Journal of Pharmaceutical Analysis 15 (2025) 101116

Contents lists available at ScienceDirect

Journal of Pharmaceutical Analysis

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

Original article

Integration of deep neural network modeling and LC-MS-based pseudo-targeted metabolomics to discriminate easily confused ginseng species

Meiting Jiang [a, b, 1], Yuyang Sha [c, 1], Yadan Zou [a, b, 1], Xiaoyan Xu [a, b], Mengxiang Ding [a, b], Xu Lian [c], Hongda Wang [a, b], Qilong Wang [a, b], Kefeng Li [c, **], De-an Guo [a, b, d, ***], Wenzhi Yang [a, b, *]

[a] *State Key Laboratory of Chinese Medicine Modernization, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China*
[b] *Haihe Laboratory of Modern Chinese Medicine, Tianjin, 301617, China*
[c] *Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, 999078, China*
[d] *Shanghai Research Center for Modernization of Traditional Chinese Medicine, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China*

ARTICLE INFO

Article history:
Received 25 June 2024
Received in revised form
19 September 2024
Accepted 23 September 2024
Available online 26 September 2024

Keywords:
Liquid chromatography-mass spectrometry
Pseudo-targeted metabolomics
Deep neural network
Species differentiation
Ginseng

ABSTRACT

Metabolomics covers a wide range of applications in life sciences, biomedicine, and phytology. Data acquisition (to achieve high coverage and efficiency) and analysis (to pursue good classification) are two key segments involved in metabolomics workflows. Various chemometric approaches utilizing either pattern recognition or machine learning have been employed to separate different groups. However, insufficient feature extraction, inappropriate feature selection, overfitting, or underfitting lead to an insufficient capacity to discriminate plants that are often easily confused. Using two ginseng varieties, namely *Panax japonicus* (PJ) and *Panax japonicus* var. *major* (PJvm), containing the similar ginsenosides, we integrated pseudo-targeted metabolomics and deep neural network (DNN) modeling to achieve accurate species differentiation. A pseudo-targeted metabolomics approach was optimized through data acquisition mode, ion pairs generation, comparison between multiple reaction monitoring (MRM) and scheduled MRM (sMRM), and chromatographic elution gradient. In total, 1980 ion pairs were monitored within 23 min, allowing for the most comprehensive ginseng metabolome analysis. The established DNN model demonstrated excellent classification performance (in terms of accuracy, precision, recall, F1 score, area under the curve, and receiver operating characteristic (ROC)) using the entire metabolome data and feature-selection dataset, exhibiting superior advantages over random forest (RF), support vector machine (SVM), extreme gradient boosting (XGBoost), and multilayer perceptron (MLP). Moreover, DNNs were advantageous for automated feature learning, nonlinear modeling, adaptability, and generalization. This study confirmed practicality of the established strategy for efficient metabolomics data analysis and reliable classification performance even when using small-volume samples. This established approach holds promise for plant metabolomics and is not limited to ginseng.

Peer review under responsibility of Xi'an Jiaotong University.

[*] Corresponding author. State Key Laboratory of Chinese Medicine Modernization, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China.

[**] Corresponding author.

[***] Corresponding author. Shanghai Research Center for Modernization of Traditional Chinese Medicine, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.

E-mail addresses: wzyang0504@tjutcm.edu.cn (W. Yang), kefengl@mpu.edu.mo (K. Li), daguo@simm.ac.cn (D.-a. Guo).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jpha.2024.101116

1. Introduction

Metabolomics, as a part of systems biology, can quantitatively evaluate variations in small-molecule metabolites across different samples, accurately reflecting the function of the metabolic network and uncovering the biological state of the system [1]. Metabolomics can link differential metabolites to molecular phenotypic changes and explore the causes of these changes [2]. Currently, the application of metabolomics is flourishing in various research fields, such as studies related to toxicology [3], pharmacology [4], functional genomics [3,5], early diagnosis of disease [6], gut microbiome [7], marine organisms [8], plants [9], and the others [10,11]. Among the general workflows involved in metabolomics analysis, the systematic profiling of metabolic features is crucial for gaining high coverage of the metabolome. This can be achieved using nuclear magnetic resonance, liquid chromatography-mass spectrometry (LC-MS), gas chromatography-MS, or capillary electrophoresis-MS. Using LC-MS, metabolic features can be recorded in the untargeted (referred to as untargeted metabolomics), targeted (targeted metabolomics), or pseudo-targeted (pseudo-targeted metabolomics) modes. Comparatively, the pseudo-targeted metabolomics approaches merge a wide linearity range typical of the targeted mode by multiple reaction monitoring (MRM) to accurately reflect the real content differences of the metabolites among different groups. Certain merits of the untargeted mode enable the simultaneous profiling of isomeric metabolites with masses of interest [12]. In most cases, MS is utilized because of its high sensitivity, broad applicability, and high mass-to-charge (*m/z*) resolution, simultaneously offering rich structural information for metabolites identification. Regardless of whether the MS^2 acquisition mode is data-independent or data-dependent, the full-scan MS^1 data were used for the multivariate statistical analysis to visualize holistic differences by untargeted metabolomics. Additionally, diverse multivariate statistical analysis tools, such as partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM), random forest (RF), and variational autoencoder, are used to discover significantly differential metabolites among different groups [13,14]. Classification, prediction, and biomarker discovery methods can be extended to other models, including logistic regression models, least absolute shrinkage and selection operator (LASSO), correlation-constrained partial least squares (CCPLS), analysis of variance-simultaneous component analysis plus (ASCA+), augmented principal component analysis plus (APCA+) (variance analysis extends to multivariate class [15]), multivariate curve resolution, neural networks, and Gaussian mixture modeling [16].

Most classic machine-learning methods, such as RF [17], SVM [18], and LASSO [19], require handcrafted features for classification or regression. However, constructing effective models for classification using omics datasets is difficult. Methods based on gradient boosting, such as extreme gradient boosting (XGBoost) [20], and light gradient boosting machine (LightGBM) [21], have been proposed. Compared to classical machine-learning methods, these approaches perform well in terms of accuracy and effectiveness; however, they still struggle with feature selection and robustness. Recently, deep-learning methods have become dominant approaches in the domains of computer vision [22], natural language processing [23], and data mining [24]. Deep neural networks (DNNs) have also been utilized in herbal metabolome analyses [25]. Deep-learning-based approaches can automatically identify the relationships between various features and demonstrate significant advantages over traditional machine-learning approaches when dealing with high-dimensional omics data. Additionally, deep-learning methods do not require feature selection, indicating that they can be developed using a fully integrated process. Moreover, deep-learning approaches can efficiently utilize advanced computer hardware, such as graphics processing units (GPUs), significantly benefiting model training and deployment.

The insights gained from plant metabolomics research can profoundly affect the development of natural products, boost agricultural productivity, and improve food quality [26]. Plants are important sources with a long medical history of use in the prevention and management of illnesses [27]. Authentication of the plant origin is particularly important for ensuring the efficacy and safety of herbal medicines. Holistic metabolic profiling is crucial for establishing chemical markers to easily differentiate confused varieties of herbal medicines. However, the exact identification of plant species requires multiple layers of evidence [28]. Thus, metabolomics has been increasingly developed and employed to distinguish similar species by providing holistic metabolome characterization of complex samples, offering a comprehensive pathway for characterizing and comparing plant metabolites, and identifying potential chemical markers crucial for species differentiation [29,30]. Plants from the *Panax* L. genus (Araliaceae) are experiencing increased global recognition for their remarkable tonifying properties. They are widely used as vital ingredients in a range of clinical applications, including healthcare products, functional foods, and cosmetic formulations [27]. Ginseng occupies a top-selling position in the global natural product market. The market value of the global ginseng industry has grown significantly. Chinese patent medicines (CPMs) containing *Panax notoginseng* and *Panax ginseng* (PG) are commonly used. However, according to the basic theory of traditional Chinese medicine, distinct varieties of *Panax* herbal medicines exhibit variations in their properties, meridian tropism, and therapeutic effects. Therefore, they should not be substitutes for each other in clinical use. For example, *Panax japonicus* (PJ) tonifies the liver and spleen meridians, disperses stasis to stop bleeding, and alleviates swelling and pain, while *Panax japonicus* var. *major* (PJvm) tonifies the liver and lung meridians, nourishes the lungs, removes blood stasis, and relieves pain. Therefore, accurate identification of the origin of *Panax* species is vital to guarantee the efficacy and clinical reliability of herbal medicines. Multiple secondary metabolites have been extracted from a variety of *Panax* plants, including saponins (well known as ginsenosides), polysaccharides, organic acids/esters, flavonoids, steroids, and phenols [31]. Additionally, ginsenosides and polysaccharides are the major bioactive ingredients among multiple *Panax* species and different parts of the same ginseng variety (e.g., root, leaf, and flower), rendering it difficult to precisely identify the ginseng varieties, especially from compound preparations [29]. To date, ginsenosides have served as exclusive chemical markers for the quality control (QC) of various ginseng varieties. Particularly in extracts and preparations (such as the formulation granules and CPMs) with destroyed appearance features and genetic information, monitoring based solely on a few ginsenoside markers (e.g., notoginsenoside R1 (noto-R1) and ginsenoside Rg1 (Rg1), ginsenoside Re (Re), and ginsenoside Rb1 (Rb1) fails to accurately distinguish among the different ginseng varieties [32]. Currently, numerous reports are available regarding the identification and differentiation of various *Panax* herbal medicines using untargeted [29,33], targeted [34], and pseudo-targeted metabolomics approaches [29].

