Zhang *et al. Journal of Wood Science* (2022) 68:38
https://doi.org/10.1186/s10086-022-02044-3

Journal of Wood Science

ORIGINAL ARTICLE Open Access

Similarity network fusion for aggregating headspace GC–MS and direct analysis in real time–mass spectrometry data from solid samples to enhance species identification efficiency of high–temperature heated wood

Maomao Zhang[1,2,3], Juan Guo[2,3], Yang Lu[2,3], Lichao Jiao[2,3], Tuo He[2,3] and Yafang Yin[2,3*]

Abstract

Pterocarpus santalinus and *Pterocarpus tinctorius* are commonly used species of the genus *Pterocarpus* in the wood trade. Although both of them have been listed in Appendix II of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) since 2019, it is still critical to identify them in terms of plant taxonomy. Currently, high-temperature heating is an accepted treatment method for high-density wood species such as *Pterocarpus* to improve dimensional stability and restore previous drying defects partially. It has proved challenging to identify the high-temperature (e.g., 120 °C) heated wood from these two species. Thus, this study approaches species identification of two *Pterocarpus* of high-temperature (e.g., 120 °C) heated solid wood samples using headspace–gas chromatography–mass spectrometry (HS–GC–MS). Besides, a computational analytical method named similarity network fusion (SNF) was proposed to aggregate data in two different types, respectively, derived from the HS–GC–MS and direct analysis in real time–mass spectrometry (DART–MS) to explore the feasibility of improving the efficiency and accuracy of wood species discrimination. The SNF exhibits more significant differences and higher predictive accuracy (100%) between *P. santalinus* and *P. tinctorius* than that based on the HS–GC–MS data (77.78%) or DART–MS (66.67%) alone. These results demonstrated the capability of the HS–GC–MS technique in the analysis of high-temperature heated solid wood and the potential of multidimensional or comprehensive data sets based on the SNF algorithm in the field of wood species identification.

Keywords: Wood identification, Headspace–gas chromatography–mass spectrometry (HS–GC–MS), Similarity network fusion (SNF), High-temperature heated wood, Direct analysis in real time–mass spectrometry (DART–MS)

Introduction

Illegal logging encourages the inefficient use of resources and results in a threat to the reputation and sustainability of the legitimate timber trade [1]. A series of measures have been taken in the international community, including enacting laws designed to discourage the trade in illegally sourced timber and prohibit or limit the trade of specific species or those from particular areas [2]. However, the enforcement of these laws relies heavily on wood identification technology. Thus, it is vital to develop and improve the wood identification technology to support further the certification and verification of timber legality.

In many cases, wood chemical analysis can provide information about wood identification, which can be

*Correspondence: yafang@caf.ac.cn

[2] Department of Wood Anatomy and Utilization, Chinese Research Institute of Wood Industry, Chinese Academy of Forestry, Beijing 100091, China
Full list of author information is available at the end of the article

difficult to determine by visual means [2]. Intra-specific variation in some species has been detected through specific chemical analyses, including mass spectrometry and near infrared spectroscopy, etc. [2–6]. Most previous studies on the chemical constituents of wood focused on odor compounds and organic solvent extracts detected by gas chromatography–mass spectrometry (GC–MS) [7–16], and direct analysis in real time–mass spectrometry (DART–MS) [3, 17–22]. Using GC–MS as an instrument method, they have a variety of injection methods. For instance, GC–MS with liquid injection was used to study the chemical components of wild and cultivated agarwood extracts [16]. Similarly, the liquid injection was also applied in the identification of volatile compounds for the ethanol–benzene extractives of *Dalbergia odorifera* and *D. stevensonii* by GC–MS [23], the extract of *Pterocarpus macrocarpus* [9], and the chemical compositions in wood and bark of *Albizia julibrissin* tree [14], etc. Compared to the complex pre-treatment process for the liquid injection method, a headspace solid-phase microextraction (HS–SPME) system requires only simple heating of the sample. It has been used for analyzing the odorous constituents of wood [15]. However, the HS–SPME injection method is more limited in terms of volatile components due to the choice of the fiber material, and this injection method cannot be easily automated [24]. As an alternative method, headspace injection (HS) eliminates the need for long distillation periods and solvent consumption and is a faster, more efficient method, which made it successfully applied for the evaluation of the separation and recognition of complex mixture compounds, such as wood. In recent years, headspace–gas chromatography–mass spectrometry (HS–GC–MS) has been employed to differentiate two species of *Asari Radix* by their odors [25], to identify *Phoebe zhennan* and *Machilus pingii* [7], and investigate the incense smoke produced by different types of agarwood powder [26], etc. In most of these studies, the wood powder was used more frequently as sample material than solid chips. However, the headspace above the solid wood chips also contains the vapor generated by the volatile compounds present in the wood and is responsible for the distinct odor of wood species, which could provide information for species identification [27]. Thus, it is vital to explore the feasibility and effectiveness of wood identification utilizing solid samples coupled with the simple and reliable HS–GC–MS.

In plant metabolomics, researchers agree that a single analytical method is seldom adequate to provide the holistic view of metabolites required for metabolic profiling [28, 29]. Similarly, this concept was also applied to plant-derived materials, such as wood. Therefore, it is well worth trying to develop a multiplatform or data integration approach, including several chemical analysis technologies for wood identification. Similarity network fusion (SNF) is a computational method for data integration, which can fuse or integrate multiple types of data to create a comprehensive descriptor of the underlying data [30]. It was proposed first to construct a sample-similarity network for each data type and then integrate them into a single similarity network by a nonlinear combination method [30]. The fused similarity networks capture both shared and complementary information from multiple types of data, making it possible for a large number of data and good robustness to noise and data heterogeneity. The SNF method has been used for integrating nine multi-molecular level omics data blocks for enabling molecular classification of chronic obstructive pulmonary disease, which included data collected from isobaric tags for relative and absolute quantitation mass spectrometry and tandem mass tag mass spectrometry [31].

Among *Pterocarpus* species, *Pterocarpus santalinus* is known for its high commercial value in furniture, crafts, dyes, and medicine. *Pterocarpus tinctorius* is famous for imitating *P. santalinus* because of its similar macroscopic and microscopic features. With the increase in illegal logging of these two species, both *P. santalinus and P. tinctorius* have been listed in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) Appendix II [32]. Concurrently, the International Union for Conservation of Nature (IUCN) listed *P. santalinus* as endangered and *P. tinctorius* as least concern [33]. Besides, the commercial market preferentially selects individual species that may have full of historical and cultural information. For instance, *P. santalinus* is particularly highly valued and popular in the international market. It is also prized due to the presence of various components, such as carbohydrates, steroids, anthocyanins, saponins, tannins, phenols, triterpenoids, flavonoids, glycosides, and glycerides [34]. However, there is a lot of controversy in the market and academia about *P. santalinus* and *P. tinctorius*. Therefore, though these two species have had the same protection level in CITES, it is still critical to differentiate them in plant taxonomy. Furthermore, wood drying and heat modification are significant in processing wood products [35]. In China, high-temperature heating is a commonly accepted treatment method for high-density timber from tree species, such as *Pterocarpus*, *Dalbergia,* and *Diospyros,* to improve dimensional stability and partially recover previous drying defects when making high value-added wooden furniture. After the high-temperature heating, the compounds existing in wood may be destroyed [36, 37]. In our previous work, although wood chips of *P. santalinus* and *P. tinctorius* under air-dried and heated at lower temperature (e.g., < 70 °C) were well differentiated

by DART–MS, it is difficult to identify the high-temperature (120 °C) heated wood using DART–MS alone [19].

