npj Science of Food

https://doi.org/10.1038/s41538-026-00748-0

Article in Press

Tracing origin and cultivation practice of *Lithocarpus litseifolius* via multi-data fusion and machine learning approaches

Received: 5 November 2025

Accepted: 2 February 2026

Cite this article as: Tang, Y., Yu, P., Xiong, F. *et al.* Tracing origin and cultivation practice of *Lithocarpus litseifolius* via multi-data fusion and machine learning approaches. *npj Sci Food* (2026). https://doi.org/10.1038/s41538-026-00748-0

Yifan Tang, Ping Yu, Feng Xiong, Zhilai Zhan, Kai Xie, Shuyan Yu, Yifan Ning, Zhanhan Zhou, Chun Wang, Weisen Qian, Xiwen Zhang, Yike Liang, Ruijiao Wang, Guoxia Han & Jian Yang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Tracing origin and cultivation practice of *Lithocarpus litseifolius* via multi-data fusion and machine learning approaches

Yifan Tang[1,5,7]*, Ping Yu[2,3,4,7], Feng Xiong[2,4], Zhilai Zhan[2,4], Kai Xie[6], Shuyan Yu[1], Yifan Ning[1], Zhanhan Zhou[1], Chun Wang[1], Weisen Qian[1], Xiwen Zhang[1], Yike Liang[1], Ruijiao Wang[1], Guoxia Han[1*], Jian Yang[2,4]*

[1] *Academy of Pharmacy, Xi'an-Jiaotong Liverpool University, Suzhou, 215123, China*

[2] *State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, PR China*

[3] *State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, PR China*

[4] *Key Laboratory of Biology and Cultivation of Herb Medicine, Ministry of Agriculture and Rural Affairs, Beijing, 100700, PR China*

[5] *Laboratory for Breeding and Processing of Lithocarpus litseifolius, Liangtian Biotech Jiangsu Co., Ltd, Changzhou, 213155, China*

[6] *Department of Thoracic and Cardiovascular Surgery, The Fourth Affiliated Hospital of Soochow University, Suzhou Dushu Lake Hospital, Medical Center of Soochow University, Suzhou, 215000, China*

[7] *These authors contributed equally to this article: Yifan Tang, Ping Yu*

* Corresponding Author:
Jian Yang, yangchem2012@163.com;
Guoxia Han, Guoxia.Han@xjtlu.edu.cn;
Yifan Tang, yifan.tang@163.com

Abstract

Lithocarpus litseifolius (sweet tea) is a medicinal and edible plant rich in flavonoids and essential nutrients, with potential as a hepatoprotective beverages and natural sweetener. Although widely cultivated across several provinces in China, the quality and consistency of its raw material remain poorly regulated. To address this, 163 samples (n ≥ 18) from 7 main producing regions were analyzed for 22 functional compounds, 4 stable isotope ratios, and 49 multi-element to discriminate cultivation practices and geographical origins. Orthogonal partial least squares discriminant analysis (OPLS-DA) successfully generated prediction models across two cultivation regions. Integrating 8 machine-learning algorithms with multi-level data fusion identified 6 key variables—caffeine, Rb, Ce, $\delta^{15}N$, Sr, and 3"-O-acetylphlorizin. Five base learners built on these variables were then combined via soft-voting ensemble learning, yielding an optimal origin classifier with 100.00% accuracy. Additionally, the study delivered the first comprehensive analysis of quality variations in sweet tea and identified seven primary influenced environmental factors, offering insights into cultivation strategies and quality formation mechanisms.

Keywords: *Lithocarpus litseifolius* (sweet tea), Geographical origin, Cultivation practice, Functional compounds, Stable isotope ratio, Multi-element, Environmental factors.

Introduction

Lithocarpus litseifolius (commonly known as “sweet tea”), a newly popular tea beverage in China and a member of the *Fagaceae* family, has gained significant attention for its distinctive bioactive and nutritional properties [1,2]. The species has a long history of use as both food and medicine, first documented during the Northern and Southern Dynasties (AD 423) [3]. At Jinyun Mountain in Chongqing, local monks have harvested wild sweet tea since the temple’s establishment, crafting it into a highly prized beverage long celebrated by visitors for its good taste [3]. Historical records in classical pharmacopoeias such as the *Dictionary of Chinese Materia Medica* (*Zhong Hua Ben Cao*) describe the health benefits of long-term consumption of sweet tea [4]. Over centuries, sweet tea has also become deeply embedded in the tea culture of ethnic communities across Hunan, Jiangxi, Yunnan, and Guizhou, where it is traditionally consumed as “Immortal Tree” [4]. Meanwhile, its derivatives—including sweet tea pastries and candies—are popular regional specialties. Modern studies have revealed that its leaves are rich in dihydrochalcones, particularly phloridzin and trilobatin, which exhibit notable antioxidant activity and potential metabolic health benefits [2]. These compounds also serve as abundant natural sweeteners, with a sweetness approximately 300 times greater than that of sucrose and only one three-hundred of its caloric content [5,6]. Owing to these properties, sweet tea, often referred to as the “Cordyceps sinensis on a tree” and “The best tea under heaven”, was officially approved as a new food ingredient by the Chinese National Health Commission in 2017, and its extracts subsequently entered clinical trials for type 2 diabetes management in 2025 [4,7]. In recent years, the diversification of downstream products and growing pharmaceutical interest have driven a sharp rise in market demand and prices. Despite this promising growth, the raw-material supply chain remains fragmented. Geographic labeling, quality grading, and pricing are often inconsistent, and the absence of standardized quality criteria has enabled fraudulent practices such as origin mislabeling, adulteration, and confusion between cultivated and wild sources. These deficiencies undermine product authenticity and erode consumer trust, underscoring the need for robust, verifiable systems to authenticate geographical origin and to discriminate cultivation practices.

In recent years, targeted metabolomics combined with machine learning has been increasingly used to determine the geographical origin of foods and agricultural products. Several analytical techniques, such as high-performance liquid chromatography (HPLC), gas chromatography–mass spectrometry (GC–MS), and ultra-performance liquid chromatography–mass spectrometry (UPLC–MS), have been developed for this purpose. For example, Bajoub *et al.* effectively traced the geographical origin of “Picholine Marocaine” olives from seven regions across Morocco by quantifying 25 phenolic compounds [8]. However, these methods primarily focus on bioactive molecules such as flavonoids, whose content is highly susceptible to annual climate, agronomic practices, and plant physiological states, exhibiting significant dynamic fluctuations. This results in single-metabolome-based traceability models potentially overlooking relatively stable "static" chemical variations shaped by regional geological history and long-term environmental stresses, which are precisely the core elements that form difficult-to-counterfeit geographical fingerprints. Meanwhile, stable isotope ratios and multi-element profiles, based on geochemical principles, have also been widely recognized as reliable indicators for origin authentication. These methods have been successfully applied to various products, including Slovenian strawberries, green coffee beans and *Tetrastigma hemsleyanum* [9-11]. Particularly, the distribution patterns of rare elements and geochemical indicators such as Sr directly reflect the parent rock composition and long-term weathering processes of the soil at the origin, forming a unique and stable imprint akin to "geological DNA," significantly increasing the difficulty of forgery [12,13]. Moreover, these methods exhibit limited resolution at small geographical scales, show weak sensitivity to cultivation practices, and provide little information related to product quality, such as sensory attributes or bioactive components. Consequently, relying on a single analytical source is often insufficient to capture the full complexity of chemical information, resulting in weaker sample differentiation, reduced model robustness, and lower classification accuracy.

To date, integrating multi-source data has therefore emerged as a promising strategy [14]. For instance, combining functional compound profiles with elemental fingerprints in *Angelica sinensis* achieved 100.00% accuracy in origin discrimination and enabled prediction of key bioactive components [15]. Similarly, fusing stable isotope, elemental, and starch composition data with ensemble learning in *Euryale ferox Salisb.* identified ten key markers and established a highly reliable origin-tracing model [16]. Such multi-source data fusion approaches, particularly when coupled with advanced machine learning algorithms, can address the challenges of high-dimensional datasets and exploit the complementary strengths of different analytical platforms. However, studies applying these multivariate data fusion methods to sweet tea remain scarce. In addition, the variables contributing most to geographical differentiation and model prediction accuracy have not yet been identified. Evaluating the relative importance of variables such as functional compounds, stable isotope ratios (SIRs), and multi-element profiles is therefore essential. Moreover, these characteristics in sweet tea are strongly influenced by environmental factors [17]. Further research is needed to elucidate how environmental drivers shape the chemical composition of sweet tea.

Hence, this study pursued the following objectives: 1) Comprehensively characterize the functional compounds, stable isotope ratios, and multi-element of sweet tea from different origins in China and build a multi-source database; 2) Discriminate cultivation practices of sweet tea and identify key agricultural factors; 3) Develop origin identification models by integrating functional compounds, stable isotope ratios, and multi-element through multi-source data fusion algorithms, and identify key discriminative variables; 4) Investigate the influence of environmental factors on key variables across different sweet tea.

Results and discussion

The analysis of functional compounds, stable isotope ratio and multi-element in sweet tea leaves

22 functional compounds (Dataset I) were used to analyze the quality differences of sweet tea

leaves from different origins. These included 3 nutritional indicators (AA, CAF, and Tp), 8 characteristic quality compounds of sweet tea leaves (Phz, Trb, Hpz_3, Gpt_2, Oapz_3, Oapz_2, Gua, and Pht), and 11 organic acids (PA, DA, GA, Dih, Esc, Tar, Cit, Fum, SuA, CA and THBA). Additionally, 4 SIRs (δ^2H, δ^{18}O, δ^{13}C, and δ^{15}N) and 49 elements (Na, Mg, Al, K, Ca, Ti, Mn, Fe, Zn, Ba, Li, Be, B, P, S, V, Cr, Co, Ni, Cu, Ga, As, Se, Rb, Sr, Y, Mo, Ag, Cd, In, Sb, Cs, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Tl, Pb, and Bi) (Dataset II) were combined to predict. The detailed contents of functional compounds, SIRs, and multi-element are provided in Tab. S3-S5.