In this study, we integrated pseudo-targeted metabolomics and DNN modeling to differentiate between easily confused medicinal plants, using PJ and PJvm as examples. The rhizomes of PJ and PJvm share rich oleanolic acid (OA)-type ginsenosides with similar compositions and contents, rendering their discrimination as extracts or formulations challenging [35]. The overall technical roadmap of this integrated strategy is illustrated in Fig. 1. First, a series of metabolic feature acquisition modes and methods

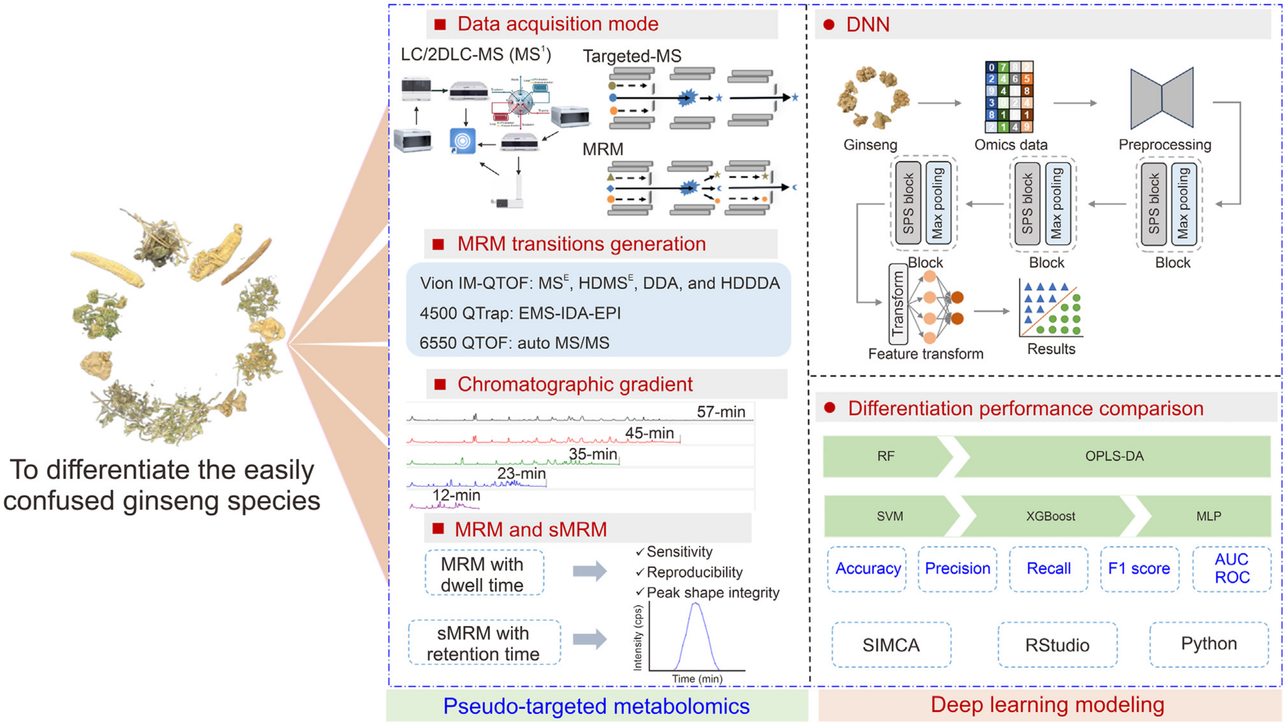

Fig. 1. Technology roadmaps for the strategy by integrating deep neural network (DNN) modeling and liquid chromatography-mass spectrometry (LC-MS)-based pseudo-targeted metabolomics to discriminate easily confused herbal medicines. 2D-LC: two-dimensional LC; MRM: multiple reaction monitoring; 1D: 1-dimensional; IM-QTOF: ion mobility quadrupole time-of-flight; MS^E: full information tandem MS; $HDMS^E$: high-definition MS^E; DDA: data-dependent acquisition; HDDDA: high-definition DDA; QTrap: triple quadrupole-linear ion trap; EMS-IDA-EPI: enhanced MS scan-information dependent acquisition-enhanced product ion scan; sMRM: scheduled MRM; SPS: split-combine structure; RF: random forest; OPLS-DA: orthogonal partial least squares-discriminant analysis; SVM: support vector machine; XGBoost: extreme gradient boosting; MLP: multilayer perceptron; AUC: area under the curve; ROC: receiver operating characteristic.

available on three LC-MS platforms were developed and compared to achieve high coverage of the ginseng metabolome: two-dimensional LC-MS (2DLC-MS) and targeted-MS on the Agilent 6550 quadrupole time-of-flight (QTOF) mass spectrometer (Agilent Technologies, Santa Clara, CA, USA); MRM on the AB SCIEX 4500 QTrap (triple quadrupole-linear ion trap) mass spectrometer (Foster City, CA, USA); and full information tandem MS (MS^E) and data-dependent acquisition (DDA) on the Waters Vion ion mobility (IM)-QTOF mass spectrometer (Milford, MA, USA). Moreover, the chromatographic elution gradient was optimized to achieve high analytical efficiency without compromising performance in species differentiation. Second, a high-coverage and efficient ultra-high performance liquid chromatography/scheduled MRM (UHPLC/sMRM) approach (targeting 1980 ion pairs) was established to cover the most comprehensive ginseng metabolome. Third, DNN model was established and its performance in differentiating between PJ and PJvm was assessed using both the entire metabolome data and feature-selection dataset. Additionally, its potential advantages over the commonly used machine- and deep-learning models were demonstrated.

2. Experimental

2.1. Chemicals and reagents

Sixty-six reference standards for ginsenosides (chemical structures shown in Fig. S1 and detailed information in Table S1) and the internal standard (IS) compound, astragaloside IV, were purchased from Chengdu Desite Biotechnology Co., Ltd. (Chengdu, China) and Shanghai Standard Biotech Co., Ltd. (Shanghai, China) with a purity >98%, as determined using high performance liquid chromatography (HPLC) coupled with an ultraviolet (UV) detector. LC-MS grade acetonitrile, methanol, and formic acid were supplied by Fisher Scientific (Fair Lawn, NJ, USA). Ultrapure water was prepared in house using a Milli-Q Integral 5 water purification system (Millipore, Bedford, MA, USA). Information on the ginseng samples (110 batches belonging to four ginseng varieties: PG, red ginseng (RG), PJ, and PJvm) are listed in Table S2. Authentication of the ginseng samples was performed by observing their appearance and comparing their LC-MS fingerprints with those reported in the literature. All specimens were deposited at the State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine (Tianjin, China).

2.2. Preparation of ginseng sample and reference standard solutions

An ultrasound-assisted extraction method was used to prepare the ginseng samples. The accurately weighed powder of each sample (500 mg) was extracted with 5 mL of 70% (V/V) aqueous methanol in a water bath at 40 °C for 1 h (power, 400 W; frequency, 40 kHz). The resultant supernatant was transferred into a 10-mL volume flask after centrifugation at 3219 *g* (4,000 rpm) for 10 min. The same extraction process was repeated by adding 3 mL of 70% (V/V) methanol to the drug residue. The supernatants from the two extractions were pooled and diluted to a constant volume (10 mL). The extraction liquid was diluted by five folds and then centrifuged at 11,481 *g* (14,000 rpm) for 10 min. The obtained supernatant was used as the test solution (concentration: 10 mg/mL). A QC sample (QC_1) was prepared by mixing test solutions of 30 batches of PG and RG at 10 mg/mL. A QC_2 sample was derived from the test solutions of 50 batches of PG and RG at 10 mg/mL. The QC_3 sample was prepared using the roots/leaves/flowers of PG, *Panax* *quinquefolius*, *Panax notoginseng*, and the rhizomes of PJ and PJvm (two batches for each ginseng variety), with the IS concentration being constant at 50 μg/mL. Additionally, the stock solutions for the reference standards were prepared by dissolving an appropriate amount of each compound in 70% (*V*/*V*) aqueous methanol, with each compound at the concentration of 20 μg/mL in the mixed standards solution.

2.3. UHPLC/IM-QTOF-MS conditions for creating ion pairs in pseudo-targeted metabolomic profiling

Several factors affecting the ultimate performance have been optimized to establish a pseudo-targeted metabolomic profiling method for the ginseng metabolome. It utilized four different MS data acquisition modes on three LC-MS platforms, and detailed information is provided in the Supplementary data and Tables S3 and S4. The conditions used for pseudo-targeted metabolomic profiling were described below.

Efficient chromatographic separation was achieved on ACQUITY UPLC I-Class/Vion IM-QTOF system (Waters) configured with a CSH C_{18} column (2.1 mm × 100 mm, 1.7 μm) kept at 30 °C. A binary mobile phase, containing 0.1% (*V*/*V*) formic acid each in water (A) and acetonitrile (B), ran according to the following gradient program: 0–1 min, 15%–20% (B); 1–6 min, 20%–30% (B); 6–11 min, 30%–31% (B); 11–13 min, 31%–35% (B); 13–15 min, 35%–40% (B); 15–17 min, 40%–95% (B); and 17–19 min, 95% (B). A flow rate of 0.3 mL/min was set, and the injection volume was 3 μL.

High-accuracy MS data were acquired using Vion™ IM-QTOF mass spectrometer coupled to UPLC I-Class system via Zspray™ electrospray ionization (ESI) source (Waters). Data acquisition was conducted using high-definition MS^E ($HDMS^E$) in negative mode. The ESI source parameters were set as follows: capillary voltage, −2.5 kV; cone voltage, −38 V; source temperature, 123 °C; desolvation temperature, 458 °C; desolvation gas flow rate (N_2), 900 L/h; and cone gas flow rate (N_2), 50 L/h. The mass analyzer scanned over a mass range of 250–1500 Da in full scan, with a scan time of 0.3 s. The low collision energy (CE) was 6 eV, and the high-energy ramp was 10–80 eV. For traveling wave IM separation, the parameters were set to the default values [36]. UNIFI™ 1.9.3.0 software (Waters) was used to acquire and process the data.

2.4. UHPLC/QTrap-MS condition

Pseudo-targeted metabolomic profiling of ginseng samples was performed using ACQUITY UPLC I-Class system (Waters) coupled with AB SCIEX QTrap 4500 mass spectrometer via an ESI source. The chromatographic conditions were the same as those described in Section 2.3 for UHPLC/IM-QTOF-MS, and the MS data were recorded in negative mode following the ion source parameters: ion spray voltage, −4500 V; source temperature, 550 °C; curtain gas, 35 psi; gas 1 (GS1) and GS2, 55 psi; declustering potential, 40 eV; and CE, 54 eV. The acquired data were processed using MultiQuant 3.0.3 software (AB SCIEX, Framingham, MA, USA).