The high stability and simple operation for HS–GC–MS, and the fast and accurate detection of DART–MS with full ions, made them suitable for the efficient and precise detection of solid wood samples. In addition, these two methods easily generate more resulting data together with do not necessitate complex pre-treatment work, which made them more suitable for the data integration in the SNF method compared with other wood identification methods. In this study, wood chips collected from *P. santalinus* and *P. tinctorius* samples heated at a high-temperature (120 °C) were analyzed by HS–GC–MS technology. Moreover, the feasibility of SNF was explored to aggregate multidimensional data sets in types of HS–GC–MS newly obtained in this work and DART–MS previously collected from the same samples [19] for proposing a new wood identifying methodology.

Materials and methods

Wood materials

The wood samples selected in this study were the same as those in our previous work [19]. In brief, 29 authorized specimens collected from the heartwood of vouchered or validated xylarium collections were randomly divided into 20 training (including nine *P. santalinus* specimens and eleven *P. tinctorius* specimens) and 9 testing samples. Wood chips less than 2 mm in thickness were obtained from each specimen and treated at a high temperature (120°C for 10 days). The treated wood ships were subsequently conditioned at 25°C and 60% relative humidity (RH) for 30 days.

Light microscopy

A sliding microtome (SM2010R, Leica, Germany) was used to prepare transverse, radial, and tangential sections of wood chips into thicknesses of 15 μm, and then sections stained with 1% aqueous safranin were observed under a microscope (BX61, Olympus, Japan).

HS–GC–MS experiments

The HS–GC–MS experiments were newly accomplished on a GC–MS (Agilent 7890A, Santa Clara, CA, USA) coupled with a 5975C mass spectrometer (Avondale, PA, USA) whose stability is ± 0.1 m/*z* mass accuracy over 48 h. The chromatographic analysis was performed on the HP–5MS capillary fused silica column (30 m × 250 μm i.d., 0.25 μm film thickness). Helium (99.999%) was used as carrier gas at a flow rate of 1 mL/min. The injection volume was 500 μL for wood chips by a Combi-PAL autosampler (CTC Analytics, Zwingen, Switzerland). The column temperature program was set as follows: 60 °C initial temperature for 2 min, then ramped to 280 °C at a rate of 10 °C/min to, and then held at 280 °C for 5 min. The GC–MS interface temperature was maintained at 260 °C. The mass spectra obtained with full scan and mass ranged from 33 to 500 m/*z*.

The HS–GC–MS data were processed and converted into NETCDF format using MS–DIAL software (v 2.74) [38]. A peak list with aligned peak area based on the full HS–GC–MS spectra was exported for the subsequent analysis.

DART–MS experiments

Mass spectral data were previously acquired using a DART–SVP ion source (IonSense, Saugus, MA, USA) coupled to a 12 T Bruker solariX XR FTICR–MS (Bruker Daltonics, Bremen, Germany) in positive mode with a resolving power of 1,000,000 full width at half maximum (FWHM). Wood chips were analyzed directly by exposing them to the open-air space between the ion source and the mass spectrometer inlet with tweezers. The details of the DART–MS test and analytical methods of wood chips were shown in the previous report [19].

SNF

Two types of data sets, including newly obtained HS–GC–MS (Data set 1, $n = 20$) and DART–MS (Data set 2, $n = 20$) from the previous study from the same samples, were aggregated. The data matrix of DART–MS contains 129 variables, and the data matrix of HS–GC–MS contains 744 variables. Three parameters mainly used in the SNF algorithm are the number of neighbors (K), hyper parameter (μ), and the number of iterations (t), which were initially performed for the ranges recommended in [30]. Integration of multidimensional data sets and subject-based clustering were performed using the R-package SNFtool (cran.r-project.org/web/packages/SNFtool) [30]. Network graphs were generated using Python 3.0 to visualize the relationship among samples, where nodes represent wood samples and edge thickness reflects the similarity degree between each pair of samples. In the SNF algorithm, the ranking of each variable can be assessed using normalized mutual information (NMI), where the *rankfeaturesbyNMI* function can help to calculate the relative contribution of each variable in different groups based on their clustering assignments.

The validation and prediction of the classification model

Orthogonal partial least squares–discriminant analysis (OPLS–DA), which improves the partial least squares–discriminant analysis (PLS–DA) approach that employs orthogonal signal correction, was used to identify the selected species based on the HS–GC–MS alone. Moreover, the OPLS–DA analyses were generated by SIMCA-P (14.1 Umetrics, Umea, Sweden) software.

In SNF analyses, the SNF matrix from each data set, and not the full matrix of the original variables, makes it possible to integrate disparate types of data with vastly different numbers of variables [30]. Leave one out cross-validation (LOOCV) was used to evaluate the performance measure for the classification model based on the SNF. Spectral clustering was used to predict the group belonging of wood samples in the testing set. The above algorithms were implemented using R version 3.3.3.

Results and discussion

In general, the anatomical features of heartwoods between *P. santalinus* and *P. tinctorius* were difficult to distinguish [39]. However, few studies have analyzed and compared their anatomical structures (especially microstructures) after high-temperature treatment. Thus, the transverse, radial, and tangential sections of two *Pterocarpus* species after high-temperature treatment were obtained, and the optical microscope was used to characterize the structure (Fig. 1). From the light micrographs, the cell arrangement and features on the vessels, axial parenchyma, fibers, and rays after high-temperature treatment, were similar to the anatomic structures of *Pterocarpus* in general.

Figure 2 shows the total ion current diagram of HS–GC–MS and the representative DART–MS spectra of *P. santalinus* and *P. tinctorius*. In the HS–GC–MS spectra, the principal compounds for these two species originated from peaks at the retention time of 11.76 min, 12.90 min, 13.69 min, 15.38 min, 15.69 min, 16.34 min, and 18.15 min. These peaks were tentatively identified by matching their mass spectra with those in the NIST 11 library and the literature. In addition, the possible compounds, respectively, included 4-hydroxybenzaldehyde ($C_7H_6O_2$, 7.9%), 3,5-dimethoxybenzaldehyde ($C_9H_{10}O_3$, 33.76%), (+)-β-selinene ($C_{15}H_{24}$, 7.09%), 2-Naphthalenemethanol, 1,2,3,4,4a,5,6,7-octahydro-α,α,4a,8-tetramethyl-, (2R-cis)- ($C_{15}H_{26}O$, 4.64%), 2-Naphthalenemethanol, decahydro-α,α,4a-trimethyl-8-methylene-, [2R-(2α,4aα,8aβ)]- ($C_{15}H_{26}O$, 11.69%), 6-Isopropenyl-4,8a-dimethyl-1,2,3,5,6,7,8,8a-octahydro-naphthalen-2-ol ($C_{15}H_{24}O$, 14.75%), and spathulenol ($C_{15}H_{24}O$, 11.06%). According to our previous inference about DART–MS, peaks at 221.19 m/*z* and 257.11 m/*z* could be assigned to spathulenol (molecular formula $C_{15}H_{24}O$) and pterostilbene (molecular formula $C_{16}H_{16}O_3$), respectively [39]. These chemical components mainly come from the extractives of heartwood. Although some chemical components seem to be detected in both HS–GC–MS and DART–MS, most substances were different during these two detection methods. DART–MS contains several high molecular weight (>500 m/*z*) compounds that cannot be detected in HS–GC–MS. Meanwhile, the existed differences in the chemical spectra of HS–GC–MS and DART–MS can be seen between *P. santalinus* and *P. tinctorius* in Fig. 2. An observation came to light during the analysis of HS–GC–MS spectra that the chemical peaks of *P. santalinus* mostly concentrated in the retention time range of 13–19 min. In comparison, the chemical peaks of *P. tinctorius* mainly appeared before the retention time of 13 min. Considering the intraspecific variation, the overall statistical analysis is essential to discriminate between the two wood species.

