.

Kuhn M (2022). _caret: Classification and Regression Training_. R package version 6.0-93, <https://CRAN.R-project.org/package=caret>.

Kutsanedzie, F. Y., Chen, Q., Sun, H., & Cheng, W. (2017). In situ cocoa beans quality grading by near-infrared-chemodyes systems. Analytical methods, 9(37), 5455-5463.

Lesnoff M (2022). _rchemo: Dimension Reduction, Regression and Discrimination for Chemometrics. *R Package Version 0.0-17.*, <https://github.com/mlesnoff/rchemo>.

Mandrile, L., Barbosa-Pereira, L., Sorensen, K. M., Giovannozzi, A. M., Zeppa, G., Engelsen, S. B., & Rossi, A. M. (2019). Authentication of cocoa bean shells by near-and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy. *Food chemistry*, *292*, 47-57.

MAPA. Ministério da Agricultura e Pecuária. Projeções do Agronegócio - Brasil 2022/23 a 2032/33 Projeções de Longo Prazo. Brasília, 2023.

Melo, T. S., Pires, T. C., Engelmann, J. V. P., Monteiro, A. L. O., Maciel, L. F., & Bispo, E. D. S. (2021). Evaluation of the content of bioactive compounds in cocoa beans during the fermentation process. *Journal of food science and technology*, *58*, 1947-1957.

Motamayor, J. C., Lachenaud, P., Da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., & Schnell, R. J. (2008). Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PloS one*, *3*(10), e3311.

Okiyama, D. C., Navarro, S. L., & Rodrigues, C. E. (2017). Cocoa shell and its compounds: Applications in the food industry. *Trends in Food Science & Technology*, *63*, 103-112.

Oliveira, M. M., Badaró, A. T., Esquerre, C. A., Kamruzzaman, M., & Barbin, D. F. (2023). Handheld and benchtop vis/NIR spectrometer combined with PLS regression for fast prediction of cocoa shell in cocoa powder. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *298*, 122807.

Oliva-Cruz, M., Mori-Culqui, P. L., Caetano, A. C., Goñas, M., Vilca-Valqui, N. C., & Chavez, S. G. (2021). Total fat content and fatty acid profile of fine-aroma cocoa from northeastern Peru. Frontiers in Nutrition, 8, 677000.

Oña, A. J. O., Grijalva, F., Proaño, K., Acuña, B., & García, M. (2020, October). Classification of Fresh Cocoa Beans with Pulp Based on Computer Vision. In *2020 IEEE ANDESCON* (pp. 1-6). IEEE.

Osborne, B. G., Fearn, T., & Hindle, P. H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*. Longman scientific and technical.

Pierna, J. F., Wahl, F., De Noord, O. E., & Massart, D. L. (2002). Methods for outlier detection in prediction. Chemometrics and Intelligent Laboratory Systems, 63(1), 27-39.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Quelal-Vásconez, M. A., Lerma-García, M. J., Pérez-Esteve, É., Arnau-Bonachera, A., Barat, J. M., & Talens, P. (2019). Fast detection of cocoa shell in cocoa powders by near infrared spectroscopy and multivariate analysis. *Food Control*, *99*, 68-72.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Santander Muñoz, M., Rodríguez Cortina, J., Vaillant, F. E., & Escobar Parra, S. (2020). An overview of the physical and biochemical transformation of cocoa seeds to beans and to chocolate: Flavor formation. *Critical reviews in food science and nutrition*, *60*(10), 1593-1613.

Senanayake, M.; Jansz, e. R.; Buckle, K. A. (1997) Effect of Different Mixing Intervals on the Fermentation of Cocoa Beans. *Journal of the Science of Food and Agriculture*, v. 74, p. 42–48.

Sentellas, S., & Saurina, J. (2023). Authentication of cocoa products based on profiling and fingerprinting approaches: Assessment of geographical, varietal, agricultural and processing features. *Foods*, *12*(16), 3120.

Strayer, L. Biochemistry. (1995) W.H.Freeman and Company/Worth Publishers, New York (4th ed.).

Teye, E., Anyidoho, E., Agbemafle, R., Sam-Amoah, L. K., & Elliott, C. (2020). Cocoa bean and cocoa bean products quality evaluation by NIR spectroscopy and chemometrics: A review. *Infrared Physics & Technology*, *104*, 103127.