2.5. Establishment and transformation of MRM transitions

In establishing a pseudo-targeted metabolomics approach, a key aspect is the selection of MRM transitions to achieve high coverage. Metabolome information was first recorded and analyzed on the Vion IM-QTOF platform, and MRM transitions were generated and further applied to the QTrap 4500 platform to acquire multi-batch data of ginseng samples. The code and script from GitHub (https://github.com/zhengfj1994/MRM-Ion_Pair_Finder) were used to obtain the MRM ion pairs and correct the retention times recorded between these two LC-MS platforms. The raw data recorded on Vion IM-QTOF were converted into a .csv file using UNIFI software to obtain the MS^1 information, and then MS Convert was used to transform it into.mgf file to obtain the MS^2 information. The “MRM_Ion_Pair_Finder” R statistical scripting language (version 3.6.1) was invoked to match between the obtained MS^1 and MS^2 information, generating a list of defined MRM transitions. The retention times of the transitions were corrected based on IS peaks.

2.6. Establishment of a DNN classification model for differentiating between PJ and PJvm

Data processing technologies can significantly affect the model performance, as demonstrated in various studies [37,38]. We incorporated multiple data processing methods to develop an accurate and robust model for metabolomics datasets. First, we used the chained equations algorithm to fill in the missing values. Next, we applied the Box-Cox algorithm to harmonize the data and reduce variations in the distribution between different institutions. Subsequently, we standardized the data using the min-max method. Finally, we addressed the data imbalance using the adaptive synthetic sampling method while maintaining a specific balancing ratio.

Model performance was typically evaluated using standard metrics, such as accuracy, precision, recall, and F1 score. The definitions of these metrics are provided in Eqs. 1–4:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} \ + \text{FN}} \qquad \text{Eq. 1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TN} + \text{FP}} \qquad \text{Eq. 2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Eq. 3}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \text{Eq. 4}$$

TP, TN, FP, and FN, represent true positives, true negatives, false positives, and false negatives, respectively.

We proposed a classification model based on DNN by learning from complex ginseng-omics data. This model comprises two procedures: a feature extraction and feature mapping modules, as illustrated in Fig. 2. Particularly, raw ginseng-omics data were normalized following a log operation. The computation can be formulated as Eq. 5:

$$x_n = \frac{x_l - \mu}{\sigma} \qquad \text{Eq. 5}$$

where x_n and x_l denote the results of the normalization and log operations, respectively. The mean and variance values are represented by μ and σ.

These normalized data were then directly fed into the proposed model without feature selection or additional processes. To alleviate the overfitting caused by high-dimensional ginseng-omics data, we introduced a feature split-combine structure (SPS) into the proposed 1D residual block (Fig. 2). The proposed SPS was formed using two parallel 1-dimensional convolution (Conv1D) layers, which forced the model to learn more distinct representations from the input sequence. SPS splits the input feature channel-wise, which has little impact on the model parameters and computational costs. We applied Mish as the activation function [39]. Mish is a smooth and self-regularized function that demonstrates better performance than rectified linear unit (ReLU) [40], Swish [41], and scaled exponential linear unit (SELU) [42]. It is defined as Eq. 6:

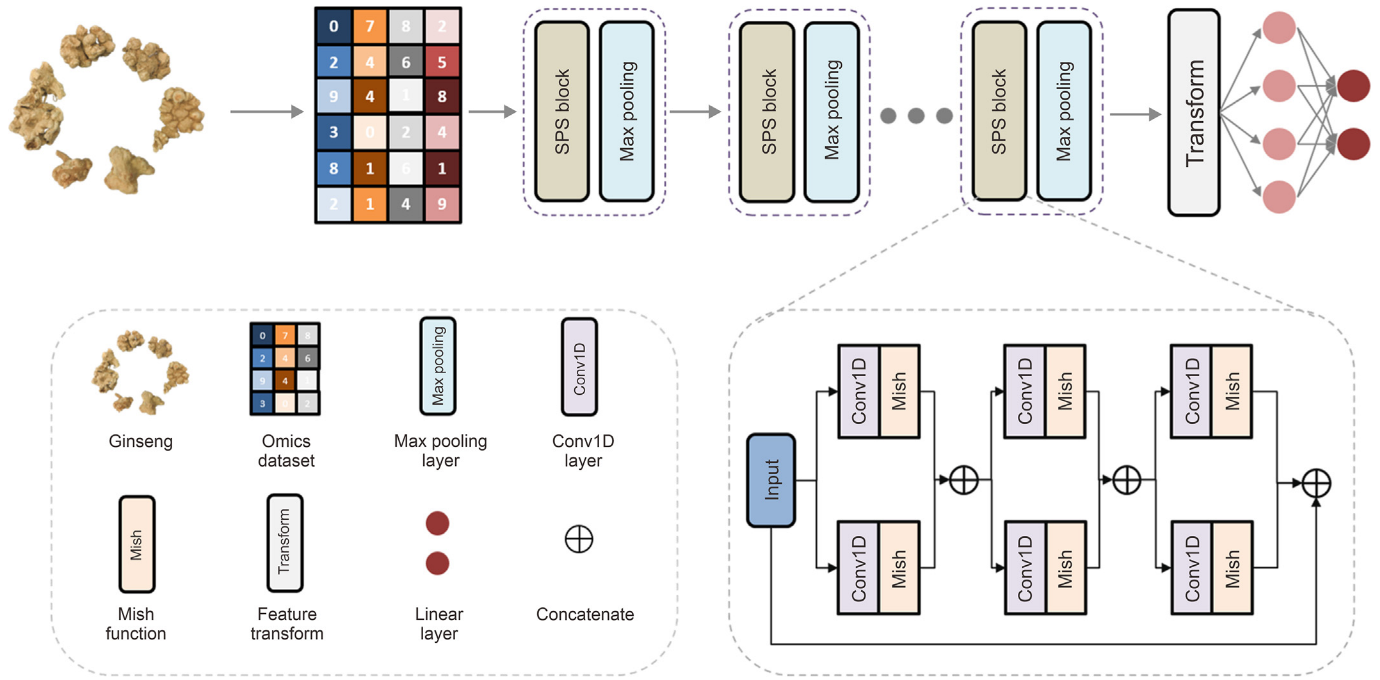

Fig. 2. Architecture of the classification model based on deep neural network (DNN). SPS: split-combine structure; Conv1D: 1-dimensional convolution.

$$f(x) = x\tanh(\ln(1+e^{x})) \qquad \text{Eq. 6}$$

where the x denotes the input feature maps.

For the feature mapping module, we employed several fully connected layers in a cascading manner that could add the output of each head to the subsequent head to progressively refine the feature representations. The predicted results are formulated as Eq. 7:

$$R_d = \Phi(\Phi(R_{d-1}) + \Delta R_d) \qquad \text{Eq. 7}$$

where, the R_d is the prediction result, and ΔR_d denotes the predicted offsets by the dth layers.

Because there were some differences between the numbers of categories in the training data, we used focal loss to optimize this challenge. Focal loss is defined as Eq. 8 [43]:

$$\mathscr{L} = -a_t(1-p_t)^{\gamma}\log(p_t) \qquad \text{Eq. 8}$$

where, a_t is the balance parameter and p_t can be calculated as Eq. 9:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{other} \end{cases} \qquad \text{Eq. 9}$$

where, $p \in [0,1]$ represents the probability of the input sample by model.

For the training stage, we applied the Adam optimizer and cosine learning rate scheduler. The total number of epochs were set to 100 with a batch size of 64. The initial learning rate was set to 1×10^{-4} with the momentum of 0.9 and weight decay of 3×10^{-4}. The model was built using PyTorch and trained from scratch on one Nvidia A100 GPU. The cross-entropy loss was employed to provide supervision information for model training.

2.7. Discovery of differential ginsenosides between PJ and PJvm using a pattern recognition pseudo-targeted metabolomics approach

Ginsenosides from PJ and PJvm were compared using the following workflows: 1) the multi-batch ginseng metabolomics data (containing 1980 metabolic features) were processed using MultiQuant software (AB SCIEX); 2) 659 robust metabolic features were screened by “80% rule” [44]; 3) the resultant data matrix was imported into SIMCA 14.1 software (Sartorius, Umea, Sweden) for pattern recognition chemometric analysis, including principal component analysis (PCA) and orthogonal partial least squares-discriminant analysis (OPLS-DA); and 4) those variables with variable importance in projection (VIP) > 1.0 were considered as potentially differential ginsenosides.

3. Results and discussion

3.1. Development of a high-coverage and efficient pseudo-targeted metabolomic profiling strategy enabling differentiation of ginseng varieties

Various ginseng varieties contain similar metabolomes (the nature and composition of different subcategories of metabolites, such as saponins [29], polysaccharides [33], and volatile oils [45]), rendering their differentiation difficult when the appearance features disappear. To achieve elaborate metabolome discrimination relying on LC-MS, high coverage on the metabolome is desirable to uncover more potential metabolite markers. However, short analysis time and uncompromised differentiation performance are beneficial for high-throughput analysis. To develop a potent analytical strategy enabling the differentiation of various ginseng varieties, the data acquisition modes and methods (e.g., comparison among full-scan MS, selective ion monitoring (SIM), MRM, and ion pairs generation in MRM) and chromatographic separation time were sequentially optimized.

3.1.1. Comparison of metabolic features acquisition modes

Considering that the untargeted full scan and targeted SIM and MRM can be utilized in the acquisition of metabolic features, four different methods, involving full scan of UHPLC/QTOF-MS (Method 1) and 2DLC-QTOF-MS (Method 2), targeted-MS of UHPLC/QTOF-MS (Method 3), and MRM of UHPLC/QTrap-MS (Method 4), were compared to evaluate their performance in discriminating the similar ginseng varieties (using the QC_1 sample containing PG and RG). The former two methods are based on full-scan MS but differ in chromatographic separation (detailed information is provided in the Supplementary data), whereas Method 4 records the MS/MS data. Method 3 is essentially the SIM mode that utilizes the targeted-MS function.