Spathulenol and pterostilbene were also elucidated as the critical compounds for the separation of the ethanol and water (EW) extracts of *P. santalinus* and *P. tinctorius* in the previous work on the chemotaxonomical discrimination using GC–MS with the liquid injection [13]. Compared with the GC–MS spectra of EW extracts, the HS–GC–MS obtained from high-temperature heated wood chips showed a significant difference in the number of peaks and retention times. The spathulenol also exists in the high-temperature heated wood chips, while the pterostilbene was not detected using HS–GC–MS. The direct analysis in real time–time-of-flight mass spectrometry paired with discriminant analysis of principal components has been successful in classifying seven *Pterocarpus* species, including *P. erinaceus*, *P. santalinus*, *P. tinctorius*, *P. indicus*, *P. macrocarpus*, *P. dalbergioides*, and *P. soyauxii* [22]. In this research, *P. santalinus* showed higher intensity at 219.1763 m/*z*, and *P. tinctorius* showed ions that were significantly reduced or missing in other species, including ions at 224.1285 and 247.1444 m/*z*. This result is different from our findings here and the previous studies on the DART–MS of air-dried wood chips [19]. However, in the survey of the original geographic region of timber of *Pterocarpus* samples by direct analysis in real time–time-of-flight mass spectrometry (DART–TOFMS), the ion at 257.115 m/*z* was also found in the species of *P. tinctorius* [40], which is consistent with our result. Therefore, it would be worth detecting the distinction of ions between these two mass spectrometers for the same species in further work.

In our previous work, the classification of these two *Pterocarpus* species based on DART–MS showed that the two species are more difficult to distinguish after high-temperature heating treatment. The classification accuracy was only 66.67% when the OPLS–DA was used [19]. Herein, a classification model between two species was also built based on the HS–GC–MS data set of the high-temperature heated wood chips coupled with OPLS–DA. Subsequently, the classification model made with a training set was applied to the testing set, consisting of records with unknown class labels. The performance of a classification model was evaluated by the counts of test records correctly and incorrectly predicted by the model. These counts results are tabulated in Table 1. Seven samples in

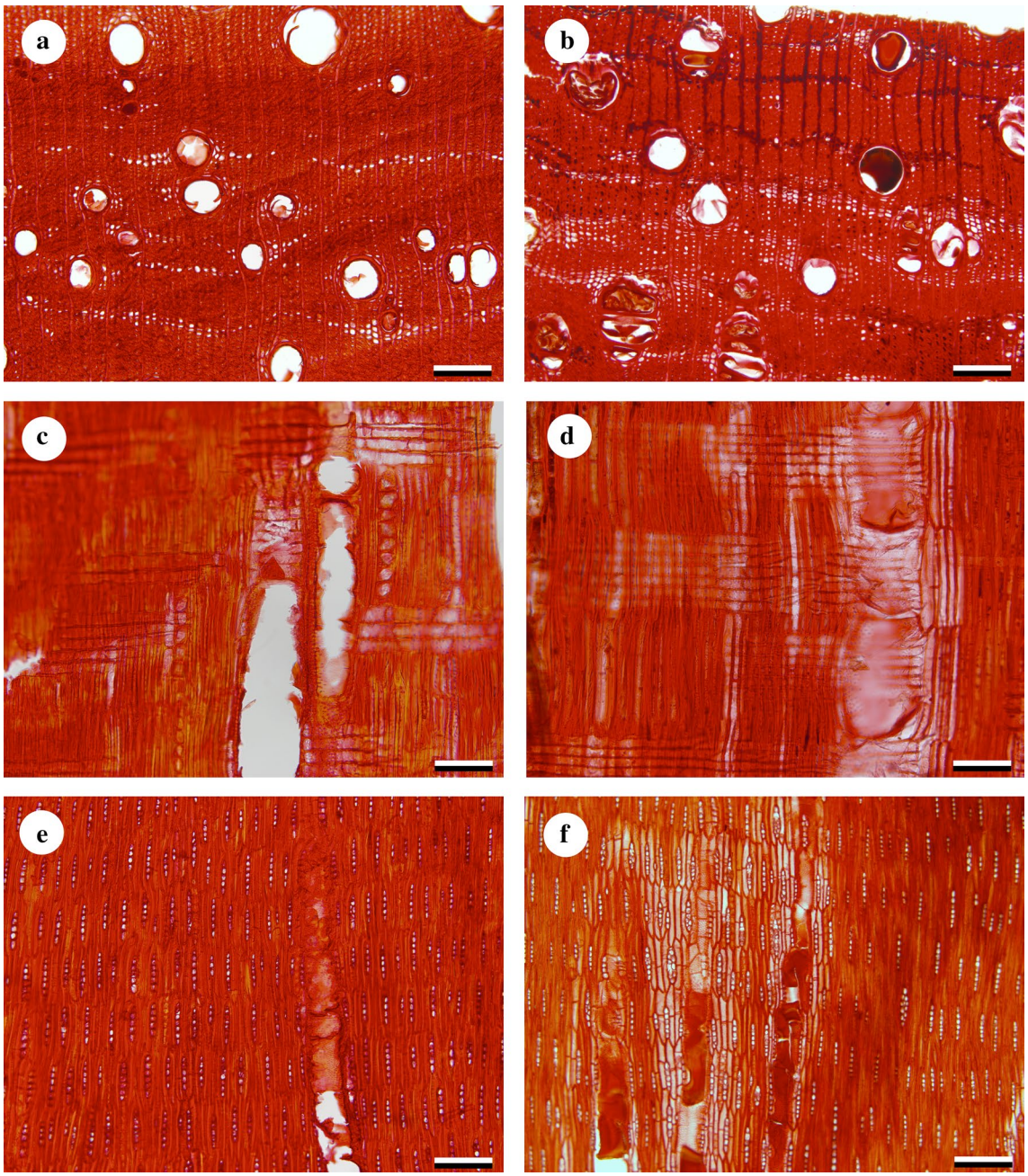

Fig. 1 Light micrographs of transverse, radial and tangential sections of heated *Pterocarpus santalinus* (**a, c, e**) and *Pterocarpus tinctorius* (**b, d, f**) wood. – Scale bars, 200 µm (**a**, **b**, **e**, **f**), 100 µm (**c**, **d**)

the testing set were identified correctly, and the accuracy was 77.78% (Table 1).

SNF can tackle the issue that the data are gathered from more than one source, because it can take advantage of the commonalities of different data types to obtain a better classification performance than a single data type. Thus, given the low classification accuracy based on DART–MS or HS–GC–MS data set alone, the SNF analyses were tried to improve differentiating performance of two *Pterocarpus* species. First, HS–GC–MS data and DART–MS data of high-temperature heated wood chips from the *P. santalinus* and *P. tinctorius* were used to construct fused similarity networks. Compared with K and t, the parameter μ has little effect on the result, so a

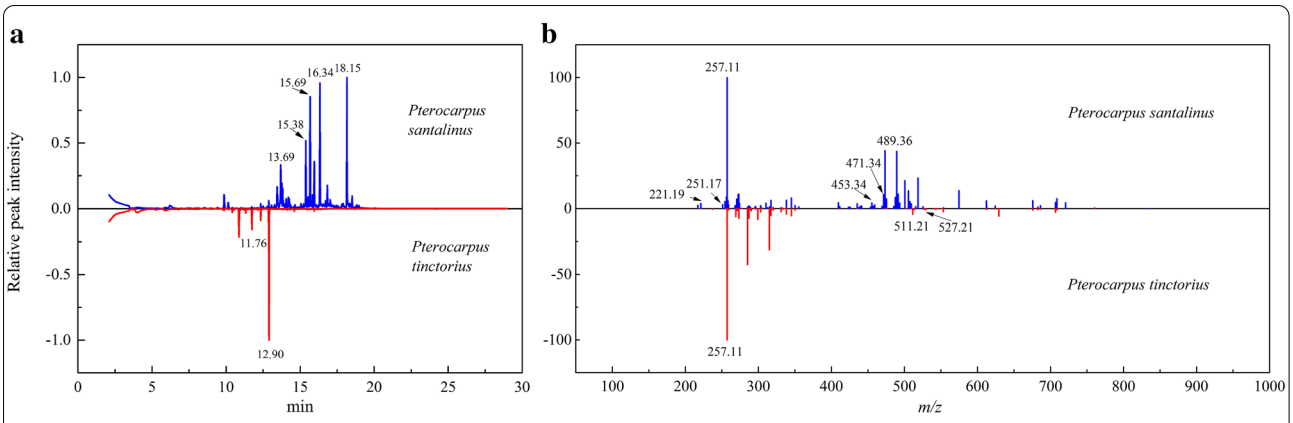

Fig. 2 Total ion current diagram of HS–GC–MS (**a**) and the representative DART–MS spectra (**b**) for *Pterocarpus santalinus* and *Pterocarpus tinctorius*