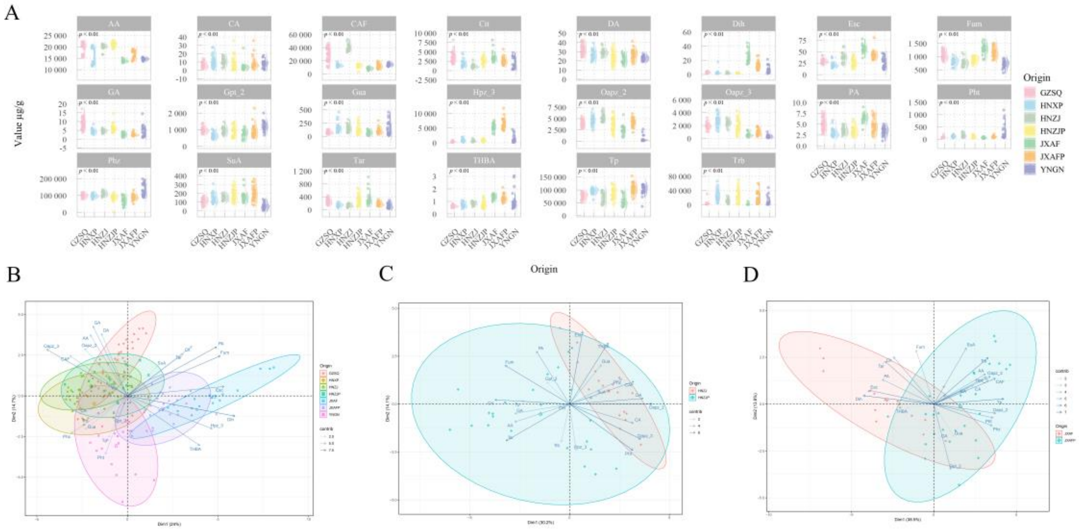

Fig. 1 Boxplots of the 22 functional compounds in sweet tea from different origins (A); PCA analyses of the samples based on functional compounds, including all origins (B), HNZJ versus HNZJP (C), and JXAF versus JXAFP (D). In plot B-D: Three PCAs were plotted along with their variable contribution. The variables are represented by arrows, with the direction, color and length of the arrows reflecting each variable's contribution to the respective principal components.

The functional compound of sweet tea leaves from different origins

Significant differences were observed in the concentrations of 22 functional compounds across the seven production regions ($p < 0.01$, Welch's test; Fig. 1A; Tab. S3). Among the three nutritional indices, AA showed the highest levels in HNZJP (21095.8 ± 1098.8 μg/g), while CAF and Tp were most abundant in HNZJ (40612.6 ± 5671.9 μg/g) and JXAFP (105293.8 ± 19746.7 μg/g), respectively. The geographical variation in secondary metabolites of sweet tea not only represents a chemical adaptation to stressful environments but may also reflect stable chemotype differentiation formed under phylogenetic constraints among distinct evolutionary lineages. This study focuses on dihydrochalcone compounds Phz and Trb, whose biosynthetic pathways may represent key traits evolved in specific *Lithocarpus* lineages [1]. The extremely high Phz accumulation observed in the YNGN population (135,166.6 ± 37,246.9 μg/g), together with the high Trb accumulation in HNZJP, likely represents two genetically distinct chemotypes. The YNGN chemotype is plausibly linked to the long-term evolutionary history of this region as a major refugium during the Quaternary glaciation. Owing to its complex topography, the Yunnan–Guizhou Plateau provided stable microenvironments during glacial periods, allowing many species to persist in geographic isolation and evolve independently [18,19]. Similarly, *Lithocarpus litseifolius* populations in this region may have experienced prolonged isolation, leading to the genetic fixation of metabolic pathways favoring high Phz synthesis. Consequently, the distinctive dihydrochalcone profile of the YNGN population is more likely to reflect long-term evolutionary history rather than direct responses to present-day climatic conditions, supporting the view that chemical diversity is shaped by phylogeographic processes [20]. In contrast, wild populations from non-refugial regions showed consistently low Trb contents (HNZJ: 10,510.4 ± 6,439.6 μg/g; JXAF: 1,945.8 ± 2,413.7 μg/g), which may reflect genetic constraints associated with populations established through post-glacial expansion from refugia [18]. Notably, all cultivated samples in this study (HNZJP and JXAFP) were derived directly from their corresponding local wild populations, with no interregional germplasm exchange and an average tree age of approximately five years. Under cultivation, Trb concentrations increased markedly by 2–10 fold (HNZJP: 23,580.4 ± 20,539.2 μg/g; JXAFP: 18,218.6 ± 16,944.4 μg/g), clearly demonstrating that cultivation practices can strongly enhance Trb biosynthesis even under the same genetic background [21]. These results indicate that the high-Trb chemotypes observed in cultivated populations primarily reflect the activation of inherent metabolic potential under optimized growing conditions. Therefore, the geographical variation of Trb arises from the combined influence of genetic background and environmental induction. Given that Trb is a high-value natural sweetener and that efficient synthetic production routes are currently unavailable [5,21], cultivation strategies tailored to local genetic resources offer a practical and traceable approach to enhancing the industrial value of regional sweet tea. Hpz_3 levels were significantly higher in JXAF (5345.9 ± 1321.1 μg/g) and JXAFP (6517.7 ± 1748.0 μg/g). Gpt_2 reached its highest concentration in YNGN (1227.3 ± 299.3 μg/g). Oapz_3 was significantly elevated in HNXP (2781.9 ± 928.0 μg/g), while Oapz_2 peaked in HNZJ (5233.4 ± 930.2 μg/g). Notably, both Oapz_3 (308.4 ± 240.6 μg/g) and Oapz_2 (541.4 ± 723.2 μg/g) showed the lowest levels in YNGN. Gua content was highest in HNXP (172.8 ± 98.5 μg/g). Pht, a key downstream metabolite of Phz and Trb [5,22,23], was also significantly higher in YNGN (291.5 ± 343.4 μg/g). Organic acids, amino acids, total phenols, and flavonoids are known to strongly influence the sensory quality of tea [24]. Among the 11 organic acids analyzed, DA and GA were significantly higher in GZSQ (32.2 ± 5.7 and 8.7 ± 3.2 μg/g, respectively), while SuA was most abundant in JXAFP (198.7 ± 77.9 μg/g). CA showed notably higher levels in HNXP and HNZJ. Seven other organic acids—PA, Dih, Esc, Tar, Cit, Fum, and THBA—were significantly enriched in JXAF (6.4 ± 1.2, 25.5 ± 9.3, 56.3 ± 9.7, 367.8 ± 249.7, 3831.3 ± 1622.0, 1278.3 ± 210.6, and 1.3 ± 0.2 μg/g, respectively). These pronounced regional differences in organic acid composition contribute to the distinct sensory characteristics of sweet tea from JXAF, underscoring its unique quality attributes.

The PCA plot on Dataset I showed weak regional separation. The first two components explained 38.7% of total variance. Within-province separation was minimal. HNXP, HNZJ and HNZJP overlapped extensively (Fig. 1B). Discrimination by cultivation practice within region was limited. All HNZJP points fell inside the HNZJ confidence ellipse. JXAF showed slightly greater separation from JXAFP, for that pair the first two components explained 52.3% of variance. Separation remained unclear (Fig. 1C, D). These findings indicate a need for richer data dimensions. Stronger machine learning approaches are required to enhance classification accuracy.

The stable isotope ratios and multi-element of sweet tea leaves from different origins

Similarly, as shown in Fig. 2, significant differences ($p < 0.01$, Welch test) were observed in 4 SIRs and 49 elemental concentrations in sweet tea leaves across the seven production regions. Regarding SIRs, the $\delta^{13}C$ values of sweet tea samples ranged from −33.550‰ to −26.877‰. Based on its taxonomic classification within the genus *Lithocarpus* and the C3 photosynthetic characteristics of *Fagaceae*, as well as its origins in humid subtropical montane forests , this species is considered a typical C3 plant [2]. The selective pressure exerted by environmental factors plays a crucial role in influencing the $\delta^{13}C$ values of sweet tea populations—higher temperatures and lower light conditions are associated with more negative $\delta^{13}C$ values [25]. Among them, the JXAF sample site has the lowest average elevation (148 m), the shortest annual sunshine duration (1,514 hours), and the highest average annual temperature (18.4°C), and its $\delta^{13}C$ value is also the most negative (-31.68 ± 0.951‰). The $\delta^{13}C$ values of wild sweet tea samples from Hunan and Jiangxi are significantly higher than those of cultivated samples (HNZJP: -28.03 ± 0.516‰ vs. HNZJ: -29.32 ± 0.767‰; JXAF: -31.68 ± 0.951‰ vs. JXAFP: -29.24 ± 0.547‰), indicating that cultivation practices such as fertilization optimize the photosynthetic structure and stomatal conductance of sweet tea by altering soil fertility [26], ultimately improving the water use efficiency (WUE) of high-Trb potential chemotype populations—a positive phenotypic plasticity response [27]. At the same time, the results also reflect that specific chemotypes exhibit different water use strategies. The $\delta^{13}C$ values of YNGN show no significant difference from those of the HNZJ population, despite the stark differences in their hydrothermal conditions (Fig. 2A; Tab. S1). This suggests that the dominant factor in their WUE differences may not be contemporary climate but rather a deeper genetic background that determines carbon-water balance strategies. This inference further supports the notion that the formation of the high-Phz chemotype in YNGN is a deep adaptation to historical stress environments. Plant $\delta^{15}N$ values responded significantly to fertilizer application frequency and type [28] and were closely associated with the biosynthesis of Phz, the key bioactive compound in sweet tea [26]. The highest $\delta^{15}N$ value was observed in the HNZJP sample (4.067 ± 1.850‰), which received two equal applications of base and topdressing fertilizers annually. This value was significantly higher than that of wild samples from the same region (HNZJ: 0.378 ± 1.502‰). In contrast, the JXAFP sample, fertilized only once per year, showed only a slight increase in $\delta^{15}N$ (−0.336 ± 1.947‰) compared to JXAF (−1.569 ± 1.185‰). The lowest $\delta^{15}N$ value was recorded at the GZSQ (−4.100 ± 0.841‰). The $\delta^{2}H$ and $\delta^{18}O$ values primarily reflect regional hydrological and environmental conditions and are independent of cultivation practices [29]. The highest $\delta^{2}H$ value was observed in JX (−82.95 ± 13.41‰), which was significantly higher than that in HN (−106.1‰ to −103.1‰), mainly due to its lower altitude, higher temperature, and greater precipitation. Unexpectedly high $\delta^{2}H$ values were recorded in YNGN (−89.87 ± 7.801‰) and GZSQ (−75.74 ± 5.498‰), despite their higher elevations and lower rainfall. This anomaly suggests a predominant influence of soil groundwater as the primary water source [29]. The spatial distribution of $\delta^{18}O$ resembled that of $\delta^{2}H$ but exhibited weaker regional variation. The highest $\delta^{18}O$ value was found in GZSQ (25.08 ± 0.739‰).

