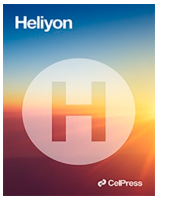
Heliyon 9 (2023) e21920

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Estimating the sensory-associated metabolites profiling of matcha based on PDO attributes as elucidated by NIRS and MS approaches

Yan Chen [a], Xiaoyao Xie [a], Zhirui Wen [a], Yamin Zuo [b,*], Zhiwen Bai [c,**], Qing Wu [a,d,e,***]

[a] Guizhou Key Laboratory of Information and Computing Science, Guizhou Normal University, 116 Baoshan North Rd, Guiyang, Guizhou, 550001, China
[b] School of Basic Medical Sciences, Hubei Key Laboratory of Wudang Local Chinese Medicine Research, Hubei University of Medicine, 30 Renmin South Rd, Shiyan, Hubei, 442000, China
[c] The Guizhou Gui Tea (Group) Co. Ltd., Huaxi District, Guiyang, Guizhou, 550001, China
[d] Guizhou Key Laboratory for Information System of Mountainous Areas and Protection of Ecological Environment, Guizhou Normal University, 116 Baoshan North Rd, Guiyang, Guizhou, 550001, China
[e] Innovation Laboratory, The Third Experiment Middle School in Guiyang, Guiyang, Guizhou, 550001, China

ARTICLE INFO

Keywords:
Metabolites profiling
Sensory analysis
Taste-active compounds
Protected designation of origin matcha
Chemometrics

ABSTRACT

Matcha has been globally valued by consumers for its distinctive fragrance and flavor since ancient times. Currently, the protected designation of origin (PDO) certified matcha, characterized by unique sensory attributes, has garnered renewed interest from consumers and the industry. Given the challenges associated with assessing sensory perceptions, the origin of PDO-certified matcha samples from Guizhou was determined using NIRS and LC-MS platforms. Notably, the accuracy of our established attribute models, based on informative wavelengths selected by the CARS-PLS method, exceeds 0.9 for five sensory attributes, particularly the particle homogeneity attribute (with a validation correlation coefficient of 0.9668). Moreover, an LC-MS method was utilized to analyze non-target matcha metabolites to identify the primary flavor compounds associated with each flavor attribute and to pinpoint the key constituents responsible for variations in grade and flavor intensity. Additionally, high three-way intercorrelations between descriptive sensory attributes, metabolites, and the selected informative wavelengths were observed through network analysis, with correlation coefficients calculated to quantify these relationships. In this research, the integration of matcha chemical composition and sensory panel data was utilized to develop predictive models for assessing the flavor profile of matcha based on its chemical properties.

1. Introduction

Matcha, derived from the plant Camellia sinensis (L.) Kuntze, cultivated under shading conditions, is a powdered green tea that

* Corresponding author.
** Corresponding author.
*** Corresponding author. Guizhou Key Laboratory of Information and Computing Science, Guizhou Normal University, 116 Baoshan North Rd, Guiyang, Guizhou, 550001, China.
E-mail addresses: 20170529@hbmu.edu.cn (Y. Zuo), bzw66@foxmail.com (Z. Bai), wuqing@gznu.edu.cn (Q. Wu).

https://doi.org/10.1016/j.heliyon.2023.e21920
Received 27 May 2023; Received in revised form 31 October 2023; Accepted 31 October 2023
Available online 3 November 2023

exclusively uses tender portions of tea leaves. Renowned for its refreshing flavor, matcha has become a highly sought-after beverage and food ingredient, second only to water in terms of commercialization. Daily matcha consumption surpasses two billion cups, with an estimated annual production of approximately 2.9 million tons. Notably, matcha production accounts for over 65 % of global tea production, underscoring the importance of high-quality products that are traceable to their origin and possess distinctive sensory attributes [1–3]. However, evaluating sensory attributes, including visual appearance (color and particle size) and interior quality (odor, liquor color, and taste) through trained panelists, is often a time-consuming, destructive, and non-representative process. Therefore, it is imperative to develop and validate analytical techniques that can accurately ascertain the origin and gauge the desired sensory quality, addressing these challenges.

The Protected Designation of Origin (PDO) label is widely recognized as an indicator of specific food quality attributes [4]. This label is designated for products that are produced within a defined geographical area, adhere to strict product specifications, and possess unique characteristics tied to their place of origin and traditional expertise. Recent research has centered on the quality and reputation of PDO products. Examples include Parmigiano Reggiano, an Italian extra-hard cheese known for being slowly matured and made from raw milk [5]; Tokaj wines, which always require noble-rotten raisins and stand as one of the most valued wines with a PDO in Europe [6]; extra virgin olive oils from Italy, renowned for their special phenolic and sterolic profiles, which are superior to other edible vegetable oils [7]; and the Rocha pear, a traditional Portuguese cultivar that has achieved PDO status [8].

In the sphere of matcha, a premium powdered tea from Japan that is shade-grown to ensure its high quality, discourse surrounding the Protected Designation of Origin (PDO) label has been scarce. Matcha's quality is gauged by its color, taste, and aroma. Its unique fruity scent and exquisite flavor, often referred to as "coverage," have seen an increased demand in recent decades, suggesting the potential for its inclusion among products with a PDO label. To improve the accuracy of quality assessment differentiation and to bolster consumer confidence in the authenticity of matcha's PDO labels, we advocate for the use of various analytical platform-based tools. Among them, near-infrared spectroscopy (NIRS) is deemed the premier food fingerprinting technique in untargeted approaches. This spectroscopic technique is valued for its speed and is apt for our objectives [9,10]. Existing literature on near-infrared spectroscopy (NIRS) attests to its significant potential for fingerprinting food sensory attributes. For instance, our prior research showcased remarkable results in assessing sensory attributes and characterizing samples through simulated spectroscopic evaluation of green tea [11]. However, research on the application of NIRS for gauging matcha's sensory attributes has been lacking. Hence, this study serves as one of the pioneering efforts to delve into matcha's sensory attributes using NIRS.

The integration of NIRS techniques and chemometric analysis is widely recognized for its myriad advantages in analyzing both the flavor and taste-related compositions of tea [12,13]. In the realm of matcha, the green analytical technique of NIRS has demonstrated its capability to correlate consumer preference or sensory panels with internal qualities, such as physicochemical indicators, and facilitate its classification [14,15]. Moreover, using NIRS to estimate matcha metabolites offers a more cost-effective phenotyping approach compared to metabolomics and boasts high-throughput capabilities. This study endeavors to develop predictive models for matcha sensory attributes by leveraging informative wavelength selection algorithms on NIRS fingerprints. In recent years, Partial least squares regression (PLSR) has garnered considerable validation due to its well-documented efficiency as a widely used factorial multivariate calibration method [16]. Additionally, the variable selection process is facilitated using the competitive adapted reweighted sampling (CARS) algorithm, which seeks to exclude variables with minimal influence on the outcomes and remove wavelengths with diminished weights [17]. By adopting these methodologies, our objective is to craft predictive models for matcha sensory attributes that can serve both industries and consumers.

Additionally, the absorption properties of various secondary structures at different wavelengths can provide crucial information related to specific functional groups and structures of molecules responsible for sensory quality. We further explored the metabolic profiles of various grades of matcha using untargeted liquid chromatography-mass spectrometry (LC-MS) to examine the correlation between secondary metabolites and sensory characteristics. Past studies have underscored the importance of secondary metabolic concentrations in matcha samples as key chemical constituents that profoundly influence overall organoleptic perception [18]. Moreover, shifts in metabolic profiles can assist in distinguishing matcha with a protected designation of origin (PDO). The chemometrics approach, particularly the multivariate calibration method, is also incorporated in the mass spectrometry analysis to effectively differentiate between sample groups. Thus, in this research, through the efficient LC-MS analysis of matcha metabolites, our foremost goal is to identify several metabolic characteristics that act as biomarkers for differentiating matcha samples based on their grade. For this purpose, the multivariate principal component analysis (PCA) and orthogonal partial least squares discriminant analysis (OPLS-DA) are applied using the LC-MS data derived from the matcha samples' fractions. The resulting values will then be used to generate a heatmap, facilitating the differentiation and visualization of grade variations in the matcha samples due to differential metabolites. A subsequent partial least squares (PLS) analysis will be performed to delve deeper into the relationships between matcha flavor intensity and chemical constituents, offering a more detailed analysis of the metabolomic data. Finally, this study aims to uncover the seldom explored interconnections between sensory quality, metabolites, and selected informative wavelengths. To deepen our understanding of this relationship, we will conduct an association study using networks to visually represent the molecular characteristics. This methodology seeks to pinpoint potential metabolites that have a robust correlation with the informative wavelengths and sensory attributes in focus. Moreover, we will evaluate the potential of employing these metabolites to develop predictive models for sensory attributes.

Thus, the objective of this study is to explore the capability of NIRS and MS methodologies in delineating the sensory characteristics of matcha and accurately defining the PDO label. The specific aims of this research are (1) to evaluate the sensory attributes of matcha using NIRS analysis, (2) to distinguish between different grades of matcha using LC-MS analysis, and (3) to establish a correlation between molecular features (metabolites) and informative wavelengths with the sensory attributes using a multivariate statistical analysis approach and network analysis.

2. Materials and methods

2.1. Matcha sample's location and collection

Approximately 210 matcha samples were sourced from Guizhou Gui Tea (Group) Co. Ltd and local supermarkets. The geographical range of the study encompassed the primary Guizhou matcha PDO regions, including Tongren (TR), Zunyi (ZY), Guiyang (GY), Qiandongnan Miao and Dong Autonomous Prefecture (QDN), and Qiannan Buyi and Miao Autonomous Prefecture (QN). All selected matcha samples were situated between 900 and 1600 m above sea level, in accordance with the register for PGI 'matcha'. These regions have an average annual temperature of 9.5 °C, receive approximately 2018.5 h of sunlight yearly, and experience a frost-free season lasting about 150 days.