Table 1 Prediction accuracy using HS–GC–MS, DART–MS, and the SNF composed of HS–GC–MS and DART–MS data, respectively

Samples	Accuracy
HS–GC–MS data of wood chips	77.78%
DART–MS data of wood chips	66.67%
SNF composed of HS–GC–MS and DART–MS data	100.00%

with random sampling was used to predict test groups and to select the value of parameters. The *K* parameter was evaluated from 2 to 19. The *t* parameter was assessed from 1 to 50. In the LOOCV test, the areas under curve (AUC) of the receiver operating characteristic (ROC) curve were used to evaluate the ability to discriminate wood species. The results are shown in Fig. 3. It can be seen from Fig. 3 that AUC was 1 when the *K* parameter was evaluated from 3 to 19, and the *t* parameter was evaluated from 1 to 50. However, when the *K* parameter was

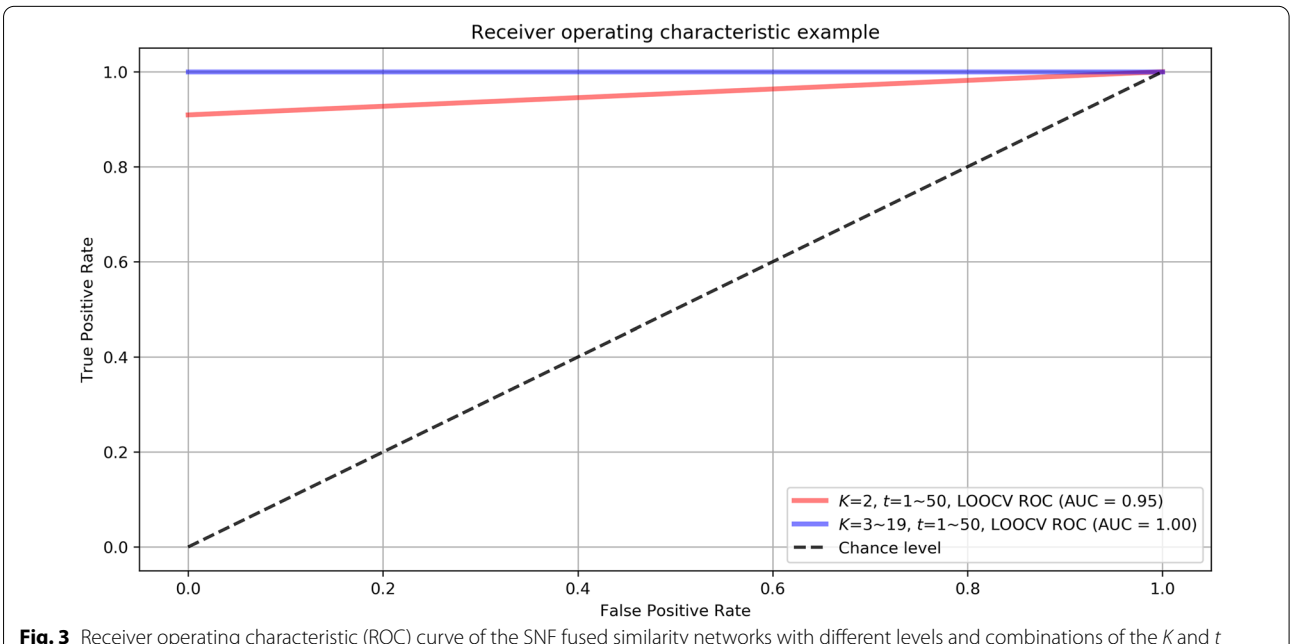

Fig. 3 Receiver operating characteristic (ROC) curve of the SNF fused similarity networks with different levels and combinations of the *K* and *t*

practical value of 0.5 was selected. Then the SNF-fused similarity networks were constructed and compared with different levels and combinations of the *K* and *t*. LOOCV set to 2 and the *t* parameter was assessed from 1 to 50, the value of AUC was only 0.95. Considering the ranges

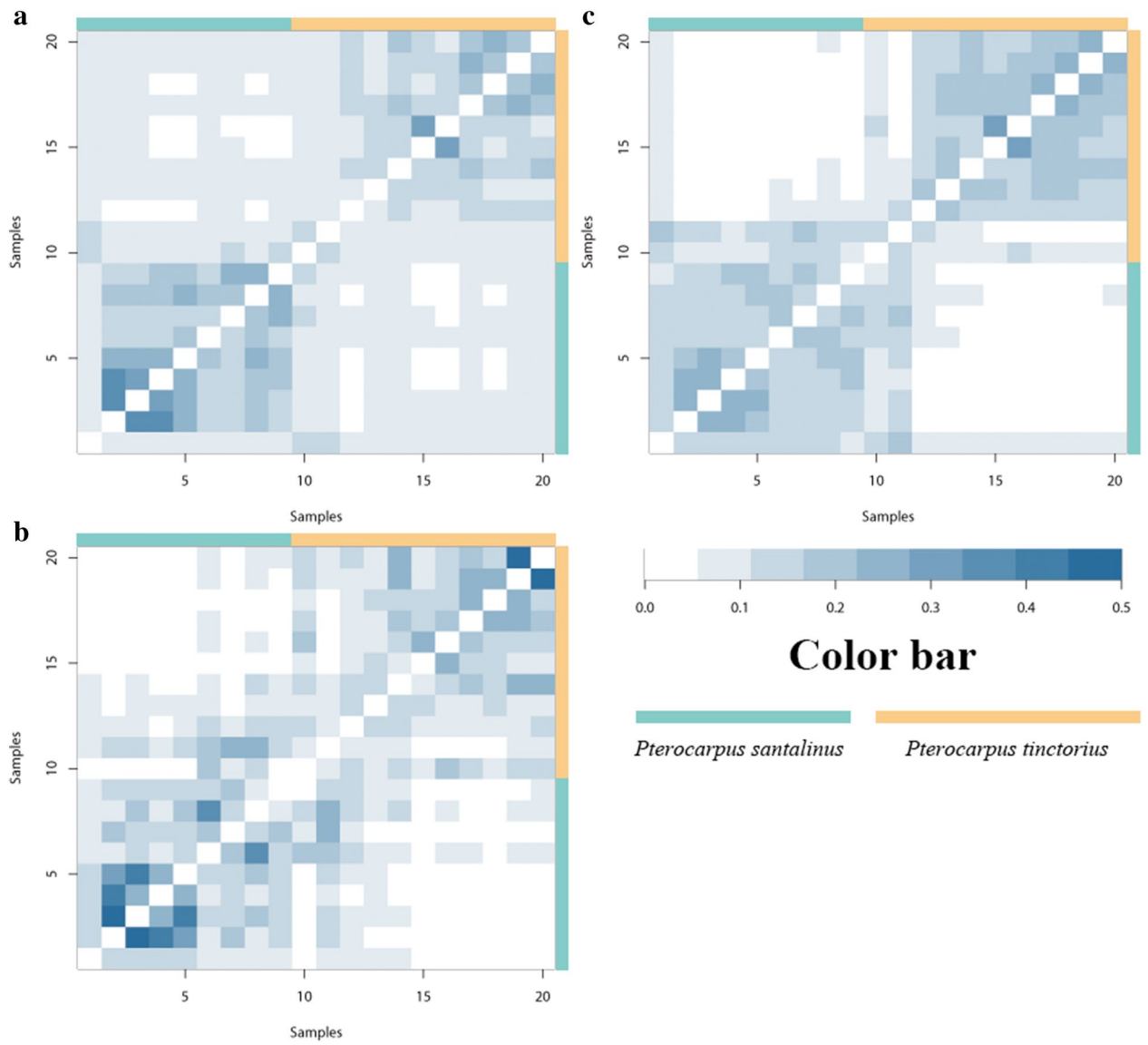

Fig. 4 Wood chips similarities for each of the data types independently compared to SNF fused similarity. HS–GC–MS data (**a**), DART–MS data (**b**), and SNF-fused similarity matrix (**c**). The first 9 samples are *Pterocarpus santalinus*, and the last 11 samples are *Pterocarpus tinctorius*

(K parameter: usually (10–30), t parameter: usually (10–50)) recommended in previous research [30], we finally selected $K=10$, $\mu = 0.5$, and $t=10$ in all further SNF analyses to reduce the amount of calculation as much as possible based on ensuring the classification accuracy.

