frontiers in Nutrition

REVIEW
published: 26 November 2021
doi: 10.3389/fnut.2021.780228

Chemometric-Guided Approaches for Profiling and Authenticating Botanical Materials

Evelyn J. Abraham[1] *and Joshua J. Kellogg*[1,2]*

[1] Intercollege Graduate Degree Program in Plant Biology, The Pennsylvania State University (PSU), University Park, PA, United States, [2] Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, United States

Botanical supplements with broad traditional and medicinal uses represent an area of growing importance for American health management; 25% of U.S. adults use dietary supplements daily and collectively spent over $9. 5 billion in 2019 in herbal and botanical supplements alone. To understand how natural products benefit human health and determine potential safety concerns, careful *in vitro*, *in vivo*, and clinical studies are required. However, botanicals are innately complex systems, with complicated compositions that defy many standard analytical approaches and fluctuate based upon a plethora of factors, including genetics, growth conditions, and harvesting/processing procedures. Robust studies rely upon accurate identification of the plant material, and botanicals' increasing economic and health importance demand reproducible sourcing, as well as assessment of contamination or adulteration. These quality control needs for botanical products remain a significant problem plaguing researchers in academia as well as the supplement industry, thus posing a risk to consumers and possibly rendering clinical data irreproducible and/or irrelevant. Chemometric approaches that analyze the small molecule composition of materials provide a reliable and high-throughput avenue for botanical authentication. This review emphasizes the need for consistent material and provides insight into the roles of various modern chemometric analyses in evaluating and authenticating botanicals, focusing on advanced methodologies, including targeted and untargeted metabolite analysis, as well as the role of multivariate statistical modeling and machine learning in phytochemical characterization. Furthermore, we will discuss how chemometric approaches can be integrated with orthogonal techniques to provide a more robust approach to authentication, and provide directions for future research.

Keywords: metabolomics, adulteration, multi-omics, dietary supplements, biochemometrics, chemometrics, botanicals, authentication

OPEN ACCESS

Edited by:
Susan Murch,
University of British Columbia, Canada

Reviewed by:
Ying Liu,
Canadian Food Inspection Agency, Canada
James J. Harynuk,
University of Alberta, Canada
Ikhlas Khan,
University of Mississippi, United States

***Correspondence:**
Joshua J. Kellogg
jjk6146@psu.edu

Specialty section:
This article was submitted to Nutrition Methodology, a section of the journal Frontiers in Nutrition

Received: *20 September 2021*
Accepted: *31 October 2021*
Published: *26 November 2021*

Citation:
Abraham EJ and Kellogg JJ (2021) Chemometric-Guided Approaches for Profiling and Authenticating Botanical Materials. Front. Nutr. 8:780228. doi: 10.3389/fnut.2021.780228

INTRODUCTION

Botanical medicines and dietary supplements represent a growing facet of personal health and medical care for Americans; the 2017 survey from the Council for Responsible Nutrition found that botanicals make up ca. 39% of total dietary supplement usage for adults in the United States (1), and US sales of herbal supplements totaled $9.6 billion in 2019, an annual increase of 8.6% (2). The use of botanical medicines and dietary supplements has come to include patients receiving

disease therapy, such as cancer (3) and chronic obstructive pulmonary disease (COPD) (4). The increase in economic and biomedical relevance of botanicals have led to a rise in research interest surrounding their potential health benefits, including support from the National Institutes of Health (5). The US National Library of Medicine's clinical trial tracker (clinicaltrials.gov) had >140 active clinical trials involving "herbal" or "botanical" preparations listed (accessed July 30, 2021) (6). However, the veracity of biomedical research, whether it is *in vitro* studies or clinical trials, is predicated on the authenticity and purity of the botanical(s) being studied. Botanical products are inherently complex chemical mixtures that can vary depending on abiotic and biotic factors during growth and post-harvest processing. Complicating this is the fact that products can be obtained from multiple producers and growers, potentially with multiple sources of raw material and processing techniques. Thus, to ensure the authenticity, efficacy, and safety of botanical dietary supplements, complex multi-faceted methods are required. This review focuses on chemometric and orthogonal methods for profiling, analyzing, and comparing botanical systems. We first provide opportunities and limitations of traditional botanical product authentication, followed by an overview of alternative chemometric approaches, then delve into a plethora of multivariate statistical approaches for botanical evaluation and present a workflow for how researchers can rationally select an analytical model based on data types and goals.

OPPORTUNITIES AND LIMITATIONS OF TRADITIONAL APPROACHES

Morphology

Plant morphology is the traditional approach to botanical product authentication, based on leaf shape and size and arrangement, color, life cycle changes, and other phenotypic factors. The combination of modern resources for plant identification and expansive collections of medicinal plant herbarium vouchers allows for fairly accurate morphological characterization (7, 8). Although trained specialists provide the most accurate identifications, guidebooks and phone applications provide a simple, inexpensive avenue for authentication. Increased accuracy results from micromorphology which allows species-specific evaluations of pollen shape, pore size, and other microscopic traits (9, 10). Recently, machine learning and image processing software have led to high-thruput identification of medicinal plants based on predefined characteristics and extensive training datasets (11, 12).

Despite its strengths, morphology-based identification is limited and often impractical, especially for rare plants. Similar environments and evolution pathways can result in unrelated plants with strong morphological resemblances but differing medicinal properties. Furthermore, important morphological information is lost when plants are dried or powdered, such as leaf shape and texture. Morphology also varies between plant parts, and recorded information for identifying plants based on below ground parts rarely exists. While certain root characteristics are useful, such as stone cells, auxiliary root angle, and rhizome length, the literature for species level identification is lacking and often contradictory between labs (13–16). Taxonomic identification is further complicated by vernacular names, which vary based on culture, location, language, and subspecies (17, 18).

Genetics

Genetic approaches, namely DNA barcoding and genome sequencing, are powerful tools for herbal product authentication. DNA can be extracted from fresh or dried tissue and is often effective with post-processed material (19). Primer-based methods are the most straightforward approach to DNA based identification: predefined primers for single genes (ITS2), a combination of genes (*matK* and *rbcL*), or chloroplast genomes (18, 20–23) amply specific fragments know to vary between species and have potential to differentiate morphologically and genetically similar species (24). Extensive sequence libraries exist which simplifies species identification; rare and understudied species are not thoroughly represented though (24). As sequencing becomes increasingly advanced and affordable, the applicability of genetic marker-based identification of a broad range of botanicals will increase.

DNA barcoding, including random amplification of polymorphic DNA (RAPD) (25) and inter-simple sequence repeats (ISSR) (26), provides a robust evaluation of genome diversity through examination of the presence/absence of more than 20 random fragments of polymorphic DNA at a time. Primer-based approaches amplify random segments of DNA to compare polymorphic variations among species. Although DNA barcoding is reliable, it is time consuming and requires meticulous method optimization for each application. Further, there is low resolution at the species or sub-species level (27). Recent advances in metabarcoding, which combines next-generation sequencing with bioinformatics, has greatly improved the ability to detect adulteration and supplementation in herbal products (28–30). Notably, the EU and other governing bodies suggest metabarcoding to evaluate the identity and safety of botanical products (31). For example, Seethapathy et al. used metabarcoding to determine that over 24% of Ayurvedic herbal products tested do not contain the botanical as labeled (28). However, metabarcoding is expensive and requires a reference DNA library and pre-defined genetic markers. So, for rare species or those without sequenced genomes, metabarcoding is ineffective as a quality control approach (32).

While genetic approaches have proven useful for botanical product quality control, there are limitations. Plant tissue is damaged and degraded during processing procedures, hindering extractions of high-quality DNA (33). Since genetic approaches do not provide quantitative data, there is limited ability to determine relative abundances of different species within a product. Thus, DNA barcoding does not allow trace contamination, as from shared equipment, to be discerned from intentional, large-scale adulteration of products. A final limitation is the inability to evaluate medicinal properties through barcoding based approaches. The medicinal value of a product is largely based on its chemical constituents. Without

detailed chemical analysis, the presence and relative abundance of specific medicinal compounds is unknown. So, while genetics may be able to detect adulteration, it cannot determine a products actual medicinal value. Thus, chemical evaluation serves to both authenticate botanicals and provide information of a product's bioactive potential.

TARGETED ANALYSIS OF BIOMARKERS

A simple and common approach to herbal product quality control is the use of small-molecule based targeted analysis. This approach uses individual and small groups of compounds specific to the botanical in question. Using a targeted analysis allows quick verification the product contains the plants as advertised. This section outlines the targeted analysis workflow, with examples and explanations of the pros and cons of targeted approaches.

Single Biomarker Approach

The first step in using small molecule chemistry to serve as biomarkers of quality and authenticity is to identify a targeted metabolite or small set of metabolites specific to the botanical in question. Since many commercially available botanical medicines and dietary supplements are fairly well characterized in scientific literature, the identification of predominant metabolites (also known as 'marker compounds') is fairly straightforward. These targeted compounds are analyzed by a chemical methodology and compared against reference standards and literature values; common analytical techniques include charged aerosol detection (CAD), ultraviolet-visible (UV/VIS) spectrophotometry, and mass spectrometry (MS), often with chromatographic separation beforehand (liquid chromatography, "LC", or gas chromatography, "GC" being the two primary forms). Nuclear magnetic resonance (NMR) is an analytical technique that has become more quantitative recently (qNMR) to facilitate comparisons between complex botanical samples (34–36).

However, axiomatic to using a defined marker compound is the knowledge of the chemistry of the system at hand and the commercial availability (or the ability to isolate and conclusively identify) of the target marker compounds. While many botanicals on the market have well-developed chemical libraries and/or have monographs detailing their chemical composition [including the German Commission E (37), US Pharmacopeia (38), and Tyler's Herbs of Choice (39)], not every botanical, nor every potential dietary supplement, is as thoroughly studied, and gaps in the literature of even well-known botanicals still exist today. The choice of marker compound also should, but doesn't necessarily, have relevance to the putative biological activity of the botanical medicine or dietary supplement. Finally, standards must be available to construct calibration curves; if they are not commercially available, researchers face the daunting task of isolating and elucidating the structure prior use as a marker compound (40).

Furthermore, tying authenticity to a single compound overlooks the broader chemical landscape present in the botanical product, and can leave products susceptible to potential adulteration. Single-point analyses can be confounded by spiking with specific compounds or mixtures that might bypass quality control procedures. One example is the discovery by Chandra et al. of adulteration in ginkgo (*Gingko biloba*) extracts spiked with either single isolated flavonoids or flavonoid-rich mixtures (41). As the broad category "flavone glycosides" was chosen by the gingko market as an authenticity marker, it was prone to spiking by flavones (e.g., quercetin, kaempferol, and isorhamnetin) to meet the quality criteria. In fact, three out of eight products analyzed in the study that were labeled to contain ginkgo extracts actually resembled those of commercial extracts from Japanese sophora (*Styphnolobium japonicum*) (41). In other cases, botanical dietary supplements have been doped with dyes or other synthetic mixtures to deceive single molecule quality control methods (42). Supplements with alleged weight loss properties were spiked with alkaloid derivatives, ephedra stimulants, or androgenic steroids (43, 44). Spiking and adulteration can also be used to bypass negative controls searching for known contaminants/adulterants; 1,3-dimethylamylamine (1,3-DMAA) is one case study. The United States Food and Drug Administration (FDA) had banned 1,3-DMAA in 2016 and pulled all products containing the stimulant from shelves because of an increased incidence of ER visits correlated with this stimulant, as well as failure to meet regulatory conditions (45). However, investigations by Cohen et al. revealed 1,3-DMAA analogs present in multiple weight loss supplements (five out of six tested), illustrating how adulteration can be used to sidestep regulatory authorities with potentially toxic constituents (45).

Molecular "Fingerprints"

Beyond single molecules for targeted biomarker detection, researchers can collect information on a range of molecules or a "chemical fingerprint" that exemplifies a more robust and nuanced representation of the botanical's metabolite profile. Using multiple components blunts the potential for metabolite spiking (as seen with single marker compound approaches) and can provide more selective and sensitive analysis for distinguishing authentic material. Lv et al. (2016) developed an HPLC-based fingerprint to differentiate species and geographical origins of *Rhizoma coptidis* using six distinct alkaloids (46), while eight organic acids were used to distinguish between *Castanea* spp. Buds (47), and Parveen et al. validated an UHPLC-UV-MS method incorporating 10 standard compounds to distinguish closely related *Tinospora* species (48). Even AOAC's official method for some botanicals incorporates multiple compounds; their method 2015.007 for investigating Ashwagandha (*Withania somnifera*) employs 10 withanolide glycosides and aglycones (49). However, multi-molecular chemical "fingerprints" are more time- and labor-intensive approaches, as they require the quantitation of multiple compounds with different linear ranges and limits of detection and quantitation (LOD and LOQ, respectively). This also does not circumvent the issue with single biomarker approaches needing reliable, commercially available standards in order to determine the overall fingerprint and quantitation for the analysis.