The different performances of these four methods could be embodied in four aspects: the number of recorded metabolic features, data repeatability, detected ion response range, and classification effectiveness. First, when viewed by their ability to detect metabolic features, Method 2 using 2DLC-MS separated and detected the most metabolites (690), followed by Method 1 using LC-MS (367), Method 3 using targeted-MS (268), and Method 4 using MRM (154). This indicated that 2DLC-MS method was more effective in acquiring the entire metabolome information than those based on one-dimensional chromatography strategies. Second, data repeatability reflecting system stability is significant in large-batch metabolomics analysis, which is usually evaluated using the coefficient of variation (CV). By observing the clustering of the QC_1 data, PCA score plot can embody data quality. The QC_1 data gathered by Method 4 using MRM were the closest, whereas those obtained by Method 2 were severely separated (Fig. 3A). Fig. 3B shows the cumulative percentage of the compounds versus the CVs measured using all four approaches. Consistently, the MRM data showed less than 10% of compounds with CV > 0.7, whereas Method 1 and Method 2 data displayed higher compound percentages with CV > 0.7. Notably, Method 3 contained more than 50% of the compounds with a CV > 0.7. Impressively, the median CV for Method 4 was approximately 0.5, and about 50% of the compounds had a CV < 0.5. Additionally, the relative standard deviation (RSD) of 35% of the compounds in Method 4 was less than 5%, and the compounds with RSD >30% accounted for 11%, demonstrating the least variation among the four methods (Fig. 3C). Third, the detected ion response range is another key parameter for establishing a robust metabolomics approach aimed at discovering untargeted markers [29]. For the same analytes using different batches of ginseng samples (data of 13 reference compounds; Table S5), a wider ion response detection range could better reflect the real content variations among the different groups. Consequently, Method 3 and Method 4 exhibited superior coverage across orders of magnitude, compared with the other two methods. Compared to Method 3, Method 4 using MRM demonstrated a similar or even more consistent performance (Fig. 3D). Fourth, the classification effect on PG and RG was evaluated using a machine-learning algorithm called RF in terms of F1 score, area under the curve (AUC), and number of features [46] (Table S6). Using the RF model in Python, we iteratively adjusted the number of decision tree branches to optimize the F1 and AUC values. Both Method 1 using LC-MS (MS^1) and Method 4 using MRM achieved F1 values of 1, which were higher than those of targeted-MS and 2DLC-MS (MS^1) methods. Moreover, the number of differential metabolic features by Method 4 MRM (17) was smaller than that of the other

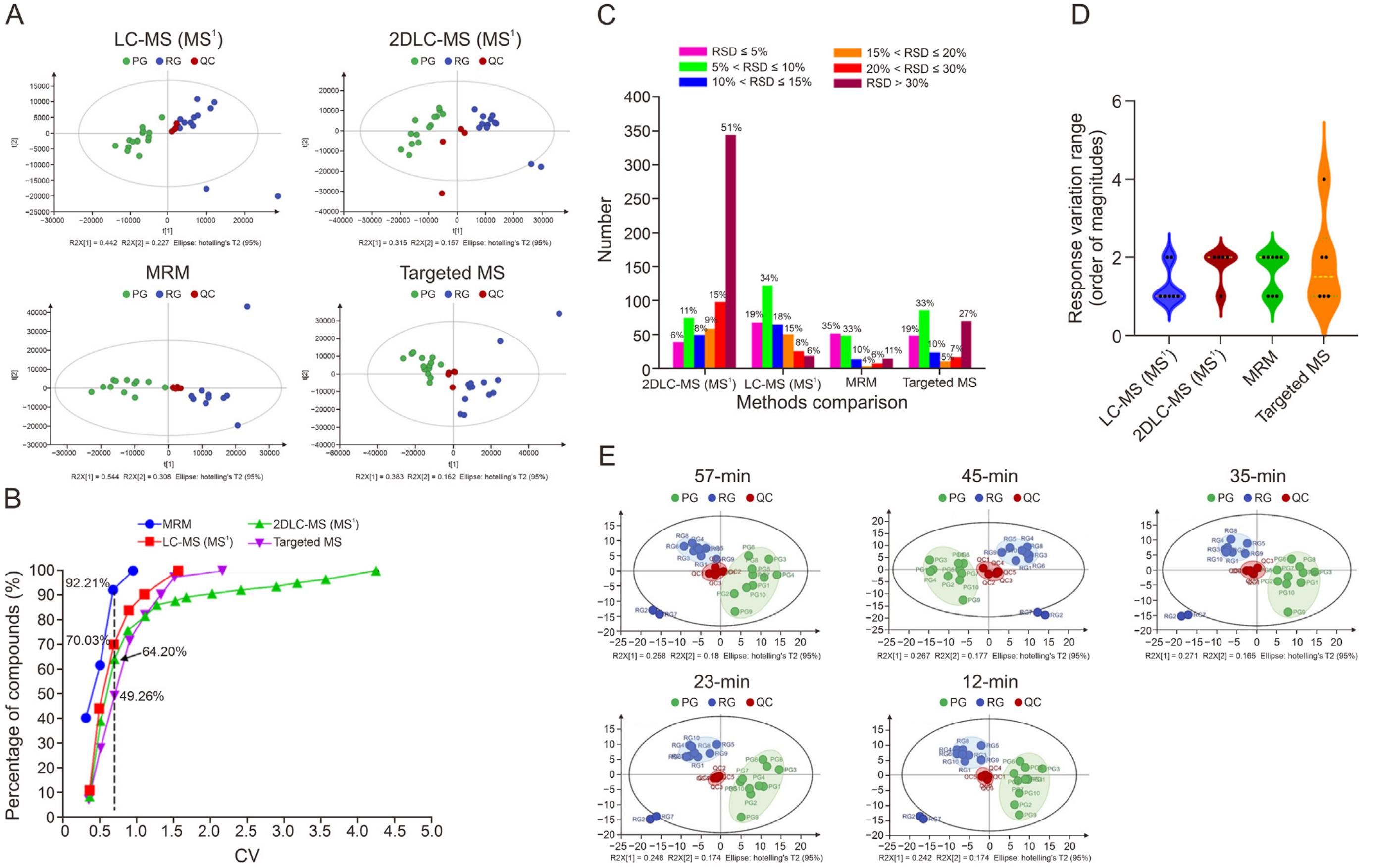

Fig. 3. Optimization and establishment of the pseudo-targeted metabolomics approach for differentiating the ginseng varieties. (A) Principal component analysis (PCA) score plots of the quality control (QC) metabolome data (QC_1) by four metabolic features acquisition modes. (B, C) QC_1 metabolome data stability obtained by four metabolic features acquisition modes: coefficient variation (CV) comparison of compounds (B) and relative standard deviation (RSD) comparison of compounds (C). (D) The response variation range of the same metabolome-QC_1 determined by four metabolic features acquisition modes. (E) PCA score plots of the QC_2 metabolome data by utilizing five chromatographic gradients. LC-MS: liquid chromatography-mass spectrometry; PG: *Panax ginseng*; RG: red ginseng; 2DLC: two-dimensional LC; MRM: multiple reaction monitoring; CV: coefficient of variation.

three methods, indicating that the MRM method could select the most relevant or informative features to effectively classify data using the minimum number of metabolic features. Thus, we concluded that the MRM method could distinguish ginseng samples more accurately and efficiently. Therefore, the MRM-based pseudo-targeted metabolomic profiling method was selected for comprehensive metabolome comparison of the ginseng varieties.

3.1.2. Comparison of the ability to generate ion pairs in establishing a high-coverage pseudo-targeted metabolomics approach

Based on the selected MRM pseudo-targeted acquisition mode, we sought to expand the composition of ion pairs to cover the most comprehensive ginseng metabolome (using the QC_3 sample composed of 12 different ginseng varieties). For this purpose, six MS^2 data acquisition approaches, available on three LC-MS platforms, were examined and assessed: 1) MS^E and $HDMS^E$ (data-independent), DDA and high-definition DDA (HDDDA) (data-dependent) on Vion IM-QTOF-MS; 2) enhanced MS scan-information dependent acquisition-enhanced product ion scan (EMS-IDA-EPI) on QTrap 4500; and 3) DDA (auto-MS/MS) on 6550 QTOF-MS. An open-access method was utilized to establish the ion pairs and align the retention times determined between the three LC-MS and QTrap 4500 platforms (used to construct the MRM approach). In this section, the number of ion pairs was utilized as the sole criterion. These six methods were ranked in the following order: MS^E (5153), $HDMS^E$ (3954), DDA (412), HDDDA (323), auto-MS/MS (219), and EMS-IDA-EPI (60). This demonstrated a much higher coverage of data-independent acquisition strategies in acquiring the MS/MS information of the ginseng metabolome. Enabling IM separation ($HDMS^E$) can significantly enhance the resolution of components, thereby largely reducing the false-positive results recorded by MS^E in inducing reliable matching between the precursor and product ions, as previously reported [47]. The balanced detection sensitivity and data quality rendered $HDMS^E$ an excellent choice for generating rich ion pairs related to the ginseng metabolome. Therefore, we chose $HDMS^E$ mode to obtain the MRM transitions to discriminate among ginseng varieties.

3.1.3. Comparison of chromatographic separation time to enable high-throughput analysis

The chromatographic conditions in LC-MS can affect the acquisition of metabolic features, thus determining the performance in differentiating ginseng varieties. According to related literature, the gradient elution program of reversed-phase HPLC used to separate the ginseng metabolome typically ranged from 30 to 70 min [34,36]. However, studies investigating the relationship between the separation time and differentiation performance in metabolomics are rare. In this study, different degrees of compression were applied to the elution gradient (45-, 35-, 23-, and 12-min gradients) based on a 57-min chromatographic gradient that could effectively resolve the major compounds. Here, we examined their differentiated ability in detecting 50 ginsenoside reference compounds (common to the ginseng varieties, 20 μg/mL for each; Table S7) and generating associated ion pairs separately from a mixed standards solution and the QC_2 sample, caused by different chromatographic separation time.