During the SNF, the *P. santalinus* and *P. tinctorius* similarities for Data set 1, Data set 2, and their SNF fused similarity were conducted. The results are visualized in Fig. 4. It can be seen that there are 20 squares in the horizontal and vertical directions, representing 20 samples, the first nine samples are *P. santalinus*, and the last eleven samples are *P. tinctorius*. The shade of each square represents the degree of similarity between the two samples. The darker the color of the square, the higher the similarity exists between the two samples. The networks built using a single data type present many different patterns of similarity between these two species. It is difficult to classify the *P. santalinus* and *P. tinctorius* based on the HS–GC–MS data (Fig. 4a) or the DART–MS data alone (Fig. 4b) alone. In comparison, the fused network using two data types shows a much clearer picture of clustering in our set of wood species (Fig. 4c).

Network visualizations are becoming more attractive for data representation, because they can clearly show

the relationship between the data as graphs. Here, network graphs between two *Pterocarpus* species based on Data set 1, Data set 2, and their SNF fused similarity are shown in Fig. 5. In the network graphs, many thin edges can be seen in the network for Data set 1 (Fig. 5a) or Data set 2 (Fig. 5b) alone, which present a weak distinctiveness between *P. santalinus* and *P. tinctorius*. The fused network gives a clearer picture of clustering, illustrated by the tightness of connectivity within clusters (Fig. 5c).

However, due to the limitation of the data and the samples, the developed model is prone to overfitting. When the overfitting exists, the model performs perfectly on the training set, while fitting poorly on the testing set. Thus, Spectral clustering was used to predict the group label of the testing set based on the similarities to all samples in the network to evaluate the performance of the classification model from multiple data sets fusion. The prediction accuracy was assessed by comparing the predicted label and the true label, and the results are listed in Table 1. As can be seen, all the testing set samples from the high-temperature heated wood chips based on the SNF-fused network of the DART–MS data set and HS–GC–MS data set were correctly classified. The prediction accuracy reached 100%, which was much higher than based on only one data type. To further understand which ion peaks or chemical compounds have a greater impact on the wood identification, the contribution of each variable was calculated by NMI. As for the HS–GC–MS, the variables with higher contribution are peaks at 11.76 min, 12.34 min, 12.90 min, 13.69 min, 15.38 min, 16.34 min, and 18.15 min, respectively. The relative peak areas of the seven peaks present significant differences between the two *Pterocarpus* species (Fig. 6). For DART–MS, the top seven ranked by NMI are ions at 251.17 m/*z*, 257.11 m/*z*, 453.33 m/*z*, 471.34 m/*z*, 487.34 m/*z*, 489.36 m/*z*, and 511.21 m/*z*, respectively. Although the relative peak intensities in several ions are not very high, a visible distinction can be observed between these two species (Fig. 7). Besides, some intra-specific variations exist among samples from the same species. These peaks in the HS–GC–MS or ions in the DART–MS provide evidence about wood identification between two *Pterocarpus* species, but whether they can be used as taxonomic markers requires further experimental validation.

It has been suggested that gas chromatography with high-resolution quadrupole time of flight mass spectrometry and the DART–TOFMS techniques are complementary to one another, each with their advantages in the study about the identification of protected *Dalbergia* timber [41]. Although there are differences in the models of mass spectrometers, our study also supports this opinion. It confirms the complementary between HS–GC–MS and DART–MS for wood identification using SNF.

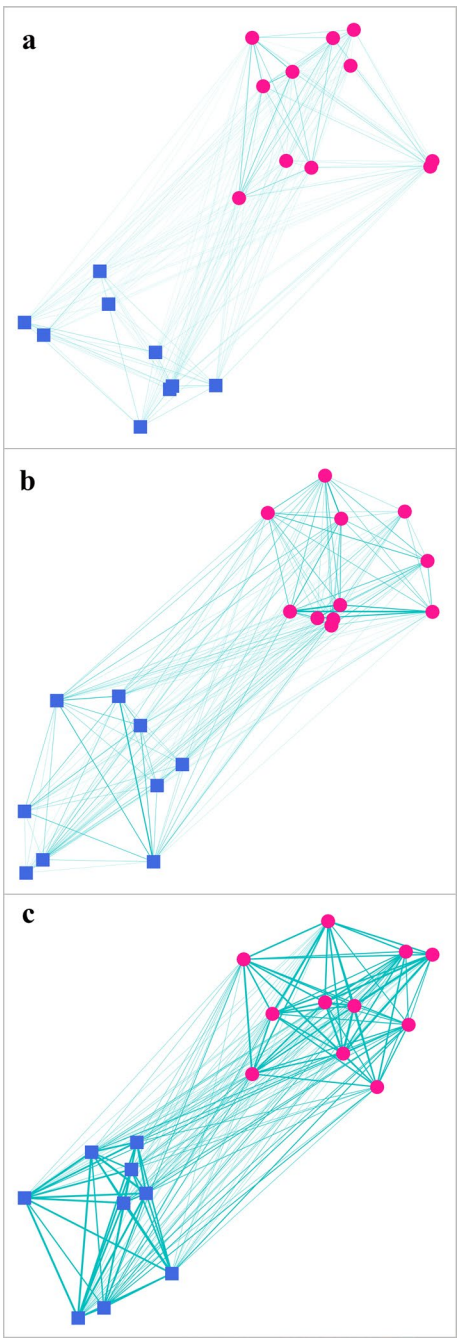

Fig. 5 Subject similarity networks of wood chips based on HS–GC–MS data (**a**), DART–MS data (**b**), and their fused SNF similarity network (**c**). Nodes represent wood samples (blue square: *Pterocarpus santalinus*; red circle: *Pterocarpus tinctorius*), edge thickness reflects the strength of the similarity

The SNF can obtain shared and complementary information from various kernel matrices so that the integrated matrix reveals HS–GC–MS/DART–MS information as much as possible. However, it should also be noted that