METABOLOMICS

The 'metabolome' is generally defined as the complete set of small molecules produced by an organism or biological sample at any given point in time. Metabolomics, therefore, is the unbiased, holistic measurement of the metabolome (though practically speaking there is no single analytical approach capable of measuring all small molecules in one experiment), and the relative areas or heights of signals within the metabolome can be employed as a basis for comparison between two or more samples. As such, metabolomics provides a powerful tool for understanding the complete chemical makeup of an herbal product, which can be used for efficient and accurate quality control and authentication. Metabolomics characterizes the chemical relationships that underlie variations based upon genotype, origin (50), climate (51), or other biotic or abiotic interactions (52–54). While a variety of analytical inputs can be used to generate metabolome data – including Fourier-Transformed infrared spectroscopy (FT-IR), charged aerosol detection (CAD), ultraviolet-visible (UV/VIS) spectrophotometry, mass spectrometry (MS), and Nuclear Magnetic Resonance (NMR) spectroscopy – the two primary analytical approaches employed for the majority of metabolomics studies are liquid chromatography coupled to mass spectrometry (LC-MS) and NMR spectroscopy. These two provide incredible sensitivity and selectivity in profiling a large fraction of the metabolome of a sample, while also offering detailed structural information crucial for metabolite annotation for the authentication of botanical dietary supplements and medicines (55, 56). The advances of metabolomics techniques is not the focus of this review, the incredible innovation and progress that has been achieved in metabolomics experiments have been discussed elsewhere (57–59).

As relative comparisons are being made across a large dataset (often hundreds to thousands of peaks in a single metabolome data matrix), the chemical identification of the peaks is not necessary at the outset of the experiment and analysis. Thus, untargeted metabolomics studies can compare complex samples with no a priori knowledge of their constituents (60) and do not require the acquisition of analytical standards to complete comparative analyses, a distinct advantage over the targeted or fingerprinting approaches described above.

CHEMOMETRIC APPROACHES FOR PATTERN RECOGNITION AND SIMILARITY DETERMINATION

While a valuable tool for authentication of herbal products, the innate complexity of metabolomic datasets can be daunting when developing novel quality control approaches. One of the major challenges facing metabolomic (or other molecular fingerprinting approaches) is not the collection of the data, but instead the processing, analysis, and interpretation of the expansive datasets that are often generated. In metabolomics, the data matrices often have more columns (independent variables, such as *m/z*-retention time pairs or NMR signal buckets) than rows (samples) and are known as "landscape" matrices. "Chemometrics" refers to the application of statistical methods to discover relevant analysis and maximize the information obtained from the chemical datasets (61). For the authentication of botanical materials, chemometric pattern recognition approaches are the most prevalent. There are a variety of multivariate mathematical–statistical methods for prediction and pattern recognition (**Figure 1**), which have disparate criteria for successful application to complex chemical datasets.

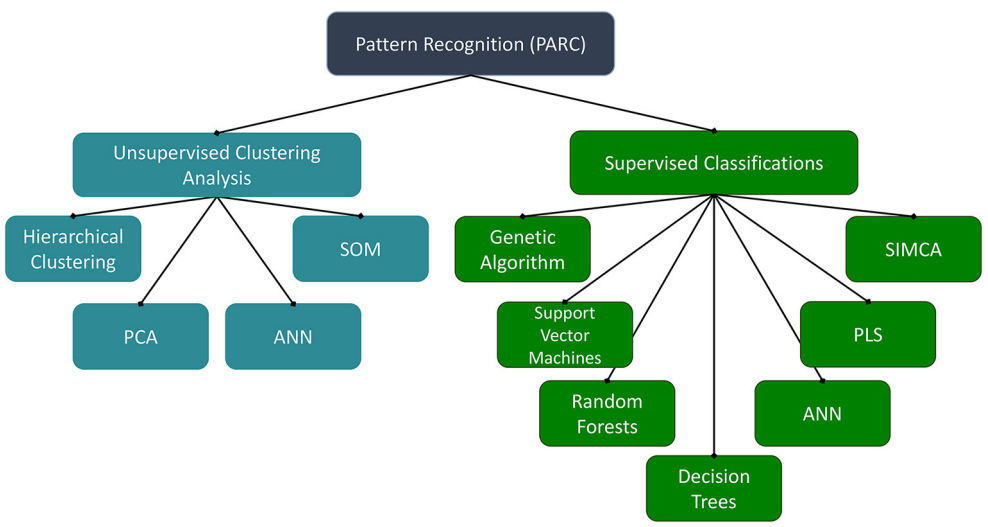

FIGURE 1 | Pattern recognition methods. ANN, artificial neural networks; PCA, principal component analysis; PLS, partial least squares; SIMCA, soft independent modeling of class analogy; SOM, self-organizing maps.

Data Preparation for Chemometric Analysis

In any statistical analysis, the robustness of the predictions and inference is limited by the quality of the data that is input into the model. For chemometric analysis, there are a number of aspects of the dataset that will contribute to the overall quality and reliability of the resulting model. One aspect of note is the reproducibility of analytical data. Variations in extraction protocol, sample handling as well as the mass spectrometer detection itself (mass analyzer, detector, and even the chromatography components) preclude facile comparisons between labs. This can potentially lead to differing raw spectral data, as well as variations in results obtained (62).

Raw spectral data, from any analytical source (LC-MS, GC-MS, NMR, FT-IR, etc.) must be processed in order for the statistics to be effective. For some spectral data (e.g., ^{1}H-NMR and FT-IR), the data is traditionally sliced into "bins" that are then used as individual features in the dataset (63, 64). Mass spectrometry data is obtained as discreet features (unique *m/z*-retention time pairs), yet still requires multi-step "preprocessing" to identify peaks and align the data. There are numerous methods and workflows to preprocess spectral data, and have been examined and reviewed exhaustively elsewhere (65–71). While most open access preprocessing software yields similar performance in detection of actual peaks ("true" features) from the data [as examined by Li et al. (68)], the abundance of parameters needed to fine tune in order to develop a robust final dataset can be challenging for researchers. The subsequent scaling, centering, and normalization of the dataset can also play a factor in the resulting statistical analysis (72, 73). Thus, careful treatment of the raw data during preprocessing is critical to downstream chemometric analyses in order to obtain reproducible and reliable interpretations of the data. The potential for variations in the processing of the data is a persuasive argument in favor of the trend in metabolomics to encourage open science by depositing the spectral data, as well as metadata associated with the preprocessing parameters used, in accordance with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles (74).

Unsupervised Approaches

Unsupervised methods are the (relatively) simplest ways of classifying large chemical datasets, designed to analyze data that can only be arranged in one matrix. These methods are "unsupervised" in the sense that no data classifications are known before the analysis; instead data structures are revealed through these pattern recognition methods. Researches should be aware of the differences between hard and soft classification techniques.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised approach which projects multivariate data (with k features/variables) onto a smaller dimensional space ($<k$-1). As such, PCA is often referred to as a projection or dimension reduction method. The metabolite profile is reduced to uncorrelated principal components (PCs) which represent the total variation present in the metabolome. The first principal component accounts for the maximum percentage of the overall variance, the second principal component (orthogonal to the first) accounts for the second largest amount of variance, and so on until all the variation in the data is accounted for or the number of principal components reaches the limit (i.e., the number of features-1) (75). The principal components are plotted in a pair-wise fashion (typically the first two, which explain the most variation) on a 2-dimensional plane – known as a "scores" plot – that demonstrates the spatial relationship between different samples. Points which cluster together have similar correlations in the PC variations, which translates to similarities in their overall chemical profile. Likewise, dissimilar samples are located further from one another in the two-dimensional graph. A second corresponding graph associated with PCA is the loadings plot, in which the features (variables) are arranged in a two-dimensional plot using the same PCs as the scores plot. The spatial representation of the loadings mirrors that of the scores, thus enabling the determination of which features are more prevalent in certain clusters of samples. Zhang et al. (76) developed an approach to authenticate juices from different berry fruits using untargeted metabolomics. Using PCA generated from LC-QTOF-MS spectra, they were able to discriminate between blueberry, cranberry, apple, and grape juice (**Figure 2**). The corresponding loadings plot yielded 18 characteristic markers that were able to categorize the juices (76). Additionally, Farag et al. differentiated ten cinnamon accessions from the main cinnamon species using ^{1}H-NMR metabolomics combined with unsupervised chemometric approaches (77). The scores plot (**Figure 3**) distinguished between *Cinnamomum cassia* and *C. verum*, with PC1 and PC2 comprising 77% of the variability in the model. The loadings plot suggested nine key metabolites which could be used to differentiate between cinnamon accessions, including cinnamaldehyde and eugenol; the exclusive presence of eugenol in *C. verum* samples suggested its potential as an authentication marker (77). Thus, PCA represents a robust and potent chemometric tool in the evaluation of different samples and their authenticity/purity.

However, while PCA can demonstrate clusters of samples based upon their chemical profile, it is not able to provide quantitative metrics around the degree of similarity between samples, nor ranking how similar samples are to one another. Furthermore, PCA relies on a subsection of the overall principal component model to visually represent similarities and differences between the samples; this is often an *ad hoc* choice of PCA components which can mask outliers or shift the overall spatial relationship between samples, leading to the possibility of specious results and subsequent conclusions. Integration of multiple PCs into a single quantitative comparison may circumvent this. Termed the composite score, it has potential to facilitate comparisons between multiple samples using the entirety (or at minimum a significant subset) of the principal component model to quantify similarity between samples (78). This approach was used recently by Wallace et al. to differentiate *Hydrastis canadensis* supplements from potential adulterants (79).

Suggestions for future use: PCA is a powerful unsupervised clustering tool with accessible computational resources to

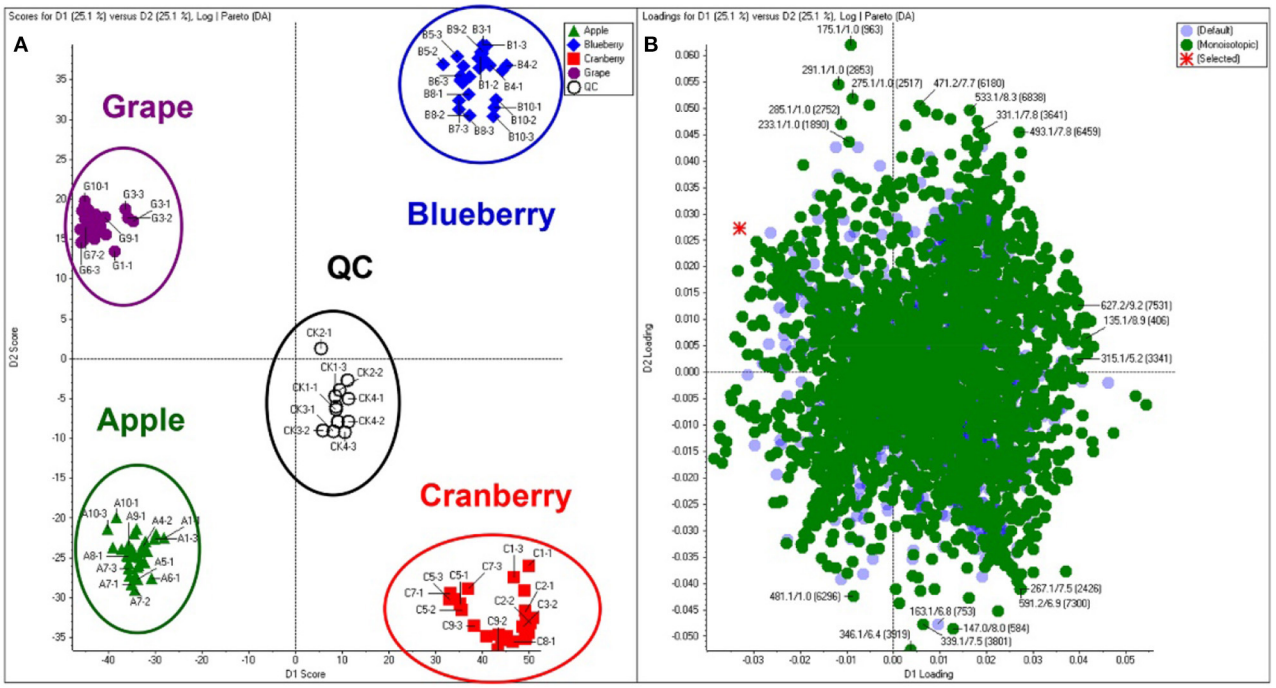

FIGURE 2 | Principal component analysis (PCA) scores **(A)** and loadings **(B)** plot demonstrating differentiation between fruit juices based upon untargeted metabolomic analysis. Reproduced with permission from Zhang et al. (76). Copyright 2018, American Chemical Society.