The beneficial elements (BEs) detected included Na, V, Se, Co, Ti, and Ni, while the essential elements (EEs) consisted of Mg, B, K, Mo, S, Mn, Fe, P, Zn, and Ca (Fig. 2B). Among the EEs, K, Ca, and Mg were the most abundant, with the highest levels observed in JXAF (12615.9 ± 1880.6 mg/kg), YNGN (9591.0 ± 7645.5 mg/kg), and HNZJ (1836.5 ± 409.8 mg/kg), respectively. Se (0.030–0.310 mg/kg) and Mo (0.016–0.063 mg/kg) showed the lowest concentrations among BEs and EEs, respectively. Samples from the red soil region [30] of JX exhibited significantly higher

Fe and Mn contents than those from other regions. The Fe content in JXAF reached 436.6 ± 119.7 mg/kg, twice that of JXAFP. A total of eight heavy metal elements (HEs) included Ag, As, Cd, Cr, Cu, Ga, Pb, and Sb were analyzed. Cu showed the highest concentration (6.9–10.9 mg/kg) among HEs but did not reach the level indicative of a hyperaccumulator [31]. It was followed by Pb (0.5 ± 0.3 – 4.1 ± 1.2 mg/kg) and Cr (0.6 ± 0.3 – 2.7 ± 0.8 mg/kg). Sb levels were consistently low across all regions, with YNGN showing the lowest value (0.049 ± 0.025 mg/kg). Samples from JX had significantly higher As, Ga, Pb, and Ag levels, with Pb in JXAF (4.1 ± 1.2 mg/kg) exceeding that of other provinces by 4-fold. However, this was not associated with significant suppression in the contents of Phz and Trb (Fig. 1). According to the US EPA safety thresholds for heavy metal concentrations [32], the average levels of all HEs in sweet tea leaves from every region were below the limits, reflecting both geological characteristics and indicating that sweet tea is a safe raw material with low heavy metal accumulation potential. 15 rare earth elements (REs) were detected, the distribution of which is largely determined by local ore deposits [33]. The highest total RE content was found in JXAFP, followed by JXAF. The RE content in sweet tea was dominated by Ce (0.070 ± 0.022 – 2.744 ± 0.913 mg/kg), La (0.060 ± 0.027 – 3.568 ± 2.357 mg/kg), Nd (0.041 ± 0.015 – 2.825 ± 1.815 mg/kg), and Y (0.035 ± 0.017 – 1.260 ± 0.603 mg/kg), while the remaining REs all showed maximum concentrations below 1 mg/kg. Results for other elements (OEs) are also shown in Fig. 2B. Rb exhibited the greatest variation among regions (11.9 ± 4.4 – 71.1 ± 19.8 mg/kg), with the highest level detected in JXAFP.

The results of the PCA based on Dataset II are presented in Fig. 2C. The first two principal components collectively explained 56% of the total variance. Samples from the same region showed highly overlapping distributions, indicating that PCA alone was unable to clearly discriminate sweet tea samples from different geographical origins. Similarly, the model exhibited limited ability to distinguish between agricultural practices (Fig. 2D, E). Therefore, further application of machine learning methods, integrated with features from additional dimensions, is necessary to construct an effective origin traceability model for sweet tea.

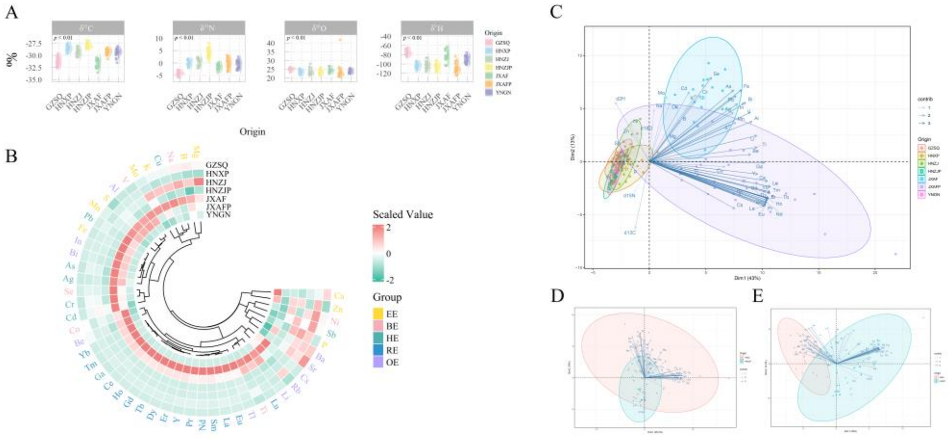

Fig. 2 Boxplots of the 4 SIRs in sweet tea from different origins (A); Cluster heatmap of 49 multi-element from different origins (B); PCA analyses based on SIRs and multi-element profiles, including all origins (C), HNZJ

versus HNZJP (D), and JXAF versus JXAFP (E). In plot B: The color gradient in the heatmap ranges from green (-2) to red (2), indicating normalized values for the relative abundance level of the factors. The shade of the color reflects the magnitude of the value, where green represents lower values and red represents higher values. In plot C-E: Three PCAs were plotted along with their variable contribution. The variables are represented by arrows, with the direction, color and length of the arrows reflecting each variable's contribution to the respective principal components.

Discrimination of sweet tea cultivation practices and identification of key agricultural factors

Therefore, OPLS-DA was employed to further discriminate sweet tea leaf samples from different cultivation practices. Generally, a Q^2(cum) value above 0.5 indicates good predictive ability, while a value exceeding 0.9 is considered excellent. Meanwhile, R^2Y should ideally be close to 1 [34]. As shown in Fig. 3A, the OPLS-DA model distinguishing HNZJ and HNZJP demonstrated excellent performance, with a Q^2(cum) of 0.895 (excellent predictive ability) and R^2Y of 0.972 (excellent explanatory power), achieving clear separation between the two groups. Similarly, the model discriminating JXAF and JXAFP also performed well, with a Q^2(cum) of 0.601 (good predictive ability) and R^2Y of 0.901 (excellent explanatory power) (Fig. 3D). To evaluate potential overfitting, 200 permutation tests were conducted for each comparison. In all cases, the regression line of R^2 remained above zero and was largely higher than that of Q^2, supporting the robustness of the model (Fig. 3C, F). A total of seven features with VIP > 1 were identified in each model (Fig. 3B, E). Most of these were functional compounds, with K being the only elemental feature ($\text{VIP}_{\text{HNZJ}} = 1.56$; $\text{VIP}_{\text{JXAF}} = 1.33$). This suggests that multi-element and SIRs are less affected by cultivation practices, whereas functional compounds show significant improvement. Key variables for cultivation practice prediction included CAF, Phz, Tp, Trb, and Oapz_2 in both regions. Additionally, Cit was important for distinguishing HNZJP, and AA contributed notably in JXAFP. Fertilization significantly influenced the levels of Phz ($\text{VIP}_{\text{HNZJ}} = 3.96$; $\text{VIP}_{\text{JXAF}} = 3.72$) and Trb ($\text{VIP}_{\text{HNZJ}} = 3.72$; $\text{VIP}_{\text{JXAF}} = 2.75$). Interestingly, Phz responded in opposite trends between HNZJ and JXAF under fertilization, whereas Trb consistently increased across both regions. This indicates that fertilization may affect Phz and Trb through distinct mechanisms and to varying extents. In summary, OPLS-DA effectively differentiated sweet tea samples from different cultivation practices within the same geographic origin. The model captures the metabolic plasticity induced by cultivation practices under the same genetic background, rather than artifacts caused by germplasm admixture.

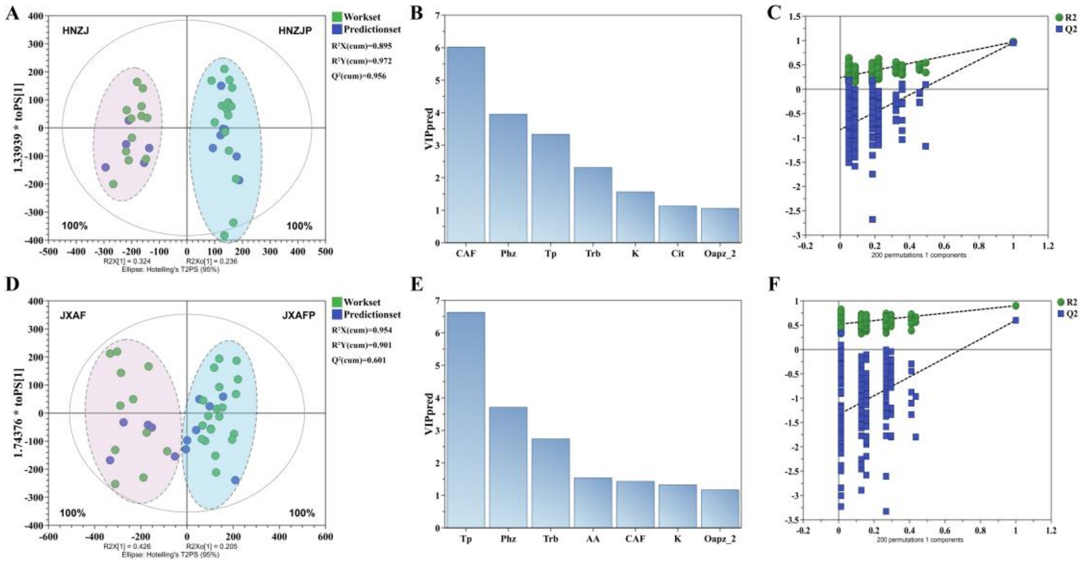

Fig. 3 OPLS-DA results of cultivation practice prediction for samples from Hunan and Jiangxi regions based on functional compounds, SIRs, and multi-element data. Training and prediction results (A), VIP score in prediction set (B) and permutation test (C) of HNZJ with HNZJP; Training and prediction results (D), VIP score in prediction set (E) and permutation test (F) of JXAF with JXAFP. In plot A and D: The dashed ellipses in this figure are not 95% confidence ellipses. They are only used to illustrate the classification of samples from different regions. Y-axis zero point divides the graph into left and right regions representing different geographical areas, while the dispersion along the X-axis indicates the within-group differences of the predicted samples. In plot C and F: Green dots and lines: Represent the model's explanatory power for the data (R2). The higher the R2, the better the model fits the data. Blue dots and lines: Represent the model's predictive ability (Q2). The higher the Q2, the stronger the model's predictive ability and the better its generalization.