As depicted in Fig. 1, the matcha samples were categorized into three groups (special, first, and second) based on the optimal harvest maturity window (from April to July) and geographical parameters (Supplementary material, Table S1). These samples were stored at room temperature, approximately 25 °C, until spectroscopic measurement and sensory evaluation took place.

2.2. Chemicals and reagents

All chemicals and solvents used in this study were of analytical grade. Methanol and formic acid (LC/MS or HPLC grade) were procured from Thermo Fisher Scientific (Waltham, MA), while ultrapure water was sourced from a Milli-Q synthesis system (Millipore). Leucine enkephalin (LC/MS grade), utilized for real-time calibration, was obtained from Waters Corp (Milford, MA).

2.3. Sensory evaluation of matcha samples

The sensory evaluation of the matcha samples was conducted using the quantitative descriptive analysis (QDA) method, also known as sensory profiling. The panel consisted of five females and five males aged between 25 and 55 years. All panelists had at least five years of sensory analysis experience and underwent a minimum of four weeks of training in accordance with DB 52/T 1358–2018 (Table 1). The attributes evaluated were (1) visual appearance (crust color, particle homogeneity) and (2) interior quality (odor attributes, liquor color, and taste attributes). Definitions for sample discriminants were clearly established to ensure understanding by the sensory team members. During the evaluation, matcha samples were scored in reference to the varying concentration evaluations of the standard taste. Each attribute was assessed for each sample using an intensity scale. The scoring criteria for the samples were as follows: 0 indicated very weak, 1–2 indicated weak, 3–4 indicated average, 5–6 indicated medium, 7–8 indicated high, and 9–10

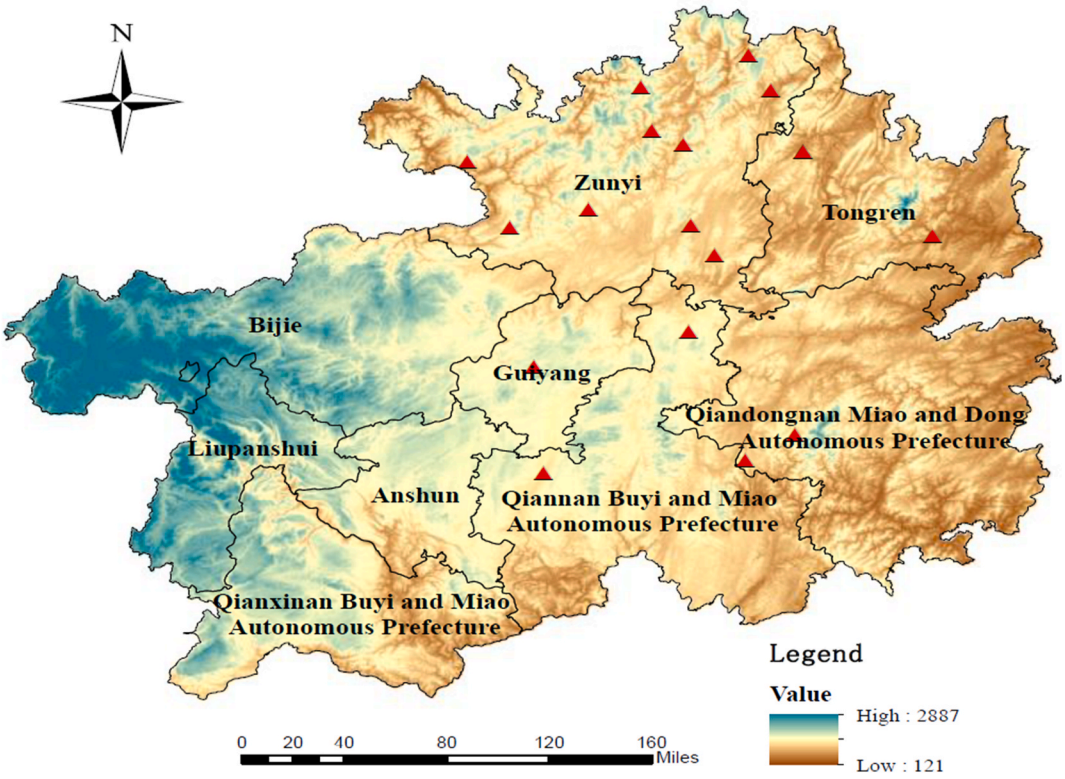


Fig. 1. Geographical distribution of PDO matcha samples from Guizhou.

Table 1
Descriptive sensory parameters selected by the assessors and the profiling of PDO matcha samples.

Parameter	Definition	Scale Criteria		
		Special grade	First grade	Second grade
Visual appearance				
Crust color	Color brightness and intensity of green	Fresh green and brightness	Emerald green and brightness	Green and brightness
Particle homogeneity	Presence or absence with particle of uniform	Soft, delicate and homogeneous	Soft, delicate and homogeneous	Delicate and homogeneous
Interior quality				
Odor attributes	Intensity of the aromatics associated with coverage	High intensity of coverage aroma	Middle intensity of coverage aroma	Low intensity of coverage aroma
Liquor color	Color tone and intensity of green	High intensity and freshness of green	High intensity of green	Low intensity of green
Taste attributes	Intensity and purity of the flavor once the product has been placed in the mouth	High intensity of fresh and mellow flavor	Middle intensity of fresh and mellow flavor	Low intensity of fresh and mellow flavor

indicated extremely high. An overall score for matcha was determined for all samples. The sensory evaluation was conducted in an area conditioned to 25 ± 2 °C with a relative humidity of 50 ± 5 %.

2.4. Acquisition of NIRS of matcha samples

The NIR spectra of the samples were collected in diffuse reflectance mode using an Antaris Nicolet FT-NIR system (Thermo Fisher Scientific Inc., USA). For this study, each reflectance spectrum was obtained by averaging 64 repetitive scans at a single location with a resolution of 16 cm^{-1} and was then transformed into an absorption log (1/R). The spectra were recorded as reference absorbance values in air. The NIR spectra from each matcha group were averaged prior to spectral preprocessing for PLSR analysis.

2.5. Prediction of matcha samples' sensory quality by NIRS

2.5.1. A PLSR-based framework to construct models of sensory attributes

The primary objective of this pilot study was to explore the potential of using PLSR-based models to establish the correlation between the sensory scores and attributes of matcha. Leveraging instrument measurements and sensory attributes, we utilized various regression approaches, including PLSR, to construct predictive models. In the PLSR analysis, the spectral intensity correlates with sample characteristics, linear regression, and conventional multiple regression analysis. Essentially, the goal of PLSR in this study was to determine the linear relationship between the X matrix (representing the independent variable) and the *Y* variables (reference value, also known as the dependent variable) using a regression coefficient matrix and a residual error matrix *E*. To ensure uniformity in the dataset from the 210 matcha samples, we partitioned our dataset into subsets for training and validation based on a random mutation of the Kennard-Stone (KS) uniform sampling algorithm. Furthermore, for the discriminant plot application, we selected 10 samples from Zunyi (ZY) as external test samples to assess their grades. Accordingly, for feature selection, the data samples were randomly divided into two groups: the calibration set (comprising 135 samples) and the validation set (comprising 65 samples). Additionally, another 10 samples were evaluated as an external test set.

To assess the relative importance of each variable in the PLSR model, the information content of each variable was compared to its variable importance in the projection (VIP). In this study, we adopted the criterion that variables with VIP scores $\geq$1 are considered especially important for the model. The data analysis procedure was as follows: we utilized raw data along with various preprocessing techniques such as multiplicative scatter correction (MSC), standard normal variate (SNV), first derivative (D1), and second derivative (D2). Subsequently, we described the model's quality using the following indices: coefficient of determination for prediction (RP^2), root mean square error of prediction (RMSEP), root mean square error of cross-validation (RMSECV), standard error of prediction (SEP), and residual predictive deviation (RPD). Latent variables (LV) were selected after building the model using the previously mentioned cross-validation method. The optimal number of latent variables was determined based on the lowest root mean square error of cross-validation (RMSECV) for each attribute, ensuring the number of LVs did not surpass the number of independent samples in the final model. After several iterations of the process, the wavelength subset yielding the smallest RMSECV in the PLS models was selected.

Within the context of the PLSR analysis, the subset of spectral features derived from optimal wavelengths was utilized in place of the full wavelengths using the CARS-PLS method. The details of the CARS-PLS method are provided in Section 2.5.2.

2.5.2. CARS-PLS method for selecting informative wavelengths

Feature selection is a pivotal step in building predictive models. As such, this study used the CARS variable selection method to identify characteristic wavelengths. CARS, a novel and effective approach, combines adaptive reweighted sampling (ARS) with partial least squares (PLS). In each sampling iteration, CARS follows four sequential steps: (1) The Monte Carlo method is employed for sampling. (2) An exponentially decreasing function (EDF) is used to filter out wavelengths with relatively small absolute regression coefficients. (3) ARS is utilized to select key wavelengths and to further remove wavelengths competitively. (4) Cross-validation is used primarily to evaluate the subset. Further details on CARS will be discussed in the subsequent sections.

As previously mentioned, the selection procedure in CARS consists of two major steps. Suppose we had *N* samplings containing *p* wavelengths performed in CARS. In the first step, an EDF defines the optimal number of selected spectral variables. The ratio of wavelengths computed is defined by the formula described below. Here, *i* represents the sampling run, and parameters *a* and *k* are two constants determined by the following two conditions: (1) in the first sampling run, all *p* wavelengths are taken as input for the model, which implies that r_i = 1; (2) in the *N*th sampling run, only two wavelengths are retained, meaning that r_N = 2/p.

$$r_i = ae^{-ki}$$

$$a = \left(\frac{p}{2}\right)^{\frac{1}{N-1}}$$

$$k = \frac{\ln\left(\frac{p}{2}\right)}{N-1}$$

As observed, the CARS algorithm selected *N* subsets of wavelengths from *N* sampling runs through iterative processes. Ultimately, the feature subset with the lowest RMSECV value was chosen as the optimal subset.