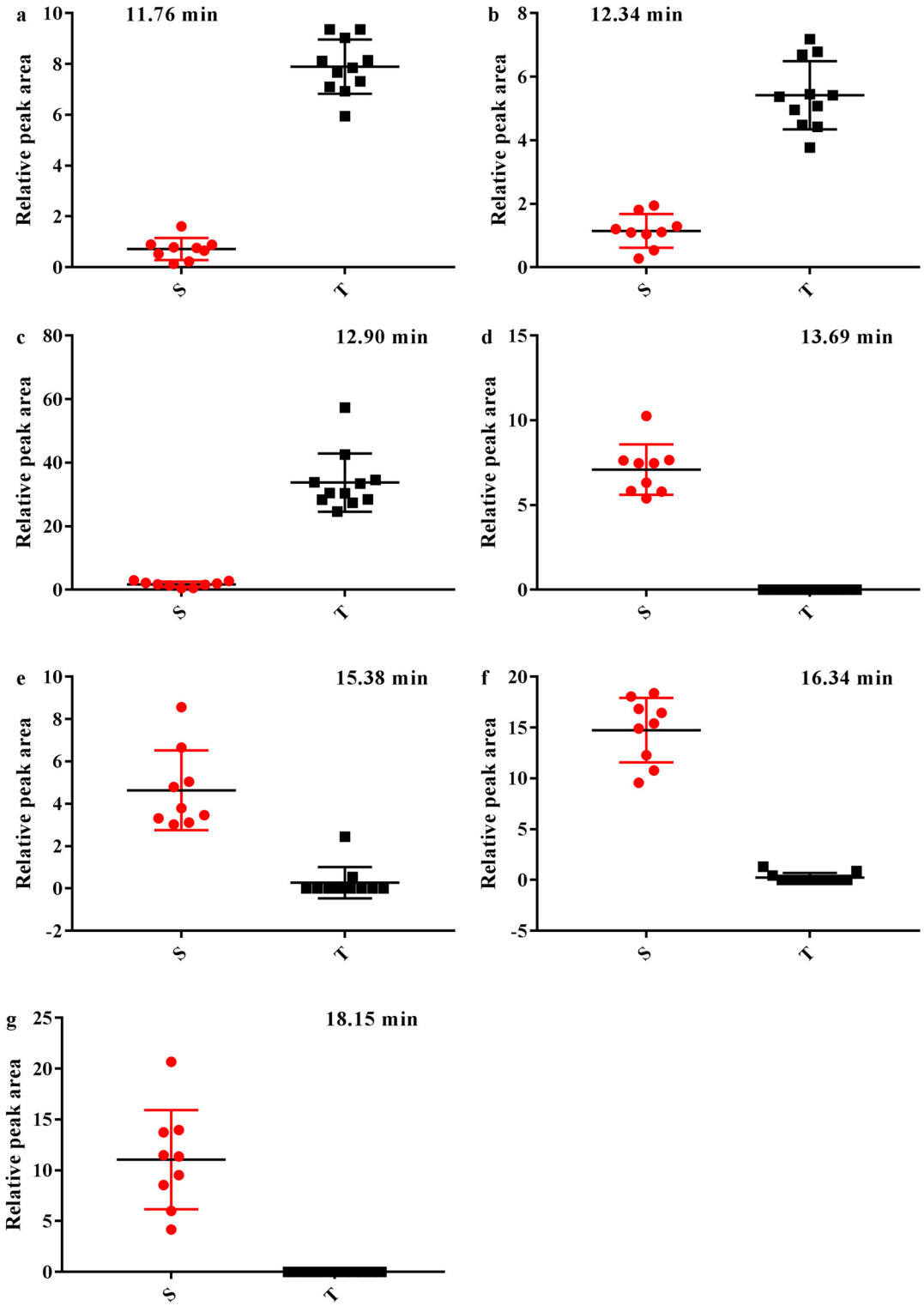

Fig. 6 Relative peak areas of seven peaks from the HS–GC–MS coupled with SNF. 11.76 min (**a**), 12.34 min (**b**), 12.90 min (**c**), 13.69 min (**d**), 15.38 min (**e**), 16.34 min (**f**), 18.15 min (**g**). *Pterocarpus santalinus* (**S**), *Pterocarpus tinctorius* (**T**)

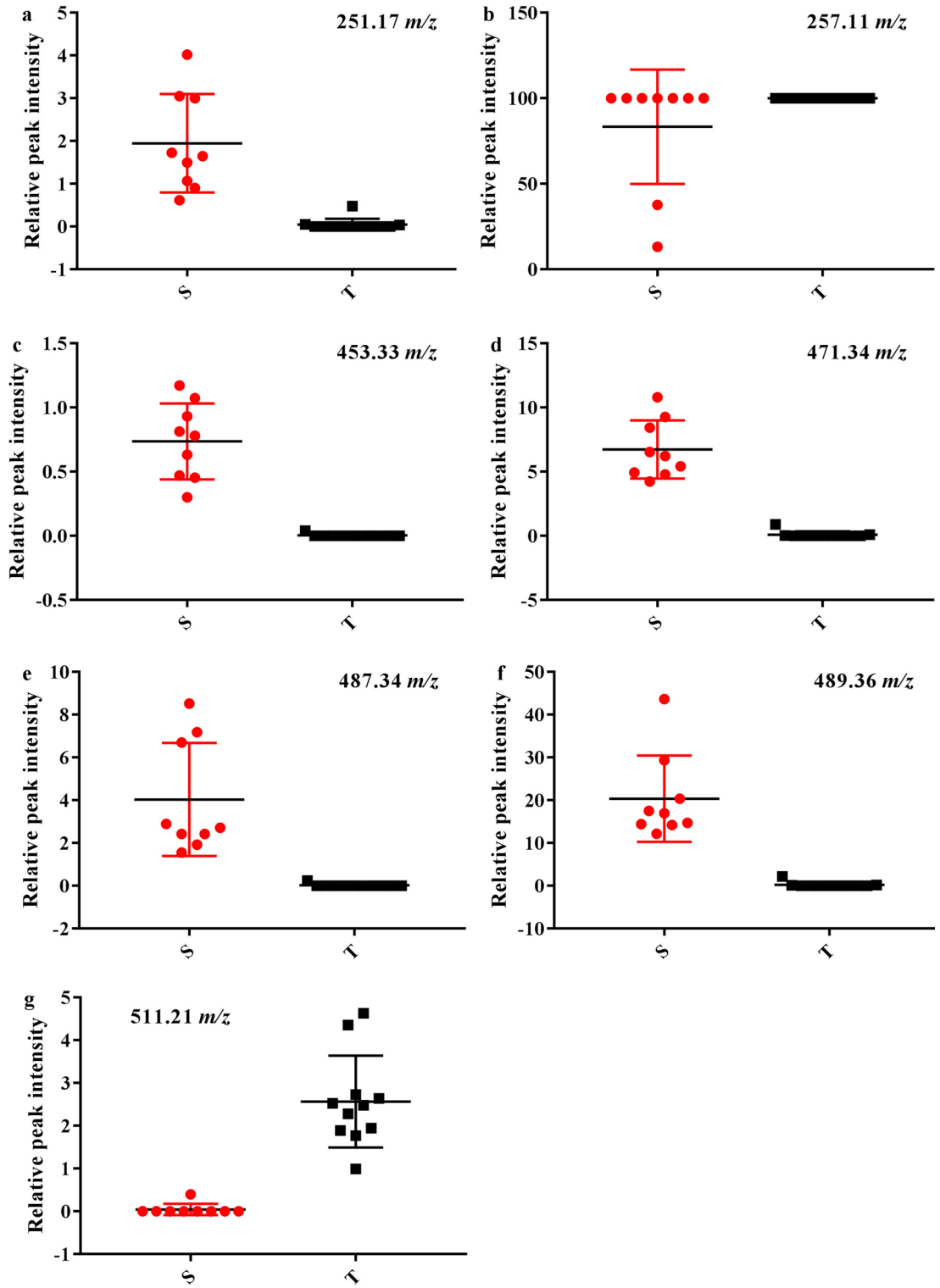

Fig. 7 Relative peak intensities of seven ions from the DART–MS coupled with SNF. 251.17 m/*z* (**a**), 257.11 m/*z* (**b**), 453.33 m/*z* (**c**), 471.34 m/*z* (**d**), 487.34 m/*z* (**e**), 489.36 m/*z* (**f**), 511.21 m/*z* (**g**). *Pterocarpus santalinus* (**S**), *Pterocarpus tinctorius* (**T**)

although the integrated matrix contains two types of chemical information, further research on the details, such as how the chemical information is related and complementary to each other in the matrix, is needed in future work.