From the perspective of mixed standard solutions, all ginsenoside compounds were detectable using 45-, 35-, and 23-min gradients. However, m-Rb3 (*m/z* 1163.5856) was not detected in the 57- and 12-min elution programs. The number of generated ion pairs was the highest for the 35-min gradient (46), followed by 23-min (44), 45-min (43), 57-min (42), and 12-min (31) gradients (Table S8). For the QC_2 sample, the average number of compounds detected through three injections were 48 for the 57- and 45-min gradients, followed by 47 for the 35-min, 46 for the 23-min, and 45 for the 12-min gradients. Owing to the complexity of ginsenosides in the QC_2 sample, the MS^2 information of the co-eluted compounds with a low response was masked, leading to a reduction in the generated ion pairs. The number of converted ion pairs were 35 for the 57- and 45-min gradients, 33 for the 35-min, 32 for the 23-min, and 26 for the 12-min gradients. The resulting metabolomic data were evaluated for comparison. PCA score plot indicated that QC_2 aggregation based on the 23-min elution was the closest (Fig. 3E). From the mixed standards sample (three parallel injections), the most stable performance was gained by the 45-, 35-, and 23-min gradients, with the ions having RSD $\leq$ 5% accounting for 98%, followed by the 57-min (94%) and 23-min (98%) (Tables S9−S13). From the QC_2 sample, the ions that exhibited the variation of RSD $\leq$ 5% were the most stable for the 57-min gradient (86%), followed by the 23-min (84%), 35-min (81%), 45-min (73%), and 12-min (59%) gradients (Tables S14−S18). Based on these results, it was concluded that the number of detected reference standards and corresponding ion pairs from the data of 23-min gradient were comparable to those obtained by longer chromatographic gradients, such as 57- and 45-min. However, the 23-min gradient significantly improved the analysis efficiency. Additionally, the stability of the compounds collected using the 23-min gradient was good, indicating that short-term chromatographic separation was more conducive to maintain the stability of the detected compounds. Following the established metabolomics data processing approach (peak picking), the quantities of extracted metabolic features were in the following order: 777 (45-min gradient) > 735 (35-min) > 708 (23-min) > 687 (57-min) > 531 (12-min). The 23-min gradient yielded the lowest proportion (23.43%) of features with RSD > 30%. RF classifier was used to compare the clustering differences resulting from the settings of the five different elution gradients (Table S19). The results indicated that the number of selected differential components based on 23-min data was fewer than that screened by the other four gradient approaches. This suggested that the 23-min method could distinguish between the two groups with fewer variables. Moreover, by selecting differential compounds with VIP >1.0, F1 values of the 23-, 45-, and 35- gradients were equal to or greater than those of the 57-min approach. This indicated a better performance of the 23-, 45-, and 35-min methods in differentiating the similar ginseng varieties. Therefore, the 23-min gradient was regarded as the most effective and was subsequently utilized to establish an MRM pseudo-targeted metabolomics approach for distinguishing the ginseng varieties.

3.1.4. Comparison between MRM and sMRM

MRM can be operated in regular and scheduled modes. sMRM has a higher capacity for ion pairs. In this mode, the spectrum for each ion pair is recorded in a predefined time window, rather than throughout the entire LC run. To encompass metabolome information covering more ginseng varieties, we analyzed the negative $HDMS^E$ data for the QC_3 sample, which generated a total of 526 ion pairs. For such a heavy detection task within 23 min, it was necessary to compare the performances of the regular MRM and sMRM.

To establish the sMRM approach on QTrap 4500 LC-MS platform, we performed a retention time correction with reference to the retention differences of 10 ginsenoside compounds to cover the pre-, mid-, and post-chromatographic gradients [40]: 20-*O*-glucosylginsenoside Rf (20-*O*-glu-Rf), Re, ginsenoside Rh1 (Rh1), ginsenoside Ra1 (Ra1), noto-S, ginsenosides Rs1, Rg4, Rk3, 20(*R*)-ginsenoside Rg3, and ginsenoside Mc. Notably, we used a ready-made function in R statistical scripting language for retention time calibration. Each MRM transition was detected within a retention time window, and the drift in retention was calibrated to prevent false negatives [48]. The final retention information for 526 ion pairs is provided in Table S20, and the retention time correction deviation met the sMRM detection time window of 60 s (retention time ± 30 s).

The principles of sMRM and MRM were elucidated and their respective representative chromatograms (using the QC_3 sample) are shown in Fig. 4. Typically, allocating a minimum 10-ms dwell time for each ion pair without compromising the reproducibility of the integrated peak in MRM is necessary [49]. The cycle time was equal to the total dwell time of all transitions plus all pause times (Fig. 4A). In the present study, up to 526 ion pairs were monitored in negative ESI mode, resulting in a calculated cycle time of 2.9 s. Therefore, it was unsatisfactory to use MRM, as approximately five data points were recorded (Fig. 4B). Notably, the data points can vary for different compounds according to their peak widths after chromatographic separation. Additionally, compared with MRM, the narrow detection window (60 s) of sMRM reduced the number of concurrent ion pairs, and the dwell time was automatically maximized without requiring a long cycle time (Fig. 4C). Approximately 12 representative chromatographic peaks were obtained using sMRM, which met the requirements for a reliable quantitative assay (Fig. 4B). Moreover, the peaks recorded by sMRM showed larger or comparable peak areas to those in regular MRM chromatograms. Meanwhile, owing to the sufficient dwell time for each ion pair, the detected noise level was lower than that in MRM. Accordingly, a pseudo-targeted ginsenoside profiling approach was developed on QTrap 4500 LC-MS platform using a 23-min gradient elution and highly specific sMRM mode.

We further assessed the quantitative performance of the established pseudo-targeted metabolomics method in terms of linearity and precision (intra- and inter-day). Notably, 1169 metabolites were obtained through the 80% rule treatment on 1980 transitions, covering 12 ginseng varieties across the entire 256-fold dilution series of QC_3. Of these, 300 metabolites showed an integral peak, which were subjected to statistical analysis to evaluate the linearity of the MRM transitions. The percentages of metabolites with $R^2 > 0.95$ and $R^2 > 0.8$ were 39% and 81%, respectively (Fig. 5A). If more than 80% of the metabolites had $R^2 > 0.8$, the results were considered to be acceptable [50]. Owing to the excessive number of ion pairs and incomplete integration of some minor or trace compounds, metabolites with complete peak integration were selected to assess the precision, which was evaluated by calculating the CV values of repeated injections on the same day (intra-day) and over three successive days (inter-day). As a result, 75.7% of the metabolites (accounting for 99.55% of the total peak area) showed RSD < 15% on the first day (Fig. 5B). For three consecutive days, 73.3% of the metabolites, accounting for 99.52% of the total peak area, exhibited variation with RSD < 15% (Fig. 5C). These results demonstrated the robustness of the established pseudo-targeted metabolomics approach for large-scale ginseng metabolome analyses.

3.2. Comparison of the differentiation performance between DNN and other models

In the practice of discriminating easily confused herbal medicines using multivariate statistical analysis, challenges often arise from the small sample size and tendency of the classification model to overfit. DNN can overcome these challenges through feature learning, fitting capabilities, and regularization. To demonstrate the potential merits of our established DNN model in discriminating between PJ and PJvm (two highly similar ginseng varieties), we evaluated the performance of DNN and other classic models