2.6. Determination of matcha metabolites by non-targeted LC-MS

2.6.1. Sample preparation for LC-MS analysis

After conducting NIRS and sensory evaluation, primary metabolites from the samples were extracted as previously described. Briefly, 500 mg of each powdered matcha sample was extracted using 20 mL of a methanol/water mixture (7/3, v/v). Ultrasonication (3500 Hz, 30 min each session) at room temperature was applied twice with a 4-h interval, followed by 8 h of standing. After centrifuging for 10 min at 12,000 g, the supernatants were collected. These supernatants were then transferred to vials for LC-MS analysis. A quality control (QC) sample was prepared by mixing equal volumes of each test sample, creating a "pooled" sample that was used to estimate a "mean" profile representing all analytes encountered during the analysis. Samples were stored at −20 °C until analyzed.

Among the 210 matcha samples, three grades from different geographical cultivars — Tongren (TR), Zunyi (ZY), Guiyang (GY), Qiandongnan Miao and Dong Autonomous Prefecture (QDN), and Qiannan Buyi and Miao Autonomous Prefecture (QN) — were analyzed. The matcha samples were produced by Guizhou Gui Tea (Group) Co. Ltd. These samples were selected because they exhibited significant variations in quality characteristics, had a known cultivation history, and were grown in a controlled environment, ensuring the expected stable quality.

2.6.2. LC-MS analysis

The samples were analyzed using an Ultra Performance Liquid Chromatography (UPLC) system (Ultimate 3000, Dionex, Sunnyvale, CA, USA) coupled with a hybrid mass spectrometer (Q-Exactive Focus, Thermo Fisher Scientific, Waltham, MA, USA).

The compound separation was conducted on a Hypersil ODS2 column (5 μm, 250 mm × 4.6 mm) with a flow rate of 1 mL/min (solvent A, H_2O with 0.1 % formic acid; solvent B, methanol with 0.1 % formic acid). Upon injecting 10 μL of the samples, the ultra-high performance liquid chromatography (UHPLC) system followed an elution pattern with a linear gradient: 0 min at 8 % B; 18 min at 19 % B; 25 min at 23 % B; 45 min at 26 % B; 70 min at 37 % B; 72 min at 70 % B; concluding at 73 min with 8 % B.

In the mass spectrometer, sample analysis was conducted in positive mode using heat electrospray ionization with a spray voltage of 3.0 kV, a sheath gas pressure of 40 arbitrary units, an auxiliary gas pressure of 10 arbitrary units, and a capillary temperature of 350 °C. The full scan MS mode, set to a resolution of 70,000, was utilized for quantitative analysis with a scan range of 100–1500 (m/z). For qualitative analysis, the MS/MS scanning plus targeted parent ions MS2 mode was configured as a data-dependent ms2 (dd-ms2) scan, boasting a resolution of 17,500. Fragmentation was carried out in the high-energy collision dissociation, set in a stepped mode with energies of 18 and 35 eV.

All samples were grouped into batches, and within each batch, the samples were injected randomly. In this study, approximately six metabolites with varying m/z values and polarities were selected for quality control/quality assurance in both ESI+ and ESI- modes. The identified metabolite, retention times, and selected mass-to-charge ratios demonstrated good system stability and repeatability.

2.7. Analysis of the feature intercorrelations between sensory quality, metabolites, and selected informative wavelengths

In this study, we aimed to explore the intercorrelations between sensory attributes, metabolites, and informative wavelengths using network analysis. To this end, a fusion dataset underwent a transformation to discern the relationship among the three data sources. First, several preprocessing methods were employed on the dataset, after which specific metabolites were selected to fit network models. The goal was to ascertain if incorporating feature-selected metabolites could enhance the classification of subjects previously excluded from the validation dataset. We then calculated the variable importance in the projection (VIP) coefficient for each metabolite. Using the VIP coefficients, networks were generated where a node represented either a molecular feature or a sensory parameter, and an edge was drawn if the VIP coefficient exceeded zero. Ultimately, the feature interaction detection framework was utilized to identify distinguishing connectivity patterns of the selected informative wavelengths.

2.8. Statistical analysis

MATLAB (R2020a) (version 9.8, MathWorks Inc., USA) equipped with PLS Toolbox version 8.8.1 (Eigenvector Research Inc., Mason, USA) was utilized for PLSR and PCA analysis. The subsequent correlation network analysis was conducted using Cytoscape software (Cytoscape 3.6.0). Correlation coefficients were determined based on Pearson's correlation.

2.9. Ethics statement

Ethical approval for this study was obtained from the Guizhou Normal University Ethics Board (approval code: GZNU-2021YF2036). All experiments were conducted in adherence to established ethical guidelines, and informed consent was secured from all participants prior to their involvement. This study adhered to all pertinent regulations, ensuring that informed consent was always procured. Participants in the sensory evaluation volunteered willingly, and all sessions were carried out using an additional questionnaire where participants directly recorded their responses.

3. Results and discussion

3.1. Overview of matcha samples' NIR spectra and transformation

The raw and average spectra of the matcha samples, obtained in the wavelength range of 4000–10000 cm^{-1}, are shown in Fig. 2a and b, respectively. For the raw spectrum of pure matcha powder, a moderate degree of variability in intensity was observed both within and between samples.

The main absorbance peaks of interest were observed in the range of 4500–5500 cm^{-1}. This range corresponds to the first overtone regions and the C–H stretch combination. This index has been identified as a critical metric for texture attributes and is used to characterize the heterogeneity of matcha samples. Additionally, the absorption at 5200 cm^{-1} is attributed to the second overtone of the carbonyl group of esters. Another peak around 5680 nm, indicative of chlorophyll content, has been used to monitor the progression of senescence and to evaluate the growth of tea leaves. Moreover, when comparing the spectra of different grade matcha samples, as shown in Fig. 2b, pronounced absorbance peaks were observed at 4620 cm^{-1} (related to N═N), 5100 cm^{-1} (related to C–H and O–H), and 5800 cm^{-1} (also related to N═N) [19–21]. Several other NIRS wavelength regions for matcha were delineated, and their corresponding biological assignments, molecular structures, and vibrational modes were investigated. These findings were instrumental in exploring and explaining the association between sensory scores and attributes such as bitterness and cleanliness.

To develop a robust and reliable model for matcha sensory evaluation, several preprocessing steps were required before extensive modeling. In this study, we explored various pretreatment techniques, including smoothing, normalization, and standard normal variate (SNV). The data obtained from these techniques was then used to construct a partial least squares (PLS) model, allowing for the analysis and assessment of the effects of the different spectral preprocessing methods. The results of this analysis are detailed in Table S2. We determined that the best-fit PLS model was achieved using SNV preprocessing (Model 4), as evidenced by the highest correlation of prediction ($R_P = 0.965$) and the lowest root-mean-square error of prediction (RMSEP = 0.075). Hence, the subsequent chemometric analysis was based on spectra processed with SNV.

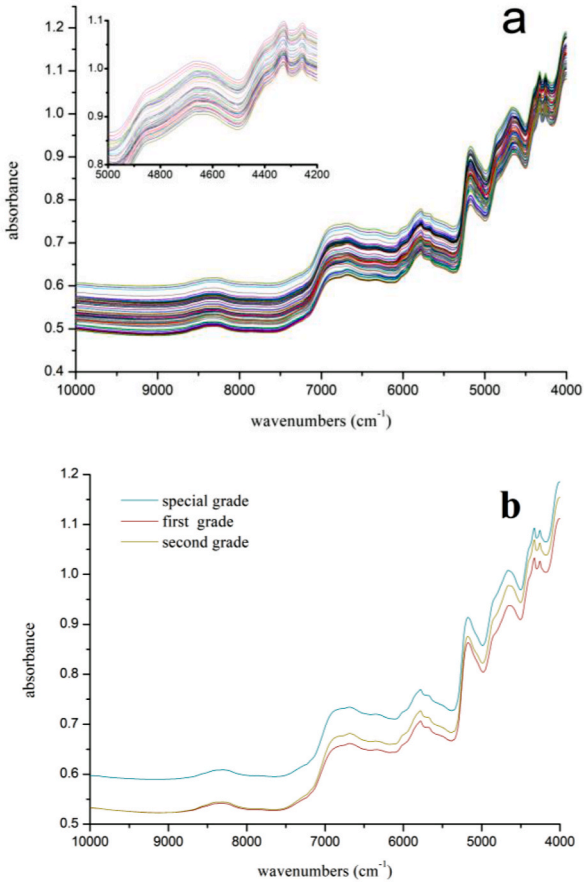


Fig. 2. The NIR spectrum of 210 matcha samples obtained from (a) raw spectra, (b) average spectra of different grades.

3.2. The sensory profile and quantitative estimation of matcha by NIRS

3.2.1. The sensory attribute prediction of matcha through PLS analysis

In this study, various PLS analyses (PLSR, PLS-DA, CARS-PLSR) were employed to investigate the sensory profile and quantitative sensory evaluation results of matcha. The range of sensory scores for each attribute provided a valuable foundation to identify markers of matcha quality and mechanisms for grading (Table S3). Since the sensory score encompassed contributions from visual appearance (crust color, particle homogeneity) and interior quality (odor attributes, liquor color, and taste attributes), we conducted PLS analysis to determine the relative significance of each attribute in predicting sensory quality. For these attributes, various spectral transformation methods were examined, and predictive models were then developed based on CARS-selected wavelengths. The optimal PLSR models were assessed based on RMSECV, RMSEP, RPD, and SEP.