649 Teye, E., Huang, X., Takrama, J., & Haiyang, G. (2014). Integrating NIR Spectroscopy and
650 Electronic Tongue Together with Chemometric Analysis for Accurate Classification of Cocoa
651 Bean Varieties. Journal of Food Process Engineering, 37, 560-566.

652 Veselá, A., Barros, A. S., Synytsya, A., Delgadillo, I., Čopíková, J., & Coimbra, M. A. (2007).
653 Infrared spectroscopy and outer product analysis for quantification of fat, nitrogen, and moisture
654 of cocoa powder. *Analytica chimica acta*, *601*(1), 77-86.

655 Samadi, Wajizah, S., & Zulfahrizal, Z. (2021). Near infrared spectra features of cocoa pod husk
656 used for feedstuff. In *IOP Conference Series: Earth and Environmental Science* (Vol. 922, No.
657 1, p. 012011).

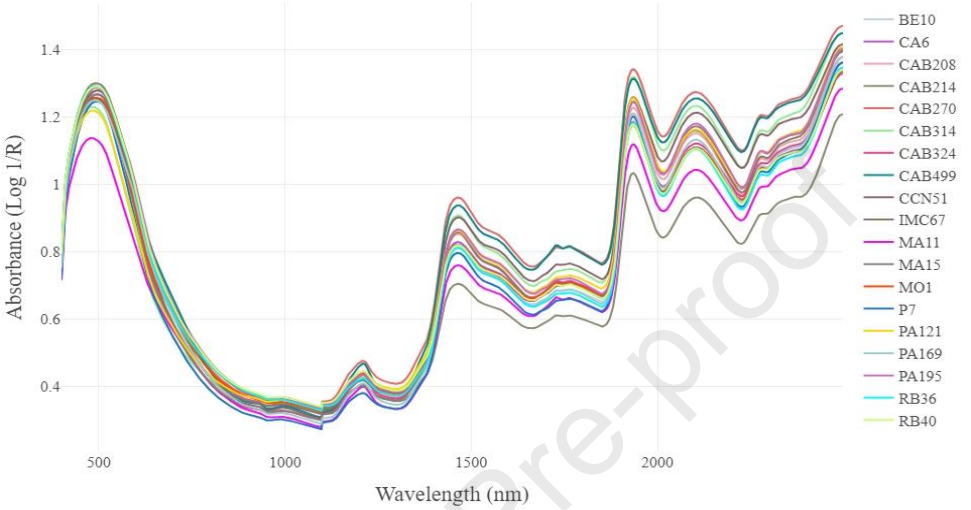
658 Sunoj, S., Igathinathane, C., & Visvanathan, R. (2016). Nondestructive determination of cocoa
659 bean quality using FT-NIR spectroscopy. Computers and Electronics in Agriculture, 124, 234-
660 242.

661 Wang, J., Zhang, X., Sun, S., Sun, X., Li, Q., & Zhang, Z. (2018). Online determination of quality
662 parameters of dried soybean protein–lipid films (Fuzhu) by NIR spectroscopy combined with
663 chemometrics. *Journal of Food Measurement and Characterization*, 12 (3), 1473–1484.
664 https://doi.org/10.1007/s11694-018-9762-z.