Individual dataset and low-level fusion in origin prediction

Datasets I and II were used as inputs for machine learning modeling, with the overall workflow illustrated in Fig. 6. Model performance metrics and data partition schemes for each fusion strategy are summarized in Tab. S6. Confusion matrices for test and training accuracies are presented in Fig. 4 and Fig. S1, respectively. Models based on individual data sources yielded suboptimal results. Dataset I showed a training error of 8.86 ± 3.13%, while Dataset II showed 5.95 ± 1.84% (Tab. 1). Corresponding test errors were 3.10 ± 1.70% and 5.68 ± 2.00%, and macro F1 scores reached 0.97 ± 0.02 and 0.94 ± 0.02, respectively—values below the desired threshold for robust traceability models. Model selection prioritized macro F1 as the key indicator of overall performance, with higher scores reflecting superior accuracy and generalization ability. SVM performed best with Dataset I. Training error measured 5.82 ± 4.53%, macro F1 achieved 0.98 ± 0.02 and test error registered 1.82 ± 2.42%. Confusion matrix analysis revealed training accuracy between 80.0 ± 3.8% and 95.4 ± 1.5% for most regions. HNZJ and YNGN showed deviations, this is likely because the sweet tea from both HNZJ and YNGN exhibits high nutritional quality as well as similar sensory characteristics. JXAF suffered severe misclassification. Test errors concentrated in HNZJP and JXAF regions, the reason is the geographical proximity of the origins. Accuracy values measured 96.0 ± 8.0% and 86.7 ± 16.3% respectively. Lasso achieved optimal performance with Dataset II. Macro F1 reached 0.97 ± 0.02. Training error measured 4.43 ± 3.48%. Test error registered 3.03 ± 1.92%. Training accuracy ranged from 80.0 ± 10.5% to 98.8 ± 2.5% across most regions. GZSQ constituted an exception. Most misclassifications involved HNZJ region. Test errors primarily affected HNXP and HNZJ regions with accuracy values measured 92.0 ± 9.8% and 85.0 ± 12.2% respectively. The predictive models based on the two datasets demonstrated complementary advantages and limitations in terms of performance, highlighting the necessity of employing data fusion strategies to enhance overall robustness and accuracy. Consequently, Dataset III was constructed through low-level fusion by directly merging Datasets I and II. Employing the optimal algorithm, LightGBM, this approach achieved a macro F1 of 1.00, with test and training errors of 0.00 ± 0.00% and 3.66 ± 4.18%, respectively, effectively enhancing both performance and stability of the geographical origin prediction model. The confusion matrix revealed misclassifications across all regions except GZSQ in the training set, with accuracy rates ranging from 87.7 ± 7.8% to 99.0 ± 2.0%, where JXAF-associated errors were most pronounced. Although Dataset III - LightGBM demonstrated promising results, the extensive feature set necessitates further optimization of key features to reduce data dependency, improve model performance, and enhance economic viability for future origin prediction applications.

Tab. 1 Classification error rate (%) and macro F1 score of best models by Dataset I ~ Dataset V

Approaches	Datasets	Best model	Training set	Testing set	Macro F1 score
Functional compounds	I	SVM	5.82 ± 4.53	1.82 ± 2.42	0.98 ± 0.02
SIRs and Multi-element	II	Lasso	4.43 ± 3.48	3.03 ± 1.92	0.97 ± 0.02
Low-level fusion	III	LightGBM	3.66 ± 4.18	0.00 ± 0.00	1.00 ± 0.00
Mid-level fusion	IV	ElasticNet	0.58 ± 1.69	0.00 ± 0.00	1.00 ± 0.00
Extracted features	V	Ridge	0.37 ± 1.23	0.00 ± 0.00	1.00 ± 0.00
High-level fusion	V	Multi-models	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00

Feature extraction and mid-level fusion in origin prediction

Therefore, we employed a combined RFE-SBS approach to obtain a refined yet discriminative feature combination at relatively low computational cost. This feature extraction workflow was independently applied to Datasets I, II, and III. The screening process involved iterative feature set reduction through RFE, selecting the subset achieving minimal classification error with the fewest features as the optimal feature set (Fig. S1). Subsequently, SBS was initiated based on this optimal set, following identical iterative optimization rules to derive the final feature subset, with corresponding screening curves presented in Fig. 4B-D. From Datasets I and II, the extracted feature combination formed Dataset IV comprising 23 variables (Mg, Zn, B, P, Rb, Sr, Y, Mo, Sb, $\delta^{13}C$, $\delta^{15}N$, $\delta^{2}H$, S, CAF, AA, Tp, Hpz_3, Oapz_3, Oapz_2, DA, Tar, Cit, SuA). Meanwhile, feature extraction from Dataset III yielded Dataset V containing 6 variables (CAF, Rb, Ce, $\delta^{15}N$, Sr, Oapz_3).

In the mid-level fusion approach, ElasticNet applied to Dataset IV achieved a macro F1 of 1.00, accompanied by a training error of 0.00 ± 0.00% and the lower error (0.58 ± 1.69%), establishing it as a high-performance prediction model (Tab. 1). The corresponding confusion matrix revealed that training misclassifications occurred exclusively in the HNZJ and JXAF regions, with accuracy rates of 98.6 ± 2.9% and 95.4 ± 6.2%, respectively. Notably, the test set demonstrated perfect classification accuracy, suggesting a potential tendency toward overfitting. For Dataset V, Ridge regression similarly attained a macro F1 of 1.00 and a test error of 0.00 ± 0.00%, while exhibiting an even lower training error (0.37 ± 1.23%) compared to the Dataset IV - ElasticNet model, thereby emerging as the top-performing approach within the mid-level fusion framework. The Dataset V - Ridge confusion matrix indicated training misclassifications solely in the HNZJ and JXAFP regions, with respective accuracy rates of 97.1 ± 3.5% and 99.1 ± 1.8%. Although the obtained performance metrics are already exceptional, advanced fusion strategies such as integrating multiple algorithm outputs through voting systems could potentially reduce the training error to zero, thereby further enhancing both the robustness and accuracy of the predictive system.

High-level fusion in origin prediction

This study implemented a soft voting ensemble learning approach, which calculates the average prediction probabilities from multiple models and selects the geographical origin with the highest mean probability as the final output. This methodology offers the advantage of balancing the bias-variance trade-off, thereby enhancing overall robustness and ultimately generating more conservative and reliable consensus probabilities. Based on feature economy considerations, the multi-algorithm fusion utilized models derived from Dataset V (6 variables) rather than Dataset IV (23 variables). Model selection was conducted with a stringent criterion of macro F1 ≥ 0.99, retaining only high-performance models including Ridge, ElasticNet, RandomForest, KNN, and SVM to avoid the introduction of noise. Based on these five models, a soft voting ensemble learning method was employed to construct the most generalized and accurate origin traceability model in this study. This model combines a lightweight structure (only 6 features), high precision (training error = 0.00 ± 0.00%), and exceptional generalization capability (test error = 0.00 ± 0.00%), demonstrating outstanding performance in geographical origin prediction applications (Tab. 1).

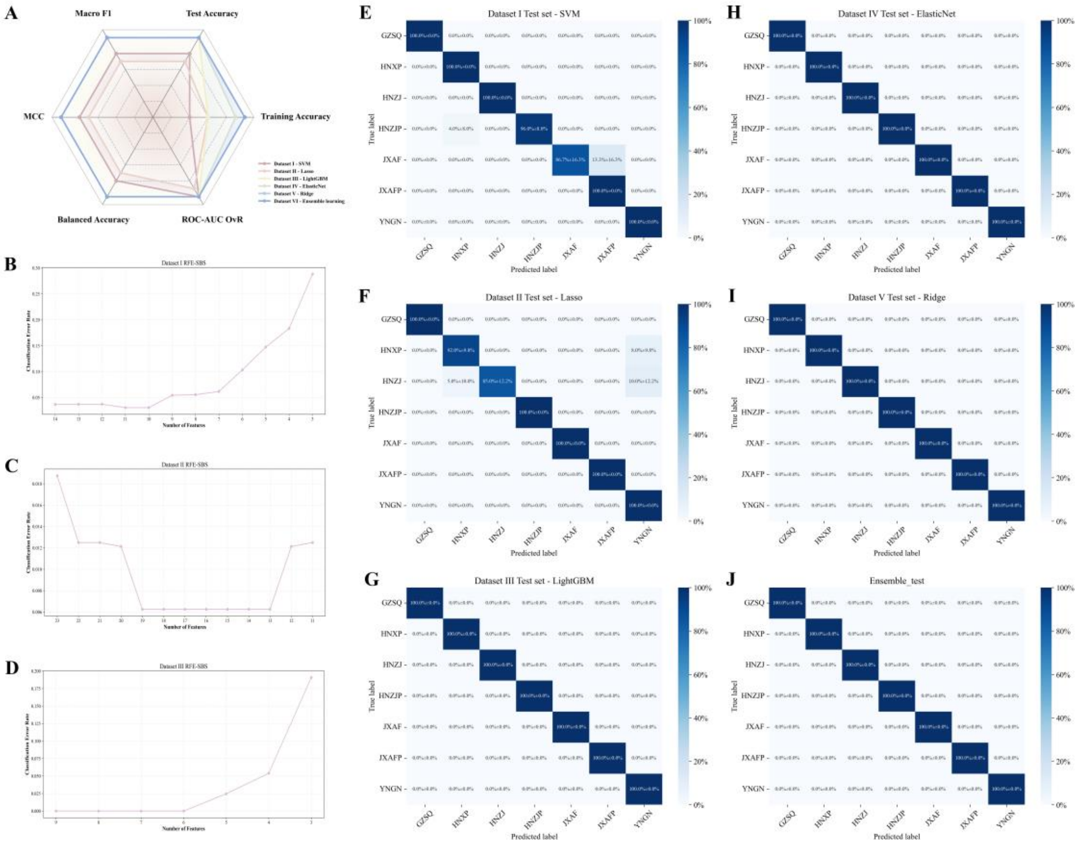

Fig. 4 Evaluation parameters of models （macro F1, test accuracy, training accuracy, MCC, balanced accuracy, and ROC-AUC (OvR)) (A). Feature extraction curves obtained using RFE-SBS for Datasets I–III (B–D), respectively. Confusion matrices for test sets: Dataset I – SVM (E), Dataset II – Lasso (F), Dataset III – LightGBM (G), Dataset IV – ElasticNet (H), Dataset V – Ridge (I), and Ensemble Test (J).