As depicted in Fig. 3, the PLSR was conducted for all sensory attributes using the first two PLSR components (t_1, t_2). This was done to explore potential interrelationships between the sensory score (independent variables, *X*) and their NIRS profiles (dependent variables, *Y*). Specifically, in this study, PLSR was utilized to establish the relationship between the visual appearance of matcha samples (Fig. 3a) and interior quality (Fig. 3b) with their respective NIRS profiles. This resulted in correlation circles with the first two PLSR components. As shown in Fig. 3, the score values of the first component exhibited a strong correlation with both the visual appearance score (R = 0.92) and the interior quality score (R = 0.93). These accounted for 94.7 % and 94.5 % of the variation in the NIR spectra of special-grade, first-grade, and second grade matcha, respectively. These findings underscore a gradual transition in matcha samples from special to second grade. Moreover, these proof-of-concept results also confirm the inherent feasibility of using NIR spectra to assess various matcha samples based on their constituent molecules.

While these correlation circle maps serve as useful indicators for the overall performance of sensory exploration, they have limitations in that they cannot specify which combination of wavelengths has a statistically significant influence on the measurement of different grades. To further address this issue and elucidate the multivariate nature of the wavelengths in relation to grade discrimination, the PLS-DA model was also employed to identify a wavelength signature that could differentiate matcha grades in correlation with the latent variable (LV).

For each PLS-DA analysis, an appropriate latent variable (LV1) and LV2 scores plot were shown in orthogonal rotation (Fig. 4a), and a new LV1 was explored to establish a better separation of wavelengths/Y-variable (Fig. 4b). The LV1-LV2 scores plot demonstrated

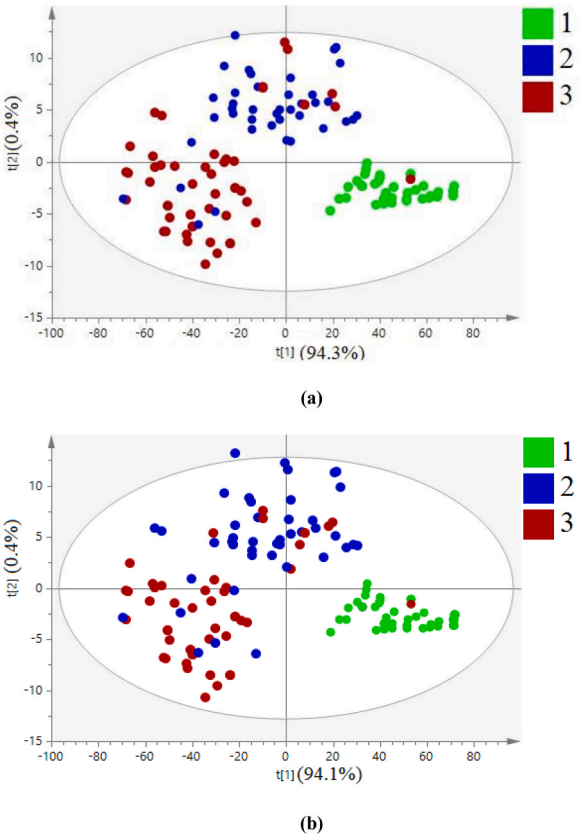


Fig. 3. Partial least squares regression (PLSR) analysis for the potential of sensory prediction for matcha. (a) Correlations between the visual appearance scores and their NIRS profiles, (b) correlations between the interior quality scores and their NIRS profiles.
(t_1, t_2) refer to the first two PLSR components, 1 (green) refer to the special grade, 2 (blue) refer to the first grade, 3 (red) refer to the second grade.

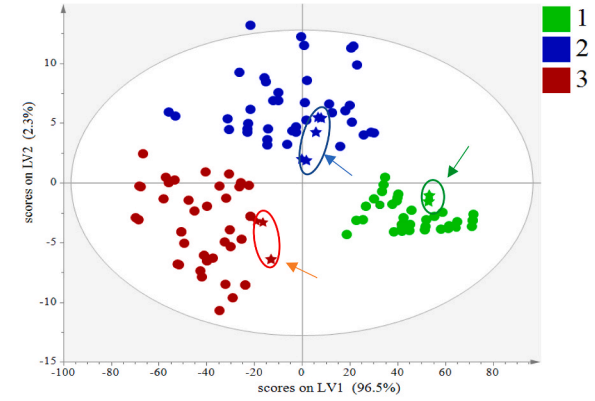
discrimination among the three grades, which exhibited superior performance compared to that obtained by examining the spectral profiles. The results indicated that the newly built PLS-DA model successfully differentiated between different grades of our matcha samples. In this context, the model's accuracy could be expressed as the percentage of correctly classified samples, which, for this PLS-DA model, was 100 %. This outcome was expected, given that external test samples were considered.

The results derived from our analysis of the spectral features of near-infrared (NIR) spectra in the latent variables (LVs) of partial least squares discriminant analysis (PLS-DA) provide significant insights into the differences between matcha grades. The scores obtained for each LV reflect the impact of the LVs on the respective NIR spectra. As such, the LVs play a pivotal role in associating NIR spectra bands with the statistical differences among matcha grades. Collectively, our findings suggest that the variations observed in our limited sample cohort are consistent with those identified in an independently derived dataset with a larger sample size.

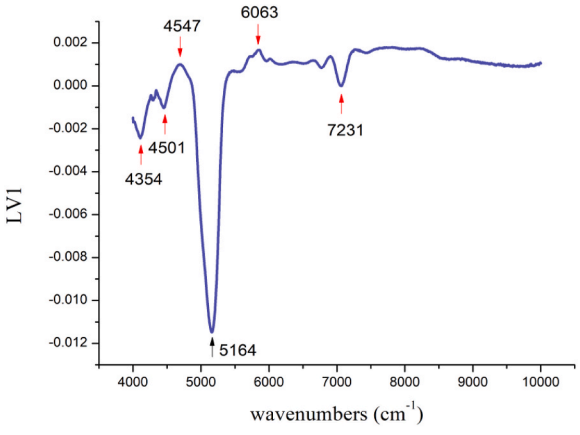
While our PLS-DA analysis successfully identified a matcha signature that could reliably distinguish between special, first, and second-grade samples, we observed significant differences between the sensory scores and wavelengths for our samples. The analysis of the regression coefficients for the wavelengths composing LV1 revealed that the coefficients for 4354, 4501, 4547, 5164, 6063, and 7231 cm^{-1} were significantly different from zero, which were primarily responsible for grade discrimination. Similar to the orthogonal rotation, the PLS-DA modeling pinpointed LVs that differentiated the various matcha grades.

3.2.2. Informative wavelength selection and quantitative model results for sensory information by the CARS-PLSR method

The CARS spectral variable selection method treats each variable as an independent entity and uses adaptive weighted sampling technology to identify the vital spectral regions for each measurement series. As a result, this data-driven algorithm effectively manages the retention rate of variables, showcasing outstanding recognition efficiency that surpasses other similar high-dimensional data approaches.



(a)



(b)

Fig. 4. Partial least squares discrimination analysis (PLS-DA) for the potential of grade prediction for matcha. (a) Correlations between the match grade and their NIRS profiles by the score plot based on the first two LVs; The projected test samples (external validation results) are symbolized by asterisks (b) score plot of LV1 and variation in contribution of wavelengths to LV1 for PLS-DA models.

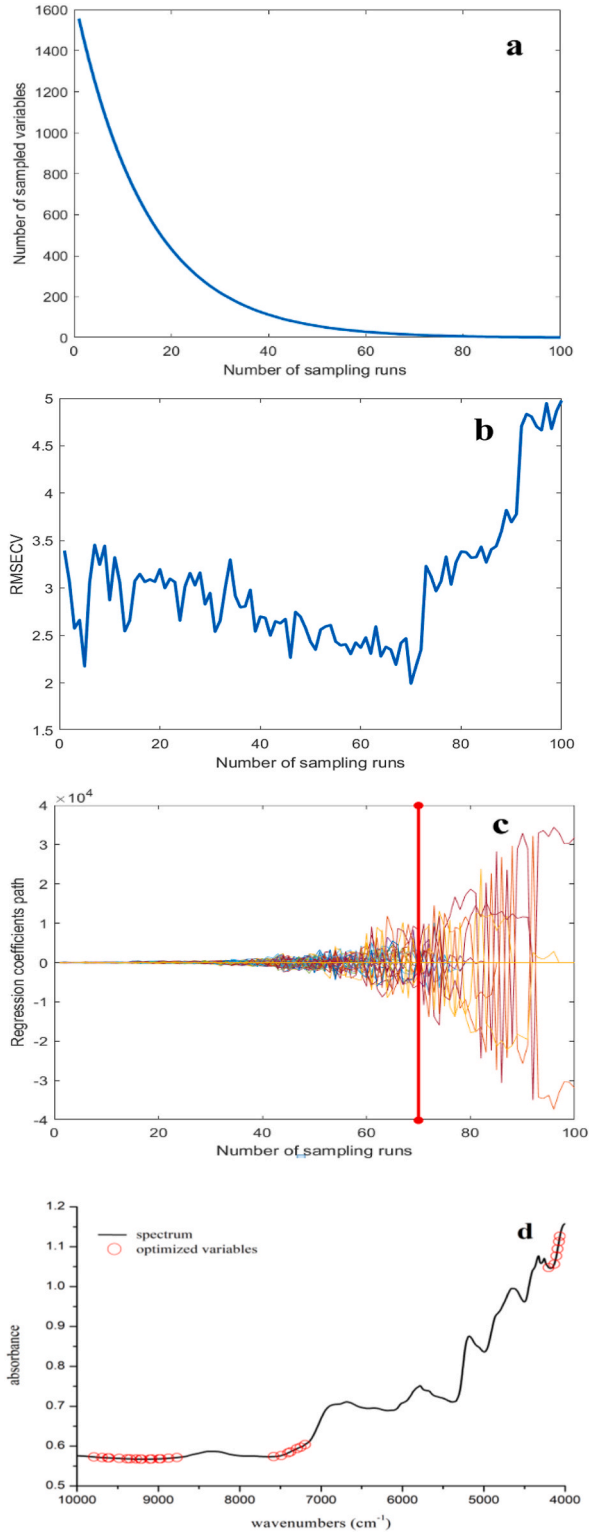


Fig. 5. The processing of effective variables selection for the matcha samples' interior quality by the CARS method (a) number of sampled variables versus number of sampling runs, (b) RMSECV versus number of sampling runs, (c) regression coefficients path versus number of sampling runs, (d) effective wavelength variables selected by CARS shown in circle markers.