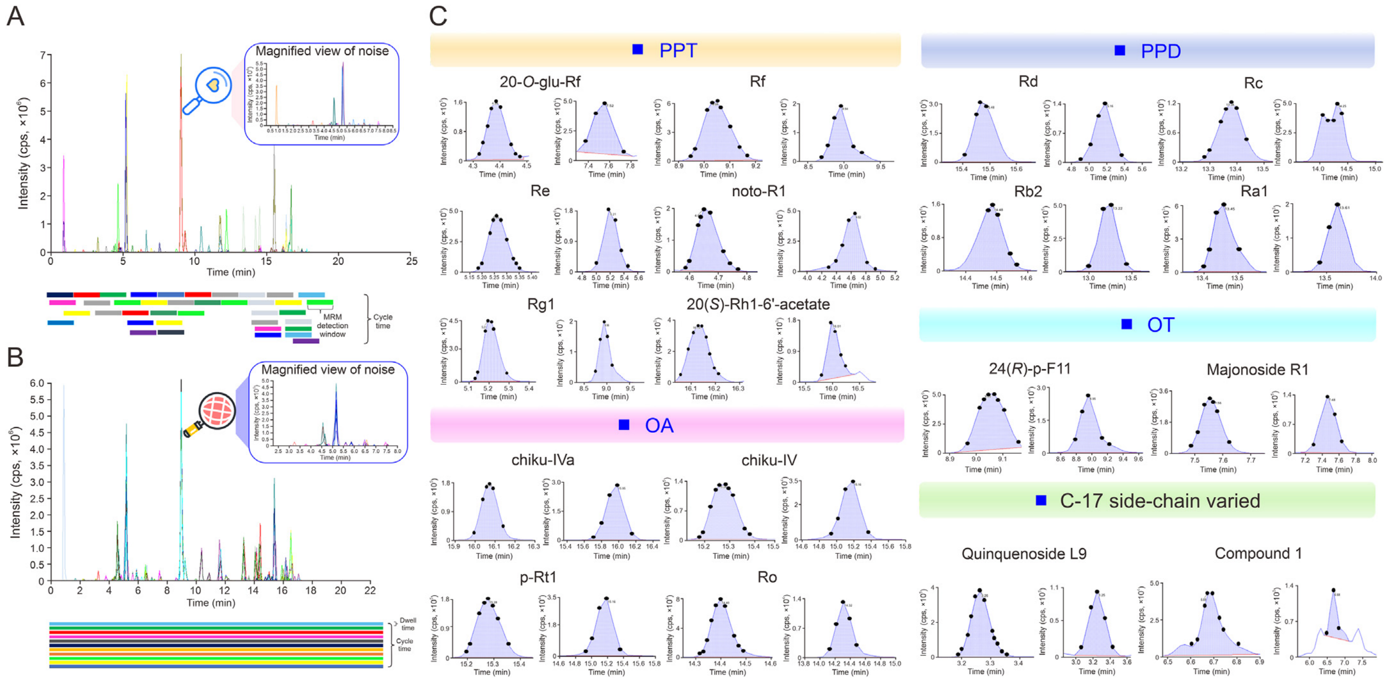

Fig. 4. Comparison of the algorithm principles and representative chromatograms acquired using multiple reaction monitoring (MRM) and scheduled MRM (sMRM). (A) Overlapped chromatograms of 526 transitions using MRM and the magnified chromatograms of the signal-to-noise ratio for the MRM method. (B) Overlapped chromatograms of all 526 transitions using sMRM and the magnified chromatograms of the signal-to-noise ratio for the sMRM method. (C) Peaks obtained using sMRM (shown in the left) and MRM (in the right) representative of the common five subclasses of ginsenosides: protopanaxatriol (PPT)-type, oleanolic acid (OA)-type, protopanaxadiol (PPD)-type, ocotillol (OT)-type, and C-17 side-chain varied. 20-*O*-glu-Rf: 20-*O*-glucosylginsenoside Rf; Re: ginsenoside Re; noto-R1: notoginsenoside R1; Rg1: ginsenoside Rg1; 20(*S*)–Rh1-6′-acetate: 20(*S*)-ginsenoside Rh1-6′-acetate; chiku-IVa: chikusetsusaponin IVa; p-Rt1: pseudoginsenoside Rt1; Ro: ginsenoside Ro; Rd: ginsenoside Rd; Rc: ginsenoside Rc; Rb2: ginsenoside Rb2; Ra1: ginsenoside Ra1; 24(*R*)-p-F11: 24(*R*)-pseudoginsenoside F11.

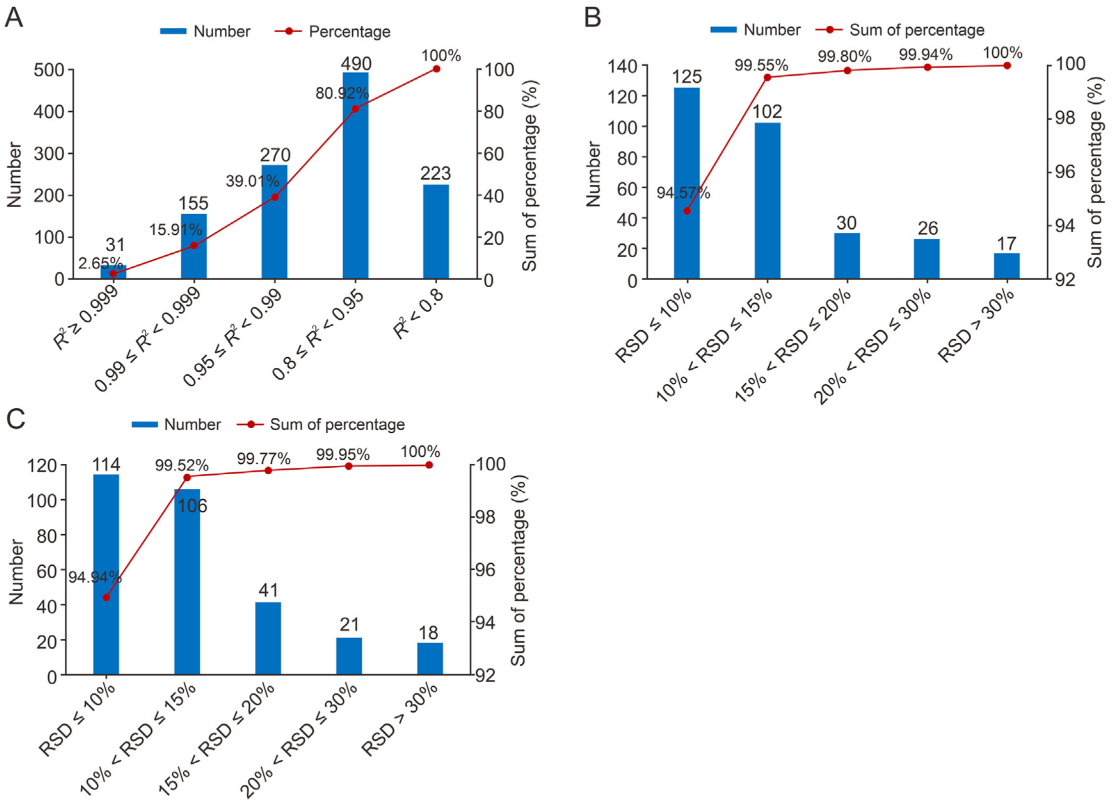

Fig. 5. Assessment on the established pseudo-targeted metabolomics approach from the quantitative assay aspects (number on the *y*-axis refers to the amount of multiple reaction monitoring (MRM) transitions giving the defined variation range among 300 transitions used for the statistical analysis). (A) Linearity scatter results. (B) Intra-day precision results. (C) Inter-day precision results. RSD: relative standard deviation.

(involving machine-learning models, such as RF, SVM, XGBoost, and common deep-learning multilayer perceptron (MLP)) separately, based on the entire metabolome dataset (original) and dataset after feature selection. To minimize the impact of random partitioning on the results, we utilized five-fold cross-validation to assess the prediction performance of the proposed model. In this section, multidimensional indicators, such as accuracy, precision, recall, F1 score, AUC, and receiver operating characteristic (ROC) were utilized for a comprehensive performance comparison. The results obtained from the original metabolome dataset and dataset after feature selection (those with VIP > 1.0 by OPLS-DA were selected) are presented in Table 1.

The DNN model evidently outperformed the existing methods with significant advantages for all examined indicators. Compared with XGBoost, our current classification system demonstrated a relative improvement of 35.6% in terms of F1 score. Notably, the DNN system achieved 100% classification accuracy without feature-selection or additional processing, which is extremely difficult to accomplish using the existing methods. Based on the feature-selection dataset, the performance of all methods was significantly improved, compared to the original metabolome dataset. For instance, XGBoost achieved relative improvements of 9.7% and 17.3% in terms of accuracy and F1 score, respectively. Despite this, existing systems (such as RF (the implementation details of RF classifier are shown in Fig. S2, SVM, XGBoost, and MLP) failed to create a flawless classification model for differentiating between PJ and PJvm. Our proposed method performed best on a feature-selection omics dataset.

The DNN method demonstrated superior performance compared to existing approaches for both the original and feature-selection datasets. This indicated that our method could learn robust representations from the complex omics datasets. Moreover, the DNN method can be easily trained and deployed end-to-end, making it a convenient solution for data analysis.

To further analyze its performance, the model size and computational cost were evaluated and compared with those of other deep-learning-based approaches such as MLP. A widely used metric to measure the computational cost is the number of floating-point operations (FLOPs), which defines the number of multiple adds. Deep-learning models are renowned for their complexity and the massive number of computations they perform [51], making the computational cost a significant consideration. In practical situations, the computational cost is closely related to the resources required. Model parameters play a crucial role in the functioning of neural networks, which form the backbone of many

Table 1
Comparison of the performance of established deep neural network (DNN) in differentiating between *Panax japonicus* (PJ) and *Panax japonicus* var. *major* (PJvm) with the other conventionally utilized methods.

Models	Accuracy (%)		Precision (%)		Recall		F1 score	
	Original[a]	FSD[b]	Original[a]	FSD[b]	Original[a]	FSD[b]	Original[a]	FSD[b]
RF	0.875	0.875	0.85	0.85	0.909	0.919	0.870	0.895
SVM	0.85	0.9	0.9	0.9	0.874	0.919	0.966	0.888
XGBoost	0.775	0.85	0.75	0.9	0.793	0.84	0.737	0.867
MLP	0.9	0.95	0.95	0.95	0.914	0.96	0.916	0.949
DNN	1	1	1	1	1	1	1	1

[a] The entire ginseng metabolome data.
[b] Feature-selection dataset (FSD) was provided by the partial least squares-discriminant analysis (PLS-DA) model with variable importance in projection (VIP) > 1.0.
RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting; MLP: multilayer perceptron.