Relationships between the key origin prediction factors with environmental factors

Environmental factors, as potential selective pressures, significantly shape the biosynthesis patterns of secondary metabolites in sweet tea populations from different geographic origins [17]. Based on 11 environmental variables from each sampling site (Tab. S1), Spearman correlation heatmap analysis (Fig. 5A) revealed that CAF exhibited strong positive correlation with T_{min} ($r = 0.61$, $p < 0.001$) and strong negative correlation with MI ($r = - 0.59$), indicating that its accumulation is primarily driven by low temperature stress, with secondary influence from water deficit, this may be a metabolic response to adapt to cold habitats. Oapz_3 demonstrated strong negative correlations with both MAP ($r = - 0.74$) and MTCO ($r = - 0.74$), suggesting that cold and rainy environments inhibit its accumulation, potentially due to reduced enzymatic conversion efficiency caused by dilution effects from enhanced transpiration [35]. Regarding elemental composition, Sr showed strong negative correlation with MAP ($r = - 0.55$). Ce exhibited strong positive correlations with MTWA, MI and T_{max} ($r = 0.53 - 0.61$), while demonstrating strong negative correlation with ALT ($r = -0.61$), its high sensitivity to ALT may reflect regional differences in the mineral deposits where populations of different geographic origins. The high sensitivity to altitude reflects regional variations in mineral deposits. Rb displayed strong negative correlation with DI (r = - 0.64). Among SIRs, $\delta^{15}N$ showed negative correlation with DI (r = - 0.63). This pattern contradicts previous cross-regional research findings and may indicate a nitrogen cycling adaptation strategy unique to the study population [36].

SFS combined with VIF analysis identified seven predictors (ALT, GP, DI, MAP, SH, MTCO, and Tmin) from eleven environmental candidates for redundancy analysis (RDA) (Fig. 5B) [37]. The RDA ranking results visually demonstrate the spatial differentiation effect of environmental selection pressure on chemical composition. Correlation coefficients with six key variables ranged from –0.74 to 0.61. Based on 999 permutations, the model was significant (adjusted R^2 = 63.74%, $p < 0.001$), with the first two axes explaining 43.1% and 27.7% of the total variance, respectively. The ranking results indicate that DI, GP, and MTCO are the dominant environmental gradients, strongly suggesting that hydrothermal conditions are the primary macro-selective pressures shaping the adaptive divergence of chemical composition in sweet tea populations of different geographic origins [13]. CAF and Oapz_3 accumulated synergistically, mainly promoted by SH, DI, and $T_{min.}$ This may reflect enhanced biosynthesis of the precursor Phz under intense light, which elevates Oapz_3 levels [38]. Conversely, low temperature may inhibit the conversion of Phz to Oapz_3, potentially due to reduced chalcone isomerase activity [22,23]. CAF accumulation was more sensitive to low temperature and drought, consistent with the stress-induced shift from growth to secondary metabolism that increases CAF biosynthesis in *Camellia sinensis* [39], this can be viewed as an adaptive metabolic phenotype to cold and dry habitats.

Among elemental and isotopic variables, Sr showed a spatial distribution pattern similar to CAF, with its bioavailability shaped by the combined influence of modern climate and deep geological history. The Sr content in YNGN (6.04 mg/kg) is significantly lower than that in JXAF/JXAFP (10.4 mg/kg), profoundly reflecting their respective geochemical backgrounds: The red soil region where JXAF/JXAFP is located develops from silicate rock parent material, and the Sr^{2+} released during weathering is relatively retained in soils with high clay content, leading to a richer source available for plants [40]. In contrast, the YNGN karst region is dominated by carbonate rocks, where intense dissolution causes Sr^{2+} and Ca^{2+} to be rapidly leached away simultaneously, resulting in an extremely low background level of available Sr in the soil [41,42]. The distribution of REs, represented by Ce, is primarily governed by regional geochemical static fingerprints, with its fundamental source being the stable soil background reservoir formed through long-term weathering [12]. Although climatic factors such as GP and MAP can regulate the bioavailability of Ce by altering soil redox conditions, accelerating the Ce^{3+}/Ce^{4+} cycle and leading to enrichment variations [13], they do not fundamentally alter this deep geological imprint. Therefore, compared to secondary metabolites like Phz and Trb, which respond rapidly to environmental changes, Ce provides a more stable and traceable geochemical signature. These signatures, derived from coupled geological-climatic timescales and purely geological timescales respectively, mark the deep environmental contexts to which different geographic populations have adapted over long periods, serving as reliable chemical indicators for deciphering their evolutionary history and geographic origins. Rb distribution was primarily influenced by local microenvironmental conditions associated with altitude, while its negative correlations with DI and T_{min} reflected topography–climate coupling effects [43]. In terms of isotope ratios, $\delta^{15}N$ showed a positive correlation with ALT and SH, but a negative correlation with the DI and T_{min}. This aligns with the general pattern that high-altitude, high-radiation conditions promote nitrogen isotope fractionation, while dry and cold conditions favor nitrogen conservation in ecosystems [44], this further supports the role of local environments related to geographic origins in shaping the nitrogen metabolism of populations [45]. Previous findings revealed that variations in $\delta^{15}N$ were closely coupled with cultivation practices and specific metabolic pathways. Specifically, in HNZJP cultivation samples fertilized twice a year, the accumulation of Trb coincided with $\delta^{15}N$ enrichment, strongly suggesting that frequent and sufficient nitrogen input not only altered the baseline $\delta^{15}N$ of the soil nitrogen pool but also likely upregulated nitrogen metabolic flux, directly driving the biosynthesis of dihydrochalcone compounds derived from phenylalanine [26]. In contrast, JXAFP samples fertilized only once a year exhibited significantly lower Trb and $\delta^{15}N$ accumulation compared to HNZJP and wild samples. This parallel trend indicates that the degree of $\delta^{15}N$ enrichment can serve as a potential "isotopic tracer" reflecting nitrogen utilization intensity and the activation state of specific secondary metabolites (Trb), providing a basis for using $\delta^{15}N$ as an indicator to assess nitrogen nutrition status and secondary metabolic potential in sweet tea. However, it must be clearly pointed out that this study is based on sampling from wild and cultivated populations, so the identified environment-chemical phenotype associations cannot strictly distinguish whether the underlying mechanism of climate factors' influence on population chemical composition is phenotypic plasticity or local adaptation. Nevertheless, the spatial distribution pattern of chemotypes—particularly the extreme, discontinuous high Phz accumulation observed in the YNGN population, and the strong spatial coupling of this pattern with Quaternary glacial refugia—strongly suggests that in historically isolated regions like YNGN, the observed chemical differentiation may more profoundly reflect a genetic adaptation background shaped by long-term evolutionary isolation and selective pressures. Future research on sweet tea could control environmental factors through ongoing common garden experiments and integrate population genomics approaches to ultimately verify the relative contributions of plasticity and genetic adaptation in environmentally driven chemical variation.

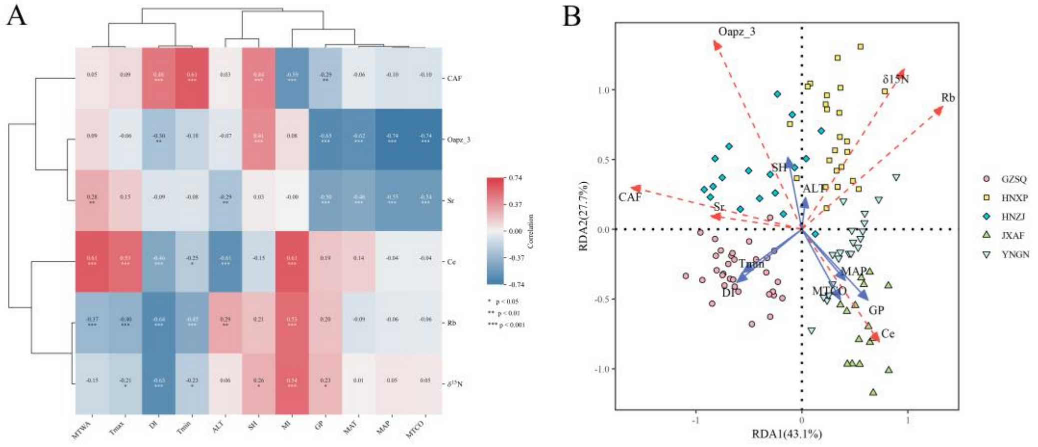

Fig. 5 Heatmap of the correlation between environmental factors and key variables (A) and RDA analysis of the key variables constrained by selective pressures from key environmental factors (B). In plot A: Red and blue represent positive and negative correlations, respectively. The darker the cell color, the larger the absolute value of the correlation coefficient, and vice versa. The symbols "*", "**", and "***" in the cells indicate significant correlations at the 0.05, 0.01 and 0.001 levels, respectively. Cells without asterisks indicate non-significant relationships. The combined lines outside the cells represent their clustering trends. In plot B: The scatter points represent the positions of samples based on the two principal components RDA1 and RDA2, with different colors indicating their geographical origin. The 6 key variable vectors selected by the previous optimal origin traceability model are plotted with red dashed lines, while the seven key environmental factor vectors are plotted with blue solid lines. The direction represents the direction of maximum contribution, and the length represents

the relative strength of contribution.

Methods

Sampling strategy and climate data origin

Between June and July 2022, a total of 163 mature leaf samples (3 months after sprouting) of sweet tea were collected from seven major production regions distributed across four Chinese provinces—Hunan, Jiangxi, Guizhou and Yunnan (23.65 – 27.55 °N, 105.44 – 114.67 °E; Fig. 6, Tab. S1). Among these, 52 samples were collected from cultivated stands managed with organic fertilizer (HNZJP, JXAFP), while the remaining 111 samples were obtained from natural or organically managed stands (HNZJ, HNXP, GZSQ, JXAF, YNGN). All samples were authenticated as *Lithocarpus litseifolius* by Prof. Yang Jian (National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China), and voucher specimens were deposited at the same institution. Following authentication, fresh leaves were washed, sliced, oven-dried at 50 °C for 36 h to inactivate endogenous enzymes, ground into fine powder (30–40 mesh), and stored in a desiccator until analysis. The bioclimate factors were obtained from the Science Data Bank and extracted by the GPS location. In total, the 11 environmental factors included ALT (altitude, m), DI (annual drought index), MI (annual moisture index), GP (growing season precipitation, mm), MTCO (mean temperature of the coldest month, ˚C), MTWA (mean temperature of the warmest month, ˚C), T_{max} (absolute maximum temperature, ˚C), T_{min} (absolute minimum temperature, ˚C), MAP (mean annual precipitation, mm), MAT (mean annual temperature, ˚C), SH (sunshine hour, hour).