The diagrams illustrating the modeling of wavelength optimization variable screening for evaluating the interior quality of matcha samples are presented in Fig. 5. In summary, as the number of sampling operations increased, we observed that the number of selected variables and corresponding wavelengths first decreased rapidly and then stabilized. After 100 iterations using CARS, it became evident that there was a coarse-to-fine feature selection of the wavelength, demonstrating the algorithm's ability to perform an initial broad selection followed by precise fine-tuning (Fig. 5A). Secondly, the results showed that the RMSECV decreased slowly during the first 0–70 sampling runs (Fig. 5B), suggesting that wavelength variables having minimal impact on the sensory evaluation of matcha interior quality were eliminated. However, there was a noticeable uptick after 70 sampling runs. This suggests that critical variables related to the interior quality evaluation of matcha were excluded. Additionally, Fig. 5C depicts the trend of the regression coefficient of each wavelength variable throughout the screening process, showcasing the evolution of the regression coefficient path. The "red line" in the figure corresponds to the moment of the lowest RMSECV value (observed during the 70th sampling run). Finally, as illustrated in Fig. 5D, when the number of samplings reached 70, 35 effective variables were identified for the assessment of matcha's interior quality. The corresponding optimal characteristic wavelengths were pinpointed, which included 4011, 4015, 4022, 4032, 4042, and 4076 cm^{-1}, among others. Using the same methodology, we also explored the wavelength optimization variable screening for the visual appearance evaluation of matcha samples, ultimately identifying 47 variables through CARS (as seen in Fig. S1).

As mentioned earlier, we advocate for the integration of complete spectral data with characteristic variables identified through the CARS variable selection method. This approach is designed to construct comprehensive PLS models for the quantitative prediction of matcha's sensory parameters associated with PDO. The results showcase the optimal scatter plots of the predicted sensory parameters for matcha as represented by the CARS-PLS models, illustrated in Fig. S2. Specifically, regarding the sensory attributes, the optimal models generated by CARS-PLS resulted in $R_P = 0.9572$, RMSEP = 0.666 for crust color; $R_P = 0.9668$, RMSEP = 0.596 for particle homogeneity; $R_P = 0.9555$, RMSEP = 0.691 for odor attributes; $R_P = 0.9542$, RMSEP = 0.626 for liquor color; and $R_P = 0.9219$, RMSEP = 0.921 for taste attributes. The results exhibited good prediction accuracy with an RPD of 1.503 for crust color and approximate prediction accuracy with RPD values of 2.013 for particle homogeneity, 1.855 for odor attributes, and 2.011 for liquor color. However, the prediction accuracy was deemed subpar for taste attributes, with an RPD of 1.055. Additionally, this study allowed for the concurrent exploration of four quality indices of matcha. It is also noteworthy that the findings here indicate a slight decrease in crust color and particle homogeneity compared to findings from previous research [22]. This variance might stem from the uneven particle distribution in the matcha samples examined. Prior research predominantly utilized the diffuse reflection method to gauge the taste attributes of matcha, potentially not capturing the comprehensive characteristics of the tea. To mitigate this, our study leveraged the transmittance spectrum of the complete matcha sample, ensuring a more holistic capture of spectral information to assess its overarching attributes. Consequently, numerous matcha samples in this research were analyzed, underscoring the potential of near-infrared transmittance spectroscopy for gauging multiple quality indices, as well as the overall quality assessment of matcha. This suggests the promising capability of NIR spectroscopy for the sensory monitoring of matcha. Ultimately, the values predicted by PLSR, informed by the CARS method, were computed. The specific wavenumbers pinpointed by CARS are highlighted with black dots on the average spectra of both models (A to E), and the corresponding optimal characteristic wavelengths were identified (Fig. S3).

Finally, the optimal PLS models for estimating all sensory attributes of matcha are presented in Table 2. These models were evaluated based on RMSECV, RMSEP, RPD, and LVs. The sensory attributes of visual appearance and interior quality were largely predicted with precision based on the optimal wavelengths identified by CARS. This is evidenced by high RMSEP values and low RMSECV values, complemented by a limited number of latent variables. The sensory attributes of crust color and particle homogeneity were adequately predicted using the optimal characteristic wavelengths spanning 21 and 29 variables, respectively. Meanwhile, the sensory attributes of odor, liquor color, and taste were effectively predicted using the optimal characteristic wavelengths spanning 17, 16, and 18 variables, respectively. Furthermore, when using the same selected spectral ranges to establish the predictability and performance of models, it's feasible to choose fewer wavelength regions. This approach can decrease measurement time and reduce instrumentation costs.

3.3. The multivariate analysis results of LC-MS-based metabolomics

3.3.1. Identification of multiple metabolic features as biomarkers responsible for the grade discrimination of matcha samples

To investigate the mechanisms behind matcha grade classification, an untargeted metabolomics analysis was undertaken. This aimed to differentiate the metabolite profiles of matcha samples and identify distinct metabolites. The raw spectrometric data were

Table 2
The PLSR analysis results of sensory attributes estimation of matcha by NIRS.

Sensory attribute	Sensory score (mean ± SD)		Spectral variables	LVs	RMSECV	RMSEP	*Rp*	RPD
	Calibration set	Validation set						
Visual appearance	12.1 ± 0.85	12.5 ± 0.76	47	6	0.012	0.527	0.9128	1.025
crust color	6.5 ± 0.81	6.8 ± 0.79	21	5	0.015	0.666	0.9572	1.503
Particle homogeneity	5.4 ± 0.79	5.8 ± 0.81	29	7	0.009	0.596	0.9668	2.013
Interior quality	17.5 ± 0.81	17.9 ± 0.77	35	5	0.025	0.485	0.9566	1.022
odor attributes	5.6 ± 0.72	5.7 ± 0.81	17	4	0.019	0.691	0.9555	1.855
liquor color	5.9 ± 0.68	6.0 ± 0.81	16	6	0.016	0.626	0.9542	2.011
taste attributes	5.8 ± 0.88	5.9 ± 0.86	18	7	0.007	0.921	0.9219	1.055

first converted to mzData (LC-MS/MS) formats using Masshunter (Agilent, US). Subsequently, the data was processed with open-source software MZmine 2.0 for peak finding, peak alignment, and peak normalization across all samples [23,24]. After normalization and standardization, the report data from MZmine 2.0 was imported into SIMCA-P software (version 11.0; Umetrics AB, Umea, Sweden) for further multivariate data analysis. During this process, all datasets were mean-centered and Pareto-scaled prior to statistical analyses.

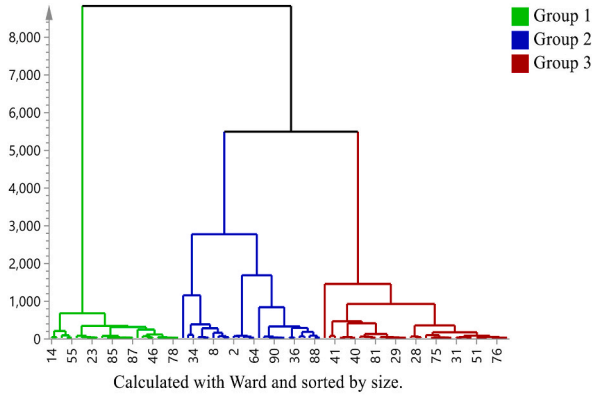
As depicted in Fig. 6, hierarchical cluster analysis (HCA) was employed to categorize the matcha samples into three distinct groups based on their features. In other words, the main component trends of the three different grades appeared similar. Subsequently, we implemented both PCA and OPLS-DA models to further visualize the distinctions among the matcha groups in this study (refer to Fig. S4). The PCA model, which was developed from the LC-MS data to discern intrinsic trends among the different matcha grades without bias, failed to clearly segregate the three distinct matcha groups (as seen in Figs. S4a and S4b). Conversely, the OPLS-DA model exhibited a clear separation between the three matcha groups (as illustrated in Fig. S4c). The parameters of R2X = 0.347 and R2Y = 0.913 indicate the model's high quality and predictive accuracy, providing a clearer differentiation among the samples.

To identify the marker compounds responsible for differentiating matcha samples of varying grades, it's essential to quantitatively evaluate the contribution of various variables to the model. This study used three distinct multivariate analyses, including Variable Importance in the Projection (VIP) values, to gauge the significance of differences in matcha sample grades. Metabolites with a potential differential impact were chosen based on variables with VIP>1, which were influential in distinguishing samples during OPLS-DA analysis. Furthermore, the Kruskal-Wallis test was applied to ascertain whether the differential metabolites identified from OPLS-DA modeling were statistically significant ($p < 0.05$) across groups at the univariate analysis level.

In this study, the tentatively identified primary marker compounds belonged to several metabolic classes, including flavan-3-ols, organic acids, purine alkaloids, oxidation products of flavan-3-ols (OPF), amino acids, fatty acids, carbohydrates, and others, as detailed in Table S4. To identify these marker compounds, we employed three analytical strategies to decipher their chemical structures. Firstly, we utilized authentic compounds that had been previously purchased and isolated by our laboratory. This method is regarded as the most reliable approach for compound identification. Secondly, we analyzed the mass fragment ions (MS/MS ions) cleavage patterns and examined the fragmentation regulation for typical compounds. For instance, specific amino acids such as L-glutamine, L-theanine, and Pyroglutamic acid were distinctly identified in this study. Lastly, the precise molecular weight, calculated from high-resolution ion *m/z*, was cross-referenced with the tea metabolome database, leading to the identification of several compounds. Throughout the identification process, these strategies were synergistically applied to confirm the structure of each marker compound.