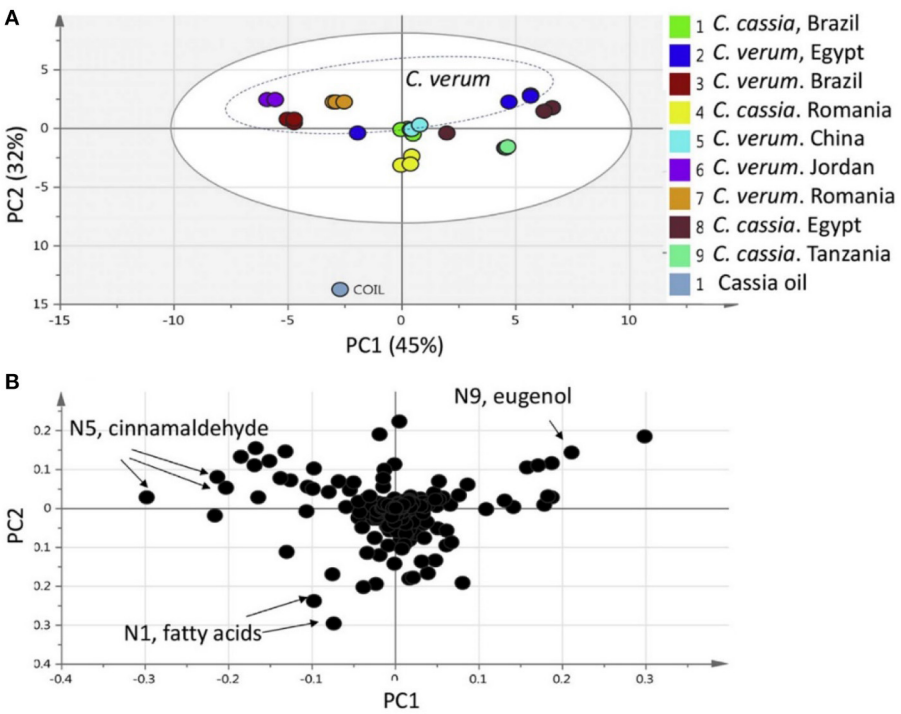

FIGURE 3 | Principal component analysis (PCA) from Cinnamomum verum and C. cassia from different geographical origins, and representative commercial oil, using ^{1}H-NMR ($n = 3$) metabolomics. The scores plot **(A)** demonstrates clusters at distinct spatial points in the PC1-PC2 scores plot, and loadings plot **(B)** highlights major contributing molecules to the separation of the samples. Reproduced with permission from Farag et al. (77). Copyright 2018, Elsevier Ltd.

simplify analysis, making it an ideal first step in any chemometric analysis. PCA can be used prior to any supervised approach to confirm expected clustering among samples, and that apparent distinctions result from true variations in sample metabolomes, not as a result of overfitting to predefined categories. Alone, PCA can be used to determine if adulterated and pure samples differ while simultaneously identifying biomarkers likely responsible for any variation. Thus, PCA has potential for a quick and easy approach to botanical authentication based on metabolite profiles. Possible sample clustering that may be identified using PCA is species proximity, cultivation procedures, or origin of plant growth. For any authentication study requiring more detailed information of how samples are related or identification of unknown with a single model, other approaches should be performed concurrently with PCA.

Hierarchical Cluster Analysis (HCA)

Hierarchical clustering analysis (HCA) uses distances between sample groupings (clusters) to organize samples into taxonomies; objects with the highest similarity cluster together, and generated clusters are treated as a new, independent feature which are clustered with the next most similar variable. Similarity is calculated as distance between variables through a variety of algorithms, including Euclidian, Mahalanobis, or city block (Manhattan); similarly, there are various linkage rules for amalgamating the cluster analysis, such as minimum or maximum similarity between variables, group average (average similarity between every possible pair of data points), or Ward's Method (sum of the squared distance between each pair of data points). Proximity matrixes are used to compare the calculated similarity of all groups. The shorter the distance, the more similar the variable, and thus more likely to be related. However, since the similarity (distance) and linkage can be calculated using different combinations of rules, the results of cluster analysis are difficult to compare between studies. In the case of sample authentication, each botanical sample is treated as a variable and clusters are formed based on similarity in peak heights (or other metabolite features) so that the most chemically similar samples group together. HCA has been used to distinguish *Cirrhosae bulbus* from common adulterants using UPLC-ELSDA fingerprinting (80). Zhou et al. demonstrated the use of HCA in discriminating between two bitter melon (*Momordica charantia*) chemotypes with different medicinal properties (81). While PCA was able to distinguish the two chemotypes, HCA allowed a deeper insight into how each variety differed within the groups (81), and the combination of PCA and HCA predicted biomarkers for easy chemotype distinction of unknown samples. NMR chemical fingerprinting of Sarsaparilla species (*Decalepis hamiltonii*, *Hemidesmus indicus*, *Pteridium aquilinum*, and *Smilax* spp.) revealed four clear clusters, which were further confirmed by patterns in the NMR spectra (**Figure 4**) (82). In addition to detection of herbal adulteration, HCA provides opportunity to detect contamination with pharmaceutical drugs; Cebi et al. used HCA to classify coffee and tea blends adulterated with sibutramine, an illegal weight-loss drug (83).

Suggestions for future use: HCA models share similar possible directions as PCAs, with additional information of how samples are related chemically within generated clusters. A potential study may evaluate differences in chemotypic and genotypic based hierarchical clustering for authentication. It is possible genetic approaches may discover adulteration by non-target species but miss contamination with synthetic compounds; chemotypic approaches may simultaneously provide species distinction and chemical authentication. Similarly, comparisons of HCAs generated via multiple analytical techniques (H-NNR vs. LC-MS) may provide a deeper understanding of sample relationships though inclusion of additional compounds.

Self-Organizing Maps

Artificial neural networks (ANN) is a collective term for several machine learning methods. The most common unsupervised ANN approach is self-organizing maps (SOM). Section 5.2.3.4 provides an overview of supervised ANN in natural product authentication.

Self-organizing maps (SOMs), sometimes referred to as Kohonen maps or Kohonen networks, is a neural network-based algorithm that reduces the input dimensionality to represent sample patterns; SOM forms a 2-dimensional map where similar samples are mapped closer together. The benefit of this approach is that SOMs account for non-linear information in the data, and each variable's importance to the model can be derived from the weights associated with each map "point" (84). Torrecilla et al. (85) employed SOM to analyze extra virgin olive oils and detect adulteration via the addition of other oils. Using random and non-random noise to simulate adulteration, the SOM was constructed which yielded a misclassification rate $<1.3\%$ (68). Using previous research, Menezes et al. generated a library of terpenes present in three tribes of Annonaceae species (521 molecules) for use in training a SOM (86). The model was able to classify unknown samples into the three predefined tribes with 80% average accuracy (86). Similar approaches have been demonstrated using diterpenes to classify Lamiaceae spp. (87) and flavonoids to classify Asteraceae spp. (88).

Suggestions for future use: Menezes et al. provide an SOM method very applicable to natural product authentication (86). Using previous metabolomics data to classify botanical samples, despite variations in analytical and collection techniques, provides an opportunity to create authentication models without extensive benchwork. This approach should be applied to commercial supplements with well-defined chemistry to develop predictive models for existing products.

Supervised Approaches

Supervised statistical methods require the data matrix have both independent and dependent variables, the latter of which can be nominal (categorical) or numerical in nature. Nominal dependent data are ideal for clustering data into pre-defined classes, such as "pure" and "adulterated," whereas numerical data can allow for the ranking, quantifying, and comparing variables against each another. Many machine learning approaches are supervised models based on training datasets. Simply, a set of samples with known dependent variables are used to train, generate, and validate a model, which subsequently predicts the classification of additional, unknown samples (or the remainder

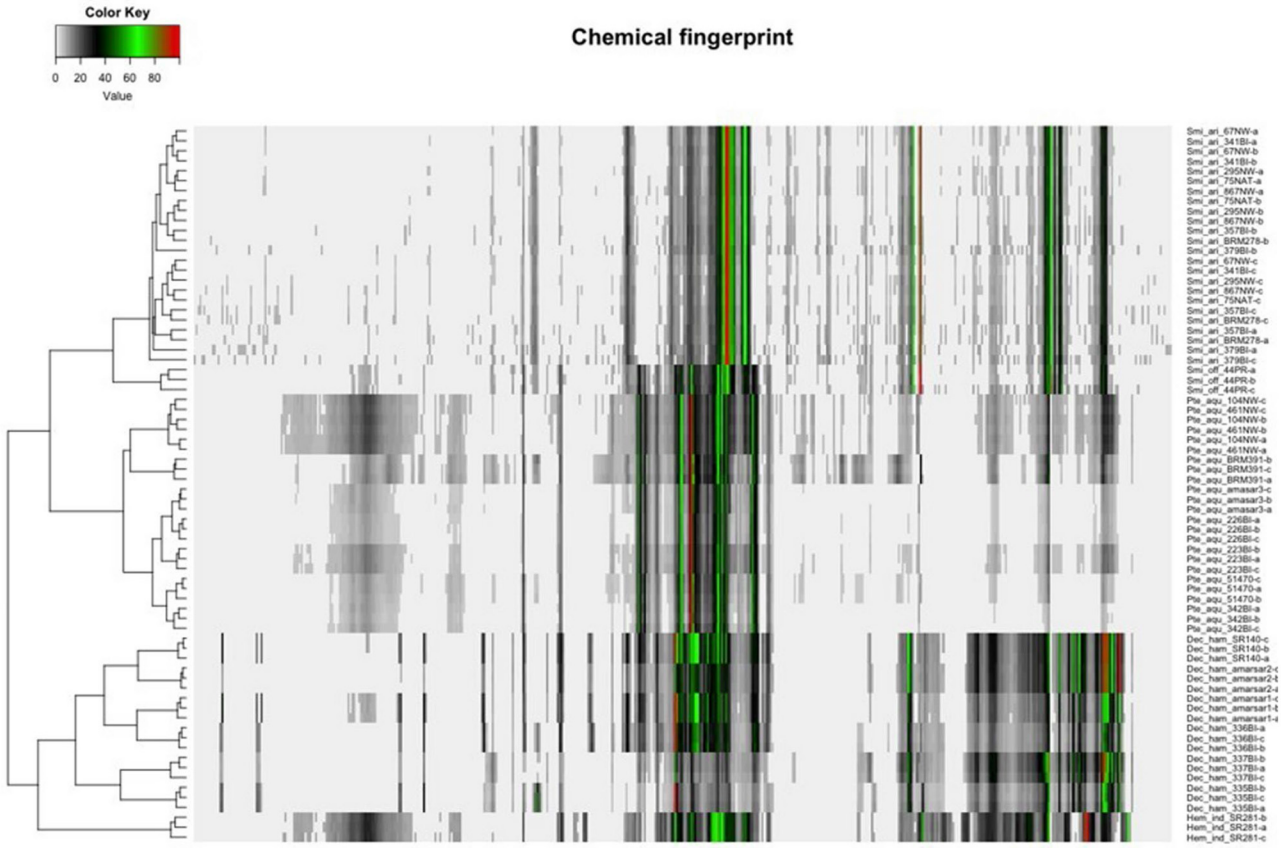

FIGURE 4 | Unsupervised clustering analysis of four species of Sarsaarilla using cheometric modeling of ^{1}H NMR data. Peak patterns are also provided to illustrate differences between four major species. Reproduced under a Creative Commons Attribution 4.0 license from Kesanakurti et al. (82).

of the data). However, as the numbers of samples in a metabolomics data set are generally fewer than the number of variables, supervised techniques are prone to overfitting the data (89); even so far as to be able to fit a model to completely random data (90). Therefore, model validation is critical before any interpretation of the model is reliable, and often quality criteria of the model are reported such as R^2 (a measure of the fit of the model) and Q^2 (the ability of the model to predict unknown samples) (91).

Partial Least Squares (PLS)

PLS is a dimension reduction tool similar to PCA. PLS condenses complex data to simpler latent variables which explain shared features between correlated samples, but with a dependent variable to supervise the construction of the model. The goals of PLS are akin to linear regression: classification of dependent variables and understanding the independent variables (metabolite features) that are predictors of this classification. A PLS model plots the latent components among the independent variables that best explain variations in dependent variables, and samples are projected onto the model space. The resulting scores plot allows simple visualization of sample clustering based on the reduced variables; the loading plot provides information about specific variables which contribute the most covariance to the model. The two primary types of PLS analyses- PLS-R and PLS-DA are defined by the nature of the dependent variable.

PLS-R

Partial least squares-regressions model variations among the independent variables to explain a numerical dependent variable. While PLS-R is uncommon for quality control of botanical products, it has been employed with biomarker identification, biochemometrics, and detection of adulteration (92). For example, PLS-R was employed to differentiate between *Hydrastis canadensis* (goldenseal) and four common adulterants using FT-NIR data (64). Following preprocessing and filtering the spectral data, PLS-R modeling successfully clustered pure goldenseal from non-target species, as well as differentiated between various goldenseal parts (roots and shoots) (64). In this case, the plot consisted of latent variables which reduced the spectral data as guided by a gradient in contamination as the dependent variable. This study also highlights the importance of preprocessing and filtering data; unprocessed data was unable to distinguish species using PLS-R (80). Partial least squares is also one of the primary predictive chemometric approaches: when there are correlations

are drawn between the dependent data set (often bioactivity or other quantitative data) and the independent chemical data from which the model is derived. This approach, known as biochemometrics when using bioactivity data, is explained more fully below.

Suggestions for future use: Some future applications of PLS-R in herbal product authentication could include evaluating products with known and specified variations in ingredients-teas with varying percentages of *Ilex paraguariensis* (yerba mate) and ashwagandha root for example. PLS-R may also be useful for discovering biomarkers to quickly differentiate between bioactive and inactive products though detailed bio-chemical analysis of commercial supplements and subsequently screening additional products for identified markers. This may bypass some typical issues with single marker analysis, as described in section Single Biomarker Approach, by using commercially available products for biomarker discovery as opposed to predetermined pure plants.