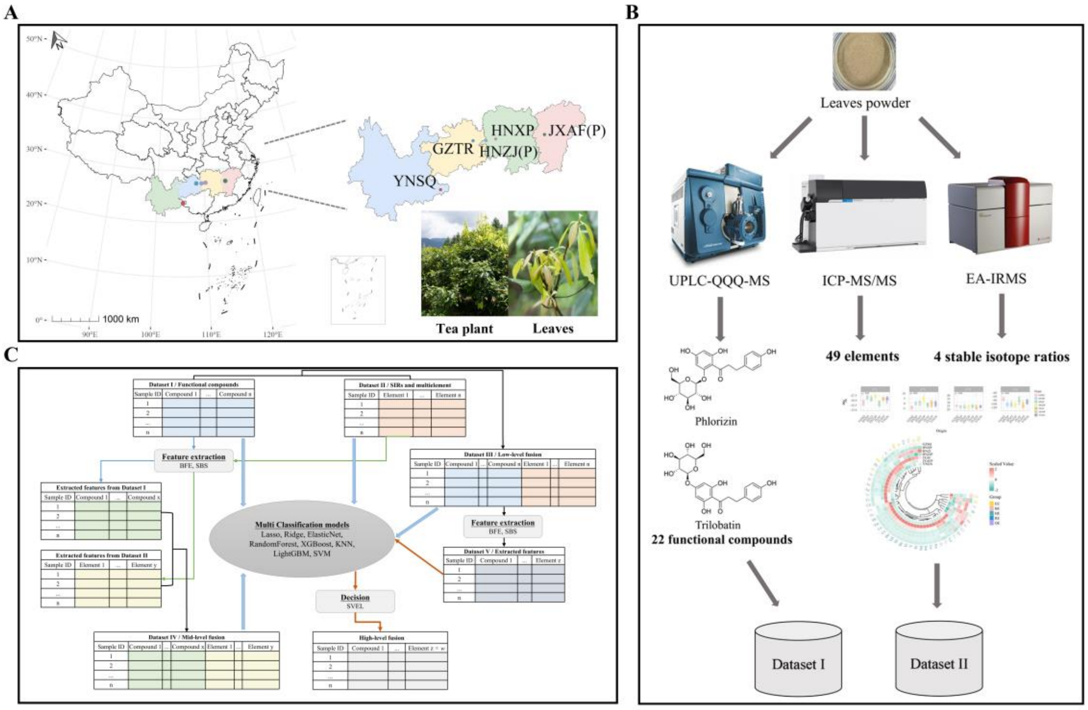

Fig. 6 Workflow. Sample collection (A), characterization (B) and data-fusion (C).
JXAF, JXAFP, Anfu County, Cultivation basement of Anfu County in Jiangxi Province; GZSQ, Shiqian County in Guizhou Province; HNZJ, HNZJP, HNXP Zhijiang County, Cultivation basement of Zhijiang County, Xupu County in Hunan Province; YNGN, Guangnan County in Yunnan Province.

Chemicals and Reagents

All elemental standards were supplied by the National Centre for Analysis and Testing of Non-ferrous Metals & Electronic Materials (China). Multi-element stock solutions (100 µg mL^{-1}) containing Ag, As, B, Ba, Be, Bi, Cd, Co, Cr, Cs, Cu, Ga, In, Li, Mn, Mo, Ni, Pb, Rb, Sb, Se, Sr, V, Zn along with rare earth element solutions (La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu) were used for calibration. Major-element standards included K, Na, Mg, P were obtained as 1 mg mL^{-1} stock solutions, while single-element standards of S, Al, Fe, Ca, and Ti were acquired at the same concentration. Additional single-element standards for Tl and Y, as well as internal standards Sc and Rh (all at 100 µg/mL), were included for quality assurance. High-purity acids included HNO_3, HF, and HCl were supplied by CNW (Shanghai Anpu Technology), and $HClO_4$ was obtained from Feichuan brand (Tianjin Xinyuan Chemical). Ultrapure water was provided by Watsons (Guangzhou). For combustion analysis, reduced copper granules, tungsten trioxide (WO_3), and glassy carbon were acquired from Elementar (Germany). Isotopic reference materials, including B2155 (δ^{13}C = –26.98 ‰, δ^{15}N = 5.94 ‰), USGS40 (δ^{15}N = – 4.52 ‰), IAEA-CH-6 (δ^{13}C = – 10.449 ‰), USGS64 (δ^{13}C = – 40.82 ‰), IAEA-N-2 (δ^{15}N = 20.3 ‰), USGS54 (δ^{2}H = – 150.4 ‰, δ^{18}O = 17.79 ‰), USGS55 (δ^{2}H = – 28.2 ‰, δ^{18}O = 19.12 ‰), USGS56 (δ^{2}H = – 44.0 ‰, δ^{18}O = 27.23 ‰) were obtained from the International Atomic Energy Agency (IAEA, Vienna). All standards and reagents were used as received from their respective suppliers. Functional compounds reference materials included phlorizin (Phz), Trilobatin (Trb), 3-hydroxy phlorizin (Hpz_3), Phloretin 2'-O-glucoside (Gpt_2), 3''-O-acetyl phloridzin (Oapz_3), 2''-O-acetyl phloridzin (Oapz_2), Guavinoside (Gua), Phloretin (Pht), Protocatechualdehyde (PA), Protocatechuic acid (DA), Gallic acid (GA), 6,7-dihydroxycoumarin (Dih), Esculin (Esc), Tartaric acid (Tar), Citric acid (Cit), Fumaric acid (Fum), Succinic acid (SuA), Chlorogenic acid (CA), 2,3,4-Trihydroxybenzoic acid (THBA) and Folin-Ciocalteu reagent were obtained from Beiterenkang Bio-Technology Co., Ltd. (Beijing, China). Methanol and acetonitrile were LC-MS grade and purchased from Fisher Scientific (Massachusetts, USA).

Multi-element Analysis

Multi-element analysis was performed using a triple-quadrupole inductively coupled plasma mass spectrometer (ICP-MS/MS, Agilent 8900). Sample digestion was carried out with the Mars 5 microwave digestion system (CEM Corp, USA) and precise weighing was achieved using an analytical balance (Sartorius BSA224S-CW). The instrumental operating parameters were set as follows: RF power at 1550 W, high-purity argon (99.999%) as carrier gas, a concentric nebulizer with a gas flow of 1.05 L min^{-1}, make-up gas at 0.15 L min^{-1}, sampling depth of 8.0 mm, and peristaltic pump speed of 0.3 rpm. For sample preparation, approximately 0.10 g of finely ground leaf powder (100 mesh) was accurately weighed into a PTFE digestion vessel. After the addition of 6.0 mL concentrated HNO_3, the open vessel was pre-digested at 120 °C for 30 minutes until the evolution of NO_x fumes ceased. Following cooling, the vessel was sealed and subjected to a four-stage microwave digestion program: 120 °C (ramp 10 min, hold 2 min), 150 °C (ramp 10 min, hold 2 min), 180 °C (ramp 10 min, hold 2 min), and 200 °C (ramp 10 min, hold 20 min). The resulting digest was evaporated at 180 °C to approximately 0.5 mL, quantitatively transferred to a 25 mL volumetric flask, diluted with 67% HNO_3 to volume, thoroughly mixed, allowed to settle, and finally filtered through a 0.45 μm syringe filter to ICP-MS/MS analysis.

Stable isotope ratio analysis

SIR analysis was conducted using an elemental analyzer coupled with isotope-ratio mass spectrometry (EA/IRMS, Elementar, Germany). Measurements of $\delta^{13}C$ and $\delta^{15}N$ were performed using a Vario Isotope cube-Biovision system, while $\delta^{2}H$ and $\delta^{18}O$ were determined using a Vario Pyro cube coupled to an Isoprime 100 IRMS (TC/EA mode). Sample weighing was carried out with an analytical balance (Mettler-Toledo XPR106DUH/AC). For $\delta^{13}C$ and $\delta^{15}N$ analysis, approximately 5 mg of plant powder was weighed into a tin capsule and introduced into the EA autosampler. Combustion occurred in a WO_3-packed tube at 1150 °C, followed by reduction in a copper-packed tube at 850 °C under a helium carrier gas (99.999% purity). The resulting CO_2 and N_2 were delivered to the IRMS, with CO_2 signal attenuation achieved using a CentrION diluter. IRMS trap currents were set to 100 μA for CO_2 and 400 μA for N_2. For $\delta^{2}H$ and $\delta^{18}O$ determination, approximately 0.5 mg of sample was weighed and loaded into the TC/EA autosampler. Pyrolysis took place in a glassy carbon-packed reactor at 1450 °C under helium carrier gas (99.999%). The produced H_2 and CO were introduced into the IRMS, with a built-in diluter moderating the CO signal. IRMS trap currents were configured to 200 μA for CO and 400 μA for H_2. Isotopic compositions were calculated using the standard delta notation: $\delta E (‰) = [(R_{sample} / R_{standard}) - 1] \times 1000$, where R represents the ratio of $^{13}C/^{12}C$, $^{15}N/^{14}N$, $^{18}O/^{16}O$, or $^{2}H/^{1}H$. All values were referenced to VPDB ($\delta^{13}C$), AIR ($\delta^{15}N$), and VSMOW ($\delta^{18}O$, $\delta^{2}H$).

Functional compounds analysis

Each 0.02 g sample of sweet tea powder was extracted with 1.5 mL of 80% methanol and weighed. The mixture was subjected to ultrasonic treatment for 40 minutes (300 W, 40 kHz). After cooling, it was weighed again, and 80% methanol was added to compensate for any loss in weight, followed by centrifugation at 13,000 rpm for 10 minutes. The resulting supernatant was filtered through a 0.22 μm membrane before injection into the liquid chromatography system. Chromatographic separation was performed on an ACQUITY UPLC™ BEH C18 column (100 mm × 2.1 mm, 1.8 μm) maintained at 40 °C. The injection volume was 1.0 μL, and the flow rate was 0.6 mL/min. The mobile phase consisted of 0.1% formic acid in acetonitrile (A) and 0.1% formic acid in water (B). The gradient program was as follows: 0–1.0 min, 5–25% A; 1.0–3.5 min, 25–40% A; 3.5–4.5 min, 40–60% A; 4.5–5.0 min, 60–5% A; 5.0–7.0 min, 5% A. Tandem mass spectrometry (MS/MS) was performed using API 6500 system (AB SCIEX, Los Angeles, CA, USA) equipped with an electrostatic ionization (ESI) source (AB SCIEX). MS analysis for flavonoid was carried out in negative ionization mode, and the operating conditions were set as follows: ion source voltage, -5500 V (ESI-); turbo spray temperature (TEM), 550 ˚C; Curtain Gas (CUR) flow, 30 L/min; Ion Source Gas (IS) flow, 55 L/min; scanning mode: Scheduled multiple reaction monitoring (Scheduled MRM). The MS parameters for flavonoid were manually optimized (Tab. S2). Data acquisition was carried out using Analyst Software 1.6.2 (AB SCIEX, Los Angeles, CA, USA), and analysis was performed using MultiQuant Software 3.0 (AB SCIEX, Los Angeles, CA, USA).