The compound with the highest VIP value identified in this study was pelargonidin, which emerged as a vital chemical marker for discriminating matcha sample grades. Its chemical content declined progressively with the grade ranking in our observations. Furthermore, pelargonidin, one of the constituents of the anthocyanin skeleton, influences floral color and is associated with the formation and accumulation of sugars in plant tissues [25,26]. It also exhibits potential health benefits such as anti-mutagenic properties and cancer prevention. Quinic acid, an organic acid, ranked second in terms of VIP value and showed a negative correlation with the grade ranking of matcha samples. This compound has been recognized for bolstering immunity and exhibiting anti-aging effects. It is also worth noting the chemical variations in amino acid compounds across different matcha samples in this study, particularly since a significant decline in amino acid content corresponding to grade differentiation has rarely been reported.

A heatmap was subsequently used to visualize the grade variations of differential metabolites in the matcha samples (Fig. 7). As depicted, a red circle indicated that a metabolite was present at levels greater than the mean in a sample, while a blue circle signified that the metabolite was at a lower level. The findings can be summarized as follows: the flavan-3-ols, organic acids (quinic acid, 3-*p*-Coumaroylquinic acid, and 4-*p*-Coumaroylquinic acid), and fatty acids (oleic acid and linoleic acid) were evidently present at higher levels in special and first-grade matcha compared to the second-grade samples. In contrast, amino acids, such as L-glutamine, L-Theanine, and pyroglutamic acid, were significantly lower in the first-grade and especially in the second-grade matcha samples. Additionally, the phenolic acids (gallic acid), carbohydrates (glucose), purine alkaloids (theobromine and caffeine), and OPF



(a)

Fig. 6. The clustering analysis of LC-MS based metabolomics data of matcha with different grades.

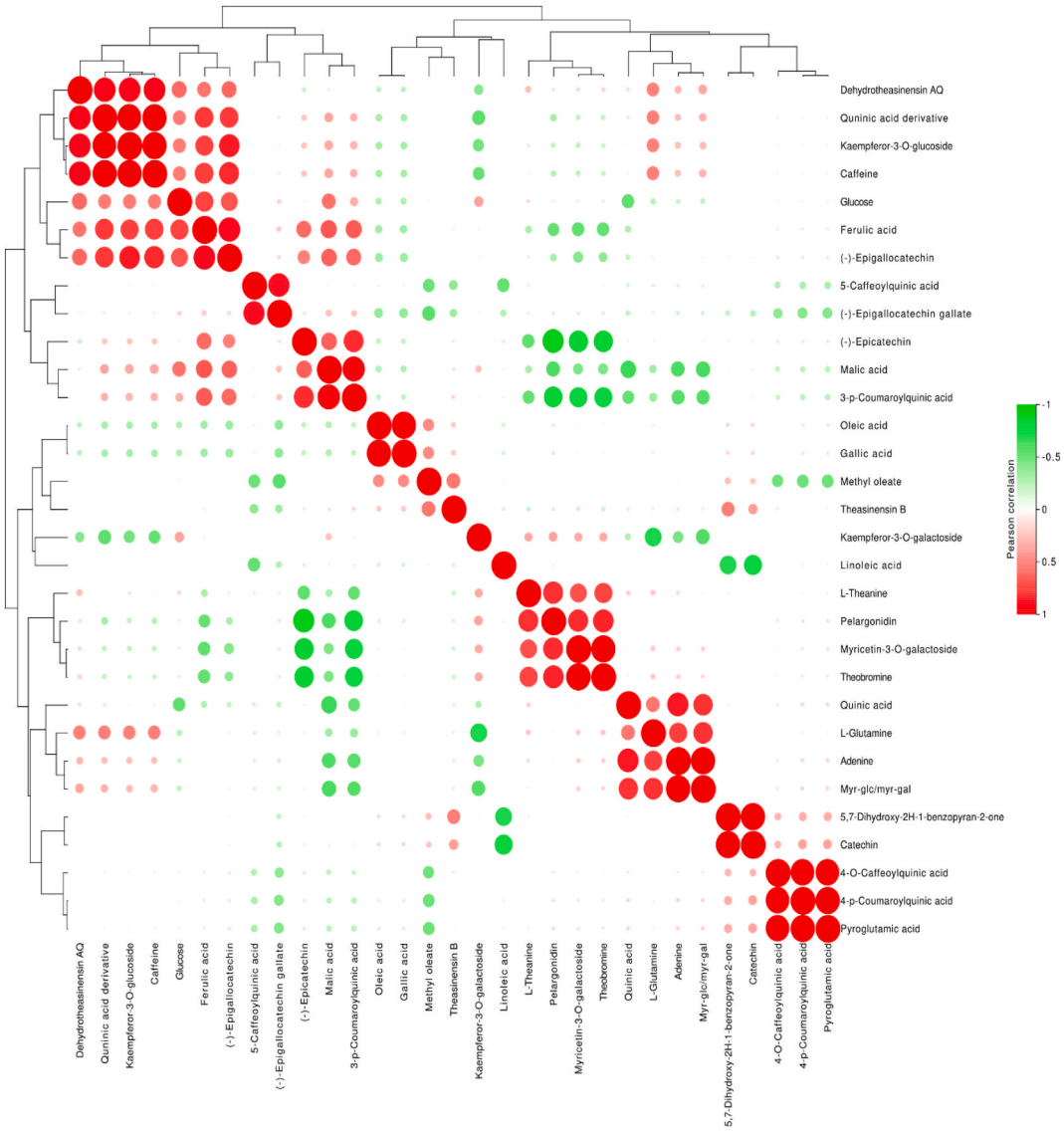


Fig. 7. The heatmap of various critical compounds responsible for the classification of different grades of matcha samples.

(dehydrotheasinensin AQ) were found to be at their lowest levels in the second-grade matcha samples.

3.3.2. Correlations between matcha flavor intensity and chemical constituents

Given that the metabolite differences in the three matcha types were previously analyzed and scored differently based on established quality evaluation criteria, we primarily constructed the Partial Least Squares (PLS) regression models to pinpoint the *m/z* features (biomarkers) closely tied to sensory attributes. For this purpose, all the samples underwent an analysis wherein the mass intensity of the metabolites was set as the X-variable, and the sensory strength of the matcha samples was set as the Y-variable. As depicted in Fig. S5, for clarity, when two variables are situated in the same quadrant, it indicates a positive correlation between them. Moreover, a metabolite positioned farther from the origin point signifies a more significant contribution to taste variations. The illustration revealed that higher contents of pelargonidin, quinic acid, adenine, L-Glutamine, L-Theanine, (−)-Epicatechin, malic acid, 3-*p*-Coumaroylquinic acid, glucose, ferulic acid, caffeine, 5-Caffeoylquinic acid, and (−)-Epigallocatechin gallate correlated strongly with the more pronounced coverage aroma found in special matcha samples. On the one hand, pelargonidin, L-Theanine, kaempferol-3-*O*-glucoside, myricetin-3-*O*-galactoside, (−)-Epicatechin, malic acid, 3-*p*-Coumaroylquinic acid, dehydrotheasinensin AQ, and glucose exhibited positive correlations with fresh and mellow sensory attributes. Conversely, quinic acid, adenine, myr-glc/myr-gal, and L-Glutamine showed negative correlations with the fresh and mellow attributes. These correlations align well with the metabolite content comparisons observed in matcha samples of different grades.

Furthermore, the expression intensity of each marker was not statistically evaluated among the different matcha grades. The

Pearson correlation coefficient between the sensory intensity and the markers is listed in Table 3. The results revealed that the chemical compound pelargonidin was a distinct biomarker for the sensory attribute of "green" in matcha samples. Notably, it had the highest correlation coefficient with green intensity, registering at 0.952 in this study. Additionally, compounds such as quinic acid, adenine, L-Glutamine, and (−)-Epicatechin were identified as typical coverage compounds in matcha, with intensity values of 0.941, 0.851, 0.906, and 0.902, respectively. These could prove pivotal in deciphering the mechanisms behind sensory perceptions. Moreover, past research has highlighted compounds like L-Theanine, (−)-Epigallocatechin gallate, and theobromine as contributing to the fresh and mellow attributes of green tea. Interestingly, these compounds have also been identified as primary sensory contributors to bitter perceptions. In this study, it was intriguing to find that while L-Theanine was uniquely associated with the mellow attribute of matcha, registering an intensity of 0.899, other organic acids and OPFs were also linked to the mellow taste.

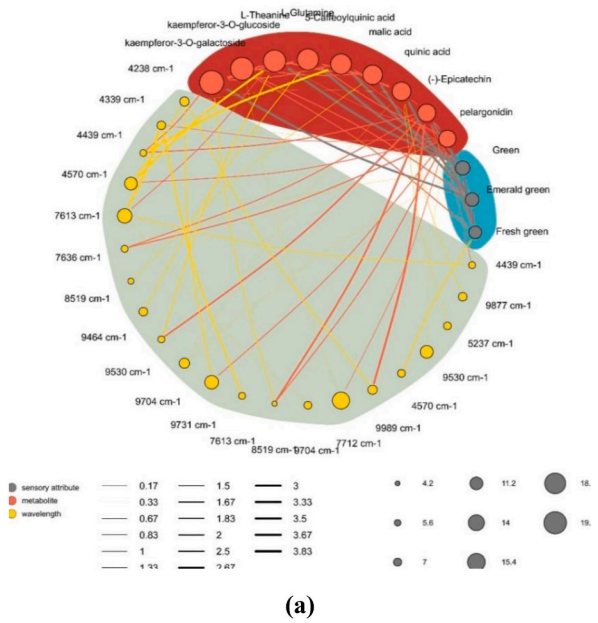
3.4. Intercorrelations between sensory-associated traits, metabolites, and selected informative wavelengths

In this study, we successfully identified a total of 32 molecular features with a VIP score exceeding 1.0, highlighting their significant contribution to the variation observed among different matcha samples. To gain a more comprehensive understanding and visual representation of the molecular features that significantly influence the specific sensory attributes of matcha samples, we conducted an association study using network visualization (Fig. 8). Within the depicted network, correlation coefficients between sensory attributes, metabolites, and informative wavelengths were calculated. Node size represents the degree of each feature, and edge width indicates the corresponding weight in the model. Furthermore, sensory attributes and molecular features were depicted as nodes of varying sizes (large and small) and were color-coded based on different compound categories. For clarity, an edge represented an association between different nodes if the corresponding VIP coefficient exceeded 1.0. The strength of the association was depicted by the thickness and color of the edge; darker and thicker lines signified stronger associations. Overall, the network analysis underscored connections both between and within the various sensory attributes.