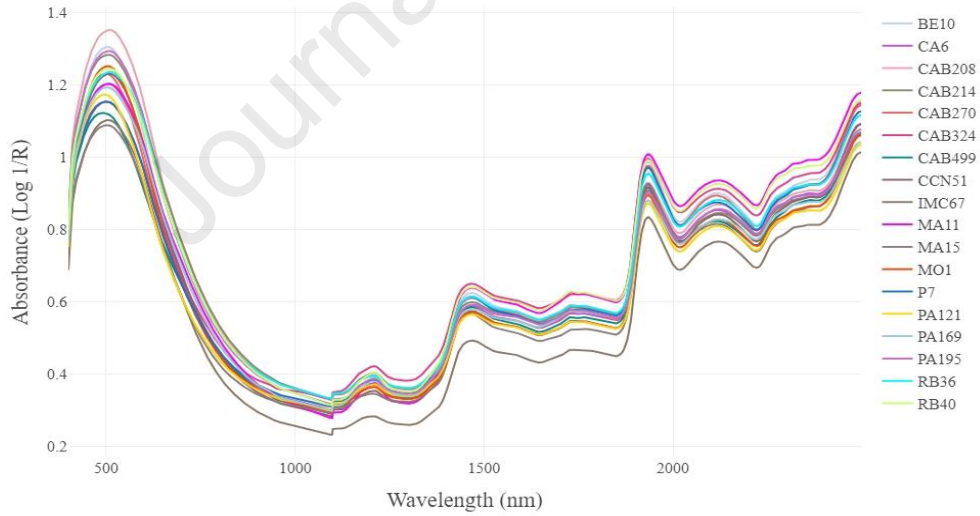
665 Wise BM, Gallagher NB, Bro R, et al (2006) PLS_Toolbox version 4.0 for use with MATLAB
666 ™

667

# SUPPLEMENTARY MATERIAL




A.




B.

**Fig. S. 1**. Mean spectra of 19 genotypes of dried cocoa beans samples from Amazonia: A. unfermented; and B. fermented.

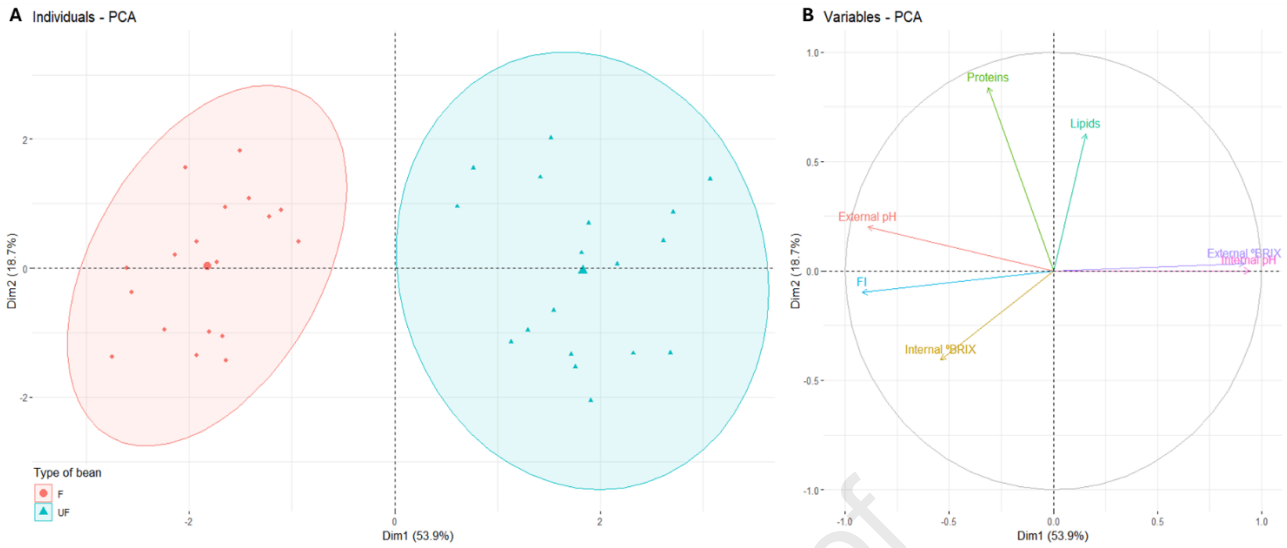



Fig. S.2 Principal Component Analysis (PCA) of cocoa beans. A: Score plot showing the distribution of fermented (F) and unfermented (UF) cocoa beans across the first two principal components. B: Loadings plot illustrating the contribution of variables (proteins, lipids, external and internal pH, internal and external total soluble solids in ºBrix) to the first two principal components.