The SNF of HS–GC–MS and DART–MS data for wood chips might achieve a complementary effect of the two data sets and ultimately improve the ability to distinguish and classify *P. santalinus* and *P. tinctorius*. It has been reported that SNF could obtain helpful information from a small number of samples [30]. *Pterocarpus* wood samples are challenging to collect, especially for the authorized wood samples from xylarium or verified specimens by DNA barcodes. The high classification accuracy in this study proved that the SNF method is appropriate for analyzing valuable and rare wood species. In addition, it is necessary to emphasize that SNF requires relatively high statistical knowledge, so it is worth trying if only one method cannot meet the requirements. Otherwise, it will increase many calculations. Wood anatomical features, DNA barcoding, wood images, physical features (color, density), and other chemical features (infrared spectrum, near-infrared spectrum, fluorescence spectrum, etc.) contain much critical information for the wood species. Therefore, SNF-fused multidimensional or comprehensive data sets consisting of DNA barcoding, anatomic features, chemical features, and physical features could be further proposed in future research on wood identification.

Conclusions

In this paper, high-temperature (e.g., 120 °C) heated solid wood samples from *P. santalinus* and *P. tinctorius* were analyzed by HS–GC–MS technology, and a novel machine-learning method SNF was applied to the wood identification of these two species. The HS–GC–MS analysis revealed the differences between chemical profiles of the volatile constituents of high-temperature heated solid wood samples of each wood species from each other. Still, the prediction accuracy of the classification model established by HS–GC–MS coupled with OPLS–DA was only 77.78%. After SNF, one fused network composed of HS–GC–MS data set and DART–MS data set was established on wood chips heated at high temperature. The developed SNF-fused network exhibits a better level to group two wood species than that based on the HS–GC–MS data set or DART–MS data set alone. Higher discrimination power (100% of prediction accuracy) was obtained when the Spectral clustering approach was used. This study demonstrates the capability of the HS–GC–MS technique in the high-temperature heated solid wood analysis and its potential as a simple, non-destructive, and rapid alternative for wood species identification. Multidimensional or comprehensive data sets based on SNF are the potential to have a broader application in the wood identification field. Furthermore, considering the possible effects on the reliability of data analysis with the limitation of the data and the samples, a larger sampling size should be adopted for the establishment of SNF-fused data sets in future research.

Abbreviations

CITES: Convention on International Trade in Endangered Species of Wild Fauna and Flora; HS–GC–MS: Headspace–gas chromatography–mass spectrometry; SNF: Similarity network fusion; DART–MS: Direct analysis in real time–mass spectrometry; GC–MS: Gas chromatography–mass spectrometry; HS–SPME: Headspace solid–phase microextraction; HS: Headspace injection; IUCN: International Union for Conservation of Nature; RH: Relative humidity; FWHM: Full width at half maximum; K: Number of neighbors; μ: Hyper parameter; t: Number of iterations; NMI: Normalized mutual information; OPLS–DA: Orthogonal partial least squares discriminant analysis; PLS–DA: Partial least squares discriminant analysis; LOOCV: Leave one out cross-validation; EW: Ethanol and water; DART–TOFMS: Direct analysis in real time–time-of-flight mass spectrometry; AUC: Areas under curve; ROC: Receiver operating characteristic.

Acknowledgements

We would like to acknowledge Mr. Shuwen Hao for his assistance to data analysis and model development, Professor Xiaomei Jiang, Professor Yongdong Zhou, Mr. Yonggang Zhang, Dr. Bo Liu of Research Institute of Wood Industry, Chinese Academy of Forestry (CAF) for their contributions to collection and preparation of wood specimens. The authors also wish to acknowledge Dr. Victor Deklerck from Royal Botanic Gardens, Kew, United Kingdom for discussions.

Author contributions

MZ: conceptualization, investigation, data curation, methodology, software, writing (original draft); JG: data curation, validation; YL: resources, data curation; LJ: resources, investigation; TH: data curation; YY: conceptualization, supervision, writing—review, and editing, funding acquisition. All authors read and approved the final manuscript.

Funding

This work was supported financially by the Project of Shaanxi Provincial Science and Technology Department (Grant No. 2021JQ-498), National Special Support Plan of China (Grant No. W02020331), and Chinese Academy of Forestry (Grant No. CAFYBB2021ZD002).

Availability of data and materials

The data sets used in this study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors have no conflicts of interest directly relevant to the content of this article.

Author details

[1]College of Art, Xi'an University of Architecture and Technology, Xi'an 710055, China. [2]Department of Wood Anatomy and Utilization, Chinese Research Institute of Wood Industry, Chinese Academy of Forestry, Beijing 100091, China. [3]Wood Collections (WOODPEDIA), Chinese Academy of Forestry, Beijing 100091, China.