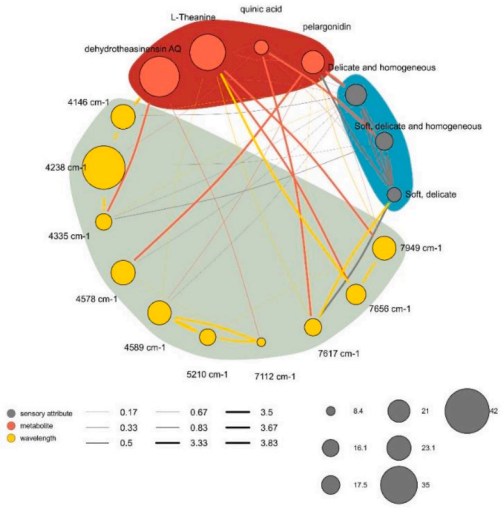
Fig. 8a and b displays the network for the PDO of matcha samples in visual appearance, which illuminates the correlations among sensory attributes (green and particle homogeneity), molecular features, and informative wavelengths. The selected informative wavelengths were positively correlated with the "green" attribute, exhibiting very high correlation coefficients ($|r| \geq 0.8$). Furthermore, the metabolites—including Flavan-3-ols (such as pelargonidin and (−)-Epicatechin), organic acids (like quinic acid, malic acid, and 5-Caffeoylquinic acid), amino acids (L-Glutamine and L-Theanine), and others (kaempferor-3-*O*-glucoside and kaempferor-3-*O*-galactoside)—were closely linked to the green attribute. These metabolites showed positive correlation coefficients with wavelengths such as 4339, 4570, 7636, 7712, 8519, 9530, 9704, and 9989 cm-1. To understand the relationship between these metabolites and the

Table 3
The Pearson correlation coefficient of marker compounds and main sensory-associated attributes' intensity.

Marker compounds	Pearson correlation coefficient			
	green	coverage	fresh	mellow
Pelargonidin	0.952	0.905	0.859	0.625
Quinic acid	0.815	0.941	−0.855	−0.896
Adenine	−0.824	0.851	−0.915	−0.854
Myr-glc/myr-gal	−0.758	0.685	−0.855	−0.895
L-Glutamine	0.855	0.906	−0.916	−0.911
L-Theanine	0.913	0.865	0.867	0.899
(−)-Epicatechin	0.859	0.902	−0.911	−0.896
Malic acid	0.845	0.762	−0.658	−0.811
3-*p*-Coumaroylquinic acid	0.758	0.633	−0.558	−0.613
Dehydrotheasinensin AQ	0.655	0.521	−0.752	−0.558
Glucose	0.512	0.869	−0.788	−0.558
Kaempferor-3-*O*-galactoside	0.426	−0.522	−0.469	−0.684
Kaempferor-3-*O*-glucoside	0.318	0.415	0.698	0.588
Ferulic acid	−0.355	0.652	−0.458	−0.678
(−)-Epigallocatechin	−0.314	−0.548	−0.598	−0.699
Quninic acid derivative	−0.185	0.485	−0.499	−0.588
Myricetin-3-*O*-galactoside	0.158	−0.396	0.457	0.599
Theobromine	0.166	−0.259	0.522	0.369
Caffeine	0.287	0.696	−0.697	−0.598
5-Caffeoylquinic acid	0.652	0.752	0.451	0.702
(−)-Epigallocatechin gallate	0.547	0.865	0.655	0.289
4-*O*-Caffeoylquinic acid	0.258	−0.825	−0.302	−0.500
4-*p*-Coumaroylquinic acid	0.287	−0.855	−0.322	−0.760
Pyroglutamic acid	0.185	0.352	−0.548	−0.688
Linoleic acid	−0.154	0.389	−0.369	0.488
Oleic acid	−0.226	0.445	0.302	0.445
Methyl oleate	0.285	−0.295	−0.588	−0.311
Theasinensin B	0.168	−0.285	−0.424	−0.320
Gallic acid	−0.287	−0.365	0.298	0.487
5,7-Dihydroxy-2H-1-benzopyran-2-one	−0.514	−0.158	−0.633	−0.450
Catechin	−0.435	0.487	−0.560	−0.621



(a)



(b)

Fig. 8. Visualization of sensory attribute, molecular feature and informative wavelengths correlations in matcha samples. Network considering visual appearance (a, green; b, particle homogeneity). Notes are as follows: sensory attributes are represented as gray circles, while molecular features are represented as red circles and wavelengths are color coded by yellow. The edge thickness indicates the strength of the association; darker and thicker lines mean stronger associations.

informative wavelengths, we observed that the selected wavelengths correspond to functional groups like carbon chains—including methyl groups (–CH3), methylene groups (–CH2), and CH═CH–CH2–CH2 fractions—or equivalent H atoms. Most of these wavelengths also relate to common functional groups, such as carboxyl, amino, hydroxyl, and aldehyde groups.

Similarly, concerning particle homogeneity in matcha, metabolites like Flavan-3-ols (pelargonidin), organic acid (quinic acid), amino acids (L-Theanine), and OPF (dehydrotheasinensin AQ) closely related to the particle homogeneity attribute, and displayed positive correlation coefficients with wavelengths like 4335, 4578, 7617, 7656, and 7949 cm-1.

Figs. S6a and S6b depict the network for the PDO of matcha samples concerning interior quality. They visualize correlations among sensory attributes (coverage, fresh, and mellow), molecular features, and informative wavelengths. The selected informative wavelengths exhibited a positive correlation with coverage, fresh, and mellow attributes, boasting very high correlation coefficients ($|r| \geq 0.8$). Additionally, metabolites—including Flavan-3-ols (such as pelargonidin and (−)-Epigallocatechin gallate), organic acids (like quinic acid, Malic acid, and 3-*p*-Coumaroylquinic acid), amino acids (L-Theanine and L-Glutamine), and carbohydrate (glucose)—were

strongly linked to the coverage aroma attribute. These metabolites showed positive correlation coefficients with wavelengths like 4513, 8199, and 8415 cm^{-1}, among others.

Similarly, regarding the fresh and mellow attributes in matcha, metabolites—including Flavan-3-ols (pelargonidin and (−)-Epigallocatechin gallate), organic acid (quinic acid), amino acids (L-Theanine), and purine alkaloids (theobromine)—were closely linked to the fresh and mellow attributes. These displayed positive correlation coefficients with wavelengths such as 4331, 4531, 8415, 8489, and 9372 cm-1.

Interestingly, we observed that the network associated with the matcha samples' aroma attributes (Fig. S6a) exhibited significantly greater connectivity compared to that for taste attributes (Fig. S6b). This suggests that the aroma-related attributes in matcha might embody a higher level of complexity compared to those associated with taste. Moreover, the comparative network analysis for matcha samples in this study hinted that molecular chemical features linked to taste attributes were largely modular in nature. In contrast, features connected with matcha samples' aroma attributes demonstrated more intricate and interwoven relationships between specific features.

Further, as depicted in Fig. S6a, the aroma-associated attributes network analysis unveiled three intriguing clusters. Within the network concerning taste-associated attributes, several specific wavelength features exhibited strong associations with these taste-related attributes. Likewise, associations were discernible between fresh and mellow-related attributes. This connection between fresh and mellow attributes was mediated through three *m/z* features, which were annotated as pelargonidin, L-Theanine, and kaempferor-3-*O*-glucoside. These observations suggest that there are correlations among matcha samples' sensory-associated traits, metabolites, and selected informative wavelengths. However, additional studies are warranted to validate these correlations further.

4. Conclusion

In this study, the efficacy of NIRS and LC-MS techniques in analyzing sensory-associated metabolite profiles in matcha based on PDO attributes was confirmed. The results suggest that these techniques complement each other. To the best of our knowledge, this is the first study to comprehensively examine the quality of matcha samples in terms of descriptive sensory attributes. Moreover, it benchmarks the NIRS system against non-targeted LC-MS metabolites.

This preliminary work paves the way for the exploration of secondary metabolites, molecular markers related to sensory attributes, and responses to NIRS informative wavelengths, as well as bioactive compounds. Firstly, the results unequivocally demonstrate that the NIRS technique, when combined with PLSR and CARS, serves as a rapid and cost-effective tool to probe both the visual and interior PDO attributes of matcha, such as the green, fresh, coverage, and mellow attributes. Through the development of predictive models for flavor perception, we have enhanced the throughput of flavor phenotyping, thereby equipping the matcha industry with novel tools for making more informed and flavorful selections.

