Current Research in Food Science 8 (2024) 100692

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Classification and characterization on sorghums based on HS-GC-IMS combined with OPLS-DA and GA-PLS

Mengjie Liu [a], Yang Yang [a], Xiaobo Zhao [a,b,*], Yao Wang [a], Meiyin Li [a], Yu Wang [a], Min Tian [a], Jun Zhou [a,b,**]

[a] Luzhou Laojiao Co. Ltd., Luzhou, 646000, China
[b] National Engineering Research Center of Solid-State Brewing, Luzhou, 646000, China

ARTICLE INFO

Handling Editor: Dr. Xing Chen

Keywords:
Sorghum
Volatile compound
Orthogonal partial least squares-discriminant analysis
Classification
Adulteration

ABSTRACT

Headspace gas chromatography-ion mobility spectrometry (HS-GC-IMS) detected 206 and 186 samples of fresh and stored sorghums respectively with three major types in Baijiu industry. The fingerprints showed the differences of volatile compounds among fresh sorghum types by qualitative analysis and artificial recognition. Organic waxy sorghums had more contents of nonanal and 2-ethyl-1-hexanol but fewer ketones. The contents of acetoin in non-glutinous sorghums and organic non-glutinous sorghums were high. On the other hand, genetic algorithm-partial least squares (GA-PLS) selected 19 and 32 characteristic volatile compounds in fresh and stored sorghums. After centering and auto scaling to unit variance, the classification models with three major types of organic waxy sorghum, non-glutinous sorghum and organic non-glutinous sorghum were established based on orthogonal partial least squares-discriminant analysis (OPLS-DA). The goodness-of-fit (R^2Y) and the goodness-of-prediction in cross-validation (Q^2) in the model of fresh sorghum types all exceeded 0.9, in stored were over 0.8, the correct classification rates of external prediction were 95 % and 100 %, which revealed good performance and prediction. On this basis, the correct classification rates reached 87 % in organic waxy sorghums adulterated over 10 % ratio. GC-IMS combined with chemometrics is applicable in practical production for rapid identification of sorghum types and adulterations.

1. Introduction

Sorghum is the fifth largest grain in the world and the most widely used raw grain in the production of high-quality Baijiu (Wu et al., 2017). It has been the first choice for brewing down the ages because of high starch content, low fat and moderate protein. Different sorghum cultivars also affect the quality of Baijiu to a large extent. First, waxy sorghum contained 90 % amylopectin with loose structure is superior to non-glutinous sorghum in Baijiu yield and quality due to strong water absorption, easy gelatinization and many enzyme sites (Zhou et al., 2008). Secondly, different contents of protein, fat and tannin in raw grain affect the flavor of Baijiu. Organic waxy sorghum has become a superior raw grain for brewing because it owns organic identification besides the feature of waxy sorghum, it is a safer and healthier raw material without synthetic fertilizers, pesticides and so on. The most striking characteristics are crimson color and small particle, however they are sometimes hard to distinguish from others through visual sense owing to some uncontrollable factors. In particularly, it shows recognizable color incompletely during harvest due to the differences in harvest time, storage period and maturity. This may cause adulteration phenomenon that will disturb the market and affect consumer trust in the quality of organic Baijiu products. In addition, sorghum stored in the warehouse also faces the same situation after harvest. Therefore, monitoring and detecting sorghum type and adulteration becomes a crucial problem for Baijiu enterprise to ensure quality and flavor.

The studies of sorghum classification are mainly in two directions, detecting traditionally physical and chemical indexes such as starch and tannin content (Boudries et al., 2009; Okoh et al., 1982) are destructive and time-consuming, with the development of non-destructive technique, Fan et al. (2021) characterized the volatile compounds of six Australian sorghum cultivars by gas chromatography-ion mobility spectrometry (GC-IMS) and showed the difference from gallery plot. The

* Corresponding author. National Engineering Research Center of Solid-State Brewing, Luzhou, Sichuan, China.
** Corresponding author. National Engineering Research Center of Solid-State Brewing, Luzhou, Sichuan, China.
E-mail addresses: zhaoxb3@lzlj.com (X. Zhao), zhouj@lzlj.com (J. Zhou).

https://doi.org/10.1016/j.crfs.2024.100692
Received 14 November 2023; Received in revised form 14 January 2024; Accepted 30 January 2024
Available online 30 January 2024

emerging discrimination method based on spectroscopy and chemometrics received satisfactory results. For sorghum discrimination, near-infrared spectroscopy (NIRS), machine vision and hyperspectral imaging (HSI) are greatly improving in practical application. Guindo et al. (2016) predicted pericarp thickness of sorghum by NIRS and partial least squares-discriminant analysis (PLS-DA). Ma et al. (2022) detected ten brewing-sorghum cultivars at single kernel sample level through machine vision system and the anti-aliased convolutional network. However, many sorghum cultivars have similar color and shape, visual recognition is limited to evaluate samples' appearance feature. Hence Bai et al. (2020) and Huang et al. (2022) applied HSI that combined internal spectrum with image technology to discriminate sorghum cultivars and adulterations, the model established by deep forest and PLS-DA separately. These researches focused on sorghum cultivars and separated into several models that each of them classified into two or three cultivars, which were hardly feasible to practical application. In Baijiu industry there are hundreds of sorghum cultivars, whether sorghum is organic or waxy can vary widely in purchasing price. The price of organic waxy sorghum is the highest, followed by organic non-glutinous sorghum and non-glutinous sorghum. Multiple cultivars are mixed into batches and categorized as a major type in order to facilitate transport and deal, which causes great challenges in sorghum classification and quality monitoring, such as cultivars with the feature of non-glutinous are categorized as a type of non-glutinous sorghum. Sorghum major types are classified based on similar prince and chemical properties (mainly amylopectin in brewing sorghum) even they are genotype differences, the divergences of cultivars in same type may affect classification between groups. The differences of several sorghum cultivars in previous studies were great limited to practical application. Therefore, it is very important to develop non-destructive and accurate method based on the actual situation of mixed sorghum cultivars.

GC-IMS formed a trend in characterizing volatile compounds of sorghum due to good separation ability and high sensitivity. It separates trace gases through the gas phase part, then characterizes chemical ionic substances according to the difference in the mobility rates of gas phase ions in the electric field (Shvartsburg, 2017). GC-IMS advances in high sensitivity, low detection limit, no sample pre-treatment and visualization of data. It's very suitable for classification owing to the samples' difference can be seen intuitively by the gallery plot after analysis. At present, GC-IMS has played a crucial role in the optimization of storage conditions and origin tracing of agricultural products, meat and so on (Martín-Gómez et al., 2022; Nie et al., 2022; Xiao et al., 2022). However, the researches of sorghum by GC-IMS mainly stayed at characterizing volatile compounds of cooked sorghum (Fan et al., 2021). In this paper, GC-IMS technology and chemometrics as a new combination to explore sorghum discrimination in types and adulterations. Orthogonal partial least squares-discriminant analysis (OPLS-DA) is a classification method in multivariate data analysis, which is an improvement of PLS-DA. The variations in X are adjusted by predictive and orthogonal components to reduce dimension. Therefore OPLS-DA can well solve samples with larger within-group divergence (Bylesjo et al., 2006). The method also has a considerable prospect on adulteration discrimination (Petrakis et al., 2015). This method is suitable for the classification of sorghum types which may have large within-group divergence. Furthermore, in order to reduce the invalid information and improve the model performance, we need to screen the characteristic volatile compounds affecting sorghum types, so the study introduced genetic algorithm-partial least squares (GA-PLS) which is a classic variable selection algorithm. The optimal number of variables is selected by plotting variables with the correlation coefficient R^2, the variables with better objective function value are retained (Hasegawa and Funatsu, 1998; Lin et al., 2012).

This paper took three major sorghum types in common (according to practical harvest and process) instead of cultivars to detect and screen characteristic volatile compounds with GC-IMS and GA-PLS. Then we established classification models with OPLS-DA for fresh and stored sorghum types and adulterations. which would ensure better quality supervision of high-quality sorghum and meet the practical application in Baijiu industry.

2. Material and methods

2.1. Materials

Three sorghum types were supplied from Luzhou Red Sorghum Modern Agriculture Development Co., Ltd (Luzhou, China) during 2022. Organic waxy sorghums (Luzhou, China), non-glutinous sorghums (Inner Mongolia and Northeast, China), organic non-glutinous sorghums (organic region in Northeast, China). Fresh sorghums harvested in 2022 and detected within a week, stored sorghum samples were stored in the warehouse under specific conditions for 4–6 months. The samples were collected at random from representative sorghum growing areas to simulate the actual sorghum collection situation.