PLS-DA

Partial least squares-discriminate analysis (PLS-DA) models the data similar to PLS-R, but with the caveat that the dependent variable be a binary descriptor (e.g., "class1" vs. "class2", "authentic" vs. "adulterated", etc.), which are coded as −1 and 1, or 0 and 1. The resulting scores plot is typically able to discriminate between the two groups, as it is guided by the classification of the samples. PLS-DA is one of the most common chemometric tools applied to chemical data for authentication and discrimination among botanical products. The study by Ismail et al. demonstrates this approach by differentiating between different grades of gaharu (agarwood, *Aquilaria malaccensis*). Using ^{1}H-NMR metabolomics, a PLS-DA model was able to differentiate between "high grade" and lower grades of gaharu (**Figure 5**), and the resulting loadings plot also highlighted aquilarone derivatives that discriminated the different quality classes (93). Windarsih et al. also employed PLS-DA analysis to differentiate between authentic *Cucuma xanthorrhiza* ("Java ginger") and samples adulterated with *C. aeruginosa*. PLS-DA yielded a robust model (R^2 and Q^2 of 0.993 and 0.986, respectively) which separated authentic from adulterated samples (94).

One of the limitations of PLS-DA is that the categorization is restricted to a binary class designation. If there are more than two main categories, the discriminant analysis requires pair-wise comparison, complicating the analysis and potentially limiting the conclusions which can be drawn. This is exemplified by Barbosa et al.'s study to differentiate and authenticate paprika grown in three different areas (La Vera and Murcia in Spain and the Czech Republic) (95). The PLS-DA classification plots were done as iterations of one region vs. the other two, to comprehensively demonstrate that the three regions were distinct from one another (a classification rate of 100%) (**Figure 6**).

Suggestions for future use: PLS-DA has excellent potential in herbal product quality control since binary categorical classes can encompass multiple facets of plant differentiation. These applications range from classifying samples based on geographic origin, plant parts, species or subspecies, or adulteration status. Although PLS-DA requires pre-determined classifications of data, the loading plots can guide discovery of biomarkers for quick screening of unclassified samples. An interesting study would model one chemical dataset for multiple classifications of the same samples to evaluate how clustering and model validation (R^2 and Q^2) change to determine the most reliable classifications for authentication.

Soft Independent Modeling of Class Analogies (SIMCA)

SIMCA is a supervised expansion of PCA: samples are grouped into predefined classes and PCA is performed on each class, so that each group is projected onto a separate PC space. To detect adulteration, there are only two classes: authentic or adulterated, so one-class PCA's for authentic or adulterated samples can be generated (79). A new, unknown sample's classification is predicted by projecting it to the PC space and calculating the Q statistic (or Q residual), quantifying the similarity of the unknown PCA to the training set's PCA (79). The Q-statistic predicts if the new sample belongs in the authentic, adulterated, both, or neither class. Thus, SIMCA distinguishes similarities among samples and unknowns rather than defining the differences between groups (96). Wallace et al. intentionally adulterated *Hydrastis canadenis* with varying concentrations of *Copits chinesis* and used untargeted metabolomics with SIMCA analysis to differentiate between pure and tainted samples. Using one-class modeling, the Q statistic of "unknown" adulterated products was calculated and found to fall above the 95% confidence interval for pure samples, successfully identifying even the lowest percentage of contamination (5%) and providing a higher resolution of differentiation than PCA alone (**Figure 7**) (79).

Suggestions for future use: SIMCA is a powerful classification tool with digestible graphic outputs. SIMCA can be used for classification problems where the output for each sample is already known, such as adulterated vs. pure. It is a straightforward tool for analysis of binary classifications but becomes more complicated as more categories are added. Thus, it should be reserved for problems focused on identifying contaminated samples when deep machine learning modeling is not necessary. A suggested approach to botanical quality control is to perform unsupervised PCA to identify and confirm a binary clustering of samples followed by SIMCA to predict the classification of unknown products.

Machine Learning Models

While easily interpretable, models such as SIMCA and PLS are inherently linear algorithms, capable of modeling only linear latent covariance. As biological data are often non-linear, it is probable that the related chemical data also has a non-linear latent structure. Thus, non-linear machine learning methods can be uniquely suited to examine relationships from metabolomic or other chemical data.

Decision Trees

Decision trees are a machine learning approach that use hierarchical decisions to determine sample classification based

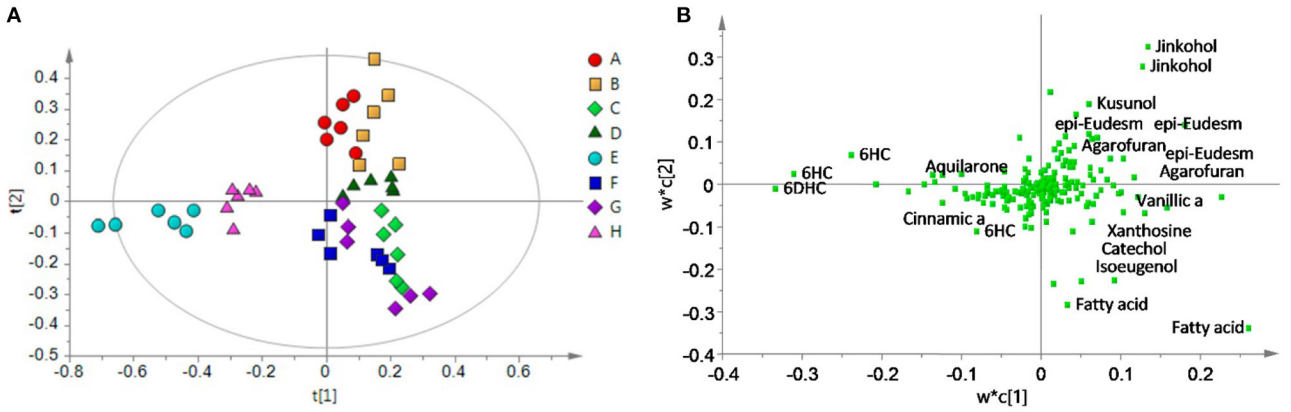

FIGURE 5 | PLS-DA analysis of gaharu (Aquilaria malaccensis) woods by ^{1}H-NMR untargeted analysis. The PLS-DA scores plot **(A)** effectively discriminated between lower grade products ("E" and "H") and higher grades. And the corresponding loadings plot **(B)** demonstrated that the lower quality products contained higher levels of aquilarone derivates. Reproduced under a Creative Commons CC BY 4.0 license from Ismail et al. (93).

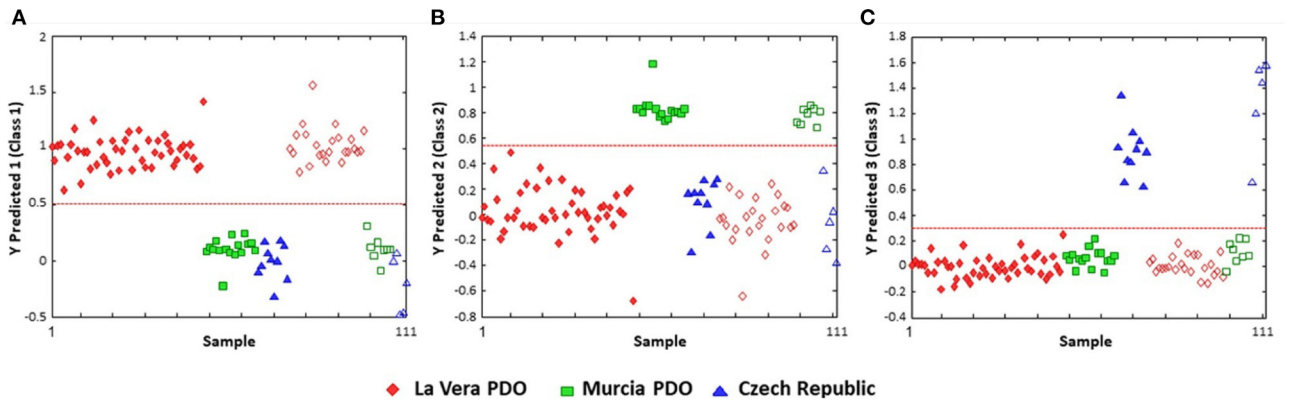

FIGURE 6 | PLS-DA discrimination plots according to the geographical region of production of paprika. **(A)** La Vera PDO vs. other regions; **(B)** Murcia vs. the other two; **(C)** Czech Republic vs. other classifications. The dashed red line indicates the classification boundary between the two designations. Open symbols represent the training data; the solid symbols are the test data. Reproduced with permission from Barbosa et al. (95). Copyright 2020, American Chemical Society.

on training data. Trees are displayed upside down, with the bulk data at the top being split based on features that best distinguish the data at each step. These distinctions are typically based on the presence or ratios of specific metabolites that separate one classifier from another. The result is a tree split into branches at decision nodes that end with leaves, or the classification groups. Decision trees are commonly referred to as classification and regression trees (CARTs) to encompass both distinct variables (classification) and numerical or continuous variables (regression). In the case of botanical product quality control, samples are classified based on species, purity, or other relevant factors. Classification trees were used to classify different cultivars of avocados based on HPLC-CAD metabolomics (97). Training data that comprised of spectra from 32 avocado samples of three varieties generated a tree which guided classification of unknown avocado oil samples into cultivar classes or no class based on specific, model generated rules (**Figure 8**) (97). A strength of decision trees is the ability to classify an unknown sample into a "no class" group to avoid overfitting or forcing a sample into a classification group. A creative application of decision trees is to predict the need of specific safety tests and evaluations for botanical products, as demonstrated by Little et al. the group used an *in silico* decision tree model to analyze the need for safety assessments of botanical products based on UHPLC with UV, CAD, and HRMS metabolomics, structure identification, consumer exposure, and existing safety evaluations (98). The developed tree used chemical data and previous records to determine if any tests are necessary for consumer safety depending on the presence of certain metabolites in the sample and database information of safety data (98). This study highlights the versatility of decision trees in quality control – they can not only identify botanical adulteration, but they can also ensure safe practices while developing botanical supplements.

Suggestions for future use: Decision trees have the appeal of a visually appealing output for a complex machine

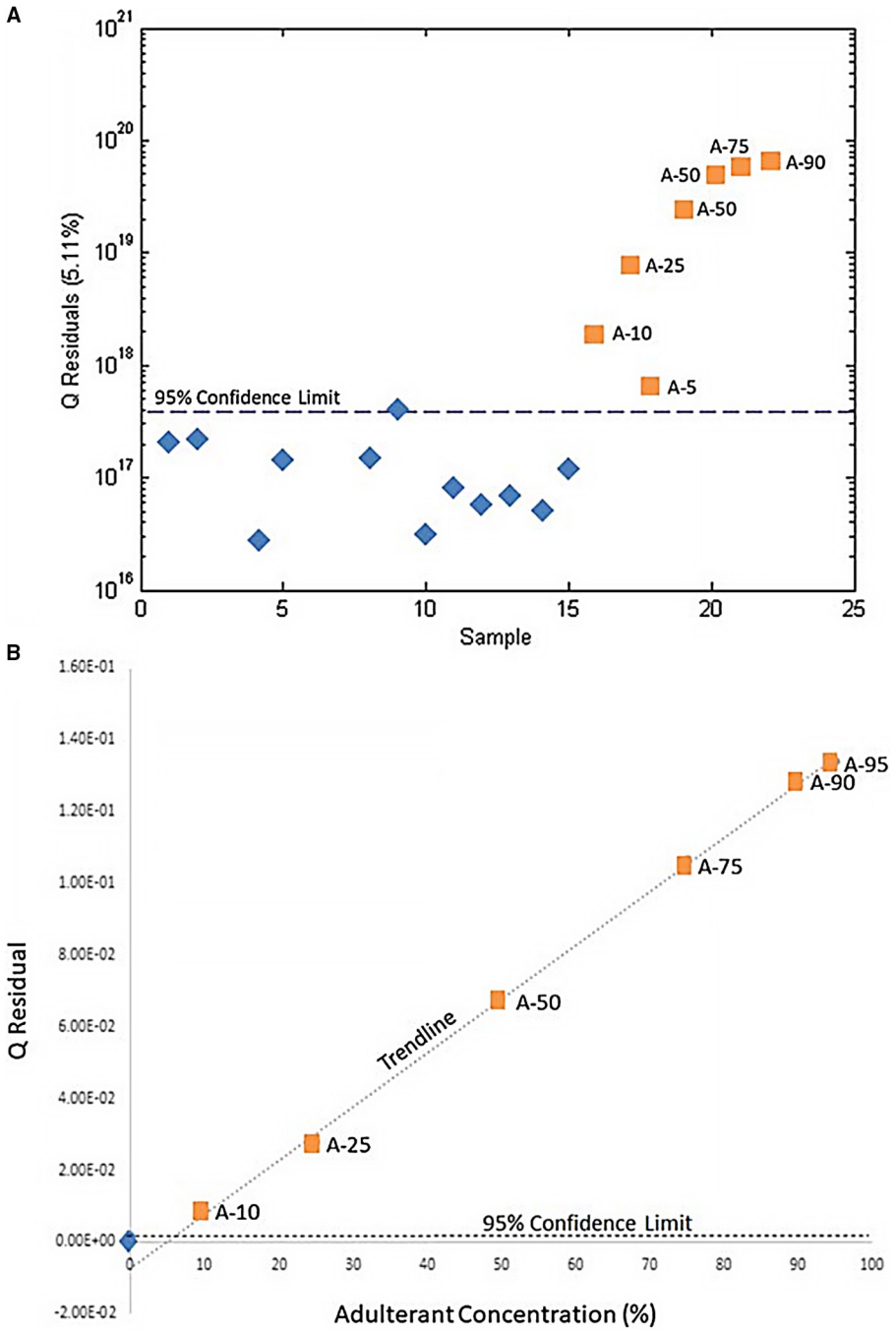

FIGURE 7 | Use of SIMCA to determine adulteration of H. canadenis by C. chinesis. **(A)** SIMCA demonstrating that pure H. candenis samples (blue diamonds) are below the 95% confidence interval and adulterated samples (orange squares) are above the 95% confidence interval. **(B)** The Q-residual of each adulterated sample. The blue diamond represents the mean Q-residual for the unadulterated H. canadenis samples. Reproduced with permission from Wallace et al. (79). Copyright 2020, Springer Nature.