Secondly, we determined that the putative biomarkers differentiating matcha samples' grades arose from variations in their abundance. These observations suggest that certain chemical compounds possess significant properties that could be harnessed for their antimicrobial, insecticidal, allelochemical, antioxidant, tumor inhibitory, and proangiogenic, among other functions. The primary components identified were Flavan-3-ols, Organic acid, Purine alkaloids, Glycosylated flavonols, Amino acids, and others. The results pinpointed key metabolites in matcha samples responsible for grade differentiation, namely Pelargonidin, Quinic acid, Adenine, Myr-glc/myr-gal, and L-Glutamine. All the OPLS-DA models demonstrated clear separations, strong interpretability, and reliable predictability across the diverse groups. Moreover, the Pearson correlation coefficient between the sensory intensity and the markers emphasized that chemical compounds such as quinic acid, adenine, L-Glutamine, and (−)-Epicatechin stood out as characteristic compounds impacting the coverage aroma in matcha, offering insights into the underlying mechanisms of sensory perception. These identified potential flavor enhancers and suppressors could prove invaluable as natural food additives in the matcha or broader food industry.

Finally, the metabolite-attribute networks were designed to pinpoint the flavor compounds and wavelength bands most influential to each flavor attribute. These analyses confirmed that the aroma-associated attributes in matcha exhibit greater complexity compared to the taste-associated attributes. A weighted correlation network analysis, based on metabolite concentrations across all matcha samples, was executed. These findings align well with our understanding of individual biosynthetic pathways and shed light on the interrelationships between these pathways. For instance, "green," representing the visual appearance of the labeled PDO matcha, exhibited a significantly stronger association than "particle homogeneity." Additionally, quinic acid demonstrated notable significance, positioning itself as a potential biomarker for matcha grade. Moreover, the wavelength bands at 9981, 9985, 9738, 9464, and 7636 cm^{-1} could be pivotal in understanding the mechanisms underlying PDO matcha.

Overall, this study offers the first comprehensive evaluation of the NIRS and MS systems as alternatives to traditional sensory analyses. This shift could streamline future laboratory quality identification processes and lay the foundation for advancements in market strategies.

In essence, this study underscores the feasibility of deciphering matcha samples' PDO attributes using the NIRS and MS tools. However, in terms of research methods, our study utilized laboratory-bench-type devices, which could potentially be enhanced by employing hand-held NIR spectrometers. Furthermore, there is a need for subsequent research to enhance the stability and accuracy of the established models, paving the way for a broader application of NIR in matcha analysis.

CRediT authorship contribution statement

Yan Chen: Formal analysis, Data curation, Conceptualization. **Xiaoyao Xie:** Validation, Visualization. **Zhirui Wen:** Project

administration, Methodology, Investigation. **Yamin Zuo:** Writing – review & editing, Writing – original draft. **Zhiwen Bai:** Software, Resources, Project administration. **Qing Wu:** Writing – review & editing, Supervision, Funding acquisition.

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.

Acknowledgements

The authors are thankful for the financial support by the NSFC-Guizhou Joint Fund for Karst Science Research Centers (U1812401), Collaborative Innovation Center of biology and Information Technology in karst area of Guizhou Province([2022]010) , the Science and Technology Support Plan of Guizhou Provincial ([2019]2778, [2022]011), the Science and Technology Support Plan of Guiyang ([2022]3–11), the Natural Science Foundation of Education Department of Guizhou Province ([2018]014, [2020]070), and the Scientific and Technological Project of Shiyan City of Hubei Province (22Y24), and they also thank the Guizhou Gui Tea (Group) Co. Ltd (Guizhou, China) for providing matcha and giving permission to publish this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e21920.

References

[1] Y. Baba, T. Kaneko, T. Takihara, Matcha consumption maintains attentional function following a mild acute psychological stress without affecting a feeling of fatigue: a randomized placebo-controlled study in young adults, Nutr. Res. 88 (2021 Apr) 44–52.
[2] F.M. Rezaeian, B.F. Zimmermann, Simplified analysis of flavanols in matcha tea, Food Chem. 373 (Pt B) (2022 Mar 30), 131628.
[3] J. Wu, Q. Ouyang, B. Park, R. Kang, Z. Wang, L. Wang, Q. Chen, Physicochemical indicators coupled with multivariate analysis for comprehensive evaluation of matcha sensory quality, Food Chem. 371 (2022 Mar 1), 131100.
[4] A.M. Machado, M.G. Miguel, M. Vilas-Boas, A.C. Figueiredo, Honey volatiles as a fingerprint for botanical origin-A review on their occurrence on monofloral honeys, Molecules 25 (2) (2020 Jan 16) 374.
[5] S. Buhler, Y. Riciputi, G. Perretti, M.F. Caboni, A. Dossena, S. Sforza, T. Tedeschi, Characterization of defatted products obtained from the parmigiano-reggiano manufacturing chain: determination of peptides and amino acids content and study of the digestibility and bioactive properties, Foods 9 (3) (2020 Mar 9) 310.
[6] O. Vyviurska, N. Koljančić, H.A. Thai, R. Gorovenko, I. Špánik, Classification of botrytized wines based on producing technology using flow-modulated comprehensive two-dimensional gas chromatography, Foods 10 (4) (2021 Apr 16) 876.
[7] C. Ingallina, A. Cerreto, L. Mannina, S. Circi, S. Vista, D. Capitani, M. Spano, A.P. Sobolev, F. Marini, Extra-virgin olive oils from nine Italian regions: an ^{1}H NMR-chemometric characterization, Metabolites 9 (4) (2019 Apr 3) 65.
[8] Y. Wang, X. Zhang, R. Wang, Y. Bai, C. Liu, Y. Yuan, Y. Yang, S. Yang, Differential gene expression analysis of 'Chili' (Pyrus bretschneideri) fruit pericarp with two types of bagging treatments, Hortic. Res. 4 (2017 Mar 8), 17005.
[9] C. Pasquini, Near infrared spectroscopy: fundamentals, practical aspects and analytical applications, J. Braz. Chem. Soc. 14 (2003) 198–219.
[10] M. Schwanninger, J.C. Rodrigues, K. Fackler, A review of band assignments in near infrared spectra of wood and wood components, J. Near Infrared Spectrosc. 19 (2011) 287–308.
[11] Y. Zuo, G. Tan, D. Xiang, L. Chen, J. Wang, S. Zhang, Z. Bai, Q. Wu, Development of a novel green tea quality roadmap and the complex sensory-associated characteristics exploration using rapid near-infrared spectroscopy technology, Spectrochim. Acta Mol. Biomol. Spectrosc. 258 (2021 Sep 5), 119847.
[12] W. Zhang, L.C. Kasun, Q.J. Wang, Y. Zheng, Z. Lin, A review of machine learning for near-infrared spectroscopy, Sensors 22 (24) (2022 Dec 13) 9764.
[13] J. Zeng, Y. Guo, Y. Han, Z. Li, Z. Yang, Q. Chai, W. Wang, Y. Zhang, C. Fu, A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition, Molecules 26 (3) (2021 Feb 1) 749.
[14] Z. Guo, A.O. Barimah, L. Yin, Q. Chen, J. Shi, H.R. El-Seedi, X. Zou, Intelligent evaluation of taste constituents and polyphenols-to-amino acids ratio in matcha tea powder using near infrared spectroscopy, Food Chem. 353 (2021 Aug 15), 129372.
[15] Q. Ouyang, Y. Rong, J. Wu, Z. Wang, H. Lin, Q. Chen, Application of colorimetric sensor array combined with visible near-infrared spectroscopy for the matcha classification, Food Chem. 420 (2023 Sep 15), 136078.
[16] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.
[17] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.
[18] C. Musial, A. Kuban-Jankowska, M. Gorska-Ponikowska, Beneficial properties of green tea catechins, Int. J. Mol. Sci. 21 (5) (2020 Mar 4) 1744.
[19] M. Santos-Rivera, A. Woolums, M. Thoresen, E. Blair, V. Jefferson, F. Meyer, C.K. Vance, Profiling Mannheimia haemolytica infection in dairy calves using near infrared spectroscopy (NIRS) and multivariate analysis (MVA), Sci. Rep. 11 (1) (2021 Jan 14) 1392.
[20] H. Sun, B. Marelli, Polypeptide templating for designer hierarchical materials, Nat. Commun. 11 (1) (2020 Jan 17) 351.
[21] H. Landari, M. Roudjane, Y. Messaddeq, A. Miled, Pseudo-continuous flow FTIR system for glucose, fructose and sucrose identification in mid-IR range, Micromachines 9 (10) (2018 Oct 13) 517.
[22] C. Musial, A. Kuban-Jankowska, M. Gorska-Ponikowska, Beneficial properties of green tea catechins, Int. J. Mol. Sci. 21 (5) (2020 Mar 4) 1744.
[23] X. Li, M. Tsuta, F. Hayakawa, Y. Nakano, Y. Kazami, A. Ikehata, Estimating the sensory qualities of tomatoes using visible and near-infrared spectroscopy and interpretation based on gas chromatography-mass spectrometry metabolomics, Food Chem. 343 (2021 May 1), 128470.
[24] S. Xu, J.J. Wang, Y. Wei, W.W. Deng, X. Wan, G.H. Bao, Z. Xie, T.J. Ling, J. Ning, Metabolomics based on UHPLC-orbitrap-MS and global natural product social molecular networking reveals effects of time scale and environment of storage on the metabolites and taste quality of raw Pu-erh tea, J. Agric. Food Chem. 67 (43) (2019 Oct 30) 12084–12093.
[25] M.S. Ursu, I. Aprodu, Ş.A. Milea, E. Enachi, G. Râpeanu, G.E. Bahrim, N. Stănciuc, Thermal degradation kinetics of anthocyanins extracted from purple maize flour extract and the effect of heating on selected biological functionality, Foods 9 (11) (2020 Nov 3) 1593.
[26] J. Sun, Z. Mei, Y. Tang, L. Ding, G. Jiang, C. Zhang, A. Sun, W. Bai, Stability, antioxidant capacity and degradation kinetics of pelargonidin-3-glucoside exposed to ultrasound power at low temperature, Molecules 21 (9) (2016 Aug 24) 1109.