2.2. HS-GC-IMS analysis

The detection conditions slightly changed in sample mass and temperature of injection needle referred to Fan's (Fan et al., 2021) method. The volatile compounds were carried out through HS-GC-IMS (FlavourSpec®, Gesellschaft für Analytische Sensorsysteme mbH, Dortmund, Germany) which equipped with an automatic headspace sampler (CTC Analytics AG, Zwingen, Switzerland). The samples in triplicates of different sorghum were weighed and placed 2.0 g in 20 mL headspace bottles (kept in the same environment during the period) and sealed with magnetic headspace caps for determining HS-GC-IMS volatile compounds. Column type: MXT-5 (15 m × 0.53 mm, 1.0 μm). The samples were incubated for 15 min at the incubator speed of 500 r/min and the incubator temperature of 60 °C. After that, injection needle heated to 80 °C automatically sucked 500 μL headspace gas and injected it into the injection port (75 °C). The samples were injected into the chromatographic column with 60 °C constantly through nitrogen (purity greater than 99.999 %). The resulting ions were taken to a migration tube (98 mm in length) with a constant temperature of 45 °C, and the drift gas was set at 150 mL/min. Each spectrum was scanned an average of 12 times. The programed flow rate was: initial 2 mL/min for 2 min, increased uniformly to 15 mL/min at 10 min, then reached to 100 mL/min at 20 min, and rose to 150 mL/min at 25 min. The analysis time was 25 min. 20 μL of n-ketones C4–C9 (Shandong Hanon Scientific Instrument Co., Ltd, China) standard solution was injected as an external parameter to calculate the volatile compound retention index.

Adulterated samples: the proportions of 2.5 %, 5.0 %, 7.5 %, 10.0 %, 20.0 %, 30.0 %, 40.0 % and 50.0 % were adulterated respectively, including organic waxy sorghum adulterated with non-glutinous sorghum, organic waxy sorghum adulterated with organic non-glutinous sorghum, organic non-glutinous sorghum adulterated with non-glutinous sorghum.

2.3. Multivariate analysis in sorghum types

GA-PLS is the program performing the variable selection for non-spectral data. Although gallery plot could directly show the differences of compounds among samples, the uniformity of volatile compounds changed due to high sensitivity and low detection limit. Screening characteristic volatile compounds manually is challenged. Visual fatigue, personnel differences and limited display will affect the selection of characteristic volatile compounds, making it unable to characterize sample information fully and draw clear conclusions. GA-PLS calculated the quantitative results of all labeled signal peak areas by algorithm then screened the characteristic volatile compounds. The data set must be a X+1 matrix, in which each line is a sample, the columns 1:X are the X variables and the last column is the Y variable.

Sorting samples in the matrix according to the Y variable. This did 100 runs with random permutations of the Y variable. The fitness function returned from these experiments is correlation coefficient R^2 in cross validation.

The data set was constructed with variables selected by GA-PLS. Sorghums were sorted in advance and the classification models of organic waxy sorghum, non-glutinous sorghum and organic non-glutinous sorghum were established through OPLS-DA after centering and auto scaling to unit variance in data. The validation of the models was carried out by seven-fold cross-validation (CV) in autofitting and permutation tests for each class. External blind samples were also used to validate the performance of models.

2.4. Statistical analysis

The data analysis software Laboratory Analytical Viewer (LAV, G.A. S., Dortmund, Germany) and three visual analysis plug-ins provided by GC-IMS detected volatile compounds. The characteristic volatile compounds of samples could be manually screened and analyzed from different angles. They were qualitatively determined by comparing the retention index and drift time of reference in the GC-IMS spectrum library, and quantitatively determined by peak area. Choosing variables

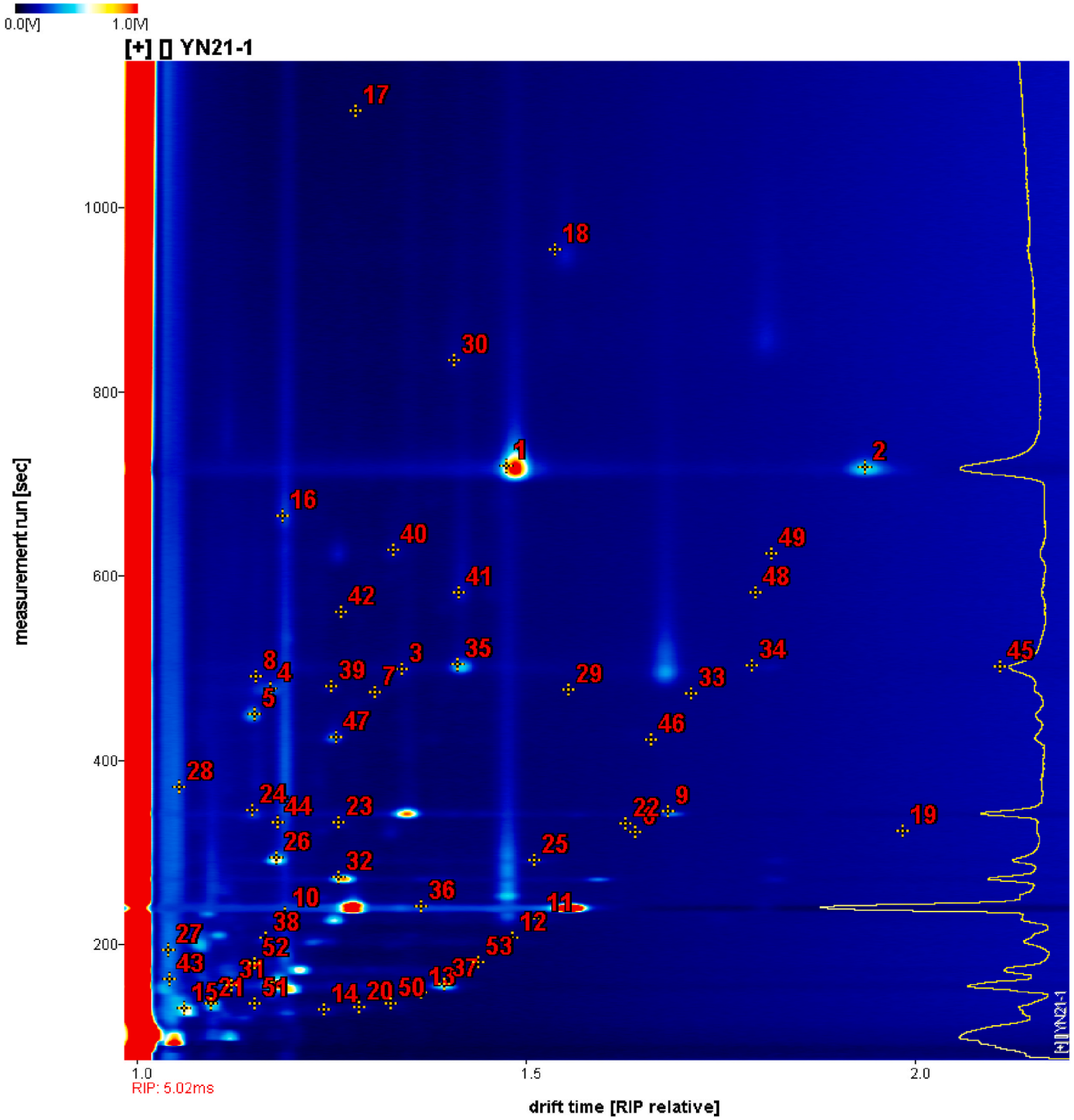

Fig. 1. Diagram of volatile compounds in fresh organic waxy sorghum.

through GA-PLS in the toolbox of MATLAB 2014a software (The Mathworks Inc., Natick, Massachusetts, USA) and constructing model with OPLS-DA by the SIMCA 14.1 software (Umetrics, Umeå, Sweden). Kruskal-Wallis test of non-parametric approach was further used to compare the significance of differences in volatile compounds among the sorghum types.

3. Results and discussion

3.1. Volatile compounds fingerprint of sorghum types

The volatile compounds of different sorghum types were detected by HS-GC-IMS. The result shows a three-dimensional false-color diagram of drift time-retention time-ionic strength (Fig. 1). The X-axis is the relative drift time of ions, and the Y-axis is the retention time of gas chromatography. The red vertical line is the reactive ion peak (RIP), which is the signal peak of water in the air ionized by radiation from the ionizing source tritium, and is often used as a reference signal to indicate the total number of all ionizable ions (Aliaño-González et al., 2019). The drift time and location of RIP were normalized for each spectrum of volatile compounds. Volatile compounds generated on the right side of RIP were different spots of circle, ellipse and water drop, which are related to the nature and concentration of the substance itself. Most of the signals distributed in drift time of 1.0-2.0 ms and retention time of 100-950 s. Most of the volatile compounds were small and medium molecules. Different colors represent different signal intensity of the substance, the brighter and redder spots, the greater substance concentration and vice versa (Li et al., 2019; Zhang et al., 2020). In addition, the ability of robbing hydronium ions in RIP is different due to substance concentration. Monomer will appear only while the substance concentration is small, and dimer, even trimer, multimer appears with the increasing concentration. The occurrence of several polymers is also determined by properties.