Total polyphenol (Tp) was determined according to GB/T 8313-2018. Weighed 0.200 g of sample. Added 5.0 mL of 70% aqueous methanol preheated to 70 °C. Stirred to ensure complete wetting. Extracted in a 70 °C water bath for 10 min with brief stirring at 5 min. Cooled to room temperature. Centrifuged at 3500 r/min for 10 min. Transferred the supernatant to a 10 mL volumetric flask. Reextracted the residue with 5.0 mL of 70% methanol under the same conditions. Combined the extracts. Diluted to 10.0 mL with the same solvent. Mixed thoroughly. Filtered through a 0.45 μm membrane to obtain the stock solution. Pipetted 1.0 mL stock into a 100 mL volumetric flask. Diluted to volume with water. Mixed to obtain the test solution. Transferred 1.0 mL of each gallic acid working solution, water blank, and test solution into separate tubes. Added 5.0 mL Folin–Ciocalteu reagent. Mixed immediately. After 3–8 min added 4.0 mL of 7.5% sodium carbonate. Brought to volume with water. Shook thoroughly. Stood at room temperature for 60 min. Measured absorbance at 765 nm. The analysis of Caffeine (CAF) was performed according to GB/T 8312-2013. Weighed 1.5 g of sample. Added 200 mL of boiling distilled water. Extracted in a boiling water bath for 45 min with shaking every 10 min. Performed hot vacuum filtration. Rinsed the residue 2–3 times with small volumes of hot water. Transferred the filtration to a 250 mL volumetric flask. Cooled to room temperature. Diluted to volume with water. Mixed well. Pipetted 10 mL of test solution into a 100 mL volumetric flask. Added 4 mL of 0.01 mol/L hydrochloric acid. Added 1 mL of basic lead acetate solution. Diluted to volume with water. Mixed thoroughly. Allowed to clarify. Filtered. Transferred 25 mL of filtrate into a 50 mL volumetric flask. Added 0.10 mL of 4.5 mol/L sulfuric acid. Diluted to volume with water. Mixed well. Allowed to clarify. Filtered. Measured absorbance at 274 nm in a 10 mm quartz cuvette using the reagent blank as reference. The analysis of Amino acids (AA) was conducted according to GB/T 8314-2013. Weighed 1.5 g of sample into a 250 mL conical flask. Added 200 mL of boiling distilled water. Extracted in a boiling water bath for 45 min with shaking every 10 min. Performed hot vacuum filtration. Rinsed the residue 2–3 times with small volumes of hot water. Transferred the filtration to a 250 mL volumetric flask. Cooled to room temperature. Diluted to volume with water. Mixed well. Pipetted 1.0 mL of test solution into a 25 mL colorimetric tube. Added 0.5 mL of pH 8.0 phosphate buffer. Added 0.5 mL of 2% ninhydrin solution. Heated in a boiling water bath for 15 min. Cooled to room temperature. Diluted to 25 mL with water. Stood for 10 min. Measured absorbance at 570 nm in a 5 mm cuvette using the reagent blank as reference.

Statistical analysis and Machine learning models evaluation

All statistical analyses were performed at R 4.4.1. Welch's ANOVA from the "onewaytests" package with Games Howell post hoc testing from "rstatix" assessed differences among origins [46]. Principal component analysis (PCA) was conducted with "FactoMineR" and "factoextra", and data wrangling and plotting used "tidyverse" [47]. Cluster heatmaps were generated with "ComplexHeatmap" and "circlize", correlation networks with "igraph", "ggraph", "tidygraph" and "ggplot2", and additional boxplots and heatmaps with "ggplot2" and "ComplexHeatmap". Supervised chemometrics employed orthogonal partial least squares discriminant analysis (OPLS-DA) in SIMCA P 14.1 with Pareto scaling, cultivation practice models were first built on all data to identify variables with VIP greater than 1, and overfitting was evaluated by 200 permutation tests [34]. Redundancy analysis (RDA) was run with the "vegan" package through "rpy2", and variance inflation factor (VIF) testing used the "variance_inflation_factor" function from "statsmodels" [37]. Multiclass modeling feature selection and visualization were implemented in Python 3.13.5 in the VSCode environment using recursive feature elimination (RFE), sequential forward selection (SFS) and sequential backward selection (SBS) from "sklearn.feature_selection" [48]. Classifiers included Elastic Net Classifier (ElasticNet), Least Absolute Shrinkage and Selection Operator (Lasso), and Ridge Classifier (Ridge) from "sklearn.linear_model"; K-Nearest Neighbors (KNN) from "sklearn.neighbors"; Random Forest Classifier (RandomForest) from "sklearn.ensemble"; Support Vector Machine (SVM) from "sklearn.svm"; and Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) from their native packages. Classification error rates and feature importance under optimal configurations were plotted with "matplotlib.pyplot". Key metrics, including macro-average F1-score, accuracy (training and test sets), Matthews correlation coefficient (MCC), balanced accuracy, and one-vs-rest ROC-AUC (ROC-AUC OvR), were computed via the "sklearn.metrics". SVEL integrated prediction probabilities using clone from "sklearn.base", and majority voting tallies used "Counter" from the collections module.

PCA served as an unsupervised screen across seven regions using 75 features. For machine learning, datasets were split into 80% training and 20% prediction sets, standardized with "StandardScaler", and partitioned with "train_test_split" using five random repetitions with stratification to preserve label representation. Training performance was assessed by 5-fold cross validation, and the test set was evaluated by leave one out cross validation (LOOCV). Data fusion adopted three levels in which low-level fusion concatenated compositional data, SIRs and multi-element fingerprints, mid-level fusion integrated features selected by RFE combined with SBS, and high-level fusion combined prediction outputs from multiple models on the optimal dataset using SVEL to produce a robust classifier [49]. 8 multiclass models including ElasticNet, KNN, Lasso, LightGBM, RandomForest, Ridge, SVM and XGBoost were compared and hyperparameters were optimized by grid search with 5-fold stratified cross validation including the regularization strength C for Lasso, Ridge and SVM, the combination of C and l1_ratio for ElasticNet, the number of estimators for RandomForest, XGBoost and LightGBM, and the number of neighbors for KNN [50]. Final performance was summarized by macro F1 and classification error rates on training and test sets [51]. In soft voting, the summed class probabilities across models were used to assign the predicted origin by the argument of the maximum (*argmax*).

Data availability

Data will be made available on request.

Code availability

The codes used to generate the results in this study are available on reasonable request from the corresponding author.

Author contributions

Y.T.: Conceptualization, data curation, formal analysis, project administration, software, visualization, writing original draft and writing—review & editing. P.Y.: Conceptualization, formal analysis, project administration, methodology, supervision, software, visualization, writing—review & editing. F.X.: formal analysis, supervision. Z.Z.: formal analysis, supervision, writing—review & editing. K.X.: Investigation, resources. S.Y.: Investigation, software. Y.N.: Investigation, data curation. Z.Z.: Investigation, resources. C.W.: Investigation, resources. W.Q.: Investigation. X.Z.: Investigation. Y.L.: Investigation. R.W.: Investigation. G.H.: Supervision, writing—review & editing. J.Y.: Supervision, funding acquisition, project administration, and writing—review & editing.

Acknowledgment

The authors would like to thank all colleagues and institutions that contributed to this work.

Funding

This study was supported by the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (Grant No. CI2023E002), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (Grant No. ZYYZDXK- 2023244), the Key Project at the Central Government Level: Capacity Building for the Sustainable Utilization of Valuable Chinese Medicine Resources (Grant No. 2060302), the China Agriculture Research System of MOF and MARA (Grant No. CARS-21), the National Key R&D Program of China (Grant No. 2020YFC1712700), Co-development Agreement on the Bioactivity Research of Flavonoids from *Lithocarpus litseifolius* (RDS10120230053).

Competing interests

The author declares no competing financial or non-financial interests.

Reference

1 Wang, Y. K. *et al.* Dihydrochalcones in Sweet Tea: Biosynthesis, Distribution and Neuroprotection Function. *Molecules* **27** (2022). https://doi.org/10.3390/molecules27248794

2 Shang, A. *et al.* Sweet tea (Lithocarpus polystachyus rehd.) as a new natural source of bioactive dihydrochalcones with multiple health benefits. *Crit Rev Food Sci Nutr* **62**, 917-934 (2022). https://doi.org/10.1080/10408398.2020.1830363

3 Ma, J. *et al.* Toxicological safety assessment of a water extract of Lithocarpus litseifolius by a 90-day repeated oral toxicity study in rats. *Front Pharmacol* **15**, 1385550 (2024). https://doi.org/10.3389/fphar.2024.1385550

4 Wei, Y. Q. *et al.* Research progress on dihydrochalcones from Lithocarpus litseifolius extracts in treatment of type 2 diabetes mellitus and its complications. *Zhongguo Zhong Yao Za Zhi* **50**, 658-671 (2025). https://doi.org/10.19540/j.cnki.cjcmm.20240827.701

5 Eichenberger, M. *et al.* Metabolic engineering of Saccharomyces cerevisiae for de novo production of dihydrochalcones with known antioxidant, antidiabetic, and sweet tasting properties. *Metab Eng* **39**, 80-89 (2017). https://doi.org/10.1016/j.ymben.2016.10.019

6 Jiang, J. *et al.* In vitro inhibitory effect of five natural sweeteners on alpha-glucosidase and alpha-amylase. *Food Funct* **15**, 2234-2248 (2024). https://doi.org/10.1039/d3fo05234f

7 Zhang, J., Yan, X., Yuan, X. & Ke, F. Research Progress on the Treatment and Regulation of Traditional Chinese

Medicine Lithocarpus litseifolius in Obese Populations. *Journal of Southwest Medical University* (2025). https://doi.org/10.3969/j.issn.2096-3351.2025.05.018

8 Bajoub, A., Carrasco-Pancorbo, A., Ajal, E. A., Ouazzani, N. & Fernandez-Gutierrez, A. Potential of LC-MS phenolic profiling combined with multivariate analysis as an approach for the determination of the geographical origin of north Moroccan virgin olive oils. *Food Chem* **166**, 292-300 (2015). https://doi.org/10.1016/j.foodchem.2014.05.153

9 Bai, L. *et al.* Combining stable isotopes and multi-elements with machine learning chemometric models to identify the geographical origins of Tetrastigma hemsleyanum Diels et Gilg. *Food Chem* **469**, 142496 (2025). https://doi.org/10.1016/j.foodchem.2024.142496

10 Sim, J., McGoverin, C., Oey, I., Frew, R. & Kebede, B. Stable isotope and trace element analyses with non-linear machine-learning data analysis improved coffee origin classification and marker selection. *J Sci Food Agric* **103**, 4704-4718 (2023). https://doi.org/10.1002/jsfa.12546