learning model. They hold promise for discerning unknown product identification, detecting adulteration of products with known contaminants, and discovering biomarkers for various classifications. Once the decision tree model is built with training data, it is relatively straightforward to feed an unknown's chemical data through the model to predict its classification. This is an exciting possibility for quality control, especially when the most common adulterants are known to base relevant decision trees around. It should be noted that the decision algorithms at each node are based on separating the data available from the

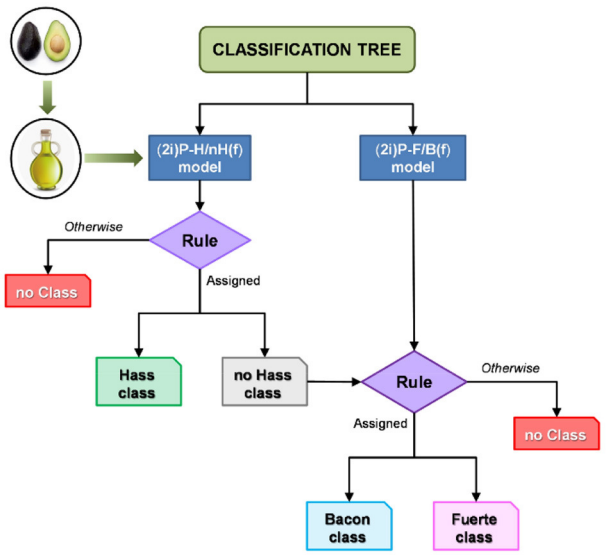

FIGURE 8 | Use of classification trees to identify unknown avocado oil samples as a specific cultivar or as no class based on HPLC-CAD metabolite profiles of a training set. Reproduced with permission from Martin-Torres et al. (97). Copyright 2019, John Wiley and Sons.

previous split, but the split progressions may not actually be the most reliable representations of divisions in the data. Random forests (described below) increase the accuracy of node splits but lose clear visual representation of the model.

Random Forests

Developed in 2001 by Breiman (99), random forest (RF) methods build an ensemble of decision trees, each of which is trained using the dependent variable(s). Each tree produces an outcome, and the aggregate outcome from all the trees (aka the forest) is reported as the outcome of the model. Random forest holds several advantages over other methods. The multiple decision trees produce more accurate classifications compared to a single decision tree algorithm, and it is less prone to overfitting than other supervised approaches (100). Additionally, one other advantage of random forest is that it can be used for both classification and regression problems. Deklerck et al. used random forests to classify heartwood samples of *Pericopsis elata*, a protected timber species. Using Direct Analysis in Real Time ionization coupled to time-of-flight mass spectrometry (DART-TofMS) on wood slivers, the random forest model using cross-validation was able to correctly predict *P. elata* samples (101). To analyze *Zanthoxylum* seed oils, Houet al. built a random forest classification model that differentiated between the two main species (*Z. bungeanum* and *Z. armatum*) with 100% accuracy from cross-validation. Even simplifying the model to only the most important chemical features, the cross-validated model still maintained 100% accuracy (102). Random forests also have the ability to be a predictive machine learning tool, and provide correlative predictions between dependent variables and the associated independent chemical dataset (103).

Suggestions for future use: Random forests can be used for the same purposes as decision trees where increased reliability of decisions at each node is necessary. This includes instances of fewer chemical data or a smaller number of training samples. Since random forests combine multiple decision trees, the computational input is much higher, so model building and training takes longer. Thus, random forests are not the best option when expecting a quick turn-around. However, there is potential application in developing random forest models for detection of adulteration of complex botanical products and mixtures. The extent of random forests in detecting contamination and purity of extremely complex samples in a high-throughput manor should be explored.

Support Vector Machine

Support vector machines (SVMs) are another supervised machine learning technique that can be employed for regression or classification analyses. The objective of the SVM algorithm is to find a plane in a *k*-dimensional space (*k* representing the number of features) that distinctly classifies the data points into groupings so that it has the maximum margin (i.e., maximum distance) between data points of both classes. Similar to other supervised machine learning or multivariate approaches of chemometrics data (where the number of features outstrips the number of samples), SVM can be prone to overfitting, so training on a smaller subset of samples, followed by cross-validation, is key to generating robust classification or predictive models.

Martín-Torres et al. (97) used SVMs to differentiate between the geographical origins as well as the botanical variety of avocados. Samples from five different countries (on three continents) representing seven avocado varieties were analyzed using normal phase HPLC-UV/VIS, after which the data was interpreted using two different multivariate approaches. The authors found that PLS-DA and HCA were unable to resolve the differences in geographical origin or between main groupings of variety. However, a three-input class SVM classification model (3iC) correlated the three different continents of origin (Africa, Americas, and Europe), as well as between the three dominant varieties ("Bacon", "Fuerte", and "Topa-Topa"); the latter having 100% correct assignments and precision and sensitivity of 1.00 (104). SVMs were also used to classify *Paris polyphylla* via fusion of Fourier-transformed infrared spectroscopy (FTIR) and UV-VIS spectroscopy data (105). Pan et al. used untargeted LC-MS metabolomics to profile five different *Uncaria* species in order to authenticate the source of Uncariae Rammulus Cum Uncis (Gou-Teng). A SVM model correctly categorized both training and test samples, and was used to classify 20 commercial Gou-Teng (GT) samples (106). The model predicted 16 of the samples were *Uncaria rhynchophylla*, while four did not match any of the *Uncaria* species. These four samples exhibited LC-MS chromatograms that were substantially different from the others, and thus it was believed that these were other *Uncaria* species or mixtures of *Uncaria* species. This represented a significant advantage over other (un)supervised techniques for discrimination purposes. And using data fusion techniques of mid-infrared (MIR) (transmission and reflection mode) and

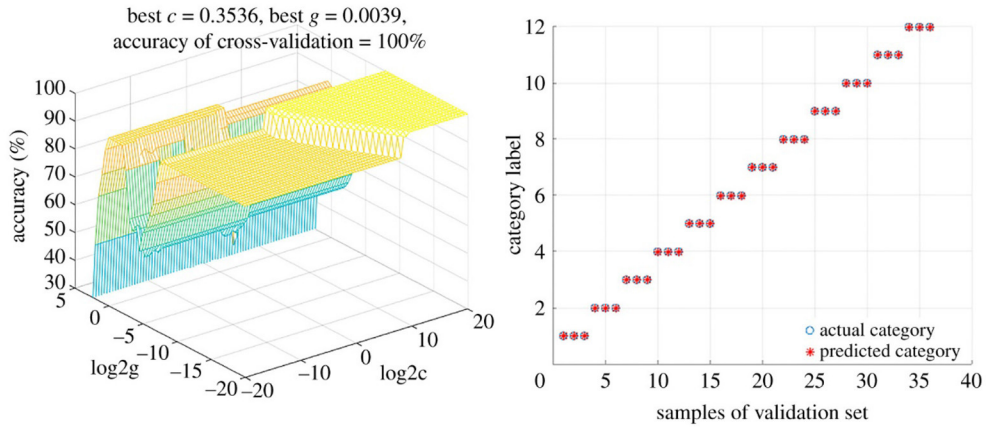

FIGURE 9 | Support Vector Machine (SVM) model for differentiating between Dendrobrium species. Using a low-level fusion strategy of MIR and NIR spectral data from 12 Dendrobrium sp. Reproduced under a Creative Commons Attribution License 4.0 from Wang et al. (107).

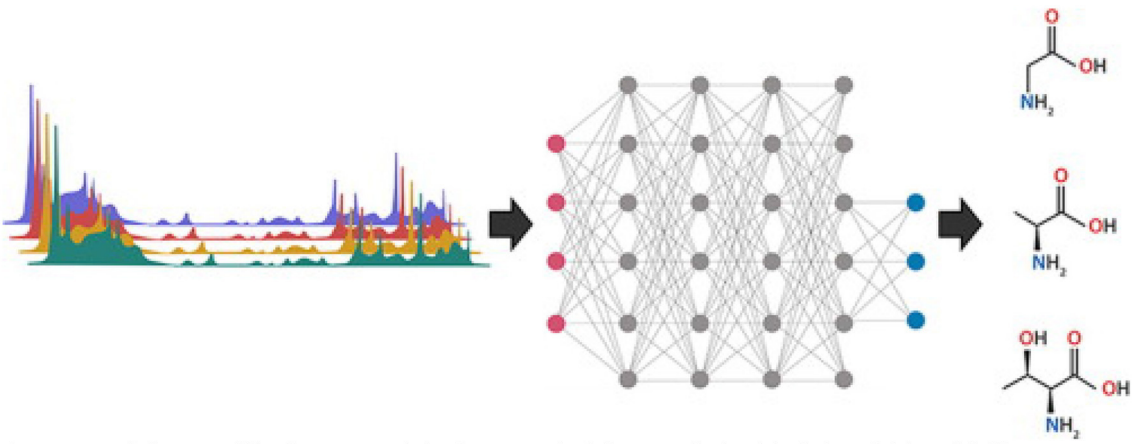

FIGURE 10 | Metabolomics workflow for ANN. Metabolite features are loaded as inputs (red circle), which are fed through hidden neuron networks (gray circles), and categorized into output categories (blue circles). Each connection (gray line) has a weight, and each neuron has a bias, which are used for the activation functions. Reproduced with permission from Pomyen et al. (112) Copyright 2020, Elsevier Ltd.

near-infrared (NIR) spectra followed by SVM analysis facilitated the discrimination of 12 different *Dendrobrium* species (107). SVM provided perfect discrimination (100% accuracy rates) for both calibration and validation sets (**Figure 9**).

Suggestions for future use: SVMs have practical applications for both classification of test data and prediction of unknown samples origin, species, or cultivar. SVMs may prove useful for distinction of genetically and chemically similar plants that cannot be differentiated by other clustering models, either supervised or unsupervised. For example, many sub-species of herbs have overlapping genotypes due to crossbreeding and PCA analysis can fail to separate the most closely related cultivars using chemical data (108); SVMs may provide a deeper level of distinction. This application can be applied to authentication of products commonly adulterated with very similar species that lack the promised medicinal output. Since SVM models can automatically handle missing data, SVMs can be used for metabolomes with variable metabolite profiles and lower resolution analytical techniques.

Genetic Algorithms

Genetic algorithms (GA) are based on the processes of evolution, including natural selection, reproduction, and mutations. These processes take place over multiple generations of increasingly accurate and simple solutions to a complex problem (109). In the case of botanical authentication and quality control, the problem may be product identification, detection of adulteration, or biomarker discovery. The solutions are the subset of metabolites and their ratios that best classify samples based on predefined classes or distinctions. As a brief example, consider Gil et al.'s study which used a GA to identify the region of rose wine origin (110). At generation 0, every combination of the 79 polyphenols present in the samples as detected by UPLC-MS were evaluated for their ability to distinguish between origin

region. Solutions with the highest fitness, or its distinguishing power as determined by linear discriminant analysis, were selected for reproduction. During reproduction, two solutions were mixed in a cross-over like process to create a new generation of unique solutions with higher fitness than the previous. The selection and reproduction processes were repeated for five generations, and each of the final combinations of polyphenols was tested for its accuracy in cross-validation tests. Those with the highest accuracy were further evaluated for their ability to discriminate wine origin regions in an unknown validation sample set. The GA model was able to discover a set of 4 polyphenols that had 86.7% accuracy (110). GA also provides the opportunity for simultaneous sample and variable selection for improved speed and accuracy for unsupervised clustering of samples and biomarker discovery (111). This bi-clustering approach opens the door for high-throughput metabolomics authentication of botanical materials.

Suggestions for future use: Potential for GA in botanical product quality control ranges from geographic identification to generation of a subset of biomarkers for subsequent analysis. The speed of GA modeling is ideal for situations requiring fast turn-around, so it is practical for developing authentication schemes for new products or products with increasing rates of adulteration. It should be noted, however, that GAs can be difficult to interpret, since the steps the model takes to combine data and reach a solution are not defined for the user and models fed the same data often reach different solutions. Thus, users should only use GA when the intermediate steps are not necessary for model validation.