HS-GC-IMS analyzed volatile compounds of 206 and 186 samples respectively in fresh and stored sorghum with three types. LAV analysis software was used to label volatile compounds as much as possible. 72 and 63 signal peaks were labeled separately in projects after eliminating interference and stray peaks. Then gallery plots were established with selected 15 samples at random (Fig. 2), where each column represents signal peaks of the same volatile compounds in different samples, each row represents all the signal peaks selected from a sample. As shown slight differences in signal peaks appeared within sorghum types. It should be noted that signal peaks of non-glutinous sorghums were inconsistent partially because it was the most common type contained many cultivars, the other two types of sorghum were relatively fewer but large yield. Within-class differences may affect classification because samples of mixed cultivars could not guarantee the consistency in the same type. 51 and 44 volatile compounds (including monomers and dimers of the same compounds) were identified respectively (Table 1 and Table 2) from marked signal peaks by comparing the NIST retention index database with the IMS drift time database, others were unidentified but different in contents. The aldehydes were the most in fresh sorghum types up to 14 kinds, followed by alcohols, ketones, esters, aromatics and heterocyclic compound. After storage, the number of alcohols and ketones increased but aldehydes had almost halved. The color depth of the signal peaks showed obvious differences in varieties and concentrations of volatile compounds of three fresh sorghum types (Fig. 2A). Kruskal-Wallis test verified significance ($p < 0.001$) further in comparing volatile compounds of sorghum types when the variance is not homogeneous. The volatile compounds of organic waxy sorghums were mainly aldehydes that showed generally fruity, sweet and astringent. And the contents of nonanal, 2-ethyl-1-hexanol were higher than other types (region a). The concentrations and varieties of volatile compounds of non-glutinous sorghums were relatively the highest, and more compounds appeared such as ketones and alcohols. The divergences of volatile compounds between non-glutinous sorghums and organic non-glutinous sorghums were inapparent but we could distinguish them from octanal, heptanal, nonanal, hexanal, benzaldehyde (region c). Aldehydes have a strong scent with a low threshold value. Lower fatty aldehydes have a pungent odor, which increases with the length of carbon chain and reaches the maximum at C8–C12. These aldehydes may be the original characteristic volatile compounds of sorghums which was consistent with the description of previous research (Fan et al., 2021). And the most significant compound which distinguished them from organic waxy sorghum was acetoin (region b). Due to the differences of volatile compounds in fresh sorghums and stored sorghums, we compared them by common compounds (Table 3). The concentrations of alcohols in volatile compounds of sorghum types increased after storage (Fig. 2B), especially n-hexanol, pentan-1-ol, 3-methylbutan-1-ol. The formation of alcohol may be related to the automatic oxidation of macromolecular substances such as oils and fats in raw grain (Zhang et al., 2009). The kinds and concentrations of volatile compounds in organic waxy sorghum increased obviously, while the concentrations of their own characteristic compounds decreased such as nonanal, benzaldehyde, acetoin. There are few studies on volatile compounds for raw sorghum detected by GC-IMS, Fan et al. (2021) concluded that the volatile compounds in six sorghum cultivars did not show significant difference but their corresponding signal intensity has difference. However, obvious differences in volatile compounds of fresh sorghum types in our study shows that they could be classified in terms of volatile compounds and corresponding signal intensity. There would be some uncertainty to distinguish characteristic volatile compounds of sorghum types by color difference merely, which needed to be further explored by chemometrics modeling.

3.2. Characteristic volatile compounds selection

Variable selection is a very important step in multivariate analysis. The chromatogram of GC-IMS can label volatile compounds with no upper limits. If all labeled compounds are modeled as variables, not only a huge amount of data will be generated, but they are profitless to maintain the model. Getting rid of variables that contain exclusive information contributes to simpler models and better predictive effects. We selected characteristic volatile compounds with GA-PLS to reduce variables marked by GC-IMS. Fig. 3A shows the variable with the greatest contribution after 100 runs was Area 49 (No. 42), which was selected 41 times. Among the stored sorghum types, nonanal (No. 4) contributed the most with 71 times of selection (Fig. 3B). Correlation coefficient R^2 (Formula 1) and root mean square error in cross validation (RMSECV, Formula 2) are the main value for choosing variables. The optimal number of variables were 65 and 55 in fresh and stored sorghum types corresponding to the minimum RMSECV. However, the number of variables in optimum may still be too more to build and maintain the OPLS-DA model. We chose the fewer variables near the optimal value of R^2 and RMSECV as alternative results (Table 4) in order to reduce the variables as much as possible and simplify the subsequent modeling on the basis of achieving the same effect. OPLS-DA modeled separately with variables and selected the final number of characteristic volatile compounds according to the stable evaluation parameters and prediction comprehensively on the premise of fewer variables.

Formula 1. A squared predictive correlation coefficient in cross validation.

$$R^2 = \left(1 - \sum_i (y_i - \hat{y}_i)^2 \Big/ \sum_i (y_i - \bar{y})^2\right) \times 100\ \%$$

y_i is the experimental value of sample i, $\hat{y}_i$ is the predicted value of sample i, $\bar{y}$ is the average of sample.

Formula 2. Root mean square error in cross validation.

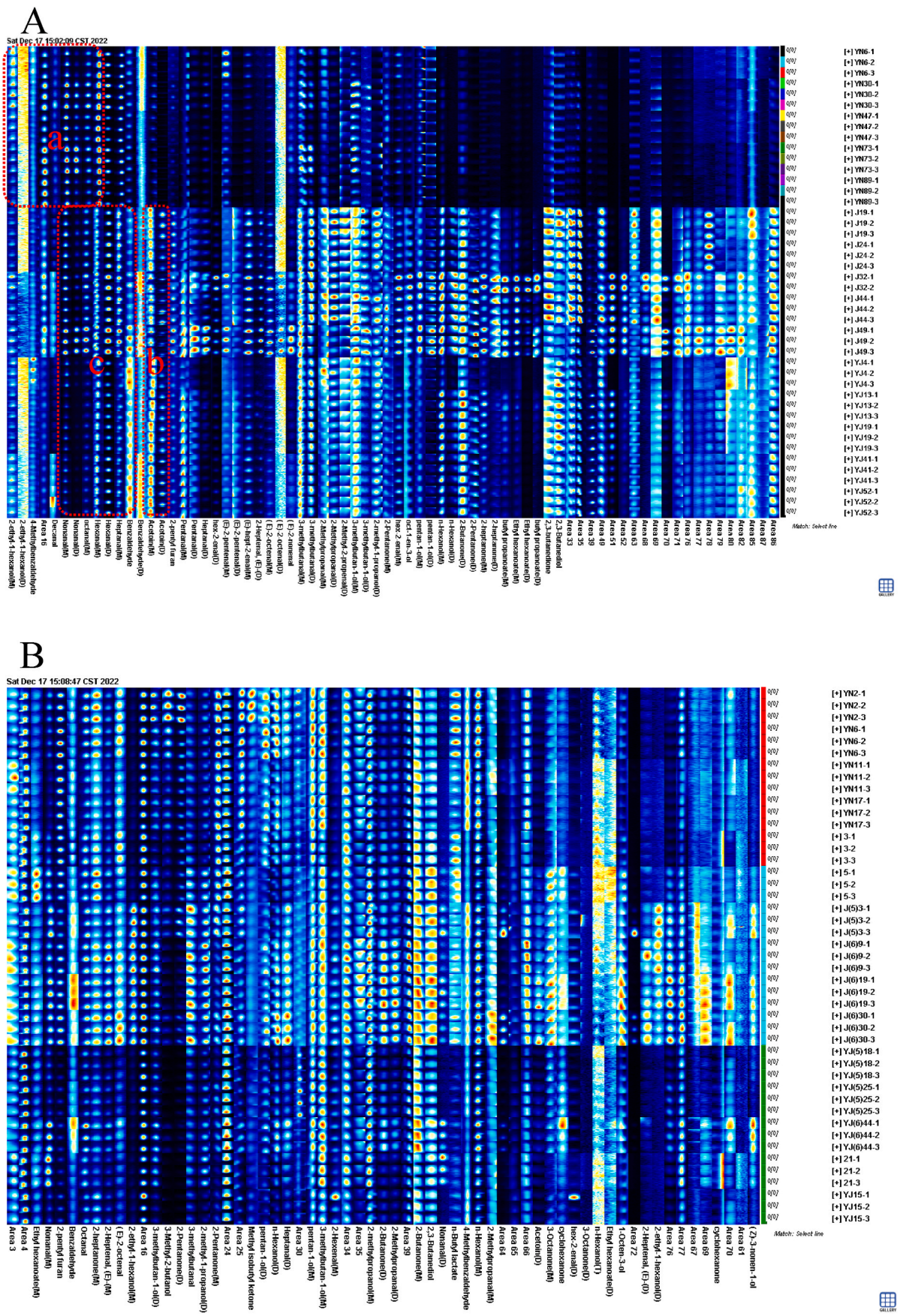

Fig. 2. Gallery plots of volatile compounds in fresh (A) and stored (B) sorghum types. 5 triplicated samples of each type with organic waxy sorghum, non-glutinous sorghum and organic non-glutinous sorghum from top to bottom.

Table 1
The identified volatile compounds in three fresh sorghum types.