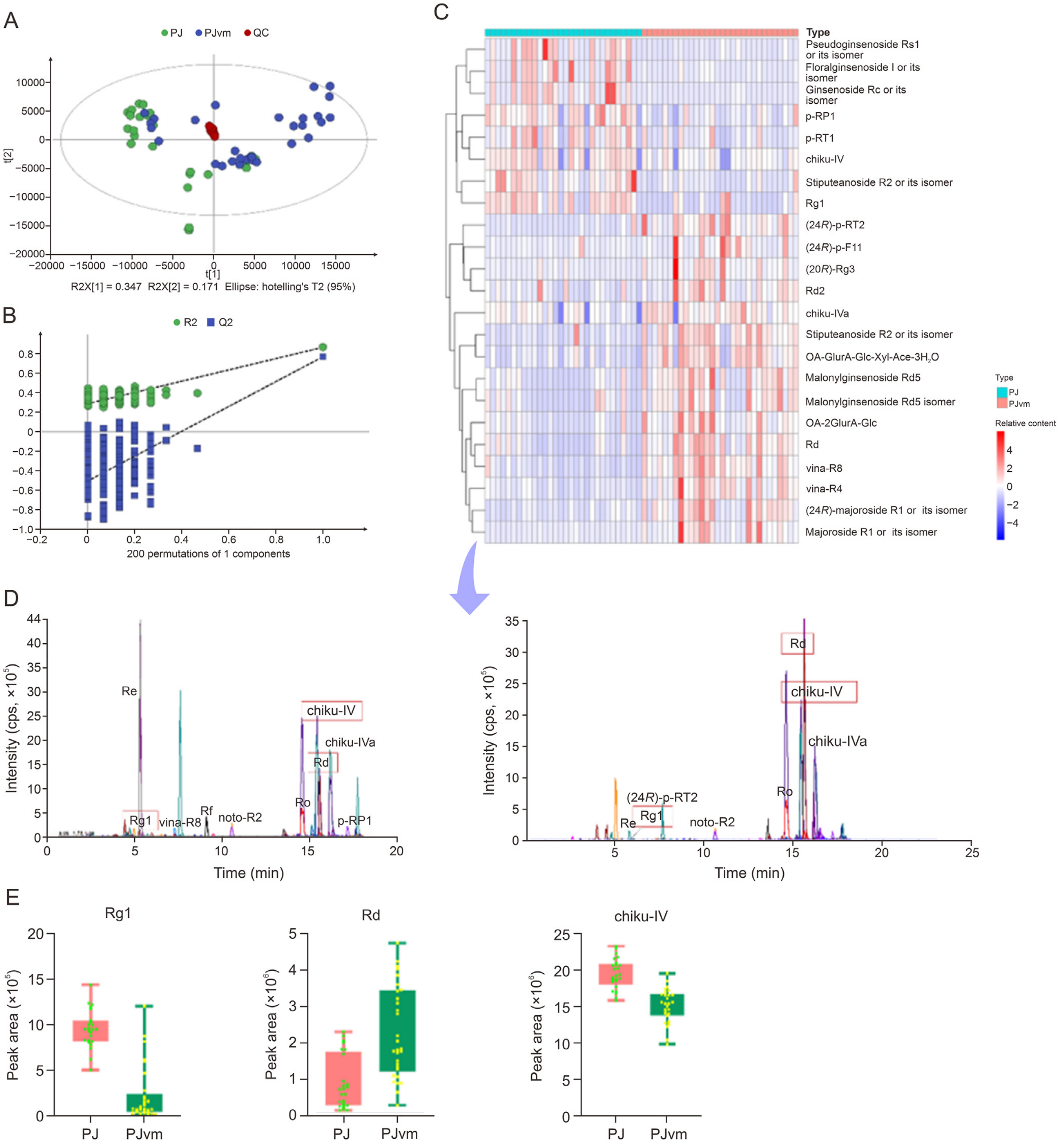

Fig. 6. Holistic comparison between two congeneric ginseng varieties (*Panax japonicus* (PJ) and *Panax japonicus* var. *major* (PJvm) based on the data of 659 metabolic features obtained by the developed pseudo-targeted metabolomics strategy. (A) Score plot of principal component analysis (PCA). (B) 200 times of permutation test plot. (C) Heatmap plotted by two ginseng varieties vs. 23 ginsenoside markers. (D) The multi-channel scheduled multiple reaction monitoring (sMRM) chromatograms to show the detection of representative ginsenoside markers. (E) Box charts illustrating the content difference for three marker compounds between two ginseng varieties. QC: quality control; p-RP1: pseudoginsenoside RP1; p-Rt1: pseudoginsenoside Rt1; chiku-IV: chikusetsusaponin IV; Rg1: ginsenoside Rg1; pRT2: pseudoginsenoside RT2; p-F11: pseudoginsenoside F11; Rd2: ginsenoside Rd2; OA-GlurA-Glc-Xyl-Ace-3H_2O: oleanolic acid-glucuronic acid-glucose-xylose-acetate-trihydrate; vina-R8: vinaginsenoside R8; Re: ginsenoside Re; Rf: ginsenoside Rf; noto-R2: notoginsenoside R2; Ro: ginsenoside Ro.

deep-learning models. During the training process, these parameters are learned, and the ability of the network to make accurate predictions is determined. Model parameters refer to the weights and biases associated with each neuron or node in the network. These parameters control the strength and direction of the connections between neurons, enabling the network to learn complex patterns and relationships in the data. Table S21 details all comparison results in terms of the model parameters and computational cost (FLOPs). Based on these findings, it is evident that the computational costs and parameter requirements of the employed deep-learning model are quite small. The MLP is a well-established deep-learning model that frequently yields favorable results for sequence data analysis tasks. However, it remains difficult to classify PJ and PJvm effectively. The proposed DNN model utilizes the benefits of Conv1D to effectively extract features. The results indicate that our proposed DNN model has only a slight disadvantage in terms of running speed and model calculation compared with the MLP method. Our method has only 6.7% of model parameters of MLP, making it compact and suitable for deployment on hardware devices, while significantly saving computing resources.

To demonstrate the effectiveness of the proposed deep-learning model, we compared it with a traditional chemometric pattern recognition method. As shown in Fig. 6A, the samples were not fully distinguishable in PCA score plot between PJ and PJvm (based on the data of 659 metabolic features). This finding indicated that the chemometric pattern recognition method had difficulty in distinguishing between these two similar ginseng varieties. A permutation plot was used to assess the integrity of OPLS-DA model (Fig. 6B), which demonstrated a satisfactory general interpretation rate ($R^2X = 0.636$ and $R^2Y = 0.874$) and predictive ability ($Q^2 = 0.787$). When the VIP cutoff value was set to 1.0, 32 ginsenosides (Table S22) were selected, which were identified by comparison with reference compounds or through integrated analysis of their negative-mode collision-induced dissociation-MS^2 data [27]. Among them, the content variations of 23 ginsenosides, including eight OA-type, four protopanaxadiol (PPD)-type, four ocotillol (OT)-type, three protopanaxatriol (PPT)-type, two malonylated, and two others, significantly contributed to the distinction between PJ and PJvm, as shown in the heatmap (Fig. 6C). Fig. 6D shows the MRM spectra of the characteristic ginsenoside markers in representative ginseng samples. Box charts displaying the differences in the content of important ginsenoside markers are shown in Fig. 6E. Importantly, the relative abundance (peak area ratio) of some ginsenoside markers may provide key identification points for PJ and PJvm. First, both PJ and PJvm were characterized by the richness of OA-type ginsenoside Ro (Ro) (*m/z* 955.4926), chikusetsusaponin (chiku)-IV (*m/z* 925.4806), and chiku-IVa (*m/z* 793.4397), whereas neutral PPD-type ginsenoside Rd (Rd) and PPT-type Rg1 were also common in these two ginseng varieties. These ginsenoside markers can be used to distinguish different ginseng varieties. Second, chiku-IV, Rg1, and Rd, showed high potential to discriminate between PJ and PJvm. PJ contained richer chiku-IV and Rg1 than PJvm, while Rd was more abundant in PJvm: 1) chiku-IV/Ro > 2 (27 out of 30 batches satisfied, 2.08−6.92) for PJ, while chiku-IV/Ro < 2 (27 out of 30 batches, 0.52−1.97) for PJvm; 2) Rg1/Ro > 0.1 (23 out of 30 batches, 0.1−0.24) for PJ, while Rg1/Ro < 0.1 (26 out of 30 batches, 0.09−0.001) for PJvm; and 3) Rd/Ro > 1 (28 out of 30 batches, 1.27−7.01) for PJvm, while Rd/Ro < 1 (17 out of 30 batches, 0.01−0.93) for PJ.

In summary, the advantages of the DNN model, compared with traditional chemometric models, lie in the performance benefits (achieving a performance comparable to the traditional methods using the feature-selection dataset), computational efficiency (suitable for modern high-performance computing hardware, such as GPU acceleration, processing multiple batches of PJ and PJvm data in only 3 s, whereas traditional pattern recognition requires complex data preprocessing and analysis), and generalization ability (capable of better handling and adapting to new data; models trained on PJ and PJvm datasets require little or no modification for the transformed application to the datasets of other ginseng varieties).

4. Conclusion

A powerful strategy was presented to differentiate easily confused herbal medicines by integrating pseudo-targeted metabolomics and DNN modeling. The established DNN model exhibited perfect classification performance in terms of accuracy, precision, recall, F1 score, AUC, and ROC. It showed renowned merits over the widely utilized RF, SVM, XGBoost, and MLP models. It also has advantages in terms of computational efficiency and generalization ability. Moreover, we successfully established a pseudo-targeted metabolomics data acquisition method, which could enable holistic metabolomic comparison among common ginseng varieties in an efficient manner. The combination of pseudo-targeted metabolomics data acquisition and DNN modeling renders it a potent vehicle for facilitating metabolomics comparison studies of easily confused medicinal herbs, such as ginseng, which is beneficial for QC in a wide variety of research fields.

CRediT authorship contribution statement

Meiting Jiang: Formal analysis, Investigation, Writing − original draft. **Yuyang Sha:** Conceptualization, Investigation, Software. **Yadan Zou:** Formal analysis, Writing − original draft. **Xiaoyan Xu:** Data curation, Investigation. **Mengxiang Ding:** Validation, Visualization. **Xu Lian:** Investigation, Software. **Hongda Wang:** Funding acquisition, Investigation, Validation. **Qilong Wang:** Funding acquisition, Validation. **Kefeng Li:** Conceptualization, Software, Validation. **De-an Guo:** Methodology, Supervision, Writing − review & editing. **Wenzhi Yang:** Conceptualization, Funding acquisition, Methodology, Writing − review & editing.

Declaration of competing interest

The authors declare that there are no conflicts of interest. As a young editorial board member, Wenzhi Yang recused himself from all review processes related to this article to ensure the fairness and objectivity of the review.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No.: 2022YFC3501805), the National Natural Science Foundation of China (Grant No.: 82374030), the Science and Technology Program of Tianjin in China (Grant No.: 23ZYJDSS00030), the Tianjin Outstanding Youth Fund, China (Grant No.: 23JCJQJC00030), and the China Postdoctoral Science Foundation-Tianjin Joint Support Program (Grant No.: 2023T030TJ).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpha.2024.101116.

References

[1] H. Pang, Z. Hu, Metabolomics in drug research and development: The recent advances in technologies and applications, Acta Pharm. Sin. B 13 (2023) 3238−3251.

[2] J. Zhou, D. Hou, W. Zou, et al., Comparison of widely targeted metabolomics and untargeted metabolomics of wild *Ophiocordyceps sinensis*, Molecules 27 (2022), 3645.