Artificial Neural Networks (ANN) With Known Outputs

ANN are the backbone of deep learning machine learning models. Mimicking the human brain and neurons allows computers to recognize complex patterns in sets of training data and predict the classification of a new dataset using the resulting model. There are three main sections of an ANN: an input layer, an output layer, and hidden layers in between (**Figure 10**) (112). Each metabolite from the complete set of samples is treated as an individual input and connections are generated randomly through multiple hidden layers to generate an output response. Hidden layers are comprised of "neurons" that connect metabolites with a random numeric weight and have a randomly assigned bias (**Figure 10**) (112). Together the weights and bias generate an activation function to determine if a neuron will be activated for use in the next hidden layer. This process of forward progression is repeated until the model predicts an output (such as adulterated or pure). Typically, the first prediction is incorrect since the weights and bias are random, so the model uses backwards progression using the prediction error to modulate weights and biases throughout the hidden layers. Through multiple rounds of forward and backward progression with a variety of inputs belonging to each output category, the model can predict the output of new data by processing the new inputs through the meticulously developed hidden layers (112, 113). The history of ANN in metabolomics, as well as an in-depth explanation of different ANN models for spectral data is reviewed by Mendez et al. (113).

Binetti et al. used ANN with merceological, NIR, and H-NMR data to classify olive oil cultivars (114). Using H-NMR spectral data, ANNs were able to classify unknown samples with >99% accuracy, despite variable environmental, harvesting, and processing conditions (114). Additionally, ANN modeling of headspace solid-phase microextraction (HS-SPME) coupled with GC-TOF-MS of 374 honey samples over two years provided 94.5% accuracy in prediction of honey origin when data from both collection years was combined (115). These studies are promising for herbal product classification - botanical material analysis is typically complicated by temporal, environmental, and procedural variations. In addition to classification and identification, ANN modeling has potential to predict the chemical and medicinal properties of supplements without extensive bioassays and robust chemical profiling (116). Using species classification and extraction procedures as inputs, Tusek et al. used an ANN to predict chemical features, including total phenolic content and extraction yield, and antioxidant potential of nine medicinal plants (116).

Suggestions for future use: ANNs hold immense potential for herbal product authentication. Since training data covers a range of environmental, temporal, and procedural variables, the predictive nature of the resulting model has very high accuracy. This is critical for commercial products that have limited information about harvest and processing procedures. An interesting study would determine if a single ANN model built on samples with a range of preparations (powdered, dried, capsules) and environmental factors can successfully classify and authenticate various types of new products. Additionally, prediction of medicinal properties using ANN should be expanded to allow confirmation of desired effects from commercial products quickly and accurately. This will take authentication a step further from identifying product constituents and increase efficacy of botanicals on the market. Users should take caution when using ANN to not over interpret their results. While ANNs are powerful classification tools for large data sets, they do not provide information on the chemical distinctions on which the model is built. Thus, the model does not allow interpretation about specific chemicals responsible for classification of samples.

Precautions for Using Classification Models

Each model describe above has benefits to the natural product community, and there are examples highlighting their usefulness in the literature. However, each model also has pitfalls. It is crucial for researchers to understand the dangers of overinterpreting their outputs. One such downfall is overfitting data, or forcing data points into a category due to the lack of a "unknown" output option within the model. Almost all of the models described in this review are prone to overfitting, but some models, like decision trees and random forests, reduce this possibility by including an unknown option or compiling the output of multiple models into the output. It is important to validate each model by withholding a sample's data as a validation set with a

known output, as well as reporting the Q^2 and R^2, as described in section Unsupervised Approaches.

An additional warning is that not every model is applicable in each situation. Despite a model seeming to fit a research goal, it is possible the type, quantity, or quality of data is not applicable to a given algorithm. Multivariate statistics and metabolomics projects require careful planning prior to data collection to ensure desired models can be used. For suggestion of models to use in different situations, see section Conclusions and Future Directions.

COMBINING ORTHOGONAL DATASETS

While modeling chemical data through chemometric approaches can leverage the immense information contained therein to investigated nuanced differences between samples, being able to differentiate samples based on their geographic origin, taxonomic relationship, or adulteration level, the chemical composition represents only one facet of potential data to be analyzed. Incorporating additional data sources, whether it is from orthogonal chemical analyses, bioactivity/toxicity data, or genetic data, has the potential to develop discriminatory models that are even more robust in authenticating botanical products. Often, combinational approaches can increase the efficacy and reliability of natural product quality control, and should be implemented when feasible.

Multiple Chemical Analyses Inputs

There is no single chemical analysis able to profile every metabolite present in a complex sample; each approach has some detractors. Ultra violet-visible spectroscopy (UV-VIS) requires a chromophore that can absorb energy within these wavelengths of light (often 180–800 nm); mass spectrometry (MS) can only monitor structures that are ionizable; nuclear magnetic resonance (NMR) is not as sensitive in detecting low-abundance compounds (55, 117). Therefore, combining different chemical investigations of a metabolome can better represent the chemical diversity present in a sample, and consequently allow for more precise modeling and differentiation between samples. These 'data fusion' approaches have been used with different botanical products to evaluate their authenticity and detect adulteration. Spiteri et al. combined 1H-NMR with LC-MS to discriminate between commercial honey. The PCA was constructed considering each technique separately, and then combining NMR and LC-MS together. The authors found that the discriminating potential increased through data fusion, allowing better separation of the four different floral origins with no misclassification observed (118). NMR and LC-MS were also combined to detect adulteration of a commercial botanical dietary supplement which had resulted in the hypotensive collapse of several consumers. The product was purported to contain the species *Crataegus oxyacantha*, *Olea europea*, *Capsella bursa-pastoris,* and *Fumaria officinalis*. However, the analysis revealed the presence of indole alkaloids belonging to the genus *Rauwolfia*, such as ajmaline, reserpine and yohimbine. Subsequent quantitative analysis determined reserpine was present in pharmacologically-relevant doses (119).

Chemometric analyses using multiple analytical inputs have also been used to elevate and extract more information from more common and less expensive analyses, such as infrared analysis (IR) and ultraviolet-visible spectroscopy [UV-VIS, often abbreviated as LC methods (HPLC or UPLC) or diode array detectors (DAD)], to provide robust data and allow clear discriminate model formation. Combining three different types of detectors: diode-array detection, evaporative light scattering detection and mass spectrometry, Deconinck et al. constructed fingerprints for three common herbal products—*Rhamnus purshiana*, *Passiflora incarnata* L. and *Crataegus monogyna*. Using unsupervised projection chemometric analyses, the researchers were able to detect the presence of these plants in three different herbal matrices as well as in commercial preparations containing multiple botanicals (120). Wu et al. reported that fusing data obtained from polyphyllin content, FTIR spectra, and UPLC chromatograms yielded correct discrimination of Paridis Rhizome samples according to botanical and geographical origins by PLS-DA modeling. The authors reported that SVM and RF provided similar results (105). Two-step fingerprints, built upon mid infrared spectroscopy (MIR) and HPLC chromatograms, were analyzed by k-nearest neighbors and SIMCA to screen for five regulated plants used in commercial dietary supplements (121). And Zhou et al. fused two different infrared technologies – Fourier transform mid-infrared (FT-MIR) and near infrared (NIR) – to detect the origin of 210 *Panax notoginseng* samples from five cities in Yunnan Province, China. Random forest was used to establish classification models, which resulted in a classification accuracy of 95.6% (122). Data fusion of orthogonal analytical approaches has the potential to cover complementary facets of chemical space, and the subsequent modeling can be seen to be more powerful in its ability to discriminate between botanical samples as a means of authentication and adulterant detection.

Biochemometrics

The ability to profile large swaths of a metabolome without iterative methods of separation and purification means metabolomics approaches have an advantage for screening for bioactive metabolites (92, 123, 124). Integrating metabolomic fingerprinting with biological activity data allows for supervised methods to statistically model correlations between variations in biological response with differences in chemical composition across samples. These methods, collectively known as "biochemometrics", have become a driver for bioactive molecule discovery. Several statistical methods have been utilized for this purpose, including hierarchical clustering analysis (125), partial least squares (92, 126, 127), and partial least squares-discriminate analysis (128, 129). Of these, PLS and PLS-DA have emerged as the foremost multivariate approaches for biochemometric analysis. These approaches utilize different variable metrics to ascribe correlation (and thus importance) to the chemical signals with the variable importance in projection (VIP) plot, the S-plot, and the selectivity ratio being among the leading metrics (92, 130–132). Biochemometrics holds great promise for botanical examination and authentication, as it could leverage

relevant biological activity to determine a targeted fingerprinting method which has relevance to the biological function of the plant (as opposed to *ad hoc* choices of metabolites). Kim et al. used a biochemometric method to evaluate 17 different species of grape myrtle (*Lagerstroemia* spp.) based on their ability to increase glucose uptake *in vitro*. From the PLS-DA model (using the glucose update as the dependent variable), the *Lagerstroemia* sp. were grouped into two clusters, and from the S-plot the authors identified three main metabolites (myricetin-3-O-β-D-rhamnoside, quercetin 3-O-β-D-rhamnoside, and corosolic acid) that predicted glucose uptake activity and could be used as a discriminatory model for identifying bioactive species from the genus (**Figure 11**) (133). The integration of biological activity as an orthogonal dataset, and as a continuous numeric dependent variable in the dataset, allows for supervised chemometric methods to provide greater interpretation of the discriminatory model creation and identification of bioactive components. This can lead to the development of fingerprinting or authentication tools that correlate with the relevant biological effects of the botanical in question.

Multi-Omics Integration for Botanical Control

While metabolomic approaches provide ample opportunity for accurate, robust, and time-efficient authentication of complex botanical products, combining chemical data with other -omics approaches may yield the most effective solutions. Most commonly, metabolomics modeling is combined with genomics data. As mentioned in section Genetics, DNA barcoding and metabarcoding can lose accuracy at the species and subspecies level. Similarly, clustering of metabolites can lose resolution of chemically similar plants. Combining DNA barcoding using *rpoC1* and LC-MS metabolomic fingerprinting allowed species-level distinction between nine Phyllanthus species (134). Integration of genetics and metabolomics is easily the most common approach to botanical product identification, as outlined in **Table 1**. There are instances when metabolomics and genetics in combination cannot differentiate between species, so additional analytic approaches are employed, such as electronic noise (141), microscopy (142, 143), high-resolution melting analysis (144), Raman spectroscopy (145), or multiple metabolomic approaches (146). **Figure 12** demonstrates that integrating multifaceted -omics approaches can be achieved in a single study to increase the power of distinction and authentication of herbal products (141).

Although genetics is most integrated with genetics, there is potential to expand to lipidomics, proteomics, and transcriptomics. Lipidomics, the study of the complete set of lipids in an organism, is analytically similar to metabolomics; different extraction and analytical instrument methods target lipids. The same research lab could seamlessly transition from metabolite to lipid analysis since the instrumentation is often the same. On its own, lipidomics has been useful for detection of adulteration of white rice – RF and SVMs were used to discriminate pure and adulterated samples using LysoPCs and lysoPEs as novel lipid biomarkers (147). Using the same UPLC-MS instrument, Anagbogu et al. combined lipid and metabolite analysis to identify 30 genotypes of coffee; joining the two approaches increased species level resolution (148). Proteomics also uses similar instruments as metabolomics, but it has more variable methods that may complicate inter-lab experimentation. Peptide analysis allowed differentiation between mountain-cultivated ginseng and cultivated ginseng with 52 variable peptides between the groups (149), and MALDI TOF-TOF/MS yielded five proteins with potential to authenticate *Ophiocordyceps sinenis*, a traditional fungal medicine (150). Given the limited successful studies utilizing integrated -omics approaches for botanical product authentication and evidence

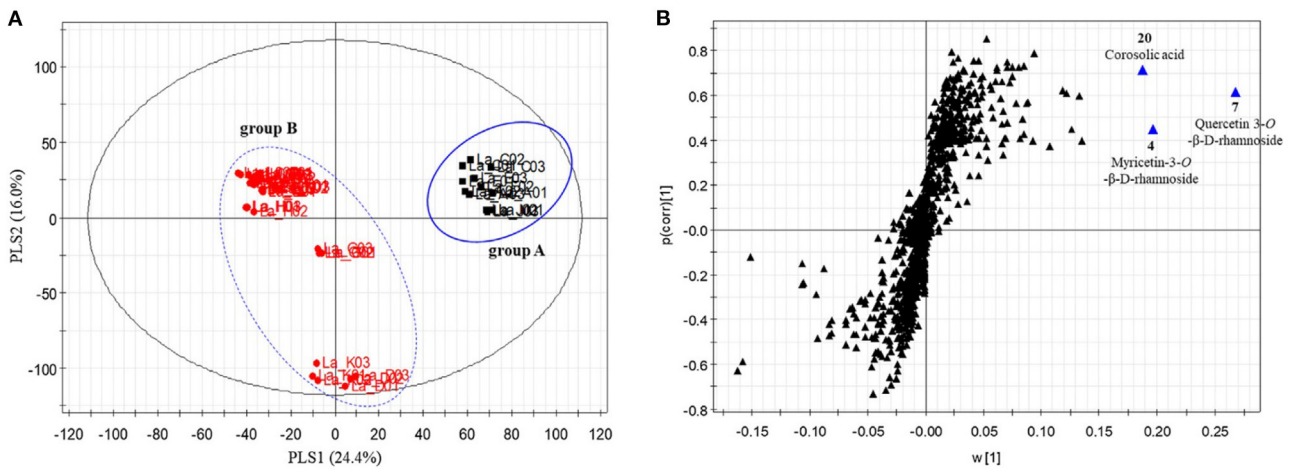

FIGURE 11 | Partial least squares-discriminate analysis of Lagerstroemia samples. Scores plot **(A)** from the PLS-DA model accounted for 40.4% of model variability, and demonstrated two distinct clusters of samples. The S-plots **(B)** revealed two flavonol glycosides (myricetin-3-O-β-D-rhamnoside and quercetin 3-O-β-D-rhamnoside) and corosolic acid as potential discriminatory biomarkers with activity in stimulating glucose uptake. Reproduced with permission from Kim et al. (133). Copyright 2020, Elsevier Ltd.