Count	Compound	CAS	Formula	RI[a]	Rt[b]/s	Dt[c]/ms
1	Nonanal M[d]	C124196	C9H18O	1104.8	719.043	1.48902
2	Nonanal D[e]	C124196	C9H18O	1103.9	717.218	1.94008
3	Octanal M	C124130	C8H16O	1001.5	502.367	1.42076
4	Heptanal M	C111717	C7H14O	894.5	340.785	1.34527
5	Heptanal D	C111717	C7H14O	894	340.098	1.68605
6	Benzaldehyde	C100527	C7H6O	963.6	443.17	1.15241
7	Decanal	C112312	C10H20O	1213.7	947.529	1.55047
8	(E)-2-Nonenal	C18829566	C9H16O	1159.8	834.272	1.4166
9	4-Methylbenzaldehyde	C104870	C8H8O	1080.2	667.395	1.1938
10	(E)-2-Octenal M	C2548870	C8H14O	1058.5	621.908	1.33413
11	2-Ethyl-1-hexanol M	C104767	C8H18O	1036.8	576.421	1.41344
12	Oct-1-en-3-ol	C3391864	C8H16O	984	473.407	1.17217
13	(E)-Hept-2-enal M	C18829555	C7H12O	950	423.038	1.25625
14	2-Heptenal, (E)- D	C18829555	C7H12O	949.1	421.679	1.65755
15	Pentan-1-ol M	C71410	C5H12O	761.5	224.621	1.25453
16	3-Methylbutanal D	C590863	C5H10O	642.1	153.482	1.39491
17	3-Methylbutanal M	C590863	C5H10O	644.3	154.412	1.18545
18	2-Pentanone M	C107879	C5H10O	691.1	174.885	1.12201
19	2-Methylpropanal D	C78842	C4H8O	588.2	130.682	1.29078
20	2-Methylpropanal M	C78842	C4H8O	593.7	133.009	1.09927
21	(E)-2-Octenal D	C2548870	C8H14O	1059.2	623.404	1.81582
22	n-Hexanol M	C111273	C6H14O	874.1	321.212	1.32745
23	2-Butanone D	C78933	C4H8O	581.9	128.02	1.24495
24	Benzaldehyde D	C100527	C7H6O	968.1	449.808	1.47426
25	2-Ethyl-1-hexanol D	C104767	C8H18O	1035.2	573.103	1.80546
26	3-Methylbutan-1-ol M	C123513	C5H12O	738.6	208.464	1.237
27	Hexanal D	C66251	C6H12O	793	248.53	1.55783
28	Hexanal M	C66251	C6H12O	793.1	248.622	1.26083
29	Pentanal M	C110623	C5H10O	691.6	175.268	1.19335
30	Pentanal D	C110623	C5H10O	693	176.214	1.41937
31	2-Pentanone D	C107879	C5H10O	681.8	170.266	1.37128
32	Ethyl hexanoate M	C123660	C8H16O2	998	494.951	1.34487
33	Ethyl hexanoate D	C123660	C8H16O2	998.6	496.197	1.81596
34	2-Heptanone M	C110430	C7H14O	892.9	338.516	1.26224
35	2-Heptanone D	C110430	C7H14O	891.5	336.743	1.63214
36	Pentan-1-ol D	C71410	C5H12O	761.7	224.755	1.51961
37	2-Methyl-1-propanol D	C78831	C4H10O	628.2	147.588	1.37151
38	2,3-Butanediol	C513859	C4H10O2	785.4	241.745	1.36669
39	Hex-2-enal D	C505577	C6H10O	846.6	296.524	1.51347
40	Hex-2-enal M	C505577	C6H10O	849.2	298.853	1.18059
41	(E)-2-Pentenal M	C1576870	C5H8O	747.5	214.712	1.1071
42	(E)-2-Pentenal D	C1576870	C5H8O	748	215.122	1.36357
43	3-Methylbutan-1-ol D	C123513	C5H12O	739.7	209.238	1.48215
44	Butyl propanoate M	C590012	C7H14O2	906.1	358.035	1.28614
45	Butyl propanoate D	C590012	C7H14O2	903.6	354.298	1.71485
46	2,3-Butanedione	C431038	C4H6O2	583	128.483	1.18425
47	2-Methyl-2-propenal	C78853	C4H6O	572.1	123.867	1.21478
48	2-Pentyl furan	C3777693	C9H14O	993.4	487.29	1.25474
49	n-Hexanol D	C111273	C6H14O	875.1	322.127	1.64109
50	Acetoin M	C513860	C4H8O2	720.2	195.448	1.06729
51	Acetoin D	C513860	C4H8O2	728.4	201.243	1.33258

[a] The retention index calculated using n-ketones C4–C9 as external standard in MXT-5 column.
[b] The retention time in MXT-5 column.
[c] The drift time relatively to RIP.
[d] M represents monomer.
[e] D represents dimer.

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

n is the number of samples in the training set, y_i is the experimental value of the sample i, $\hat{y}_i$ is the predicted value of the model except the sample i (Hasegawa et al., 1997; Lin et al., 2012).

3.3. Classification models in sorghum types

The classification models of fresh and stored sorghum types were established through OPLS-DA under different variables selected by GA-PLS, and we optimized the model effects by adjusting predictive and orthogonal components after auto-fitting. The classification results provided the predicted Y value for the dummy variables (0 or 1) in each class used to direct the projection. Membership of a class depends upon matching the value of the dummy variable, so a value close to one indicates fitted membership. The discrimination value of 0.5 is often used as a practical threshold in order to classify whether the sample belongs to one class. The models were evaluated by the variance in X explained by the model (R^2X), the goodness-of-fit (R^2Y), and the goodness-of-prediction in cross-validation (Q^2) (Rivera-Pérez et al., 2022). R^2Y and Q^2 values more than 0.5 demonstrate accepted model fitting, further close to 1 indicate the excellent performance (Rubert et al., 2016; Triba et al., 2015). Components is a comprehensive variable, which is used for variable reduction and data visualization (Gu et al., 2020). Usually, the top several components can reflect most of the information of the original variables and reduce information redundancy. Taking fewer variables as a prerequisite, the optimal models with 19 and 32 volatile

Table 2
The identified volatile compounds in three stored sorghum types.

Count	Compound	CAS	Formula	RI[a]	Rt[b]/s	Dt[c]/ms
1	Nonanal M[d]	C124196	C9H18O	1100.1	719.467	1.4756
2	Nonanal D[e]	C124196	C9H18O	1099.1	717.439	1.93775
3	Ethyl hexanoate	C123660	C8H16O2	998.2	498.173	1.34146
4	Benzaldehyde	C100527	C7H6O	966	449.581	1.15173
5	n-Hexanol D	C111273	C6H14O	870.8	321.214	1.64171
6	n-Hexanol M	C111273	C6H14O	871.3	321.645	1.32511
7	3-Octanone M	C106683	C8H16O	982	473.13	1.30685
8	1-Octen-3-ol	C3391864	C8H16O	993.6	490.111	1.1543
9	Heptanal D	C111717	C7H14O	894.2	343.992	1.6834
10	Pentan-1-ol M	C71410	C5H12O	769	229.654	1.25491
11	Pentan-1-ol D	C71410	C5H12O	766	227.444	1.51814
12	3-Methylbutan-1-ol M	C123513	C5H12O	737.6	206.858	1.24055
13	3-Methylbutan-1-ol D	C123513	C5H12O	737.2	206.581	1.48463
14	2-Methylpropanol M	C78831	C4H10O	640.2	151.709	1.17293
15	2-Methyl-1-propanol D	C78831	C4H10O	629	147.175	1.36607
16	2-Butanone D	C78933	C4H8O	584.6	129.194	1.24158
17	2-Butanone M	C78933	C4H8O	584.8	129.282	1.06204
18	4-Methylbenzaldehyde	C104870	C8H8O	1075	665.068	1.18809
19	n-Hexanol	C111273	C6H14O	872.7	322.888	1.98646
20	2-Methylpropanal D	C78842	C4H8O	589.7	131.282	1.28729
21	2-Methylpropanal M	C78842	C4H8O	597.8	134.556	1.09662
22	2-Heptanone M	C110430	C7H14O	883.2	332.686	1.26108
23	Cyclohexanone M	C108941	C6H10O	895.1	345.372	1.14986
24	Hex-2-enal D	C505577	C6H10O	838.3	290.962	1.51264
25	2-Hexenal M	C505577	C6H10O	841.3	293.703	1.18056
26	3-Octanone	C106683	C8H16O	981.7	472.605	1.71434
27	Ethyl hexanoate D	C123660	C8H16O2	1000.3	502.619	1.79245
28	Octanal	C124130	C8H16O	1000.9	503.912	1.41343
29	2,3-Butanediol	C513859	C4H10O2	784.6	240.999	1.36741
30	Acetoin D	C513860	C4H8O2	725.3	197.919	1.33014
31	2-Pentanone D	C107879	C5H10O	684.3	169.563	1.36569
32	2-Pentanone M	C107879	C5H10O	685	169.837	1.12095
33	3-Methyl-2-butanol	C598754	C5H12O	701.8	180.862	1.44114
34	3-Methylbutanal	C590863	C5H10O	650.4	155.84	1.39692
35	Methyl isobutyl ketone	C108101	C6H12O	737.7	206.956	1.16622
36	2-Pentyl furan	C3777693	C9H14O	986.7	480.038	1.25177
37	(E)-2-Octenal	C2548870	C8H14O	1058.2	628.466	1.33129
38	2-Ethyl-1-hexanol	C104767	C8H18O	1036.6	581.617	1.41462
39	n-Butyl lactate	C34451199	C7H14O3	1027	560.681	1.26397
40	2-Heptenal, (E)- M	C18829555	C7H12O	947.4	422.194	1.66242
41	2-Heptenal, (E)- D	C18829555	C7H12O	949.2	424.817	1.25662
42	2-Ethyl-1-hexanol	C104767	C8H18O	1036.4	581.165	1.79607
43	Cyclohexanone D	C108941	C6H10O	892.5	341.504	1.45296
44	(Z)-3-Nonen-1-ol	C10340235	C9H18O	1155.5	839.981	1.41414

[a] The retention index calculated using n-ketones C4–C9 as external standard in MXT-5 column.
[b] The retention time in MXT-5 column.
[c] The drift time relatively to RIP.
[d] M represents monomer.
[e] D represents dimer.

compounds (Table 6) in fresh and stored sorghum types were chosen under comprehensive consideration in evaluation parameters and predictions (Table 4). For choosing excellent model, a larger R^2Y is a necessary condition but insufficient. In fresh sorghum types, model with 65 compounds revealed similar or poorer predictive performance compared with the model of 19 compounds even with a larger R^2Y. In the fresh sorghum model with 19 compounds, the evaluation values ($R^2X = 0.922$, $R^2Y = 0.913$, $Q^2 = 0.902$) all exceeded 0.9, which explained excellent performance by all extracted components. Furthermore, the values ($R^2X = 0.873$, $R^2Y = 0.856$, $Q^2 = 0.843$) over 0.8 in the stored sorghum model proved good effect even the small divergence between types. We accessed whether models were overfitting through permutation tests. The intercept of Q^2 regression line in cross-validation less than zero indicated the models were fitting (Abreu et al., 2019) and the model validations were effective. Fewer variables were used to achieve similar or better discrimination effect than others in both models. In particularly, fewer variables were applied in the model of fresh sorghum due to the larger differences of volatile compounds between types. Fig. 4 are the OPLS-DA score plots of the fresh and stored sorghum, which shows the relative positions between samples. The score plots draw the tolerance ellipse which is defined as the 95 % critical limit based on Hotelling's T2 (Huang et al., 2018). The sorghum types were classed separately into three clusters and the overall classification appearance was well. The samples of non-glutinous sorghum presented a decent shape of aggregation among types even though they contained the largest number of cultivars compared with others. Especially for fresh samples, their diverse concentrations of characteristic volatile compounds leaded to large divergences within the class. OPLS-DA well eliminated substances unrelated to classification and revealed the differences indeed. The relatively small divergences of non-glutinous and organic non-glutinous sorghums also leaded to a small part of adhesions between samples. For the stored sorghums (Fig. 4B), the distributions of samples in each type were slightly dispersed due to the varieties and concentrations of volatile compounds were more similar, but a clear distinction made them distinguish from each other. These results were agreed with the manual recognition through gallery plots. The concentrations of characteristic volatile compounds of three fresh sorghum types were high, they could be identified well by fingerprint and modeled with the fewer compounds, especially the high-quality organic waxy sorghum. After storage, the volatile compounds of sorghum types

Table 3
The results in peak areas of compounds in both fresh and stored sorghum types.

Compound	Peak area (fresh)/a.u.			Peak area (stored)/a.u.		
	OG[d]	N	ON	OG	N	ON
Nonanal M[e]	5656 ± 1028^{a}	4012 ± 909^{b}	2497 ± 564^{c}	1397 ± 592^{b}	1626 ± 374^{a}	1222 ± 580^{b}
Nonanal D	2194 ± 1039^{a}	1216 ± 542^{b}	488 ± 143^{c}	223 ± 124^{ab}	232 ± 70^{a}	194 ± 102^{b}
Octanal M	777 ± 112^{b}	943 ± 283^{a}	450 ± 43^{c}	515 ± 94^{a}	496 ± 72^{a}	403 ± 89^{b}
Heptanal D	209 ± 56^{b}	880 ± 647^{a}	233 ± 77^{b}	85 ± 34^{a}	109 ± 59^{a}	77 ± 21^{a}
Hex-2-enal D	187 ± 34^{c}	600 ± 419^{a}	309 ± 66^{b}	47 ± 13^{c}	100 ± 37^{a}	59 ± 26^{b}
2-Heptenal, (E) - D	236 ± 65^{c}	482 ± 188^{a}	279 ± 63^{b}	29 ± 11^{b}	54 ± 35^{a}	28±8^{b}
Benzaldehyde	520 ± 99^{c}	1205 ± 433^{b}	2278 ± 258^{a}	192 ± 60^{b}	281 ± 67^{a}	265 ± 127^{a}
2,3-Butanediol	340 ± 45^{b}	957 ± 355^{a}	1057 ± 129^{a}	256 ± 44^{c}	357 ± 68^{a}	313 ± 69^{b}
Acetoin D	1112 ± 106^{b}	9529 ± 6432^{a}	9358 ± 976^{a}	323 ± 102^{c}	1374 ± 556^{a}	672 ± 219^{b}
4-Methylbenzaldehyde	951 ± 113^{a}	820 ± 170^{b}	801 ± 224^{b}	440 ± 174^{a}	520 ± 164^{a}	506 ± 115^{a}
2-Methylpropanal M	331 ± 94^{c}	571 ± 156^{a}	448 ± 91^{b}	160 ± 77^{a}	143 ± 42^{a}	124 ± 38^{a}
Ethyl hexanoate M	322 ± 52^{c}	1741 ± 992^{a}	895 ± 169^{b}	242 ± 106^{a}	210 ± 69^{a}	223 ± 65^{a}
Ethyl hexanoate D	185 ± 17^{b}	1048 ± 1056^{a}	196 ± 18^{b}	29 ± 10^{a}	28±6^{a}	26±4^{a}
2-Methyl-1-propanol D	237 ± 103^{b}	940 ± 383^{a}	1006 ± 77^{a}	987 ± 500^{c}	2354 ± 673^{a}	1682 ± 660^{b}
n-Hexanol M	328 ± 117^{b}	2069 ± 573^{a}	1755 ± 437^{a}	5053 ± 934^{a}	4828 ± 1349^{a}	4603 ± 1044^{a}
Pentan-1-ol M	759 ± 136^{b}	1025 ± 365^{a}	665 ± 79^{c}	1698 ± 194^{a}	1525 ± 224^{a}	1554 ± 199^{a}
3-Methylbutan-1-ol M	606 ± 197^{b}	1324 ± 326^{a}	1195 ± 102^{a}	1954 ± 418^{b}	2248 ± 358^{a}	2340 ± 364^{a}
3-Methylbutan-1-ol D	165 ± 43^{c}	461 ± 267^{a}	245 ± 44^{b}	1257 ± 477^{b}	1802 ± 729^{a}	1518 ± 638^{ab}
2-Pentanone M	765 ± 177^{b}	1101 ± 305^{a}	783 ± 74^{b}	1351 ± 180^{a}	894 ± 238^{b}	761 ± 262^{c}
2-Pentanone D	235 ± 44^{b}	2396 ± 871^{a}	2146 ± 477^{a}	3741 ± 1529^{a}	2040 ± 921^{b}	1137 ± 764^{c}
2-Ethyl-1-hexanol M	633 ± 158^{a}	510 ± 121^{b}	467 ± 83^{b}	576 ± 428^{b}	1894 ± 991^{a}	411 ± 78^{b}
2-Ethyl-1-hexanol D	122 ± 17^{a}	127±9^{a}	124±7^{a}	70 ± 19^{b}	161 ± 100^{a}	62±3^{b}
Pentan-1-ol D	346 ± 59^{b}	666 ± 348^{a}	285 ± 47^{c}	725 ± 163^{a}	547 ± 139^{b}	529 ± 331^{c}
2-Methylpropanal D	80 ± 17^{b}	589 ± 286^{a}	412 ± 40^{a}	268 ± 85^{b}	498 ± 196^{a}	290 ± 76^{b}
2-Heptanone M	188 ± 28^{b}	813 ± 392^{a}	583 ± 171^{a}	251 ± 46^{a}	290 ± 67^{a}	207 ± 39^{b}
2-Butanone D	139 ± 47^{b}	694 ± 179^{a}	723 ± 138^{a}	357 ± 169^{c}	784 ± 291^{a}	474 ± 166^{b}
2-Pentyl furan	194 ± 24^{c}	539 ± 138^{a}	308 ± 20^{b}	438 ± 163^{a}	484 ± 160^{a}	330 ± 181^{b}

abc: the data with different letters are significantly different ($p < 0.001$) in Kruskal-Wallis test and samples were performed in triplicates.
[d] OG, N, ON represents organic waxy, non-glutinous and organic non-glutinous sorghum types respectively.
[e] M and D represent monomer and dimer respectively.

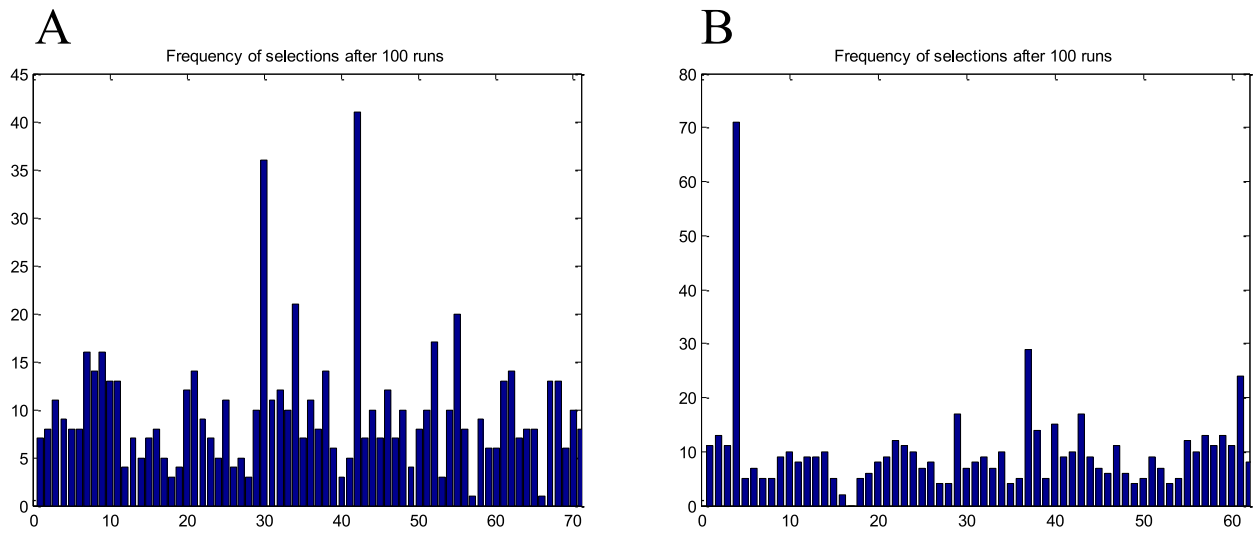

Fig. 3. Histogram of frequency of selections in fresh (A) and stored (B) sorghum types.

trended to close due to oxidation and degradation. The model could distinguish sorghum types through digging out data information in depth to avoid the drawback in gallery plots.

In addition to using validation set to evaluate model performance through R^2Y and Q^2, the study further evaluated classification effect of the types through external blind samples. 15 % of fresh and stored sorghums (32 and 27 respectively) were randomly selected as blind samples for independent testing which were not involved in the modeling. The classification list exhibited a column of predicted value for each sorghum type (Table 5). The value of decidable column concentrated on 0.8 to 1.1 (the threshold value 0.5 was the boundary line), while the other two columns were closed to 0 because they were completely unrelated types. The correct classification rates of blind samples in fresh and stored sorghum types were 95 % and 100 % with 95 % confidence interval respectively, which proved an excellent predictivity of OPLS-DA models and the practicality of the classifiers.

In addition, potential volatile compounds affecting sorghum types classification were selected according to the variable importance in projection (VIP) analysis, and VIP score≥1 was taken as the condition for screening (Table 6). For the classification of fresh sorghum types

Table 4
Parameters and prediction results of OPLS-DA models in sorghum with different variables.

Models	The number of variables	R^2/ % [a]	Components[b]	R^2X	R^2Y	Q^2	CCR[c]/ %	CCR-A[d]/ %	CCR-B[e]/ %
Fresh sorghum	65	99.00	3 + 6+0	0.867	0.965	0.96	100	76	93
	30	98.41	3 + 4+0	0.918	0.919	0.909	94	89	100
	19	97.96	3 + 4+0	0.922	0.913	0.902	95	87	100
	7	96.91	3 + 1+0	0.899	0.872	0.866	96	84	93
Stored sorghum	55	83.71	3 + 8+0	0.828	0.891	0.875	99	-[f]	–
	32	81.95	3 + 8+0	0.873	0.856	0.843	100	–	–
	24	79.59	3 + 8+0	0.906	0.826	0.809	95	–	–

[a] The correlation coefficient parameter in GA-PLS.
[b] Components of predicted, orthogonal X and orthogonal Y.
[c] Correct classification rates of predictive samples.
[d] Predictive adulterated samples of organic waxy sorghum adulterated with 10.0–50.0 % ratio of non-glutinous sorghum.
[e] Predictive adulterated samples of organic waxy sorghum adulterated with 10.0–50.0 % ratio of organic non-glutinous sorghum.
[f] No data due to the model is unsuitable for the discrimination of adulterated sorghum after storage.

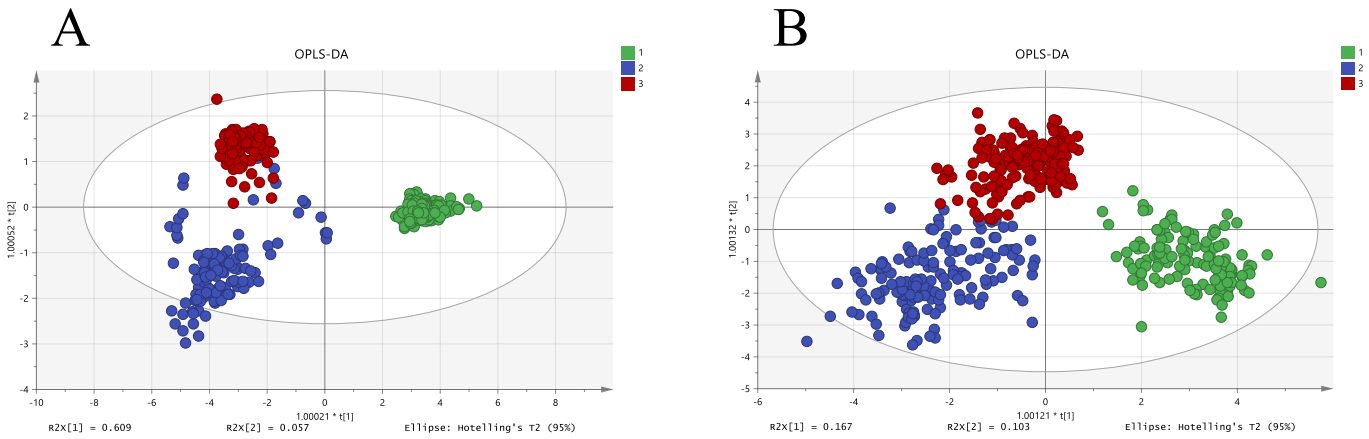

Fig. 4. Score plots of fresh (A) and stored (B) sorghum types based on OPLS-DA. Legend 1–3 represent samples of organic waxy sorghum, non-glutinous sorghum and organic non-glutinous sorghum.

Table 5
The average prediction results of the optimal OPLS-DA model in fresh sorghum samples.

Count	actual class[a]	Pred-v (1)[b]	Pred-v (2)	Pred-v (3)	Count	actual class	Pred-v (1)	Pred-v (2)	Pred-v (3)
1	1	0.96	0.02	0.02	17	2	0.24	0.28	0.49
2	1	1.04	0.01	−0.05	18	2	−0.06	0.90	0.16
3	1	1.03	−0.04	0.01	19	2	−0.03	0.99	0.04
4	1	1.05	0.09	−0.14	20	2	0.30	0.50	0.19
5	1	0.96	0.02	0.01	21	2	0.08	0.93	−0.01
6	1	0.93	0.08	−0.01	22	2	0.07	0.91	0.02
7	1	0.91	0.05	0.04	23	2	0.13	0.84	0.03
8	1	0.96	0.01	0.03	24	3	0.08	0.11	0.80
9	1	1.00	0.05	−0.05	25	3	0.07	0.21	0.72
10	1	0.98	−0.08	0.10	26	3	−0.03	0.05	0.98
11	1	0.96	0.07	−0.03	27	3	0.03	0.11	0.87
12	1	0.97	0.01	0.01	28	3	0.07	0.10	0.83
13	1	1.05	−0.02	−0.03	29	3	−0.06	0.02	1.03
14	1	1.03	0.00	−0.03	30	3	0.00	0.13	0.88
15	1	0.98	0.09	−0.07	31	3	−0.06	0.03	1.03
16	2	−0.06	1.12	−0.06	32	3	−0.01	0.16	0.86

[a] Class 1–3 represents organic waxy sorghum, non-glutinous sorghum and organic non-glutinous sorghum respectively.
[b] Pred-v (1–3) means the predicted value if the sample belongs to class 1, class 2, class 3 respectively.

combined with the significance ($p < 0.001$) in Kruskal-Wallis test, the potential volatile compounds were benzaldehyde, (E)-hept-2-enal, acetoin, 2-methylpropanal. And for stored sorghum types, the volatile compounds were 3-methyl-2-butanol, 2-heptenal (E), n-butyl lactate, 2-ethyl-1-hexanol, methyl isobutyl ketone, 2-butanone, (E)-2-octenal which could not be selected through fingerprint. Different from the volatile compounds selected visually by fingerprint with pairwise comparison, the potential compounds were used to classify three sorghums types and selected by algorithms with greater reliability.

Many investigations confirmed that OPLS-DA has been widely studied and tested in rice, quinoa flour (Li et al., 2023; Yang et al., 2022). However, few studies conduct on the identification of sorghum types. OPLS-DA can well solve samples with larger within-group divergence and obtain delightful classification effect (Bylesjo et al., 2006). In

Table 6
Volatile compounds and VIP scores in fresh and stored sorghum models.

Count	Compound	VIP[a] score	Count	Compound	VIP score	Count	Compound	VIP score
Fresh sorghum								
1	Benzaldehyde	1.41*	8	Area 49	0.98*	15	2-Methyl-2-propenal	0.90*
2	(E)-Hept-2-enal M	1.21*	9	2-Butanone D	0.97*	16	Area 80	0.86*
3	Acetoin M	1.14*	10	2-Methyl-1-propanol D	0.97*	17	2-Ethyl-1-hexanol M	0.85*
4	Area[b] 71	1.07*	11	2-Pentanone D	0.97*	18	Area 82	0.81*
5	Area 35	1.05*	12	Pentanal M	0.95*	19	Area 39	0.81*
6	2-Methylpropanal M	1.04*	13	(E)-2-Pentenal D	0.92*			
7	3-Methylbutanal M	0.98*	14	3-Methylbutan-1-ol M	0.91*			
Stored sorghum								
1	3-Methyl-2-butanol	1.33*	12	Area 67	1.09*	23	Nonanal M	0.87*
2	Area 76	1.30*	13	Area 65	1.08*	24	Pentan-1-ol D	0.86*
3	2-Heptenal, (E)- M	1.28*	14	Area 66	1.03	25	n-Hexanol D	0.81
4	n-Butyl lactate	1.28*	15	Area 77	1.00*	26	n-Hexanol M	0.77
5	Area 16	1.26*	16	Area 69	0.94*	27	Nonanal D	0.76*
6	2-Ethyl-1-hexanol D	1.19*	17	Heptanal D	0.92	28	Ethyl hexanoate M	0.73
7	Methyl isobutyl ketone	1.19*	18	Area 64	0.91*	29	Ethyl hexanoate D	0.73
8	2-Butanone D	1.15*	19	3-Methylbutan-1-ol D	0.87*	30	Area 70	0.70
9	Area 39	1.14*	20	2-Methylpropanal M	0.87	31	Cyclohexanone D	0.69
10	(E)-2-Octenal	1.13*	21	Area 34	0.87*	32	Cyclohexanone M	0.59*
11	Area 4	1.10*	22	Area 3	0.87*			

*: represents significantly different ($p < 0.001$) in Kruskal-Wallis test and samples were performed in triplicates.
[a] VIP represents variable importance in projection.
[b] Area represents volatile compounds unidentified.

previous study, accuracy of the anti-aliased convolutional network with machine vision obtained 89.15 % for ten cultivars (Ma et al., 2022). Bai et al. (2020) obtained 96 % accuracy of the model with HSI and PLS-DA. The study achieved the similar or even higher classification accuracy with fresh and stored sorghum in the application of algorithm. More importantly, we achieved the sorghum types' classification and solved samples with larger within-cultivars divergence through OPLS-DA, which would be feasible in practical harvest to solve the situation of cultivars combination. More than 200 representative samples applied for model establishment verified the practicability and accuracy. There are hundreds of sorghum cultivars and many of them mix into a major type for practical production which causes research limitation on classification and practical application. In this study, combined with the situation of sorghum cultivars mixed into batches, three sorghum types in common that can cover most sorghum cultivars were applied for classification in the actual sorghum collection process of enterprises. They can directly obtain the sorghum type and adulteration of this batch without pre-treatment. The study confirmed that classification of sorghum type is achievable through volatile compounds and OPLS-DA algorithm, and the divergence within cultivars in same sorghum type can be reduced.

3.4. Adulteration discrimination

The classification models of sorghum types can also be used to identify sorghum adulteration based on the change of discrimination value. This paper focused on the adulteration of fresh organic waxy sorghum. As for the stored sorghums, the similar volatile compounds of different sorghum types were inapplicable to discriminate adulteration. We interpreted the discrimination value from another angle, for organic waxy sorghum, the predicted value less than 0.5 or over 1.5 could be judged as adulteration, and the rule was consistent with other types. The discrimination value decreased and gradually deviated from the standard value 1 with the increasing adulterated proportion, until the threshold of 0.5 was broken to reach the correct discrimination. When the adulterated proportion of non-glutinous sorghum and organic non-glutinous sorghum were 2.5 %–50.0 %, correct classification rates of organic waxy sorghum adulteration were just 64 % and 83 %, while the adulterated proportion increased to 10.0 %–50.0 %, the classification rates reached to 87 % and 100 % respectively (Table 4). The model discrimination effect was better when adulterated ratio was greater than 10.0 %. The model discriminated inadequately in the case of the adulterated ratio less than 10.0 % because the characteristic volatile compounds concentration of the adulterated sorghum types could not reach the degree of being distinguished. However, the discrimination in organic non-glutinous sorghum adulterated with non-glutinous sorghum basically failed due to the differences between volatile compounds of these two sorghum types were inapparent, which made it more difficult to identify adulteration in the form of adulteration. The discrimination value in the column of the adulterated sample was usually in the middle range, not too close to either 1 or 0, because only part of its properties belonged to this type. The discrimination value was applied to discriminate both sorghum types and adulterations, yet the value size and meaning were distinguishable.

In a previous study Bai et al. (2020) achieved the identification accuracy of 91 % for the adulterated sorghums through HSI, Ma et al. (2022) focused on several sorghum cultivars with one adulterated ratio and obtained the number of misclassified sorghum kernels through machine vision system. Our accuracy rate of adulteration discrimination is slightly lower but the samples were much more complicated. We designed three major types instead of simply several sorghum cultivars, that is, the samples' difference in same sorghum type would be larger than the same cultivar. To sum up, the adulteration discrimination of sorghum types is effective and delightful. Moreover, there are few studies apply discrimination value to adulteration based on characteristic volatile compounds. Considering possible adulterated situations from the perspective of price such as organic waxy sorghum (higher price) adulterated with non-glutinous sorghum (lower price), we designed adulteration discrimination to protect high-quality sorghum and guarantee Baijiu quality. In addition, the paper involves a wider range of adulterated ratios from 2.5 % to 50 %, which evaluated the actual possible adulterated situation more comprehensively. We could find the adulteration tendency from the degree of deviation that changed regularly with the adulterated ratio, which could be used as the preliminary basis for adulteration discrimination and further verification. The intuitional discrimination value can be used for discrimination on both sorghum type and adulteration, which is very convenient and applicable to meet the practical application demands for Baijiu industry.

4. Conclusions

In this study, accurate classification of sorghum types and

adulterations could be carried out according to the actual sorghum harvest mode, which solved the challenge of cultivars combination in practical harvest. 19 and 32 volatile compounds separately were chosen to build the optimal classification model by OPLS-DA and GA-PLS in fresh and stored sorghum types designed from harvest process, the correct classification rates of prediction all exceeded 95 %,in adulteration of organic waxy sorghum reached 87 % with the ratio over 10 %, which proved excellent effect. The classification models are well applicable and predictable, and they have a good prospect for sorghum types protection and production monitoring. The composition of volatile compounds of each type of sorghum would change with storage time and other external environment. Studying the variation of characteristic volatile compounds of different types with storage time can further improve the performance and application scope of the model. In addition, it is also important to classify imported sorghum types based on large yield. GC-IMS combined with chemometrics is a non-destructive, efficient and sensitive discriminant method for sorghum types and adulterations in practical application.

CRediT authorship contribution statement

Mengjie Liu: Investigation, experimental design and operation, formal analysis, validation, writing – original draft, visualization. **Yang Yang:** experimental design and operation, data processing. **Xiaobo Zhao:** project administration, experimental design, formal analysis, writing – review & editing. **Yao Wang:** experimental design and operation, data processing. **Meiyin Li:** experimental design and operation, data processing. **Yu Wang:** experimental design and operation, data processing. **Min Tian:** formal analysis, writing-review. **Jun Zhou:** conceptualization, writing – review & editing, project acquisition, supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

Thanks Ziwen Li provided guidance in software application.

References

Abreu, A.C., Molina-Miras, A., Aguilera-Saez, L.M., Lopez-Rosales, L., Ceron-Garcia, M.D. C., Sanchez-Miron, A., Olmo-Garcia, L., Carrasco-Pancorbo, A., Garcia-Camacho, F., Molina-Grima, E., Fernandez, I., 2019. Production of amphidinols and other bioproducts of interest by the marine microalga amphidinium carterae unraveled by nuclear magnetic resonance metabolomics approach coupled to multivariate data analysis. J. Agric. Food Chem. 67 (34), 9667–9682. https://doi.org/10.1021/acs.jafc.9b02821.

Aliaño-González, M.J., Ferreiro-González, M., Barbero, G.F., Palma, M., 2019. Novel method based on ion mobility spectrometry sum spectrum for the characterization of ignitable liquids in fire debris. Talanta 199, 189–194. https://doi.org/10.1016/j.talanta.2019.02.063.

Bai, Z., Hu, X., Tian, J., Chen, P., Luo, H., Huang, D., 2020. Rapid and nondestructive detection of sorghum adulteration using optimization algorithms and hyperspectral imaging. Food Chem. 331, 127290 https://doi.org/10.1016/j.foodchem.2020.127290.

Boudries, N., Belhaneche, N., Nadjemi, B., Deroanne, C., Mathlouthi, M., Roger, B., Sindic, M., 2009. Physicochemical and functional properties of starches from sorghum cultivated in the Sahara of Algeria. Carbohydr. Polym. 78 (3), 475–480. https://doi.org/10.1016/j.carbpol.2009.05.010.

Bylesjo, M., Rantalainen, M., Cloarec, O., Nicholson, J.K., Holmes, E., Trygg, J., 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J. Chemometr. 20 (8–10), 341–351. https://doi.org/10.1002/cem.1006.

Fan, X., Jiao, X., Liu, J., Jia, M., Blanchard, C., Zhou, Z., 2021. Characterizing the volatile compounds of different sorghum cultivars by both GC-MS and HS-GC-IMS. Food Res. Int. 140, 109975 https://doi.org/10.1016/j.foodres.2020.109975.

Gu, S., Chen, W., Wang, Z., Wang, J., Huo, Y., 2020. Rapid detection of Aspergillus spp. infection levels on milled rice by headspace-gas chromatography ion-mobility spectrometry (HS-GC-IMS) and E-nose. Lebensm. Wiss. Technol. 132, 109758 https://doi.org/10.1016/j.lwt.2020.109758.

Guindo, D., Davrieux, F., Teme, N., Vaksmann, M., Doumbia, M., Fliedel, G., Bastianelli, D., Verdeil, J.-L., Mestres, C., Kouressy, M., Courtois, B., Rami, J.-F., 2016. Pericarp thickness of sorghum whole grain is accurately predicted by NIRS and can affect the prediction of other grain quality parameters. J. Cereal. Sci. 69, 218–227. https://doi.org/10.1016/j.jcs.2016.03.008.

Hasegawa, K., Funatsu, K., 1998. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. J. Mol. Struct.: THEOCHEM 425 (3), 255–262. https://doi.org/10.1016/S0166-1280(97)00205-4.

Hasegawa, K., Miyashita, Y., Funatsu, K., 1997. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. J. Chem. Inf. Comput. Sci. 37 (2), 306–310. https://doi.org/10.1021/ci960047x.

Huang, B.-M., Zha, Q.-L., Chen, T.-B., Xiao, S.-Y., Xie, Y., Luo, P., Wang, Y.-P., Liu, L., Zhou, H., 2018. Discovery of markers for discriminating the age of cultivated ginseng by using UHPLC-QTOF/MS coupled with OPLS-DA. Phytomedicine 45, 8–17. https://doi.org/10.1016/j.phymed.2018.03.011.

Huang, H., Hu, X., Tian, J., Peng, X., Luo, H., Huang, D., Zheng, J., Wang, H., 2022. Rapid and nondestructive determination of sorghum purity combined with deep forest and near-infrared hyperspectral imaging. Food Chem. 377, 131981 https://doi.org/10.1016/j.foodchem.2021.131981.

Li, M., Yang, R., Zhang, H., Wang, S., Chen, D., Lin, S., 2019. Development of a flavor fingerprint by HS-GC-IMS with PCA for volatile compounds of Tricholoma matsutake Singer. Food Chem. 290, 32–39. https://doi.org/10.1016/j.foodchem.2019.03.124.

Li, Z., Sun, X., Xu, T., Dai, W., Yan, Q., Li, P., Fang, Y., Ding, J., 2023. Insight into the dynamic variation and retention of major aroma volatile compounds during the milling of Suxiang japonica rice. Food Chem. 405, 134468 https://doi.org/10.1016/j.foodchem.2022.134468.

Lin, P., Chen, Y., He, Y., 2012. Identification of geographical origin of olive oil using visible and near-infrared spectroscopy technique combined with chemometrics. Food Bioprocess Technol. 5 (1), 235–242. https://doi.org/10.1007/s11947-009-0302-z.

Ma, S., Li, Y., Peng, Y., Nie, S., Yan, S., Zhao, X., 2022. An intelligent and vision-based system for Baijiu brewing-sorghum discrimination. Measurement 198, 111417. https://doi.org/10.1016/j.measurement.2022.111417.

Martín-Gómez, A., Segura-Borrego, M.P., Ríos-Reina, R., Cardador, M.J., Callejón, R.M., Morales, M.L., Rodríguez-Estévez, V., Arce, L., 2022. Discrimination of defective dry-cured Iberian ham determining volatile compounds by non-destructive sampling and gas chromatography. Lebensm. Wiss. Technol. 154, 112785 https://doi.org/10.1016/j.lwt.2021.112785.

Nie, S., Li, L., Wang, Y., Wu, Y., Li, C., Chen, S., Zhao, Y., Wang, D., Xiang, H., Wei, Y., 2022. Discrimination and characterization of volatile organic compound fingerprints during sea bass (Lateolabrax japonicas) fermentation by combining GC-IMS and GC-MS. Food Biosci. 50, 102048 https://doi.org/10.1016/j.fbio.2022.102048.

Okoh, P.N., Obilana, A.T., Njoku, P.C., Aduku, A.O., 1982. Proximate analysis, amino acid composition and tannin content of improved Nigerian sorghum varieties and their potential in poultry feeds. Anim. Feed Sci. Technol. 7 (4), 359–364. https://doi.org/10.1016/0377-8401(82)90005-0.

Petrakis, E.A., Cagliani, L.R., Polissiou, M.G., Consonni, R., 2015. Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by 1H NMR metabolite fingerprinting. Food Chem. 173, 890–896. https://doi.org/10.1016/j.foodchem.2014.10.107.

Rivera-Pérez, A., Romero-González, R., Garrido Frenich, A., 2022. A metabolomics approach based on 1H NMR fingerprinting and chemometrics for quality control and geographical discrimination of black pepper. J. Food Compos. Anal. 105, 104235. https://doi.org/10.1016/j.jfca.2021.104235.

Rubert, J., Lacina, O., Zachariasova, M., Hajslova, J., 2016. Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. Food Chem. 204, 201–209. https://doi.org/10.1016/j.foodchem.2016.01.003.

Shvartsburg, A.A., 2017. Ion mobility spectrometry (IMS) and mass spectrometry (MS). In: Lindon, J.C., Tranter, G.E., Koppenaal, D.W. (Eds.), Encyclopedia of Spectroscopy and Spectrometry, third ed. Academic Press, Oxford, pp. 314–321. https://doi.org/10.1016/B978-0-12-803224-4.00012-1.

Triba, M.N., Le Moyec, L., Amathieu, R., Goossens, C., Bouchemal, N., Nahon, P., Rutledge, D.N., Savarin, P., 2015. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. Mol. Biosyst. 11 (1), 13–19. https://doi.org/10.1039/c4mb00414k.

Wu, G., Johnson, S.K., Bornman, J.F., Bennett, S.J., Fang, Z., 2017. Changes in whole grain polyphenols and antioxidant activity of six sorghum genotypes under different irrigation treatments. Food Chem. 214, 199–207. https://doi.org/10.1016/j.foodchem.2016.07.089.

Xiao, Y., Huang, Y., Chen, Y., Xiao, L., Zhang, X., Yang, C., Li, Z., Zhu, M., Liu, Z., Wang, Y., 2022. Discrimination and characterization of the volatile profiles of five Fu brick teas from different manufacturing regions by using HS–SPME/GC–MS and HS–GC–IMS. Curr. Res. Food Sci. 5, 1788–1807. https://doi.org/10.1016/j.crfs.2022.09.024.

Yang, X., Xing, B., Guo, Y., Wang, S., Guo, H., Qin, P., Hou, C., Ren, G., 2022. Rapid, accurate and simply-operated determination of laboratory-made adulteration of quinoa flour with rice flour and wheat flour by headspace gas chromatography-ion

mobility spectrometry. Lebensm. Wiss. Technol. 167, 113814 https://doi.org/10.1016/j.lwt.2022.113814.

Zhang, K., Wong, J.W., Hayward, D.G., Sheladia, P., Krynitsky, A.J., Schenck, F.J., Webster, M.G., Ammann, J.A., Ebeler, S.E., 2009. Multiresidue pesticide analysis of wines by dispersive solid-phase extraction and ultrahigh-performance liquid chromatography-tandem mass spectrometry. J. Agric. Food Chem. 57 (10), 4019–4029. https://doi.org/10.1021/jf9000023.

Zhang, Q., Ding, Y., Gu, S., Zhu, S., Zhou, X., Ding, Y., 2020. Identification of changes in volatile compounds in dry-cured fish during storage using HS-GC-IMS. Food Res. Int. 137, 109339 https://doi.org/10.1016/j.foodres.2020.109339.

Zhou, K., Chen, M., Zheng, M., 2008. Occurrence of main disease and pest and its controlling countermeasures in organic sorghum for maotai wine. Guizhou Agricultural Sciences 36 (4), 90–91.