[3] Wurihan, Aodungerle, Bilige, et al., Metabonomics study of liver and kidney subacute toxicity induced by garidi-5 in rats, Chin. Herb. Med. 14 (2022) 422–431.

[4] Y. Zhu, F. Wang, J. Han, et al., Untargeted and targeted mass spectrometry reveal the effects of theanine on the central and peripheral metabolomics of chronic unpredictable mild stress-induced depression in juvenile rats, J. Pharm. Anal. 13 (2023) 73–87.

[5] Y. Wei, J. Zhang, K. Qi, et al., Combined analysis of transcriptomics and metabolomics revealed complex metabolic genes for diterpenoids biosynthesis in different organs of *Anoectochilus roxburghii*, Chin. Herb. Med. 15 (2023) 298–309.

[6] X. Meng, Z. He, L. Guo, et al., *OSCA*-finder: Redefining the assay of kidney disease diagnostic through metabolomics and deep learning, Talanta 264 (2023), 124745.

[7] D. Ye, J. Huang, J. Wu, et al., Integrative metagenomic and metabolomic analyses reveal gut microbiota-derived multiple hits connected to development of gestational diabetes mellitus in humans, Gut Microbes 15 (2023), 2154552.

[8] L.M. Bayona, N.J. de Voogd, Y.H. Choi, Metabolomics on the study of marine organisms, Metabolomics 18 (2022), 17.

[9] Y. Zou, M. Ding, H. Wang, et al., Integration of ion-mobility high-resolution liquid chromatography/mass spectrometry-based untargeted metabolomics and desorption electrospray ionization-mass spectrometry imaging to unveil the ginsenosides variation induced by steaming for *Panax ginseng*, *P. quinquefolius* and *P. notoginseng*, Arab. J. Chem. 17 (2024), 105781.

[10] W. Jin, J. Bi, S. Xu, et al., Metabolic regulation mechanism of *Aconiti Radix Cocta* extract in rats based on ^{1}H-NMR metabonomics, Chin. Herb. Med. 14 (2022) 602–611.

[11] B. Jin, X. Pang, Q. Zang, et al., Spatiotemporally resolved metabolomics and isotope tracing reveal CNS drug targets, Acta Pharm. Sin. B 13 (2023) 1699–1710.

[12] Y. Li, Q. Ruan, Y. Li, et al., A novel approach to transforming a non-targeted metabolic profiling method to a pseudo-targeted method using the retention time locking gas chromatography/mass spectrometry-selected ions monitoring, J. Chromatogr. A 1255 (2012) 228–236.

[13] D. Amaratunga, J. Cabrera, Y.S. Lee, Enriched random forests, Bioinformatics 24 (2008) 2010–2014.

[14] L.E. Broughton-Neiswanger, S.M. Rivera-Velez, M.A. Suarez, et al., Urinary chemical fingerprint left behind by repeated NSAID administration: Discovery of putative biomarkers using artificial intelligence, PLoS One 15 (2020), e0228989.

[15] M. Thiel, B. Féraud, B. Govaerts, ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs, J. Chemom. 31 (2017), e2895.

[16] M. Efimenko, A. Ignatev, K. Koshechkin, Review of medical image recognition technologies to detect melanomas using neural networks, BMC Bioinformatics 21 (2020), 270.

[17] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[18] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.

[19] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.

[20] W. Li, Y. Wen, K. Wang, et al., Developing a machine learning model for accurate nucleoside hydrogels prediction based on descriptors, Nat. Commun. 15 (2024), 2603.

[21] G. Ke, Q. Meng, T. Finley, et al., Proceedings of the Neural Information Processing Systems (NeurIPS) Conference, December 4–9, 2017, Long Beach, California, USA, 2017, pp. 3146–3154.

[22] M. Guo, C. Lu, Z. Liu, et al., Visual attention network, Comput. Vis. Medium. 9 (2023) 733–752.

[23] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, arXiv. 2017. https://arXiv.org/abs/1706.03762.

[24] Z. Wang, W. Yan, T. Oates, in: 2017 International Joint Conference on Neural Networks (IJCNN), May 14–19, 2017, Anchorage, AK, USA, 2017, pp. 1578–1585.

[25] T. Bai, H. Guan, S. Wang, et al., Traditional Chinese medicine entity relation extraction based on CNN with segment attention, Neural Comput. Appl. 34 (2022) 2739–2748.

[26] Y. Okazaki, K. Saito, Recent advances of metabolomics in plant biotechnology, Plant Biotechnol. Rep. 6 (2012) 1–15.

[27] X. Li, J. Liu, T. Zuo, et al., Advances and challenges in ginseng research from 2011 to 2020: The phytochemistry, quality control, metabolism, and biosynthesis, Nat. Prod. Rep. 39 (2022) 875–909.

[28] Z. Li, F. Zhang, C. Fan, et al., Discovery of potential Q-marker of traditional Chinese medicine based on plant metabolomics and network pharmacology: Periplocae Cortex as an example, Phytomedicine 85 (2021), 153535.

[29] X. Wang, M. Jiang, J. Lou, et al., Pseudotargeted metabolomics approach enabling the classification-induced ginsenoside characterization and differentiation of ginseng and its compound formulation products, J. Agric. Food Chem. 71 (2023) 1735–1747.

[30] D. Yoon, W.-C. Shin, S.-M. Oh, et al., Integration of multiplatform metabolomics and multivariate analysis for geographical origin discrimination of *Panax ginseng*, Food Res. Int. 159 (2022), 111610.

[31] R. Li, W. Duan, Z. Ran, et al., Diversity and correlation analysis of endophytes and metabolites of *Panax quinquefolius* L. in various tissues, BMC Plant Biol. 23 (2023), 275.

[32] Q. Lou, T. Xin, W. Xu, et al., TaqMan probe-based quantitative real-time PCR to detect *Panax notoginseng* in traditional Chinese patent medicines, Front. Pharmacol. 13 (2022), 828948.

[33] J. Liu, H. Wang, F. Yang, et al., Multi-level fingerprinting and cardiomyocyte protection evaluation for comparing polysaccharides from six *Panax* herbal medicines, Carbohydr. Polym. 277 (2022), 118867.

[34] C. Zhang, X. Wang, Z. Lin, et al., Highly selective monitoring of in-source fragmentation sapogenin product ions in positive mode enabling group-target ginsenosides profiling and simultaneous identification of seven *Panax* herbal medicines, J. Chromatogr. A 1618 (2020), 460850.

[35] R. Ji, T.A. Garran, Y. Luo, et al., Untargeted metabolomic analysis and chemometrics to identify potential marker compounds for the chemical differentiation of *Panax ginseng, P. quinquefolius, P. notoginseng, P. japonicus*, and *P. japonicus* var. *major*, Molecules 28 (2023), 2745.

[36] H. Wang, L. Zhang, X. Li, et al., Machine learning prediction for constructing a universal multidimensional information library of *Panax* saponins (ginsenosides), Food Chem. 439 (2024), 138106.

[37] Y. Gao, S. Zeng, X. Xu, et al., Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: A retrospective, multicentre, diagnostic study, Lancet Digit. Health 4 (2022) e179–e187.

[38] Y. Sha, W. Meng, G. Luo, et al., MetDIT: Transforming and analyzing clinical metabolomics data with convolutional neural networks, Anal. Chem. 96 (2024) 2949–2957.

[39] D. Misra, Mish: A self-regularized non-monotonic activation function, arXiv. 2019. https://arXiv.org/abs/1908.08681.

[40] A. Young, H. Röst, B. Wang, Tandem mass spectrum prediction for small molecules using graph transformers, Nat. Mach. Intell. 6 (2024) 404–416.

[41] R. Prajit, Z. Barret, V.L. Quoc, Searching for activation functions, arXiv. 2017. https://arXiv.org/abs/1710.05941.

[42] K. Günter, U. Thomas, S.M. Andrea, et al., Self-normalizing neural networks, Adv. Neural Inf. Process Syst. (2017) 971–980.

[43] T.Y. Lin, P. Goyal, R. Girshick, et al., Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 318–327.

[44] S. Bijlsma, I. Bobeldijk, E.R. Verheij, et al., Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation, Anal. Chem. 78 (2006) 567–574.

[45] S. Wang, X. Li, M. Jiang, et al., Headspace solid-phase micro-extraction gas chromatography/mass spectrometry (HS-SPME-GC/MS)-based untargeted metabolomics analysis for comparing the volatile components from 12 *Panax* herbal medicines, Phyton Int. J. Exp. Bot. 91 (2022) 1353–1364.

[46] J. Wang, X. Gong, H. Chen, et al., Causative classification of ischemic stroke by the machine learning algorithm random forests, Front. Aging Neurosci. 14 (2022), 788637.

[47] H. Wang, H. Wang, X. Wang, et al., A novel hybrid scan approach enabling the ion-mobility separation and the alternate data-dependent and data-independent acquisitions (HDDIDDA): Its combination with off-line two-dimensional liquid chromatography for comprehensively characterizing the multicomponents from Compound Danshen Dripping Pill, Anal. Chim. Acta 1193 (2022), 339320.

[48] T. Huan, Y. Wu, C. Tang, et al., DnsID in MyCompoundID for rapid identification of dansylated amine- and phenol-containing metabolites in LC-MS-based metabolomics, Anal. Chem. 87 (2015) 9838–9845.

[49] Q. Song, Y. Song, N. Zhang, et al., Potential of hyphenated ultra-high performance liquid chromatography-scheduled multiple reaction monitoring algorithm for large-scale quantitative analysis of traditional Chinese medicines, RSC Adv. 5 (2015) 57372–57382.

[50] F. Zheng, X. Zhao, Z. Zeng, et al., Development of a plasma pseudotargeted metabolomics method based on ultra-high-performance liquid chromatography-mass spectrometry, Nat. Protoc. 15 (2020) 2519–2537.

[51] N. Shlezinger, Y.C. Eldar, S.P. Boyd, Model-based deep learning: On the intersection of deep learning and optimization, IEEE Access 10 (2022) 115384–115398.