TABLE 1 | Orthogonal approaches to integrate genomics and metabolomics data analysis of botanicals.

Botanicals	Product type	Genetic approach	Metabolomic approach	Modeling	Author	Year (Ref.)
Qin jiao	Dried root powder	ITS2 barcoding	Q-TOF-MS H-NMR	ANOVA PCA OPLS-DA	Li et al.	2020 (135)
***Hypericum* taxa**	Essential oil Dried leaf	ITS2 barcoding ITS1 barcoding	GC-MS LC-HRMS LC-DAD-MS HPLC-DAD	PCA Biplots Mantel test	Zeliou et al.	2020 (136)
Salvia* subg *Perovskia	fresh leaf and root	trnH-psbA barcoding ITS2 barcoding	UHPLC-QTOF-MS	PCA	Bielecka et al.	2021 (137)
***Hypericum* spp.**	Cultured leaf	ITS1 barcoding ITS2 barcoding Chromosome number genome size	HPLC-DAD	PCA HCA	Brunakova et al.	2021 (138)
Sarsaparilla	Dried root	rbcL barcoding matK barcoding genome skimming DNA probe	H-NMR	HCA	Kesanakurti et al.	2020 (82)
***Glycyrrhiza* spp.**	Dried root powder Dried root stick Dried root capsule	rbcL barcoding matK barcoding ITS barcoding trnH-psbA barcoding	H-NMR UHPLC-UV	PCA SIMCA CDA	Simmler et al.	2015 (139)
***Echinacea* spp.**		Genome skimming metabaroding matK barcoding rbcL barcoding	HPLC-UV		Handy et al.	2021 (140)

that each approach has potential to identify adulteration, there is a gap in the botanical products community developing methods and statistical approaches for combined datasets. This is not a trivial undertaking; often the data sets generated for genomics, proteomics, and metabolomics experiments are very different, and their integration can be a challenge. The wide variety of expertise required to generate high quality data is also a factor in the wider implementation of a multi-omics approach to botanical authentication; these disparate techniques have different methodological proficiencies and even reagents and laboratory setups, necessitating broad proficiency in a single lab or a reliable collaboration between different laboratory groups. For data integration, the R tool mixOmics (including PCA, PLS, and PLS-DA tools) may prove useful for combined biomarker discovery and species identification (**Figure 13**) (151).

WHEN TO USE WHAT

This review highlights the number of chemometric techniques that can be applied to datasets in order to help authenticate botanical materials or detect adulteration. However, the diversity of approaches that are possible can be daunting to researchers unfamiliar with chemometric analysis and multivariate analysis/machine learning. While there is a bit of trial and error in selecting a chemometric approach, there are some points to consider in determining which technique to employ in analyzing a dataset. The decisions and chemometric options available to a researcher analyzing data are summarized in the following workflow (**Figure 14**).

First, is there response data collected with the chemical information? This could take the form of classification identifiers (e.g., "pure" vs. "adulterated"), control or QC datasets, taxonomic identification, quantitative data (e.g., temperature, geographic coordinates, elevation), or bioactivity data (inhibitory studies, cell studies, *in vivo* experiments, toxicological data, etc.). For datasets which do not contain any dependent information (only chemical input from FTIR, UV-VIS, MS, NMR, etc.), unsupervised analyses are recommended to understand the shape and relationships between samples without any guiding variables or observations. For a hierarchical analysis, where the similarity relationship between samples is ranked by distance, hierarchical cluster analysis (HCA) is the foremost choice. For examining

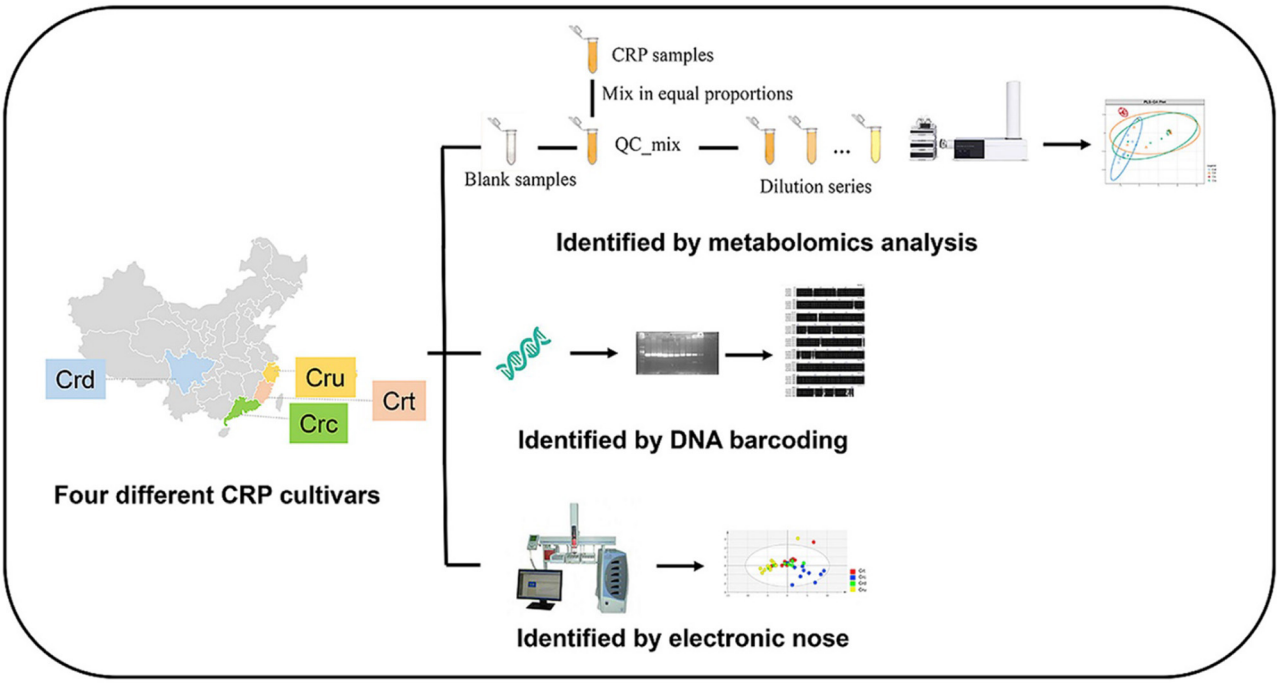

FIGURE 12 | Integration of metabolomics, DNA barcoding, and electronic noise increases accuracy of Citri Reticulatae Pericarpium cultivar distinction compared to each method alone. Reproduced with permission from Li et al. (135) Copyright 2020, Springer Nature.

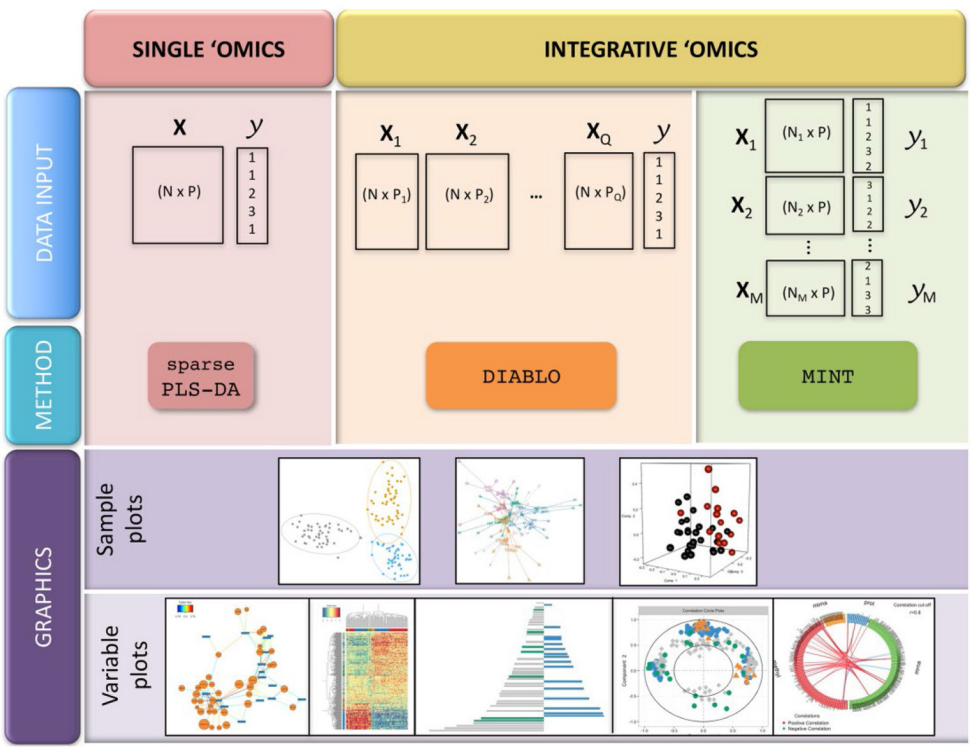

FIGURE 13 | Integration of multiple-omics datasets and potential outputs using the mixOmics R package. Reproduced under a Creative Commons CC BY 4.0 license from Rohart et al. (151).

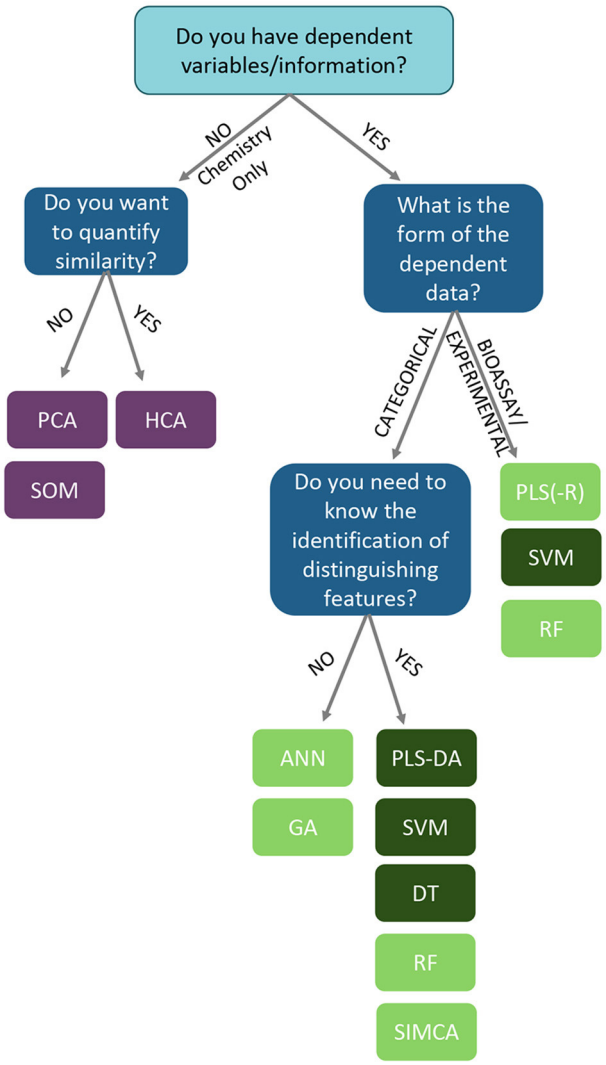

FIGURE 14 | A decision tree to establish which analysis is more appropriate to analyzing complex chemical data. Based upon the presence of response data, and how the overall analysis needs to be structured/interpreted. ANN, artificial neural networks; DT, decision trees; GA, genomic algorithms; HCA, hierarchical clustering analysis; PCA, principal component analysis; PLS, partial least squares; PLS-DA, partial least squares-discriminate analysis; RF, random forests; SIMCA, soft independent modeling of class analogies; SOM, self-organized maps; SVM, support vector machines. Light green boxes represent "soft" classification techniques. Dark green boxes represent "hard" classification techniques.

similarities between samples without a hierarchy, principal component analysis (PCA), self-organizing maps (SOMs), and k-means clustering are viable options.