**Table S.1** Characteristics and performances of the Linear Discriminant Analysis (LDA) models at visible (472 nm and 636 nm) and infrared (2096 nm and 2078 nm) wavelengths of raw spectra.

| Parameters | Visible | NIR |
|---|---|---|
| Sensitivity | 0.937 | 1 |
| Specificity | 0.982 | 1 |
| Accuracy | 0.958 | 1 |

The optimization criterion for the two-class PLS-DA model is the Area Under the Receiver Operating Characteristic Curve (AUC). The AUC value represents the model's overall ability to correctly rank predictions, with the ranking being derived from the prediction scores and therefore reflecting the model's ability to classify. This metric is sensitive to class imbalance (difference in the number of samples of each class) and has been proven more efficient than the accuracy for binary classifiers. The AUC is calculated using Eq. (A.1).

$$AUC = \frac{S_f - n_f(n_u + 1)/2}{n_f n_u} \qquad Eq.(A.1)$$

With:

$S_f = Sum\ of\ ranks\ of\ the\ fermented\ samples$
$n_f = Numbers\ of\ fermented\ samples$
$n_u = Numbers\ of\ unfermented\ samples$

The three metrics obtained from this are the Specificity, the Sensitivity and the Accuracy.

The Sensitivity represents the models' ability to detect the positive cases among the samples that are actually positive. It is derived from the Eq. (A.2).

$$Sensitivity = TP/(TP + FN) \quad Eq.(A.2)$$

With:

$TP = Number\ of\ True\ Positives$
$FN = Number\ of\ False\ Negatives$

The Specificity indicates the models' ability to detect negative cases among the samples that are actually negative. It is obtained with the Eq. A.3.

$$Specificity = TN/(TN + FP) \quad Eq.(A.3)$$

With:

$TN = Number\ of\ True\ Negatives$
$FP = Number\ of\ False\ Positives$

The Accuracy represents the model's ability to correctly classify a sample and is calculated using Eq. (A.4).

$$Accuracy = n_A/n_T \quad Eq.(A.4)$$

With:

$n_A = Number\ of\ samples\ assigned\ to\ their\ actual\ class\ (equal\ to\ TP + TN\ for\ binary\ classifiers)$

$n_T = Total\ number\ of\ samples\ (equal\ to\ TP + TN + FP + FN\ for\ binary\ classifiers)$

For multi-class models, where calculating the AUC is computationally expensive, accuracy was chosen as the optimization criterion. This metric, along with the confusion matrices, was also employed to evaluate the models' performances.

Reference : (Hossin et M.N 2015)

Biblio: Hossin, Mohammad, et Sulaiman M.N. 2015. « A Review on Evaluation Metrics for Data Classification Evaluations ». *International Journal of Data Mining & Knowledge Management Process* 5 (mars): 01-11.

725 **Table S.2** Confusion matrix of the PLS-DA model built on SG data for the
726 discrimination of the genotypes of fermented cocoa beans.

| Ref<br>Pred | BE10 | CA6 | CAB208 | CAB214 | CAB270 | CAB324 | CAB499 | CCN51 | IMC67 | MA11 | MA15 | MO1 | P7 | PA121 | PA169 | PA195 | RB36 | RB40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BE10 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CA6 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB208 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB214 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CAB270 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB324 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB499 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCN51 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| IMC67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MA11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MA15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MO1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| P7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| PA121 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 |
| PA169 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| PA195 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| RB36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| RB40 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

727 .

728 **Table S.3.** Confusion matrix of the PLS-DA model built on SG data for the discrimination of the genotypes of unfermented cocoa
729 beans.

| RefPred | BE10 | CA6 | CAB208 | CAB214 | CAB270 | CAB314 | CAB324 | CAB499 | CCN51 | IMC67 | MA11 | MA15 | MO1 | P7 | PA121 | PA169 | PA195 | RB36 | RB40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BE10 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CA6 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB208 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| CAB214 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| CAB270 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB314 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB324 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAB499 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCN51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| IMC67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MA11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| MA15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| MO1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| P7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| PA121 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| PA169 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| PA195 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| RB36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| RB40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

730

## Highlights

Analysis of 19 cocoa genotypes from the Brazilian germplasm bank;

The visible range is sufficient to discriminate between fermented and unfermented beans, as well as an LDA with two wavelengths in both the visible range (472 nm and 636 nm) and the infrared range (2096 nm and 2278 nm)

Genetic information captured by NIR was more pronounced in unfermented beans.

**Declaration of interests**

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: