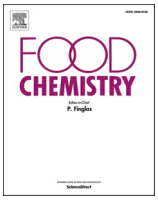
Food Chemistry 426 (2023) 136610

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem



Analysis and classification of coffee beans using single coffee bean mass spectrometry with machine learning strategy

Jia-Jen Tsai [a,1], Che-Chia Chang [b,1], De-Yi Huang [a], Te-Sheng Lin [b,c,*], Yu-Chie Chen [a,d,*]

[a] Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[b] Department of Applied Mathematics, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[c] National Center for Theoretical Sciences, National Taiwan University, Taipei 10617, Taiwan
[d] International College of Semiconductor Technology, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan

ARTICLE INFO

Keywords:
Coffee
Ambient ionization
Neural network
Machine learning
Palm civet coffee
Mass Spectrometry

ABSTRACT

Coffee is a daily essential, with prices varying based on taste, aroma, and chemical composition. However, distinguishing between different coffee beans is challenging due to time-consuming and destructive sample pretreatment. This study presents a novel approach for directly analyzing single coffee beans through mass spectrometry (MS) without the need for sample pretreatment. Using a single coffee bean deposited with a solvent droplet containing methanol and deionized water, we generated electrospray to extract the main species for MS analysis. Mass spectra of single coffee beans were obtained in just a few seconds. To showcase the effectiveness of the developed method, we used palm civet coffee beans (kopi luwak), one of the most expensive coffee types, as model samples. Our approach distinguished palm civet coffee beans from regular ones with high accuracy, sensitivity, and selectivity. Moreover, we employed a machine learning strategy to rapidly classify coffee beans based on their mass spectra, achieving 99.58% accuracy, 98.75% sensitivity, and 100% selectivity in cross-validation. Our study highlights the potential of combining the single-bean MS method with machine learning for the rapid and non-destructive classification of coffee beans. This approach can help to detect low-priced coffee beans mixed with high-priced ones, benefiting both consumers and the coffee industry.

1. Introduction

From the busy streets in a big city to the quiet mountainous regions in the countryside, coffee is a common beverage that fuels the daily lives of millions. It is no wonder why it is one of the most profitable products in the world, with coffee exports propelling economic development in many countries in Latin America, Africa, and Asia (Utrilla-Catalan et al., 2022). The price of coffee varies greatly and depends on various factors, such as species, taste, aroma, flavors, and processing methods (Do Carmo et al., 2020). Even the degree of roasting determines the taste and aroma of coffee, with longer roasting times leading to more bitterness (Münchow et al., 2020). Although the exact reasons for bitterness remain to be fully understood, the destruction of sugar and the appearance of chlorogenic acid lactones and phenylindanes (Münchow et al., 2020; Gigl et al. 2021) have been considered as possible causes.

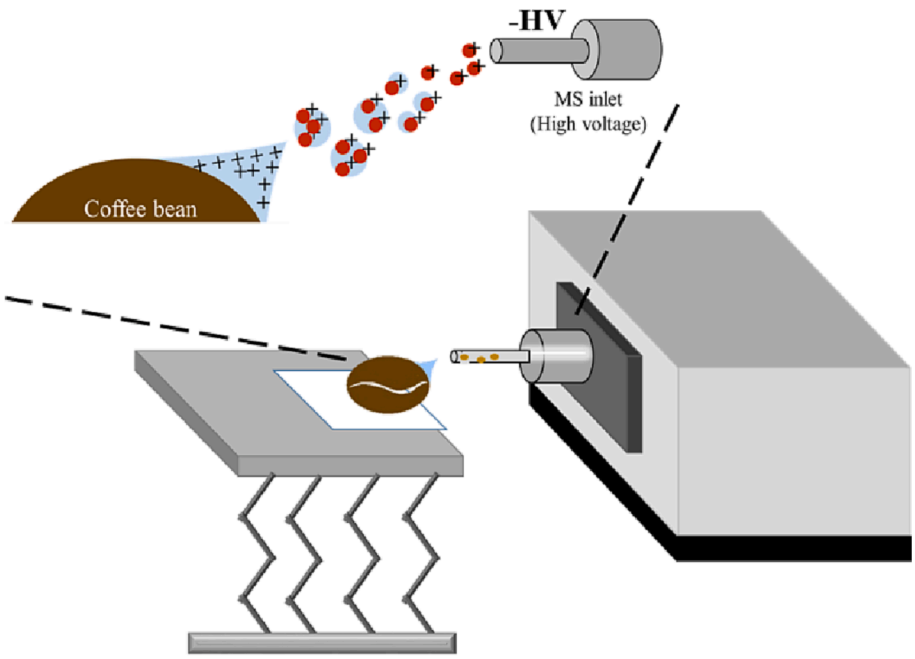
However, with the high prices of some coffee beans, it is not surprising to hear about adulteration with lower-priced ones (Cheah and Fang, 2020). For example, palm civet coffee or *kopi luwak*, which are processed by the Asian palm civet through its intestinal tract and collected from its defecations, is one of the most expensive coffees in the world (Muzaifa et al., 2019). Because coffee beans look quite similar after processing at a similar roast degree, distinguishing beans based on their outlooks is difficult, and experts have difficulty identifying high-priced coffee that is adulterated with a limited number of low-priced ones. The taste and aroma of coffee are generally derived from the chemical composition of its beans, which is the basis of its classification according to bean quality. Suitable analytical methods are helpful as they provide the chemical information of different coffee beans. Thus,

Abbreviations: DESI, desorption electrospray ionization; EASI, easy ambient sonic spray ionization; MS, mass spectrometry; PI-ESI, polarization-induced electrospray ionization; SHAP, SHapley Additive exPlanations.

* Corresponding authors at: Department of Applied Mathematics, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (T.-S. Lin). Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (Y.-C. Chen).
E-mail addresses: tslin@math.nctu.edu.tw (T.-S. Lin), yuchie@nycu.edu.tw (Y.-C. Chen).

[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.foodchem.2023.136610
Received 10 March 2023; Received in revised form 18 April 2023; Accepted 10 June 2023
Available online 12 June 2023



Scheme 1. Cartoon illustration of the setup of the single coffee bean MS method.

analytical methods such as ultraviolet–visible absorption spectroscopy (Belay et al., 2008), electrochemical methods (Wada et al., 2021; Tomac et al., 2020), chromatographic methods (Yashin et al., 2017), liquid chromatography coupled with mass spectrometry (MS) (Gigl et al. 2021; Farag et al., 2023), gas chromatograph coupled with MS (Farag et al., 2023), paper spray ionization MS (Garrett et al., 2013), and nuclear magnetic resonance spectroscopy (Gigl et al. 2021; Farag et al., 2023) have been used to characterize and classify coffee beans. However, most methods require time-consuming and labor-intensive extraction procedures with destruction methods prior to detection by analytical tools (Garrett et al., 2013; Münchow et al., 2020; Gigl et al. 2021; Farag et al., 2023). If individual coffee beans can be examined one-by-one without the need to perform time-consuming and tedious sample pretreatment steps, informative chemical information from single coffee beans can be rapidly obtained. This should make it possible to find low-priced coffee beans mixed with high-priced ones. Therefore, exploring analytical methods for single-coffee bean analysis could be invaluable for quality control.

Ambient ionization MS allows for the direct analysis of samples in their native environment, with or without minimal sample pretreatment steps (Takats et al., 2004; Hiraoka et al., 2007; Hsieh et al., 2011; Garrett et al., 2014; Wleklinski et al., 2015; Meher and Chen, 2015a; Meher and Chen, 2015b; Rosa et al., 2016; Wu et al., 2017; Huang et al. 2022). For instance, desorption electrospray ionization (DESI)-MS and easy ambient sonic spray ionization (EASI) have been used to directly analyze intact green Arabica coffee beans, conducting *in situ* extraction and ionization for MS analysis (Garrett et al., 2014; Rosa et al., 2016). Additionally, polarization-induced electrospray ionization (PI-ESI) has been demonstrated as a straightforward ionization method, which requires neither a high-voltage power supply nor gas for assisting the ionization of analytes (Meher and Chen, 2015a; Meher and Chen, 2015b). Using PI-ESI, a droplet of solvent deposited on an intact coffee bean should be able to extract main compositions from the coffee bean and generate electrospray for instantaneous MS analysis. Although coffee beans are poor dielectric materials, their dielectric features can be raised to a certain extent after being deposited with a small droplet of solvent. This method should allow the rapid and informative chemical analysis of individual coffee beans, making it invaluable for quality control.

For classification strategies, the emergence of processing MS data using machine learning approaches (Zhou and Zare, 2017; Kantz et al., 2019; Xie et al., 2020; Hung et al., 2021; Lassen et al., 2021; Yang et al., 2021; Bonetti et al., 2022; Caporaso et al., 2022; Gebreyes GG 2021) has recently attracted considerable attention. One strategy is the ensemble tree model, which has been successfully applied to classify individual cells (Xie et al., 2020) and discover personal information from latent fingerprints (Zhou and Zare, 2017). Given that the Universal Approximation Theorem (Cybenko, 1989) guarantees the existence of an optimal solution, this study uses a one-hidden-layer neural network as the classification model. Based on the results obtained from single coffee bean MS analysis, neural network-based machine learning was used to accelerate the coffee bean classification. For the trained network model, deep SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) was used to seek important features in explaining the classification strategy of the network.

2. Experimental section

2.1. Reagents and materials

Arabica coffee beans containing 8 different sources were purchased from illy (Italy). Non-arabica coffee beans were obtained from Central and South America. Palm civet coffee beans, i.e. *kopi luwak*, were obtained from Indonesia. Potassium chloride was purchased from Fluka (St. Gallen, Switzerland). Sodium formate was purchased from J. T. Baker (Phillipsburg, NL, USA). Methanol was purchased from Aencore (Box Hill, Australia), whereas deionized water was obtained from Taisun (Taiwan).

2.2. Instrumentation

Single coffee bean mass spectra were obtained using a Bruker Daltonics AmaZon SL ion trap mass spectrometer (Bremen, Germany). Bruker Daltonics micrOTOF Q II mass spectrometer (Bremen, Germany) was used to obtain the exact masses of target ions. The inner diameter of the ion transfer capillary in ion trap mass spectrometer and micrOTOF Q II mass spectrometer was ~ 0.024 in. (= ~0.061 cm) and 0.0194 in. (= ~0.049 cm), respectively. The outer diameter of the ion transfer

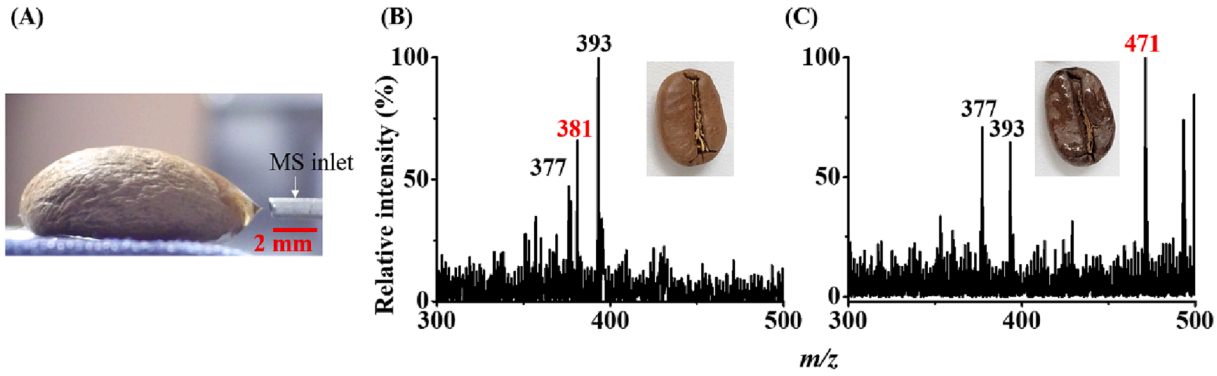


Fig. 1. (A) Photograph of the single coffee bean-based MS analysis. Mass spectra of the single coffee bean obtained from the (B) light and (C) dark roast non-Arabica coffee beans. The insets in Panels (B) and (C) show the corresponding photographs of the coffee beans.

capillary was ~ 0.256 in. (= ~0.657 cm). The length of the metal extension tube (inner diameter: ~ 1.0 mm; outer diameter: ~1.5 mm) adapted to the orifice of the mass spectrometer was ~ 4 cm. Photographs were obtained using a Sage Vision SG-210X camera (New Taipei City, Taiwan).

2.3. *In situ extraction and ionization*

A coffee bean was placed close to (~1 mm) the inlet of the mass spectrometer first followed by switching on the mass spectrometer. Scheme 1 shows the cartoon illustration of the setup. Coffee beans were randomly selected from the coffee-bean containers. A droplet of solvent (~10 μL) containing methanol and deionized water (3:1, v/v) was deposited on the surface of the coffee bean. Mass spectra were acquired immediately after the deposition of the solvent droplet. The voltage applied on the orifice of the mass spectrometer was set to − 4500 V and operated at the positive ion mode. The temperature of the ion transfer capillary was set to 200 °C. The number of the ions set at the ion charge control was 70,000, whereas the maximum acquisition time was set to 100 ms when conducting a single coffee bean analysis.

2.4. *Data pre-processing*

The intensity of the mass spectrum of each coffee bean was pre-processed, first, normalized so that the mean was equal to unity, and second, a max-pooling procedure was applied to get a representative value for each integer intensity between 200 and 500. The details were described in Supporting Information (SI) Appendix 1.1.

2.5. *Classification using neural network*

We use a one-hidden-layer neural network as the model. The input layer consists of 301 neurons, the hidden layer has 50 neurons, and the number of neurons in the output layer is equal to the number of classification classes. The output vector of the network is converted using the softmax function to a vector of values between 0 and 1, which presents the probability of each class. The network is trained using the cross-entropy loss, and the training process is accomplished by the Adam algorithm. To verify the robustness of the proposed algorithm, we applied the repeated random sub-sampling validation, namely, we validated the accuracy of the results by averaging over 20 experiments. In each experiment, we randomly took four-fifths of the samples as the training set and the rest as the testing set. After a model was trained, we used deep SHAP to find the important features. These found features were further validated by constructing again a one-hidden-layer neural network (with 50 neurons in the hidden layer) that has the found features as input, and we then checked the classification performance of such a model. SI Appendix 1.2–1.5 shows the details of classification of our MS data using neural network.

2.6. *Identification of the features by MS/MS analysis*

To know the identity of the features found from the machine learning results, coffee beans were extracted off-line. That is, five coffee beans were vortex-mixed in the solvent (2 mL) containing of methanol and deionized water (3:1, v/v) for 10–20 s followed by removing the coffee beans and adding the other five coffee beans to the same solvent. The resultant solvent was directly injected to the conventional electrospray ion source coupled with the ion trap mass spectrometer for MS/MS analysis by a syringe operated by a syringe pump with a flow rate of 0.5 mL/h. The voltage applied to the orifice of the mass spectrometer was set to − 4500 V, whereas the temperature of the ion transfer capillary was set to 200 °C. The number of ions set at the ion charge control was 100,000, whereas the maximum acquisition time was set to 200 ms. Target ions were selected with a mass window of ± 0.6 amu. The exact masses of the discovered features were obtained by using the qTOF. When operating in positive ion mode, the voltage applied to the orifice of the mass spectrometer was set to − 4500 V. The temperature of the ion transfer capillary was set at 200 °C. Sodium formate was used as the internal calibration standard for obtaining the exact masses of target ions.

3. Results and discussion

3.1. *Single coffee bean MS analysis*

To analyze a single coffee bean by MS, we used a droplet of micro-sized solvent to extract the main composition from a single coffee bean *in situ*. A Taylor cone was initialized when putting the single coffee bean deposited with a droplet solvent close to the inlet of the mass spectrometer applied with high voltage. Methanol/deionized water (3:1, v/v; 10 μL) was used as the *in situ* extraction solvent and electrospray solvent. Fig. 1**A** shows the photograph of a coffee bean with light roast deposited with a droplet of the solvent placed close to the inlet of the mass spectrometer applied with − 4.5 kV. Apparently, a Taylor cone was generated and clearly observed in the photograph therein. Fig. 1**B** shows the resulting mass spectrum. The inset shows the photograph of the representative photograph of the coffee bean with pale brown. The peaks at *m/z* 377 and 393, derived from the sodium and potassium adducts of caffeoylquinic acid (Moon et al., 2009; Li et al., 2020; Montis et al., 2022), while the peak at *m/z* 381 was derived from the potassium adduct of disaccharide (e.g., sugar) (Portillo and Arévalo, 2022), which dominated the mass spectra. Unsurprisingly, there were ions derived from caffeoylquinic acid and disaccharide since they are the main compositions in the coffee beans (Moon et al., 2009; Li et al., 2020; Montis et al., 2022). Given that the coffee bean used in this experiment was a light roast, the disaccharide could remain its chemical structure. A dark-roast coffee bean (inset photograph in Fig. 1**C**) was also used as the sample. Due to dark-roasting, the coffee bean became dark brown and

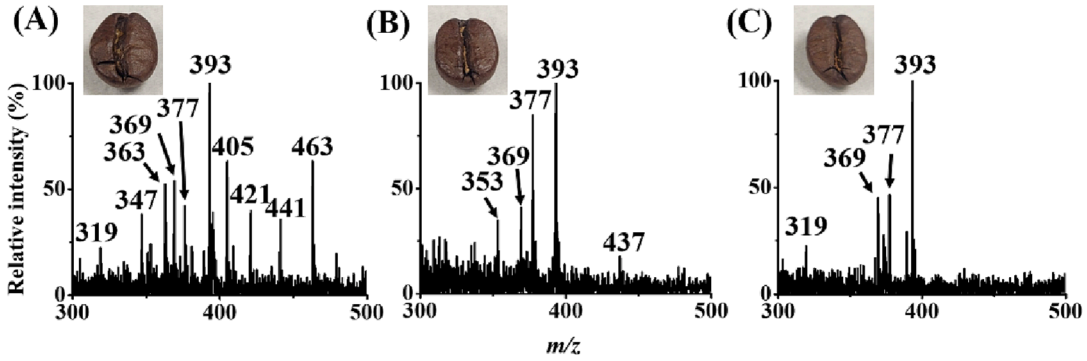


Fig. 2. Single coffee bean mass spectra of (A) palm civet coffee bean, (B) non-Arabica coffee bean, and (C) Arabica coffee bean. All the coffee beans were roasted in the medium degree. The insets in Panel (A)-(C) show the corresponding photographs of the coffee beans.

looked oily. Fig. 1**C** shows the resultant mass spectrum. The ions at *m/z* 377 and 393 were still observed in the same mass spectrum. However, the ion at *m/z* 381 derived from the potassium adduct of disaccharide disappeared. Because the coffee bean was dark roasted, the disaccharide structure was destroyed due to the high roast temperature. Moreover, a new peak at *m/z* 471, which was derived from the protonated $^{\beta}N$-behenoyl-5-hydroxytryptamide [Chemspider], was observed in the resultant mass spectrum. These results indicate that ions derived from the main composition of single coffee beans can be readily observed in the resulting mass spectra using the developed approach. Accordingly, *in situ* extraction followed by ionization can be carried out on a single coffee bean simultaneously, with the coffee bean itself used as the ionization source. The approach used herein shows that light and dark roast coffee beans can be easily distinguished through its ions at *m/z* 381 and 471, respectively. The appearance of the peak at *m/z* 381 in the mass spectrum of the coffee bean indicated that it was not highly roasted. The peak at *m/z* 471 was observed in the mass spectrum of the dark-roast coffee bean using our single coffee bean MS analysis approach.

3.2. Analysis of palm civet coffee beans and non-palm civet coffee beans

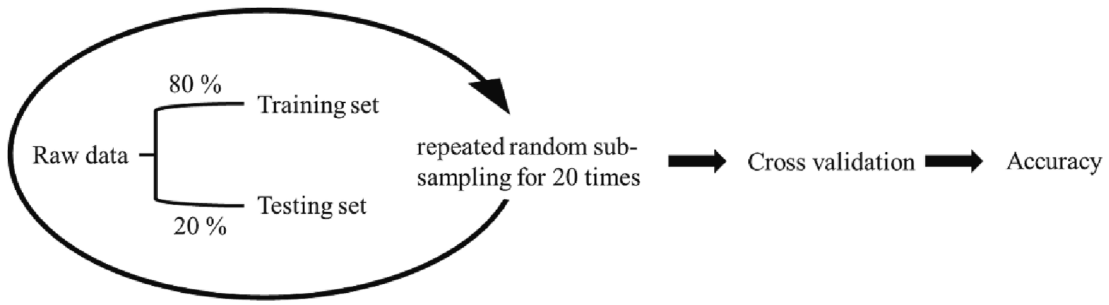
Abovementioned results show the feasibility of using our developed method to distinguish coffee beans with different roast degrees. Nevertheless, dark-roast coffee beans can be easily distinguished based on their outlooks according to their brownness level and oily surface (*cf.* inset in Fig. 1**C**). To further demonstrate the usefulness of our approach, palm civet coffee (*kopi luwak*) and non-palm civet coffee were used as the model samples. All the coffee beans were roasted in a medium level. Fig. 2**A** shows the representative mass spectrum of a single palm civet coffee bean (medium roast): the ions at *m/z* 319, 347, 363, 369, 377, 393, 405, 421, 441, and 463 dominated the mass spectrum of the bean. Fig. 2**B** shows the representative mass spectrum of a single non-Arabica coffee bean: here, the mass spectrum was dominated by the peaks at *m/z* 353, 369, 377, and 393. Fig. 2**C** shows the representative mass spectra of single Arabica coffee bean: the mass spectrum was dominated by the peaks at *m/z* 319, 369, 377, and 393. The peaks at *m/z* 377 and 393, derived from the sodium and potassium adducts of caffeoylquinic acid (Moon et al., 2009; Li et al., 2020; Montis et al., 2022), respectively, were observed in all the mass spectra of these coffee beans. These two peaks had also been previously observed in Fig. 1. Because caffeoylquinic acid is a main component of coffee beans and is involved in the bitterness of the beverage (Moon et al., 2009; Münchow et al, 2020; Li et al., 2020; Montis et al., 2022), these peaks appearing in the mass spectra of these coffee beans with medium roast were unsurprising. More peaks, especially at *m/z* > 400, were observed in the mass spectrum of the single palm civet coffee beans. This was understandable given that the coffee bean was treated by a palm civet. To accelerate the classification of palm civet and non-palm civet coffee, a one-hidden-layer neural network was further used for classification.

Table 1A
Model confusion matrix of the accumulated testing sets in the cross-validation.

	Predicted luwak	Predicted non-luwak
Luwak	237 (a)	3 (b)
Non-luwak	0 (c)	480 (d)

3.3. Classification of MS data by the neural Network-based Machine learning strategy

Palm civet coffee beans and two non-palm civet coffee beans (Arabica and non-Arabica) as the model samples, applying beans from three different sources for classification. Sixty coffee beans from each coffee bean source were analyzed using the single coffee bean MS analysis approach. The total number of mass spectra was 180. Supporting Information (SI) lists the details of the machine learning algorithm that was applied to classify palm civet and non-palm civet coffee beans. 80% of the mass spectral results were randomly sampled and trained on the training set, while the rest of the results (20%) were used as a testing set (Scheme 2). The data was randomly sub-sampled 20



Scheme 2. Illustration of the random sub-sampling validation process.

Table 1B
Model accuracy/sensitivity/specificity of the accumulated testing sets in the cross-validation.

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Cross validation	99.58 (717/720)	98.75 (237/240)	100 (480/480)

*Accuracy (%) = a + d/a + b + c + d; Sensitivity (%) = a/a + b; Specificity (%) = d/c + d.

times for cross-validation using the model to determine the accuracy. Results showed that the accuracy for identifying palm civet coffee beans was at 98.75% (Table 1A), whereas it was 100% for identifying non-palm civet coffee beans. The accuracy, sensitivity, and selectivity of cross-validation were 99.58%, 98.75%, and 100%, respectively (Table 1B). Results therefore indicated that palm civet coffee can be identified using the developed strategy yielding high accuracy, sensitivity, and selectivity.

3.4. Characterization of the discovered features

Deep SHAP was then used to find important features from the data. Fig. 3 shows the important features in Deep SHAP. The top features used to distinguish palm civet coffee from non-palm civet coffee were *m/z* 463, 405, 441, 421, 347, 409, and 363. These discovered features are rare peaks found in the mass spectra of coffee beans. qTOF was then used to further determine the exact masses of the top 7 features (463, 405, 441, 421, 347, 409, and 363). Table 2 shows the list of the experimental exact mass, the possible molecular formulae, theoretical *m/z*, and the mass error from theoretical values. Some of the molecular formulae contain alkali metal ions, i.e. sodium and potassium ions. Expectedly, we found that most of these ions possessed hydroxyl groups based on our MS/MS analysis (Fig. 4). The MS/MS spectra were mainly dominated by the fragments (marked blue) with a loss of a water from the target ions. SI **Figs. S2–S8** show the MS/MS spectra of these discovered features, their possible chemical structures following Chemspider [https://www.chemspider.com/], and the possible fragments. However, we are unable to identify the exact chemical structures due to a few possibilities for each feature based on the results shown in Fig. 4, Table 2, and SI **Figs. S2–S8**. Moreover, the information related to the metabolites from palm civet coffee beans is very limited, especially for the molecular weights higher than 300 Da. Some of the possible molecular formulae contain sulfur (SI **Figs. S3-S6**). Sulfur containing molecules are commonly found in aroma molecules in fruit peels (McGorrin, 2011; Cannon and Ho, 2018). Because palm civets eat coffee cherries and digest these cherry peels completely in their intestinal tract, these molecules were present in the mass spectra of palm civet coffee beans. Further efforts must hence be devoted to determine the chemical structures of these features. Nevertheless, the classification of palm civet coffee from regular ones using the developed single coffee bean MS

Table 2
List of the possible molecular formulae and their theoretical masses of the discovered features.

Experimental exact mass (*m/z*)	Possible molecular formula	Theoretical mass (*m/z*)	Mass error to the theoretical values (ppm)
463.2774	$C_{21}H_{34}N_8O_4 + H^+$	463.2781	−1.51
	$C_{23}H_{40}N_2O_6 + Na^+$	463.2784	−2.15
405.2441	$C_{22}H_{38}O_3S + Na^+$	405.2439	0.49
441.1588	$C_{22}H_{24}N_4O_4S + H^+$	441.1596	−1.81
421.2159	$C_{22}H_{38}O_3S + K^+$	421.2178	−4.5
	$C_{22}H_{32}N_2O_4S + H^+$	421.2161	−0.47
347.2041	$C_{13}H_{26}N_6O_5 + H^+$	347.2043	−0.57
	$C_{21}H_{30}O_2S + H^+$	347.2045	−1.15
409.3196	$C_{25}H_{42}N_2O + Na^+$	409.3195	0.24
	$C_{22}H_{46}N_2O_2 + K^+$	409.3196	0
363.1726	$C_{22}H_{28}O_2 + K^+$	363.1726	0

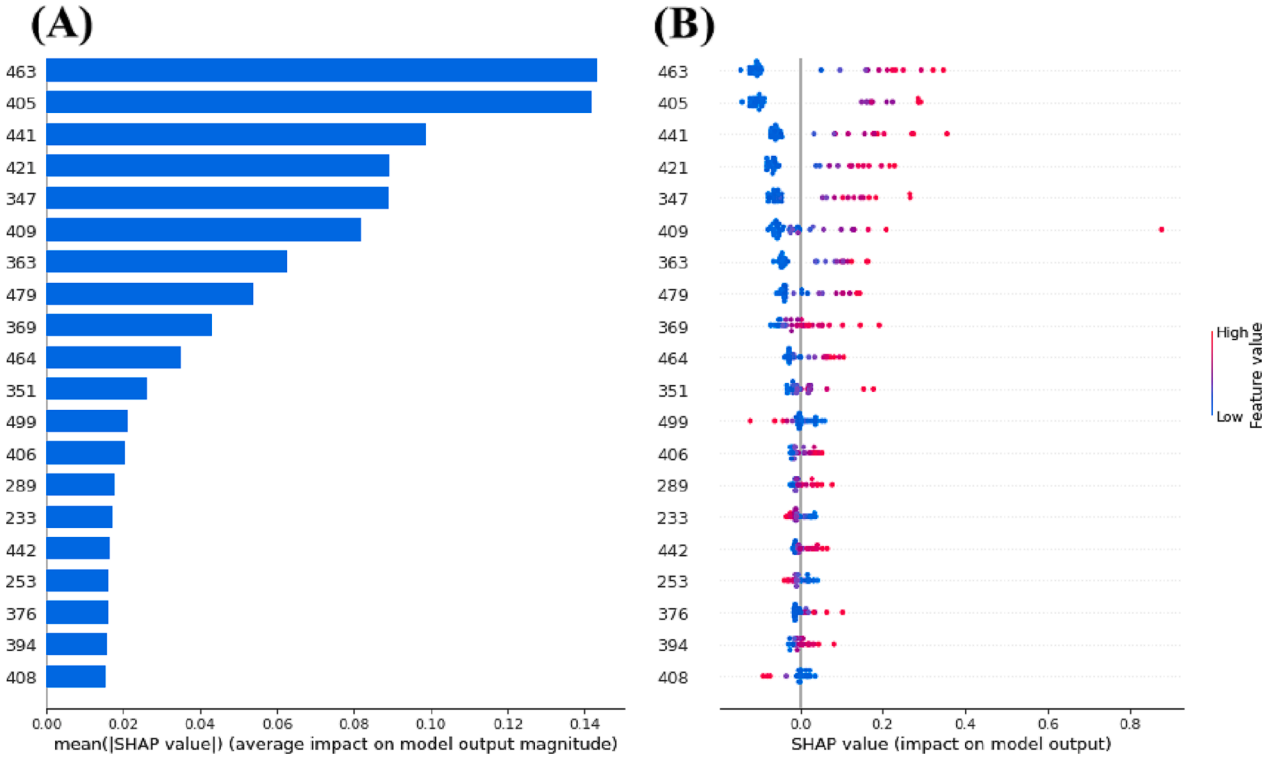


Fig. 3. Deep SHAP results. (A) Mean absolute SHAP values and (B) SHAP values impact on model output.

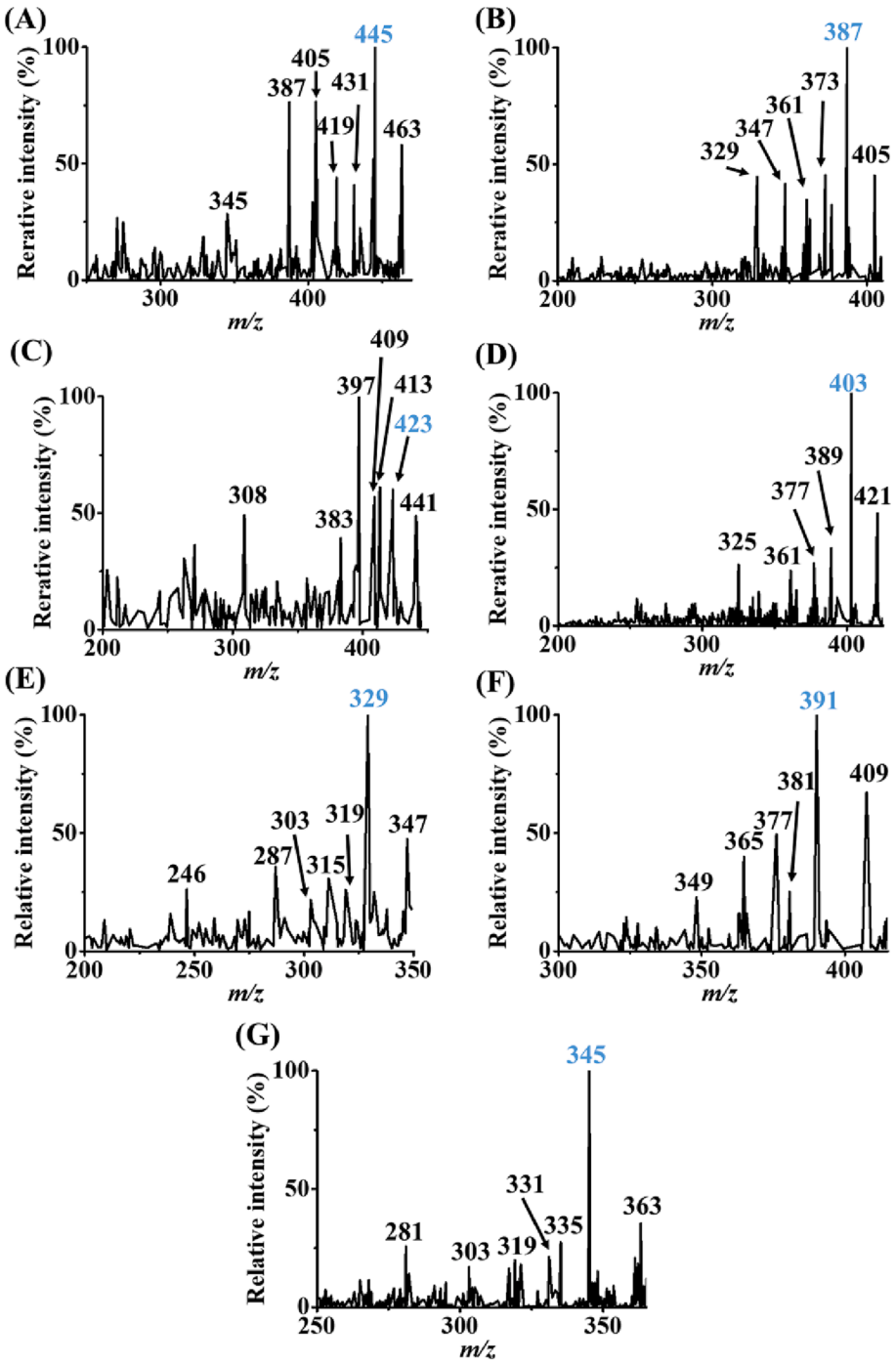


Fig. 4. MS/MS spectra of the peaks at (A) *m/z* 463, (B) *m/z* 405, (C) *m/z* 441, (D) *m/z* 421, (E) *m/z* 347, (F) *m/z* 409, and (G) 363 derived from the extracts of palm civet coffee beans.

analysis with the machine learning technique approach proved to be highly successful.

4. Conclusions

To distinguish high quality coffee beans from low quality ones without performing any sample pretreatment remains a challenge. Most existing analytical methods require destructive extraction methods using time-consuming steps prior to the analysis. This study demonstrates a novel method using single coffee bean MS analysis to analyze intact coffee beans one-by-one without performing any destructive extraction steps. This allows it to be both simple and straightforward. The setup of the ionization method is easy: a single coffee bean deposited with a droplet of solvent is directly used as the ionization source, where the solvent is then used to simultaneously extract the main composition of the coffee bean and facilitate the formation of the Taylor cone to generate electrospray containing analytes for MS analysis. The results show the feasibility of using the MS method to characterize single coffee beans, one-by-one, in just a few seconds. Additionally, a machine learning strategy was employed to classify MS data. Among various available strategies, the neural network model was chosen because of its strong expressibility and its ability to guarantee the optimal solution. Owing to the speed of MS analysis and fast data processing, this approach can be potentially used for high-throughput analysis.

Currently, the quality of coffee beans mainly depends on the judgment of baristas. For simplicity and speed, our approach may be useful in ensuring the quality of coffee beans in the coffee industry.

In addition, we believe that the developed single coffee bean MS approach can be extended to the analysis of single objects of interest based on the similar concept demonstrated in this work.

CRediT authorship contribution statement

Jia-Jen Tsai: Formal analysis, Investigation, Methodology, Validation, Visualization. **Che-Chia Chang:** Formal analysis, Investigation, Methodology, Validation, Visualization. **De-Yi Huang:** Formal analysis, Investigation, Visualization. **Te-Sheng Lin:** Supervision, Funding acquisition, Conceptualization, Writing – original draft, Writing – review & editing. **Yu-Chie Chen:** Supervision, Funding acquisition, Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

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.

Acknowledgements

We thank the National Science and Technology Council, Taiwan (108–2113-M-009–018-MY3, 111-2113-M-A49-019-MY3, and 111-2628-M-A49-008-MY4) for the financial support of this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2023.136610.

References

Belay, A., Ture, K., Redi, M., & Asfaw, A. (2008). Measurement of caffeine in coffee beans with UV/vis spectrometer. *Food Chemistry, 108*, 310–315.

Bonetti, J. L., Samanipour, S., & van Asten, A. C. (2022). Utilization of machine learning for the differentiation of positional NPS isomers with direct analysis in real time mass spectrometry. *Analytical Chemistry, 94*, 5029–5040.

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2022). Prediction of coffee aroma from single roasted coffee beans by hyperspectral imaging. *Food Chemistry, 371*.

Cannon, R. J., & Ho, C. T. (2018). Volatile sulfur compounds in tropical fruits. *Journal of Food and Drug Analysis, 26*, 445–468.

Cheah, W. L., & Fang, M. (2020). HPLC-based chemometric analysis for coffee adulteration. *Foods., 9*, 880.

Chemspider, http://www.chemspider.com/ (Accessed Aug. 30th, 2022).

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2*, 303–314.

do Carmo KB, do Carmo JCB, Krause MR, Moreli AP, Lo Monaco PAV, Quality of arabic coffee under different processing systems, drying methods and altitudes. Biosci. J. 2020, 36, 1116-1125.

Farag, M. A., Mohamed, T. A., El-Hawary, E. A., & Abdelwareth, A. (2023). Metabolite profiling of premium civet luwak bio-transformed coffee compared with conventional coffee types, as analyzed using chemometric tools. *Metabolites, 13*, 173.

Garrett, R., Rezende, C. M., & Ifa, D. R. (2013). Coffee origin discrimination by paper spray mass spectrometry and direct coffee spray analysis. *Analytical Methods, 5*, 5944–5948.

Garrett, R., Schwab, N. V., Cabral, E. C., Henrique, B. V. M., Ifa, D. R., Eberlin, M. N., et al. (2014). Ambient mass spectrometry employed for direct analysis of intact arabica coffee beans. *J. Brazil Chem. Soc., 25*, 1172–1177.

Gebreyes, G. G. (2021). *Image based coffee bean classification using deep learning technique.* Ethiopia: Debre Berhan University. PhD dissertation.

Gigl, M., Frank, O., Barz, J., Gabler, A., Hegmanns, C., & Hofmann, T. (2021). Identification and quantitation of reaction products from quinic acid, quinic acid lactone, and chlorogenic acid with strecker aldehydes in roasted coffee. *Journal of Agricultural and Food Chemistry, 69*, 1027–1038.

Hiraoka, K., Nishidate, K., Mori, K., Asakawa, D., & Suzuki, S. (2007). Development of probe electrospray using a solid needle. *Rapid Communications in Mass Spectrometry, 21*, 3139–3144.

Hsieh, C.-H., Urban, P. L., & Chen, Y.-C. (2011). Capillary Action-supported contactless atmospheric pressure ionization for the combined sampling and mass spectrometric analysis of biomolecules. *Analytical Chemistry, 83*, 2866–2869.

Huang, D. Y., Tsai, J. J., & Chen, Y. C. (2022). Direct mass spectrometric analysis of semivolatiles derived from real samples at atmospheric pressure. *ACS Omega, 7*, 10256–10261.

Hung, Y. C., Lee, F. S., & Lin, C. I. (2021). Classification of coffee bean categories based upon analysis of fatty acid ingredients. *Journal of Food Processing & Preservation, 45*.

Kantz, E. D., Tiwari, S., Watrous, J. D., Cheng, S., & Jain, M. (2019). Deep neural networks for classification of LC-MS spectral peaks. *Analytical Chemistry, 91*, 12407–12413.

Lassen, J., Nielsen, K. L., Johannsen, M., & Villesen, P. (2021). Assessment of XCMS optimization methods with machine-learning performance. *Analytical Chemistry, 93*, 13459–13466.

Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions, NIPS, 2017, 4765–4774.

Montis, A., Souard, F., Delporte, C., Stoffelen, P., Stévigny, C., & Van Antwerpen, P. (2022). Targeted and untargeted mass spectrometry-based metabolomics for chemical profiling of three coffee species. *Molecules, 27*, 3152.

Moon, J. K., Yoo, H. S., & Shibamoto, T. (2009). Role of roasting conditions in the level of chlorogenic acid content in coffee beans, correlation with coffee acidity. *Journal of Agricultural and Food Chemistry, 57*, 5365–5369.

Li, N., Dong, J., Dong, C., Han, Y., Liu, H., Du, F., & Nie, H. (2020). Spatial distribution of endogenous molecules in coffee beans by atmospheric pressure matrix-assisted laser desorption/ionization mass spectrometry imaging. *Journal of the American Society for Mass Spectrometry, 31*, 2503–2510.

McGorrin, R. J. (2011). The significance of volatile sulfur compounds in food flavors. *ACS Symposium Series, 1068*, 3–31.

Meher, A. K., & Chen, Y.-C. (2015). Polarization induced electrospray ionization mass spectrometry for the analysis of liquid, viscous and solid samples. *Journal of Mass Spectrometry, 50*, 444–450.

Meher, A. K., & Chen, Y.-C. (2015). Tissue paper assisted spray ionization mass spectrometry. *RSC Advances, 5*, 94315–94320.

Münchow, M., Alstrup, J., & Steen, I. (2020). D Giacalone, Roasting conditions and coffee flavor: A multi-study empirical investigation. *Beverages., 6*, 29.

Muzaifa, M., Hasni, D., Rahmi, F., & Syarifudin. (2019). What is kopi luwak? A literature review on production, quality and problems. *IOP Conf. Series, Earth and Environ. Sci., 365*.

Portillo, O. R., & Arévalo, A. C. (2022). Coffee's carbohydrates. A critical review of scientific literature. *Revis Bionatura, 7*, 11.

Rosa, J. S., Freitas-Silva, O., Rouws, J. R. C., Moreira, I. G. S., Novaes, F. J. M., Azevedo, D. A., Schwab, N., Godoy, R. L. O., Eberlin, M. N., & Rezende, C. M. (2016). Mass spectrometry screening of Arabica coffee roasting: A non-target and non-volatile approach by EASI-MS and ESI-MS. *Food Research International, 89*, 967–975.

Takats, Z., Wiseman, J. M., Gologan, B., & Cooks, R. G. (2004). Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science, 306*, 471–473.

Tomac, I., Šeruga, M., & Labuda, J. (2020). Evaluation of antioxidant activity of chlorogenic acids and coffee extracts by an electrochemical DNA-based biosensor. *Food Chemistry, 325*.

Utrilla-Catalan, R., Rodríguez-Rivero, R., Narvaez, V., Díaz-Barcos, V., Blanco, M., & Galeano, J. (2022). Growing Inequality in the Coffee Global Value Chain. *A Complex Network Assessment. Sustainability., 14*, 672.

Wada, R., Takahashi, S., Muguruma, H., & Osakabe, N. (2021). Electrochemical Analysis of Coffee Extractions at Different Roasting Levels Using a Carbon Nanotube Electrode. *Analytical Sciences, 37*, 377–380.

Wleklinski, M., Li, Y., Bag, S., Sarkar, D., Narayanan, R., Pradeep, T., et al. (2015). Zero-volt paper spray ionization and its mechanism. *Analytical Chemistry, 87*, 6786–6793.

Wu, M.-L., Chen, T.-Y., & Chen, Y.-C. (2017). Carbon fiber ionization mass spectrometry for the analysis of analytes in vapor, liquid, and solid phases. *Analytical Chemistry, 89*, 13458–13465.

Xie, Y. R., Castro, D. C., Bell, S. E., Rubakhin, S. S., & Sweedler, J. V. (2020). Single-cell classification using mass spectrometry through interpretable machine learning. *Analytical Chemistry, 92*, 9338–9347.

Yang, S., Li, C., Mei, Y., Liu, W., Liu, R., Chen, W., Han, D., & Xu, K. (2021). Determination of the Geographical Origin of Coffee Beans Using Terahertz Spectroscopy Combined with Machine Learning Methods. *Frontiers in Nutrition, 8*.

Yashin, A., Yashin, Y., Xia, X., & Nemzer, B. (2017). Chromatographic methods for coffee analysis: a review. *Journal of Food Research, 6*, 60.

Zhou, Z. P., & Zare, R. N. (2017). Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning. *Analytical Chemistry, 89*, 1369–1372.

Supporting Information

Analysis and Classification of Coffee Beans Using Single Coffee Bean Mass Spectrometry with Machine Learning Strategy

Jia-Jen Tsai,[1#] Che-Chia Chang,[2#] De-Yi Huang,[1] Te-Sheng Lin,[2,3*], and Yu-Chie Chen[1,4*]

[1]Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[2]Department of Applied Mathematics, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[3]National Center for Theoretical Sciences, National Taiwan University, Taipei 10617, Taiwan
[4]International College of Semiconductor Technology, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan

[#]The authors contributed equally to this work.

[*]Corresponding authors
T.-S. Lin
E-mail: tslin@math.nctu.edu.tw
Tel: +886-3-5712121 ext: 56422

Y.-C. Chen
E-mail: yuchie@nycu.edu.tw
Tel: +886-3-5131527
Dax: +886-3-5723764

1. Appendix - Classification with machine learning

1.1. Data preprocessing

The intensity of the mass spectrums of each coffee bean was preprocessed, first, normalized so that the mean was equal to unity, and second, a max-pooling procedure was applied to obtain a representative value for each integer intensity between 200 and 500. Fig. S1 shows an example.

(A) (B)

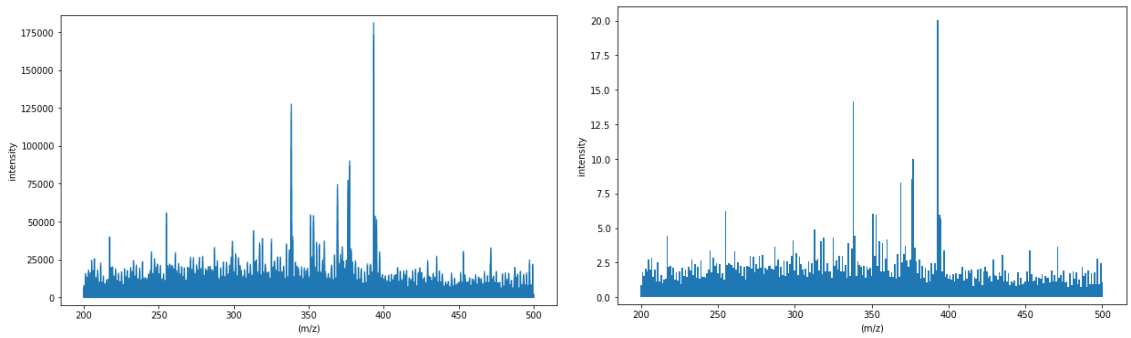


Fig. S1. Preprocessing of the intensity of mass spectra. (A) Original and (B) processed data.

We then went through the detail of how the preprocessing was done step by step, with adding a process of removing the background *m/z* values. Given a data $X = \{(l_i, x_i)\}_{i=1}^{d}$, where l_i are the *m/z* values and x_i are the intensity values, we first calculate the mean of the data by

$$\mu_{\text{old}}(X) = \frac{1}{d}\sum_{i=1}^{d} x_i.$$

Suppose that the background values occur at $i = i_1, i_2, \cdots, i_m$, then the identity values are assigned to this mean value, that is

$$x_{i_j} \leftarrow \mu_{\text{old}}(X), \quad \text{for } j = 1,2,\cdots,m.$$

Next, we calculated the mean after replacing the background intensity values by

$$\mu(X) = \frac{1}{d}\sum_{i=1}^{d} x_i.$$

Then, we did a normalization on the intensities such that the mean is equal to unity, that is

$$x_i \leftarrow \frac{x_i}{\mu(X)}, \quad i = 1,2,\cdots,d.$$

Finally, we did a max-pooling such that we obtained the intensity value for 200 to 500. Let's say the new data is $\hat{X} = \{(j, \hat{x}_j)\}_{j=200}^{500}$. Define round$(\cdot)$ to be the function that does *round half to even* (*bankers' rounding* or *Gaussian rounding*), and let $S_j = \{i \mid \text{round}(l_i) = j\}$, each $\hat{x}_j$ is defined as

$$\hat{x}_j = \begin{cases} \max\limits_{i \in S_j} \{x_i\}, & \text{if } S_j \neq \emptyset, \\ 0, & \text{if } S_j = \emptyset, \end{cases}$$

for $\text{j} = 200,201,\cdots,500$.

1.2. Binary classification using neural networks

1.2.1. Model

We used a one-hidden-layer neural network (with 50 neurons in the hidden layer) as the model, given as

$$Z = W_2\sigma(W_1X + b_1) + b_2,$$

where $X \in \mathbb{R}^{301}$ is the input intensity, $W_1 \in \mathbb{R}^{50\times301}$ and $W_2 \in \mathbb{R}^{2\times50}$ are the weight matrices, $b_1 \in \mathbb{R}^{50}$ and $b_2 \in \mathbb{R}^2$ are the bias vectors, and σ is the sigmoid activation function. The output of the neural network $Z = [z_0, z_1]^T$ is then pass into a softmax function to obtain a single value p as

$$p = \frac{e^{z_0}}{e^{z_0} + e^{z_1}}.$$

One should note that p is a value between 0 and 1 that presents the probabilities of the input intensity X being in class 0. The probability of X being in class 1 is simply $1 - p$.

1.2.2. Training

We used the cross-entropy loss function:

$$\text{Loss} = \frac{1}{N}\sum_{j=1}^{N}\left(t^j \log\left(p^j\right) + \left(1 - t^j\right)\log\left(1 - p^j\right)\right),$$

where N is the number of training data, $t^j \in \{0,1\}$ and p^j are the ground truth label and the model prediction, respectively, for the jth data. The training process is accomplished by Adam algorithm.

1.3. Validation

To verify the robustness of the proposed algorithm, we applied the repeated random sub-sampling validation; we then randomly took four-fifths of the sample as the training set and the rest as the testing set in each experiment. We validated the accuracy of the results by averaging over 20 experiments.

1.4. Feature importance

After a model was trained, we used Deep SHAP [Lundberg, Scott M and Lee, Su-In, A Unified Approach to Interpreting Model Predictions, NIPS (2017), pp.4765–4774] to find the important features. The mean absolute SHAP value showed the importance of each feature, while the sign of SHAP value indicated which class the feature was most important for.

1.5. Validation of the important features

Using the results of the Deep SHAP analysis shown in Figure 3, we identified the important features according to their mean absolute SHAP values. To validate the effectiveness of the found features, we constructed again a one-hidden-layer neural network model but with its input to be some of those found features, not the whole intensity data. Table S1 lists the results. With just 8 feature inputs, we achieved a high accuracy of 99.44, which validated the validity of the features found.

Table S1. Testing accuracy of the model.

selected feature inputs	accuracy (%)
463,405,441,421,347,409,363	99.31
463,405,441,421,347,409,363,479	99.44

2. Additional Figures

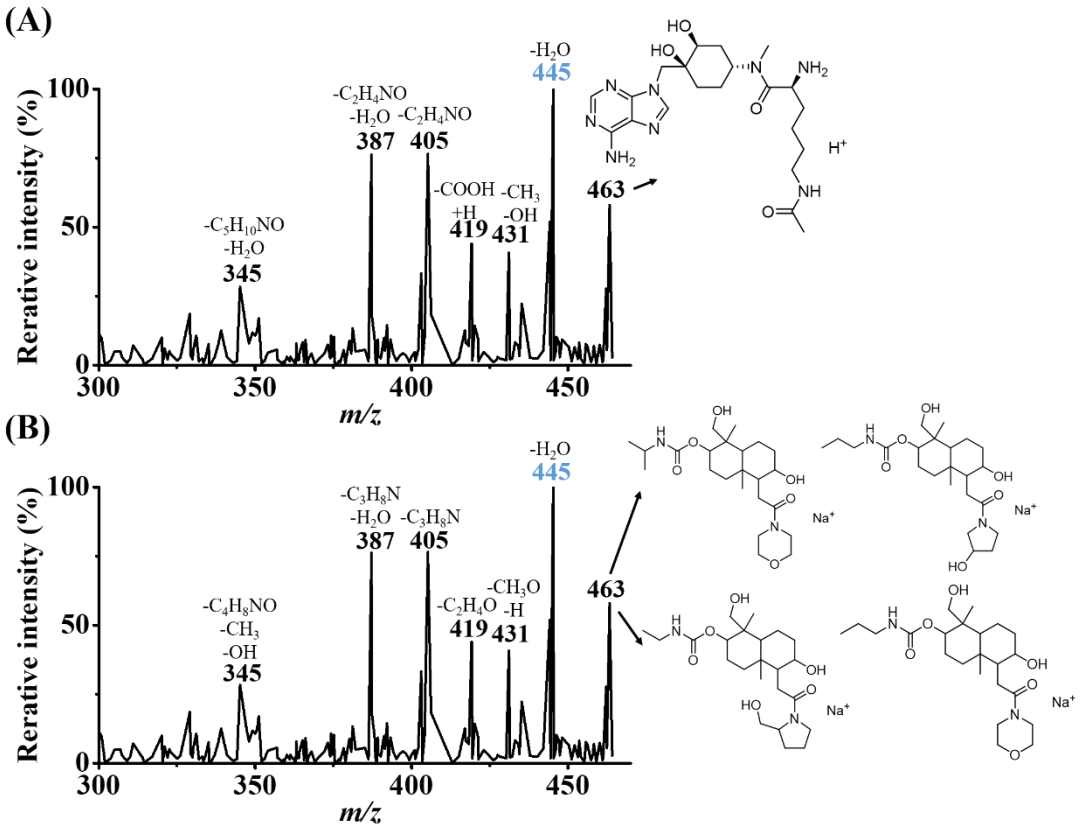


Fig. S2. (A)-(B) MS/MS spectra of the feature at *m/z* 463 derived from palm civet coffee beans and its possible chemical structures.

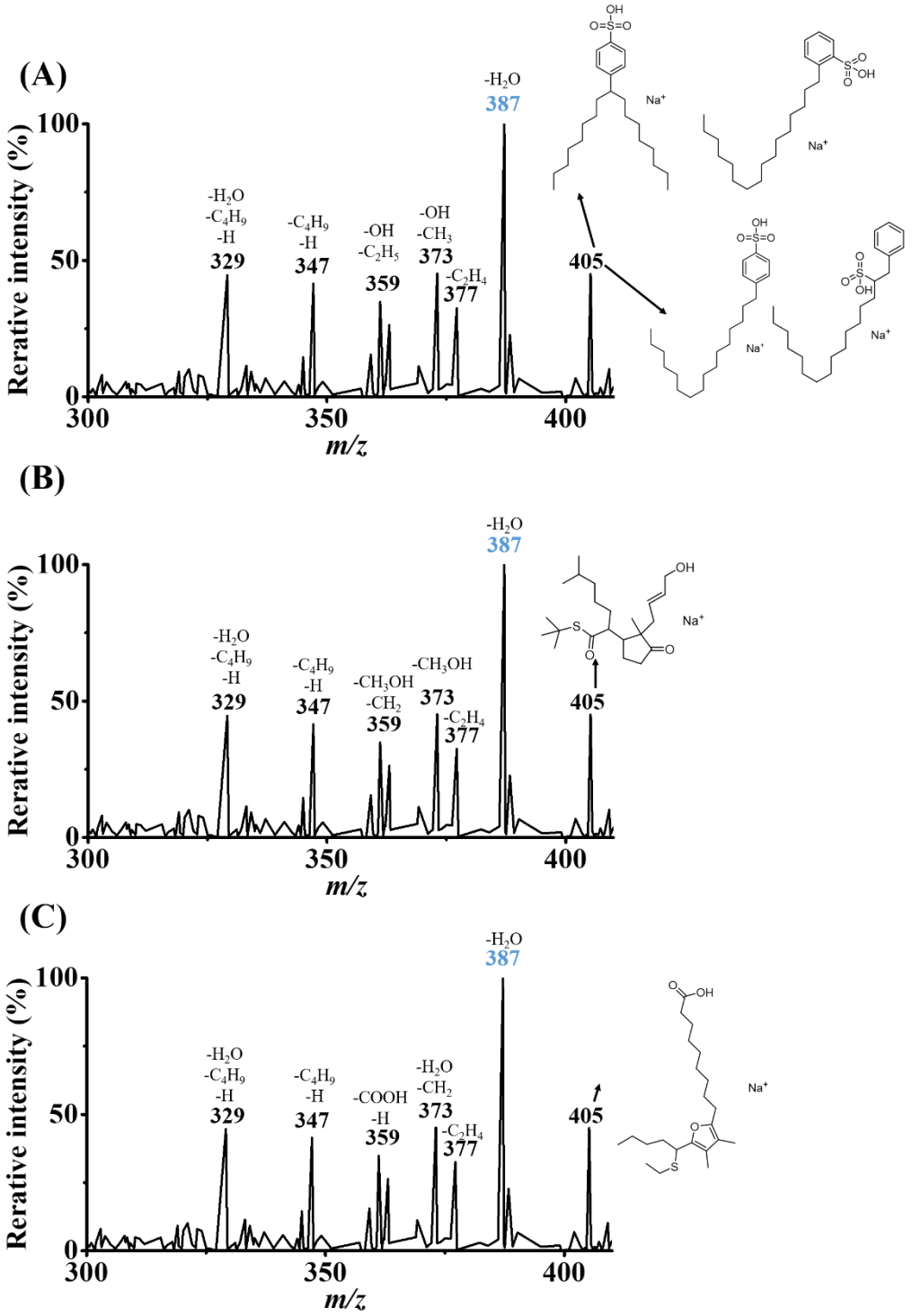


Fig. S3. (A)-(C) MS/MS spectra of the feature at *m/z* 405 derived from palm civet coffee beans and its possible chemical structures.

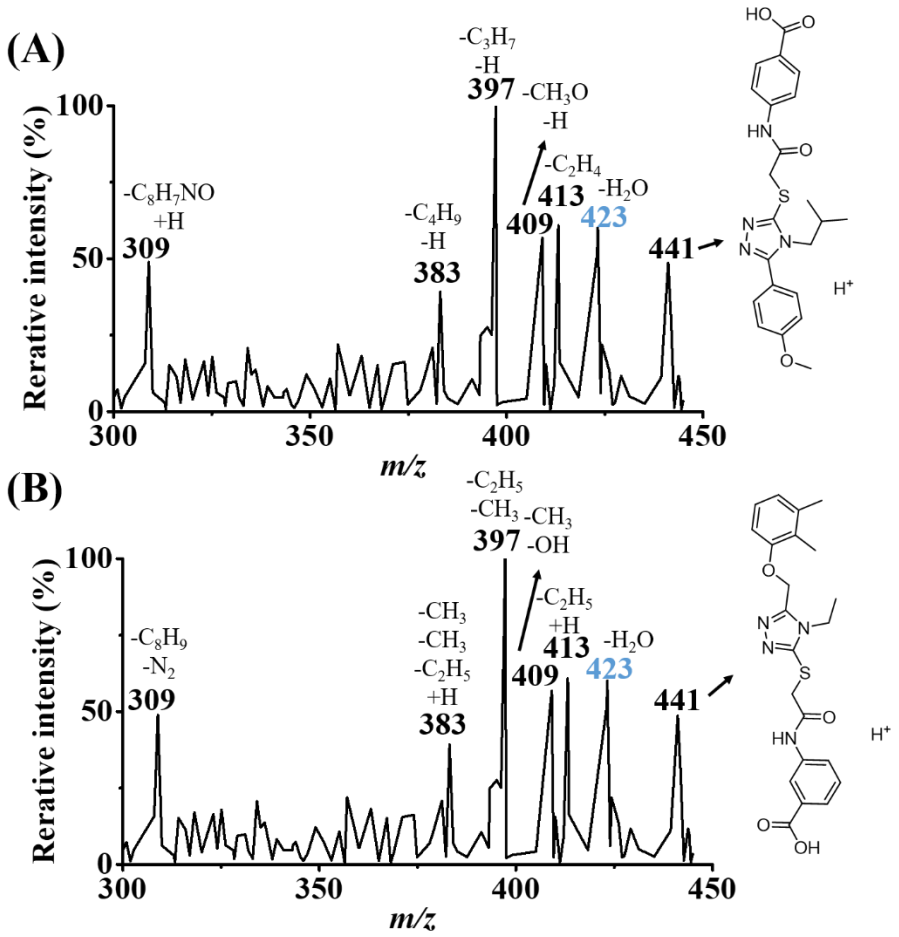


Fig. S4. (A)-(B) MS/MS spectra of the feature *m/z* 441 derived from palm civet coffee beans and its possible chemical structures.

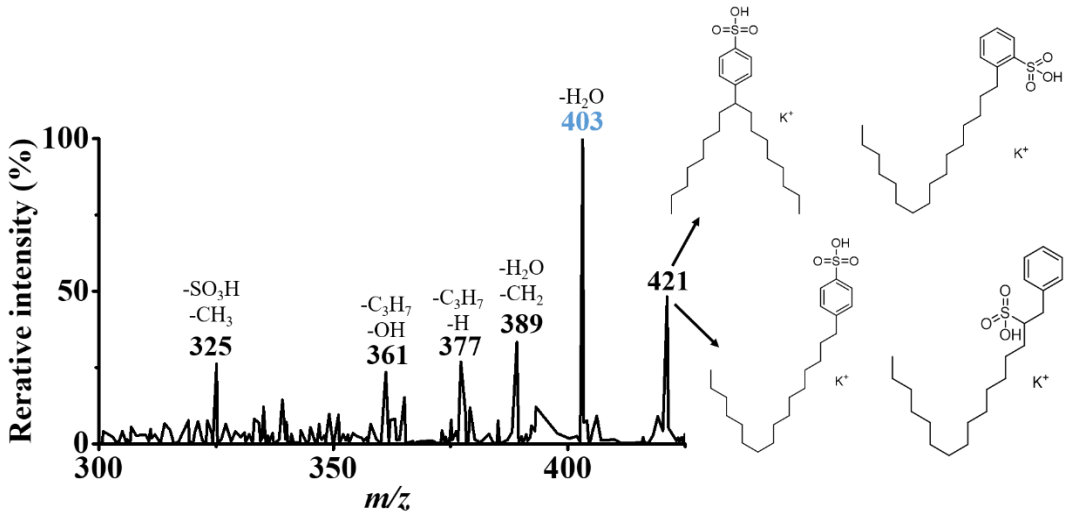


Fig. S5. MS/MS spectrum of the feature *m/z* 421 derived from palm civet coffee beans and its possible chemical structures.

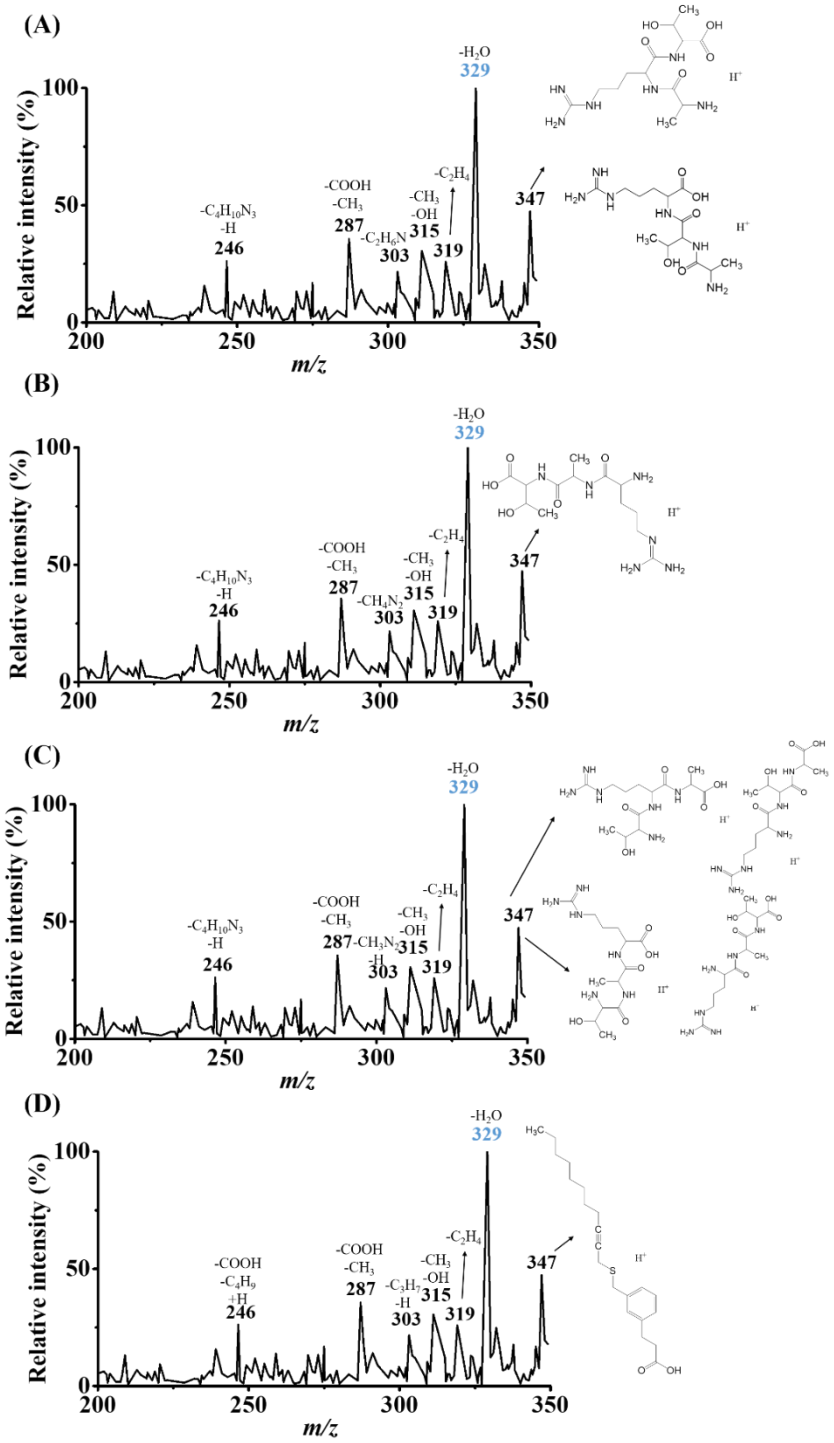


Fig. S6. (A)-(D) MS/MS spectra of the feature at *m/z* 347 derived from palm civet coffee beans and its possible chemical structures.

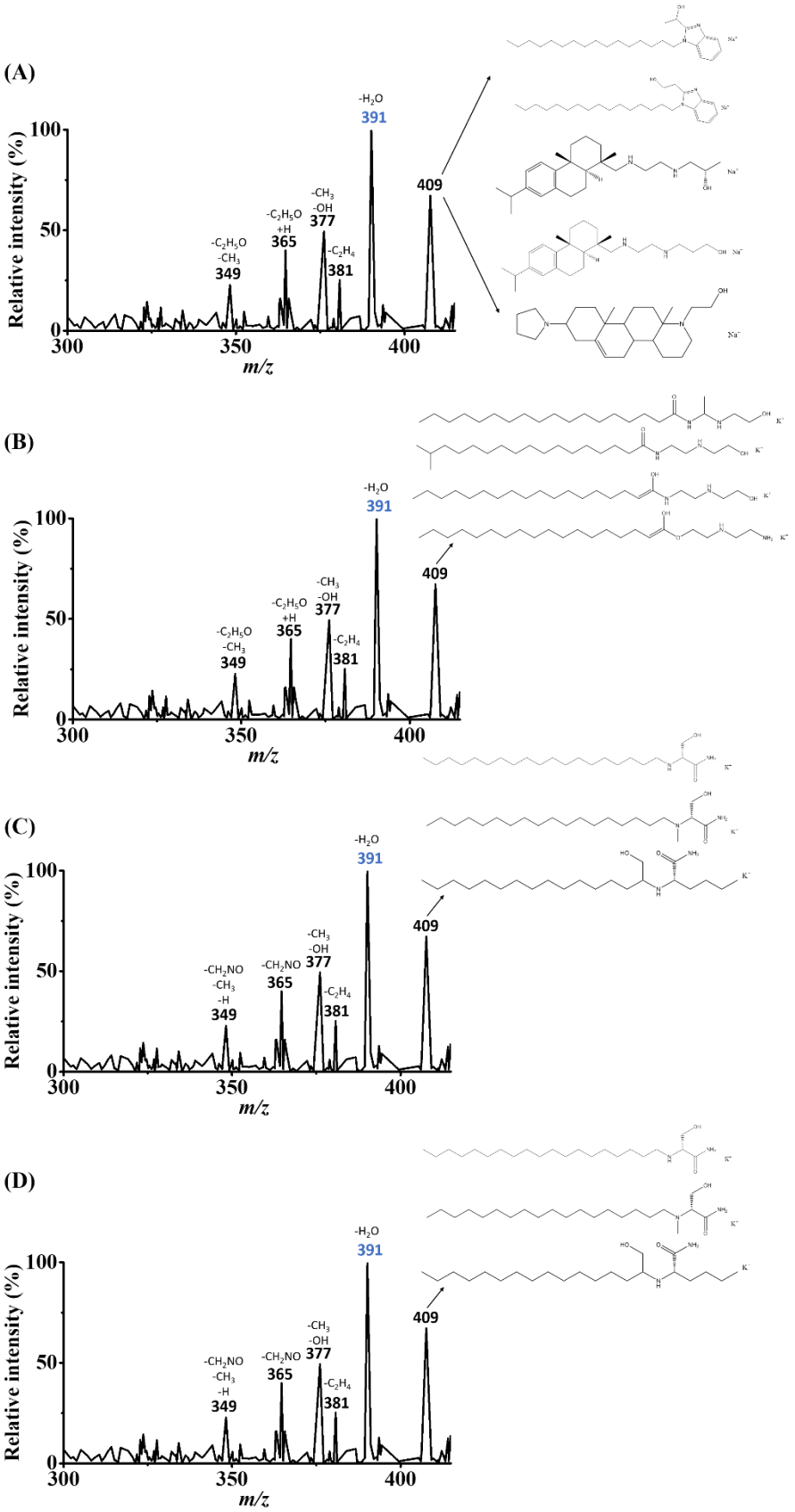


Fig. S7. (A)-(D) MS/MS spectra of the feature at *m/z* 409 derived from palm civet coffee beans and its possible chemical structures.

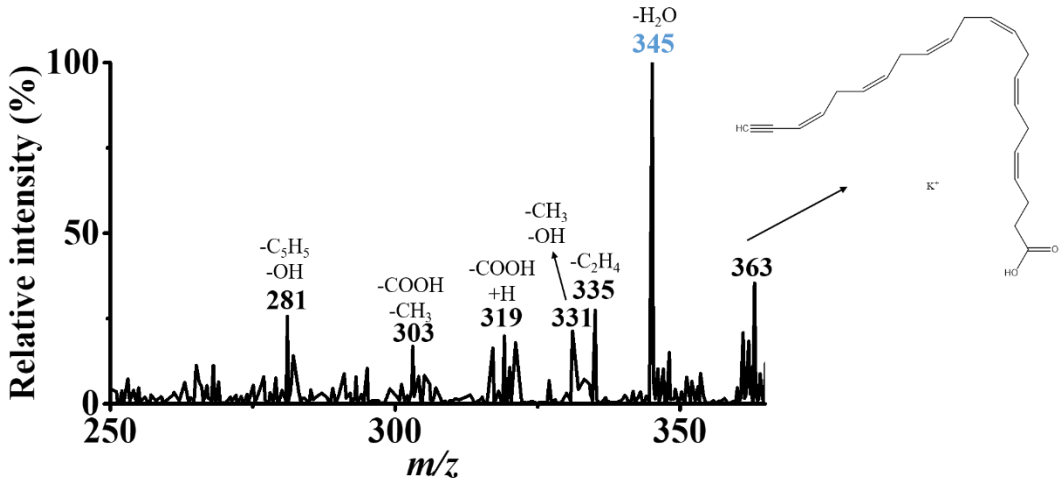


Fig. S8. MS/MS spectrum of the feature at *m/z* 363 derived from palm civet coffee beans and its possible chemical structure.