Received: 19 January 2022 Accepted: 6 June 2022
Published online: 21 June 2022

References

1. Scotland N, Ludwig S (2002) Deforestation, the timber trade and illegal logging. In: EC workshop on Forest Law, enforcement, Governance and Trade. Brussels, Belgium
2. Dormontt EE, Boner M, Braun B, Breulmann G, Degen B, Espinoza E, Gardner S, Guillery P, Hermanson JC, Koch G (2015) Forensic timber identification: it's time to integrate disciplines to combat illegal logging. Biol Conserv 191:790–798. https://doi.org/10.1016/j.biocon.2015.06.038
3. Carmona RJ, Wiemann MC, Baas P, Barros C, Chavarria GD, McClure PJ, Espinoza EO (2020) Forensic identification of CITES appendix I *Cupressaceae* using anatomy and mass spectrometry. IAWA J 1:720–739. https://doi.org/10.1163/22941932-bja10002
4. Deklerck V, Fowble KL, Coon AM, Espinoza EO, Beeckman H, Musah RA (2021) Opportunities in phytochemistry, ecophysiology and wood research via laser ablation direct analysis in real time-imaging mass spectrometry. New Phytol 234(1):319–331. https://doi.org/10.1111/nph.17893
5. Horikawa Y, Mizuno-Tazuru S, Sugiyama J (2015) Near-infrared spectroscopy as a potential method for identification of anatomically similar Japanese diploxylons. J Wood Sci 61(3):251–261. https://doi.org/10.1007/s10086-015-1462-2
6. Hwang SW, Horikawa Y, Lee WH, Sugiyama J (2016) Identification of *Pinus* species related to historic architecture in Korea using NIR chemometric approaches. J Wood Sci 62:156–167. https://doi.org/10.1007/s10086-016-1540-0
7. Xu B, Zhu T, Li J, Liu S (2013) Identification of wood between *Phoebe zhennan* and *Machilus pingii* using the gas chromatography-mass spectrometry direct injection technique. Eur J Mass Spectrom 19(3):187–193. https://doi.org/10.1255/ejms.1226
8. Gao X, Xie M, Liu S, Guo X, Chen X, Zhong Z, Wang L, Zhang W (2014) Chromatographic fingerprint analysis of metabolites in natural and artificial agarwood using gas chromatography-mass spectrometry combined with chemometric methods. J Chromatogr B 967:264–273. https://doi.org/10.1016/j.jchromb.2014.07.039
9. Liu X, Xu D, Yang Z, Zhang N (2017) Chemical composition of essential oils from the heartwood of *Pterocarpus macrocarpus* by different extraction methods in southern China. J Essent Oil-Bear Plant JEOP 20(1):110–115. https://doi.org/10.1080/0972060X.2016.1278183
10. Liu R, Wang C, Huang A, Lv B (2018) Characterization of odors of wood by gas chromatography-olfactometry with removal of extractives as attempt to control indoor air quality. Molecules 23(1):203. https://doi.org/10.3390/molecules23010203
11. Linda S, Patrick B, Andrea B (2018) Resolving the smell of wood-Identification of odour-active compounds in scots pine (*Pinus sylvestris* L.). Sci Rep 8(1):8294–8294. https://doi.org/10.1038/s41598-018-26626-8
12. Liu Y, Zhu X, Qin X, Wang W, Hu Y, Yuan D (2020) Identification and characterization of odorous volatile organic compounds emitted from wood-based panels. Environ Monit Assess 192(6):1–10. https://doi.org/10.1007/s10661-019-7939-5
13. Zhang M, Zhao G, Guo J, Liu B, Jiang X, Yin Y (2019) A GC-MS protocol for separating endangered and non-endangered *Pterocarpus* wood species. Molecules 24:799. https://doi.org/10.3390/molecules24040799
14. Ghavidel A, Bak M, Hofmann T, Hosseinpourpia R, Vasilache V, Sandu I (2021) Comparison of chemical compositions in wood and bark of Persian silk tree (*Albizia julibrissin Durazz*). Wood Mater Sci Eng. https://doi.org/10.1080/17480272.2021.1953141
15. Li T, Li G, Li J, Li X, Li M, Li Y (2021) HS-SPME and GC-MS for the analysis of odorous constituents from heat-treated rubberwood and the chemical change of heat-treated rubberwood by XPS analysis. Wood Sci Technol 55(2):361–378. https://doi.org/10.1007/s00226-020-01253-7
16. Yao C, Qi L, Zhong F, Li N, Ma Y (2022) An integrated chemical characterization based on FT-NIR, GC-MS and LC-MS for the comparative metabolite profiling of wild and cultivated agarwood. J Chromatogr B 1188:123056. https://doi.org/10.1016/j.jchromb.2021.123056
17. Cody RB, Dane AJ, Dawson-Andoh B, Adedipe EO, Nkansah K (2012) Rapid classification of white oak (*Quercus alba*) and northern red oak (*Quercus rubra*) by using pyrolysis direct analysis in real time (DART™) and time-of-flight mass spectrometry. J Anal Appl Pyrol 95:134–137. https://doi.org/10.1016/j.jaap.2012.01.018
18. Espinoza EO, Lancaster CA, Kreitals NM, Hata M, Cody RB, Blanchette RA (2014) Distinguishing wild from cultivated agarwood (*Aquilaria spp.*) using direct analysis in real time and time of-flight mass spectrometry. Rapid Commun Mass Spectrom 28:281–289. https://doi.org/10.1002/rcm.6779
19. Zhang M, Zhao G, Guo J, Wiedenhoeft AC, Liu CC, Yin Y (2019) Timber species identification from chemical fingerprints using direct analysis in real time (DART) coupled to Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS): comparison of wood samples subjected to different treatments. Holzforschung 73:975–985. https://doi.org/10.1515/hf-2018-0304
20. Giffen JE, Lesiak AD, Dane AJ, Cody RB, Musah RA (2017) Rapid species-level identification of *Salvias* by chemometric processing of ambient ionisation mass spectrometry-derived chemical profiles. Phytochem Anal 28:16–26. https://doi.org/10.1002/pca.2639
21. Deklerck V, Lancaster CA, Van Acker J, Espinoza EO, Van den Bulcke J, Beeckman H (2020) Chemical fingerprinting of wood sampled along a pith-to-bark gradient for individual comparison and provenance identification. Forests 11:107. https://doi.org/10.3390/f11010107
22. Price ER, Miles-Bunch I, Gasson PE, Lancaster CA (2021) *Pterocarpus* wood identification by independent and complementary analysis of DART-TOFMS, microscopic anatomy, and fluorescence spectrometry. IAWA J 42(4):397–418. https://doi.org/10.1163/22941932-bja10064
23. Wang F, Huang A, Yin X, Wang W, Chen J (2018) Multilevel profiling and identification of *Dalbergia odorifera* and *Dalbergia stevensonii* by FTIR, NMR and GC/MS. Chin Chem Lett 29:1395–1398
24. Trine A, Federico C, Henrik S (2015) Optimization of biochemical screening methods for volatile and unstable sesquiterpenoids using HS-SPME-GC-MS. Chromatography 2(2):277–292. https://doi.org/10.3390/chromatography2020277
25. Li C, Xu F, Cao C, Shang MY, Zhang CY, Yu J, Liu GX, Wang X, Cai SQ (2013) Comparative analysis of two species of *Asari Radix* et Rhizoma by electronic nose, headspace GC–MS and chemometrics. J Pharm Biomed Anal 85:231–238. https://doi.org/10.1016/j.jpba.2013.07.034
26. Chen Y, Yan T, Zhang Y, Wang Q, Li Q (2020) Characterization of the incense ingredients of cultivated grafting kynam by TG-FTIR and HS-GC-MS. Fitoterapia 142(1):104493. https://doi.org/10.1016/j.fitote.2020.104493
27. Kalaw JM, Sevilla IIIF (2019) Differentiation of wood species using headspace fingerprinting through fourier-transform infrared spectroscopy. Acta Manilana 67:31–38
28. Okazaki Y, Saito K (2012) Recent advances of metabolomics in plant biotechnology. Plant Biotechnol Rep 6:1–15. https://doi.org/10.1007/s11816-011-0191-2
29. Wei L, Lin M, Han B, Deng X, Hou W, Liao Q, Xie Z (2016) The comparison of cinnamomi cortex and *Cinnamomum burmannii* blume using 1H NMR and GC-MS combined with multivariate data analysis. Food Anal Methods 9:2419–2428. https://doi.org/10.1007/s12161-016-0418-5
30. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11:333. https://doi.org/10.1038/nmeth.2810
31. Li CX, Wheelock CE, Sköld CM, Wheelock ÅM (2018) Integration of multi-omics datasets enables molecular classification of COPD. Eur Respir J 51:1701930. https://doi.org/10.1183/13993003.01930-2017
32. CITES (2019) Decisions made on proposals to amend appendices I and II at CoP18. https://www.cites.org/eng/updates_decisions_cop18_species_proposals. Accessed 12 September 2019.
33. Planned red list updates (2022) International Union for Conservation of Nature (IUCN). https://www.iucnredlist.org/assessment/updates. Accessed 12 May 2022
34. Narayan S, Devi RS, Devi CSS (2007) Role of *Pterocarpus santalinus* against mitochondrial dysfunction and membrane lipid changes induced by ulcerogens in rat gastric mucosa. Chem Biol Interact 170:67–75. https://doi.org/10.1016/j.cbi.2007.07.005
35. Herrera-Díaz R, Sepúlveda-Villarroel V, Pérez-Peña N, Salvo-Sepúlveda L, Salinas-Lira C, Llano-Ponte R, Ananías RA (2018) Effect of wood drying and heat modification on some physical and mechanical properties of

radiata pine. Drying Technol 36:537–544. https://doi.org/10.1080/07373937.2017.1342094
36. Esteves B, Videira R, Pereira H (2011) Chemistry and ecotoxicity of heat-treated pine wood extractives. Wood Sci Technol 45:661–676. https://doi.org/10.1007/s00226-010-0356-0
37. Yang Y, Zhan TY, Lu JX, Jiang JH (2015) Influences of thermo-vacuum treatment on colors and chemical compositions of alder birch wood. Bio Resour 10:7936–7945
38. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods 12(6):523–526. https://doi.org/10.1038/nmeth.3393
39. Zhang M, Zhao G, Liu B, He T, Yin Y (2019) Wood discrimination analyses of *Pterocarpus tinctorius* and endangered *Pterocarpus santalinus* using DART-FTICR-MS coupled with multivariate statistics. IAWA J 40(1):1–16. https://doi.org/10.1163/22941932-40190224
40. Price ER, Miles-Bunch I, Gasson PE, Lancaster CA (2021) Inference of origin of *Pterocarpus* timber by chemical profiling of ambient ionization mass spectra. Forensic Sci Intern: Anim Env 1:100032. https://doi.org/10.1016/j.fsiae.2021.100032
41. Shang D, Brunswick P, Yan J, Brunoa J, Duchesneb I, Isabelb N, VanAggelena G, Kimc M, Evans PD (2020) Chemotyping and identification of protected *Dalbergia* timber using gas chromatography quadrupole time of flight mass spectrometry. J Chromatogr A 1615:460775. https://doi.org/10.1016/j.chroma.2019.460775

Publisher's Note