11 Strojnik, L. *et al.* Geographical identification of strawberries based on stable isotope ratio and multi-elemental analysis coupled with multivariate statistical analysis: A Slovenian case study. *Food Chem* **381**, 132204 (2022). https://doi.org/10.1016/j.foodchem.2022.132204

12 Sanematsu, K. Adsorption of REE in Weathered Granite and Its Importance for Resources. *Journal of the Clay Science Society of Japan (in Japanese)* **50**, 128-134 (2012). https://doi.org/10.11362/jcssjnendokagaku.50.3_128

13 Cao, M. *et al.* Optimistic contributions of plant growth-promoting bacteria for sustainable agriculture and climate stress alleviation. *Environ Res* **217**, 114924 (2023).

https://doi.org/10.1016/j.envres.2022.114924

14 Bai, R. *et al.* Deep learning-based fusion of color and spectral features from hyperspectral imaging for the origin identification of Salvia miltiorrhiza. *Science of Traditional Chinese Medicine* **3**, 250-258 (2025). https://doi.org/10.1097/st9.0000000000000079

15 Yu, P. *et al.* Unveiling the origin and quality traits of Angelica sinensis: Hyperspectral imaging combined with chemometrics and information fusion strategies. *Journal of Food Composition and Analysis* **147** (2025). https://doi.org/10.1016/j.jfca.2025.108089

16 Yu, D. *et al.* Interpretable AI-driven multidimensional chemical fingerprints for geographical authentication of Euryales Semen. *NPJ Sci Food* **9**, 133 (2025). https://doi.org/10.1038/s41538-025-00510-y

17 Liu, G. *et al.* Temporal dynamics of bioactive compounds in sweet tea (Lithocarpus litseifolius (Hance) Chun): Linking harvest stages to flavor and health benefits. *Food Research International* **218** (2025). https://doi.org/10.1016/j.foodres.2025.116918

18 Qiu, Y. X., Fu, C. X. & Comes, H. P. Plant molecular phylogeography in China and adjacent regions: Tracing the genetic imprints of Quaternary climate and environmental change in the world's most diverse temperate flora. *Mol Phylogenet Evol* **59**, 225-244 (2011). https://doi.org/10.1016/j.ympev.2011.01.012

19 Fan, L., Zheng, H., Milne, R. I., Zhang, L. & Mao, K. Strong population bottleneck and repeated demographic expansions of Populus adenopoda (Salicaceae) in subtropical China. *Ann Bot* **121**, 665-679 (2018). https://doi.org/10.1093/aob/mcx198

20 Chen, X. *et al.* Biogeographic and metabolic studies support a glacial radiation hypothesis during Chrysanthemum evolution. *Hortic Res* **9**, uhac153 (2022). https://doi.org/10.1093/hr/uhac153

21 Liu, H. Y. *et al.* Phenolic Content, Main Flavonoids, and Antioxidant Capacity of Instant Sweet Tea (Lithocarpus litseifolius [Hance] Chun) Prepared with Different Raw Materials and Drying Methods. *Foods* **10** (2021). https://doi.org/10.3390/foods10081930

22 Dare, A. P. *et al.* Overexpression of chalcone isomerase in apple reduces phloridzin accumulation and increases susceptibility to herbivory by two-spotted mites. *Plant J* **103**, 293-307 (2020). https://doi.org/10.1111/tpj.14729

23 Zhang, X. *et al.* Identification of UDP-glucosyltransferase involved in the biosynthesis of phloridzin in Gossypium hirsutum. *Plant J* **121**, e17248 (2025). https://doi.org/10.1111/tpj.17248

24 Das, P. R., Kim, Y., Hong, S. J. & Eun, J. B. Profiling of volatile and non-phenolic metabolites-Amino acids, organic acids, and sugars of green tea extracts obtained by different extraction techniques. *Food Chem* **296**, 69-77 (2019). https://doi.org/10.1016/j.foodchem.2019.05.194

25 Gilbert, A., Silvestre, V., Robins, R. J., Remaud, G. S. & Tcherkez, G. Biochemical and physiological determinants of intramolecular isotope patterns in sucrose from C(3), C(4) and CAM plants accessed by isotopic (1)(3)C NMR spectrometry: a viewpoint. *Nat Prod Rep* **29**, 476-486 (2012). https://doi.org/10.1039/c2np00089j

26 Zeng, S. *et al.* Integrated transcriptome and metabolome analysis reveals the regulation of phlorizin synthesis in Lithocarpus polystachyus under nitrogen fertilization. *BMC Plant Biol* **24**, 366 (2024).

https://doi.org/10.1186/s12870-024-05090-9

27 Li, L. *et al.* Enhanced carbon use efficiency and warming resistance of soil microorganisms under organic amendment. *Environ Int* **192**, 109043 (2024). https://doi.org/10.1016/j.envint.2024.109043

28 Hayashi, N. *et al.* Annual variation of natural 15N abundance in tea leaves and its practicality as an organic tea indicator. *J Agric Food Chem* **59**, 10317-10321 (2011). https://doi.org/10.1021/jf202215z

29 Roy, A. *et al.* Unravelling 30 ka recharge history of an intensely exploited multi-tier aquifer system in North West India through isotopic tracers - Implications on deep groundwater sustainability. *Sci Total Environ* **807**, 151401 (2022). https://doi.org/10.1016/j.scitotenv.2021.151401

30 Jiang, F. *et al.* Selenium levels in soil and tea as affected by soil properties in Jiangxi Province, China. *BMC Plant Biol* **24**, 1130 (2024). https://doi.org/10.1186/s12870-024-05844-5

31 Fu, L. *et al.* Differences in Copper Absorption and Accumulation between Copper-Exclusion and Copper-Enrichment Plants: A Comparison of Structure and Physiological Responses. *PLoS One* **10**, e0133424 (2015). https://doi.org/10.1371/journal.pone.0133424

32 MEE. (Ministry of Ecological Environment, 2020).

33 Dutta, T. *et al.* Global demand for rare earth resources and strategies for green mining. *Environ Res* **150**, 182-190 (2016). https://doi.org/10.1016/j.envres.2016.05.052

34 Fu, H. *et al.* Combining stable C, N, O, H, Sr isotope and multi-element with chemometrics for identifying the geographical origins and farming patterns of Huangjing herb. *Journal of Food Composition and Analysis* **102** (2021). https://doi.org/10.1016/j.jfca.2021.103972

35 Minsat, L. *et al.* Sustainable and Scalable Enzymatic Production, Structural Elucidation, And Biological Evaluation of Novel Phlorizin Analogues. *ChemSusChem* **18**, e202401498 (2025). https://doi.org/10.1002/cssc.202401498

36 Elrys, A. S. *et al.* Global gross nitrification rates are dominantly driven by soil carbon-to-nitrogen stoichiometry and total nitrogen. *Glob Chang Biol* **27**, 6512-6524 (2021). https://doi.org/10.1111/gcb.15883

37 Kim, J. H. Multicollinearity and misleading statistical results. *Korean J Anesthesiol* **72**, 558-569 (2019). https://doi.org/10.4097/kja.19087

38 Wei, Z. *et al.* Melatonin increases the performance of Malus hupehensis after UV-B exposure. *Plant Physiol Biochem* **139**, 630-641 (2019). https://doi.org/10.1016/j.plaphy.2019.04.026

39 Kfoury, N. *et al.* Striking changes in tea metabolites due to elevational effects. *Food Chem* **264**, 334-341 (2018). https://doi.org/10.1016/j.foodchem.2018.05.040

40 Jiang, J., Xu, R.-k. & Zhao, A.-z. Comparison of the surface chemical properties of four soils derived from Quaternary red earth as related to soil evolution. *Catena* **80**, 154-161 (2010). https://doi.org/10.1016/j.catena.2009.11.002

41 Liu, W.-J. *et al.* Elemental and strontium isotopic geochemistry of the soil profiles developed on limestone and sandstone in karstic terrain on Yunnan-Guizhou Plateau, China: Implications for chemical weathering and parent materials. *Journal of Asian Earth Sciences* **67-68**, 138-152 (2013). https://doi.org/10.1016/j.jseaes.2013.02.017

42 Gunadasa, S. G., Tighe, M. K. & Wilson, S. C. Arsenic and cadmium leaching in co-contaminated agronomic soil and the influence of high rainfall and amendments. *Environ Pollut* **316**, 120591 (2023). https://doi.org/10.1016/j.envpol.2022.120591

43 Ritter, A., Regalado, C. M. & Aschan, G. Fog reduces transpiration in tree species of the Canarian relict heath-laurel cloud forest (Garajonay National Park, Spain). *Tree Physiol* **29**, 517-528 (2009). https://doi.org/10.1093/treephys/tpn043

44 Schaeffer, S. M., Sharp, E., Schimel, J. P. & Welker, J. M. Soil-plant N processes in a High Arctic ecosystem, NW Greenland are altered by long-term experimental warming and higher rainfall. *Glob Chang Biol* **19**, 3529-3539 (2013). https://doi.org/10.1111/gcb.12318

45 Liu, Y. *et al.* Genomic basis of geographical adaptation to soil nitrogen in rice. *Nature* **590**, 600-605 (2021). https://doi.org/10.1038/s41586-020-03091-w

46 Midway, S., Robertson, M., Flinn, S. & Kaller, M. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ* **8**, e10387 (2020). https://doi.org/10.7717/peerj.10387

47 Seredin, P. *et al.* A Study of the Association between Primary Oral Pathologies (Dental Caries and Periodontal Diseases) Using Synchrotron Molecular FTIR Spectroscopy in View of the Patient's Personalized Clinical Picture (Demographics and Anamnesis). *Int J Mol Sci* **25** (2024). https://doi.org/10.3390/ijms25126395

48 Zhou, R. *et al.* Hybrid wavelength selection strategy combined with ATR-FTIR spectroscopy for preliminary exploration of vintage labeling traceability of sauce-flavor baijiu. *Spectrochim Acta A Mol Biomol Spectrosc* **321**, 124691 (2024). https://doi.org/10.1016/j.saa.2024.124691

49 Liu, C. *et al.* Metabolomics for origin traceability of lamb: An ensemble learning approach based on random forest recursive feature elimination. *Food Chem X* **29**, 102856 (2025). https://doi.org/10.1016/j.fochx.2025.102856

50 Ashiq, W. *et al.* Roman urdu hate speech detection using hybrid machine learning models and hyperparameter optimization. *Sci Rep* **14**, 28590 (2024). https://doi.org/10.1038/s41598-024-79106-7

51 Lyu, J. *et al.* Generative Adversarial Network-based Noncontrast CT Angiography for Aorta and Carotid Arteries. *Radiology* **309**, e230681 (2023). https://doi.org/10.1148/radiol.230681