For experimental sets which contain dependent variables, chemometric options include numerous supervised analyses, which require response or dependent variables to train models. Generally, an unsupervised approach (PCA) should be applied to the metabolomics data set to ensure clustering occurs without predefined categories before delving into supervised analysis. Within the supervised approaches, the chemometric options vary depending on whether the response data are categorical or numerical in nature. Categorical dependent data, such as class assignments (e.g., "authentic" and "unknown") enable supervised analysis to generate models that maximizes differences between the two classes. When choosing a classification methodology, one can consider whether the particular chemometric approach is "soft" or "hard." These designations relate to a method's rigidness in assigning an unknown to a particular class. A "soft" classification rule estimates the probability associated with each class and subsequently provides a class prediction based on the largest estimated probability. In comparison, "hard" classification delivers a final class prediction without probabilistic reasoning behind the classification. Of the reviewed approaches, SIMCA, PLS, random forest (RF), genomic algorithms (GA), and artificial neural networks (ANN) are generally considered "soft" computational approaches (**Figure 14**, light green boxes) (152), while other techniques, such as PLS-DA, decision trees (DT), and support vector machines (SVM) (153) are "hard" methodologies (**Figure 14**, dark green boxes).

At this point, the last decision is the degree of interpretability the model will have for the researcher. A highly interpretable algorithm means that one can easily understand how any individual predictor (variable) is associated with the response, so it's easier to relate the final classification structure back to specific variables contributing to model responses. Techniques like partial least squares-discriminant analysis (PLS-DA), support vector machines (SVM), decision trees (DT), soft independent modeling of class analogies (SIMCA), and random forests (RF) are able to provide interpretable models. If interpretation is not essential (a "black box" approach), and only the final classification of the data is important, models like artificial neural networks (ANN) or genetic algorithms (GA)are prime options.

Numeric dependent variables are frequently obtained from biological activity experiments, and thus enable the use of prediction algorithms to correlate the dependent variable with variations in the chemical information. For the biochemometric analysis of these orthogonal datasets, partial least squares approaches (PLS, PLS-R) are most common in teasing out these relationships (92). However, newer machine learning approaches, such as SVM and RF, have the ability to provide predictive capabilities and understand relationships with input variables. As an example, Deng et al. employed random forests to provide geographical classification of green teas (which outperformed several other chemometric techniques), but also were able to correlate the geography with several isotopic indicators (103).

As with data-collection, where multiple orthogonal techniques facilitate a greater coverage of the overall chemical composition of the samples, multiple data analysis techniques are often utilized to gain a more comprehensive perspective of the data structure and relationship between samples. It is common to begin with unsupervised approaches (e.g., PCA) to glean a preliminary understanding of how samples are relating to one another, then followed up with supervised or machine learning methods to further classify the samples and obtain information about potential biomarkers or bioactive constituents. Zhang et al., in authenticating berry juices, first used PCA to identify clusters of juices by origin, then followed with PLS-DA to

determine relevant biomarkers (76). PCA and HCA were employed to reveal well-differentiated clusters for black peppers, then followed by supervised PLS-DA for a prediction model for additional unknown samples (154). Thus, merging chemometric methods, when possible, offers researchers a potentially more rigorous analysis of their botanical data, which is essential to draw relevant and robust conclusions about authentication and adulteration questions.

CONCLUSIONS AND FUTURE DIRECTIONS

As the demand for botanical medicines and dietary supplements grows, in terms of relevance to human health as well as economic importance, ensuring reliable determination of starting materials for research, safety, and production considerations remains a challenge. Plant-based formulations pose a particularly unique hurdle due to their inherent chemical complexity as well as their variability. Non-targeted chemical fingerprinting techniques, including metabolomics, hold immense potential for describing the chemical composition of botanicals. However, organizing that highly complex information and deducing relevant conclusions from it can represent a major obstacle for researchers. This review has sought to address this hurdle by presenting examples of major chemometric techniques that can be employed to distill complex chemical data into models for authentication and classification of unknown samples. The adaptation of statistical models to wrangle large, complex datasets represents a significant advancement in modeling botanical chemical data. While the chemometric analysis methods profiled in this review are the most common, and some of the most powerful, approaches in use for botanical authentication, it is by no means an exhaustive list. Other variations of unsupervised and supervised techniques have been reported, and there is considerable research being undertaken to advance the capabilities of these statistical and machine learning approaches. And the combination of complementary methods (e.g., biological data and metabolomics, chemical profiling and genomics, or multi-omics techniques) has the potential to provide even more efficient and robust tools to advance authentication and discovery efforts.

AUTHOR CONTRIBUTIONS

JK and EA conceived, wrote and reviewed the manuscript, and secured funding. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the USDA National Institute of Food and Agriculture and Hatch Appropriations (PEN04772), and the Garden Club of America's Anne S. Chatham Fellowship in Medicinal Botany.

ACKNOWLEDGMENTS

We would like to thank Diane Weatherspoon (orcid.org/0000-0001-6724-4841) for her insightful comments and helpful discussions on the development of the manuscript.

REFERENCES

1. Vogtman H. Dietary supplement usage increases, says new survey. The Council for Responsible Nutrition. Available online at: https://www.crnusa.org/newsroom/dietary-supplement-usage-increases-says-new-survey (accessed July 29, 2021).
2. Smith T, May G, Eckl V, Reynolds CM. US Sales of herbal supplements increase by 8.6% in 2019. *HerbalGram.* (2019) 127:54–69. Available online at: http://cms.herbalgram.org/herbalgram/issue127/hg127-mktrpt-2019.html
3. Li C, Hansen RA, Chou C, Calderón AI, Qian J. Trends in botanical dietary supplement use among US adults by cancer status: The National Health and Nutrition Examination Survey, 1999 to 2014. *Cancer.* (2018) 124:1207–15. doi: 10.1002/cncr.31183
4. Sun Y, Wang R, Tang W, Li C, Huo N. Trends and factors of botanical dietary supplement use among US adults with COPD from 1999 to 2016. *PLoS ONE.* (2020) 15:e0239674. doi: 10.1371/journal.pone.0239674
5. Kuszak AJ, Hopp DC, Williamson JS, Betz JM, Sorkin BC. Approaches by the US National Institutes of Health to support rigorous scientific research on dietary supplements and natural products. *Drug Test Anal.* (2016) 8:413–7. doi: 10.1002/dta.1931
6. US National Library of Medicine. Clinical Trial Database. Available online at: https://clinicaltrials.gov/ct2/home (accessed July 30, 2021).
7. Towns AM, Quiroz D, Guinee L, de Boer H, van Andel T. Volume, value and floristic diversity of Gabon's medicinal plant markets. *J Ethnopharmacol.* (2014) 155:1184–93. doi: 10.1016/j.jep.2014.06.052
8. Applequist, Wendy. *The Identification of Medicinal Plants: A Handbook of the Morphology of Botanicals in Commerce*. Austin, TX: American Botanical Council (2006). p. 209.
9. Ahmed SN, Ahmad M, Zafar M, Rashid S, Yaseen G, Sultana S, et al. Comparative light and scanning electron microscopy in authentication of adulterated traded medicinal plants. *Microsc Res Tech.* (2019) 82:1174–83. doi: 10.1002/jemt.23266
10. Ayaz A, Zaman W, Ullah F, Saqib S, Jamshed S, Bahadur S, et al. Systematics study through scanning electron microscopy a tool for the authentication of herbal drug *Mentha suaveolens* Ehrh. *Microsc Res Tech.* (2020) 83:81–7. doi: 10.1002/jemt.23391
11. Kan HX, Jin L, Zhou FL. Classification of medicinal plant leaf image based on multi-feature extraction. *Pattern Recognit Image Anal.* (2017) 27:581–7. doi: 10.1134/S105466181703018X
12. Wäldchen J, Mäder P. Machine learning for image based species identification. *Methods Ecol Evol.* (2018) 9:2216–25. doi: 10.1111/2041-210X.13075
13. Selvam ABD. Presence or absence of stone cells in the roots of indian aconites: an aid to identification of species. *AJPRD.* (2018) 6:4. doi: 10.22270/ajprd.v6i3.377
14. de Boer HJ, Ouarghidi A, Martin G, Abbad A, Kool A, DNA. Barcoding reveals limited accuracy of identifications based on folk taxonomy. *PLoS ONE.* (2014) 9:e84291. doi: 10.1371/journal.pone.0084291
15. Ouarghidi A, Martin GJ, Powell B, Esser G, Abbad A. Botanical identification of medicinal roots collected and traded in Morocco and comparison to the existing literature. *J Ethnobiol Ethnomed.* (2013) 9:59. doi: 10.1186/1746-4269-9-59
16. Rewald B, Meinen C, Trockenbrodt M, Ephrath JE, Rachmilevitch S. Root taxa identification in plant mixtures – current techniques and future challenges. *Plant Soil.* (2012) 359:165–82. doi: 10.1007/s11104-012-1164-0

17. Dauncey EA, Irving JTW, Allkin R, A. review of issues of nomenclature and taxonomy of *Hypericum perforatum* L. and Kew's Medicinal Plant Names Services. *J Pharm Pharmacol.* (2018) 71:4–14. doi: 10.1111/jphp.12831
18. Veldman S, Ju Y, Otieno JN, Abihudi S, Posthouwer C, Gravendeel B, et al. DNA barcoding augments conventional methods for identification of medicinal plant species traded at Tanzanian markets. *J Ethnopharmacol.* (2020) 250:112495. doi: 10.1016/j.jep.2019.112495
19. Aboul-Maaty NA-F, Oraby HA-S. Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method. *Bull Natl Res Cent.* (2019) 43:25. doi: 10.1186/s42269-019-0066-1
20. Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, et al. DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnol J.* (2016) 14:8–21. doi: 10.1111/pbi.12419
21. Zhang C, Liu T, Yuan X, Huang H, Yao G, Mo X, et al. The plastid genome and its implications in barcoding specific-chemotypes of the medicinal herb Pogostemon cablin in China. *Chiang T-Y, editor PLoS ONE.* (2019) 14:e0215512. doi: 10.1371/journal.pone.0215512
22. Gao T, Yao H, Song J, Liu C, Zhu Y, Ma X, et al. Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *J Ethnopharmacol.* (2010) 130:116–21. doi: 10.1016/j.jep.2010.04.026
23. Sucher N, Carles M. Genome-based approaches to the authentication of medicinal plants. *Planta Med.* (2008) 74:603–23. doi: 10.1055/s-2008-1074517
24. Techen N, Parveen I, Pan Z, Khan IA. DNA barcoding of medicinal plant material for identification. *Curr Opin Biotechnol.* (2014) 25:103–10. doi: 10.1016/j.copbio.2013.09.010
25. Shinde VM, Dhalwal K, Mahadik KR, Joshi KS, Patwardhan BK, RAPD. Analysis for Determination of Components in Herbal Medicine. *Evidence-Based Complem Alternat Med.* (2007) 4:21–3. doi: 10.1093/ecam/nem109
26. Hadipour M, Kazemitabar SK, Yaghini H. Dayani S. Genetic diversity and species differentiation of medicinal plant Persian Poppy (Papaver bracteatum L) using AFLP and ISSR markers. *Ecol Genet Genom.* (2020) 16:100058. doi: 10.1016/j.egg.2020.100058
27. Besse P. *Molecular Plant Taxonomy: Methods and Protocols*. New York, NY: Springer US (2021). (Methods in Molecular Biology vol. 2222). Available online at: http://link.springer.com/10.1007/978-1-0716-0997-2 (accessed September 10, 2021).
28. Seethapathy GS, Raclariu-Manolica A-C, Anmarkrud JA, Wangensteen H, de Boer HJ. DNA metabarcoding authentication of ayurvedic herbal products on the european market raises concerns of quality and fidelity. *Front Plant Sci.* (2019) 10:68. doi: 10.3389/fpls.2019.00068
29. Raclariu AC, Heinrich M, Ichim MC, de Boer H. Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication: DNA barcoding and metabarcoding in herbal product authentication. *Phytochem Anal.* (2018) 29:123–8. doi: 10.1002/pca.2732
30. Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, et al. Advances in DNA metabarcoding for food and wildlife forensic species identification. *Anal Bioanal Chem.* (2016) 408:4615–30. doi: 10.1007/s00216-016-9595-8
31. EMA E. *Guideline on specifications: test procedures and acceptance criteria for herbal substances, herbal preparations and herbal medicinal products/traditional herbal medicinal products.* (2011) p. 25.
32. Dormontt EE, van Dijk K, Bell KL, Biffin E, Breed MF, Byrne M, et al. Advancing DNA barcoding and metabarcoding applications for plants requires systematic analysis of herbarium collections—an australian perspective. *Front Ecol Evol.* (2018) 6:134. doi: 10.3389/fevo.2018.00134
33. de Boer HJ, Ichim MC, Newmaster SG. DNA barcoding and pharmacovigilance of herbal medicines. *Drug Saf.* (2015) 38:611–20. doi: 10.1007/s40264-015-0306-8
34. Carvalho KR, Souza ASQ, Alves Filho EG, Silva LMA, Silva EO, Rita de Cássia AP, et al. NIR and 1H qNMR methods coupled to chemometrics discriminate the chemotypes of the gastroprotective herb Egletes viscosa. *Food Res Int.* (2020) 138:109759. doi: 10.1016/j.foodres.2020.109759
35. Huang T, Li H, Zhang W, Numata M, Mackay L, Warren J, et al. Advanced approaches and applications of qNMR. *Metrologia.* (2020) 57:014004. doi: 10.1088/1681-7575/ab336b
36. Ohtsuki T, Matsuoka K, Fuji Y, Nishizaki Y, Masumoto N, Sugimoto N, et al. Development of an HPLC method with relative molar sensitivity based on 1H-qNMR to determine acteoside and pedaliin in dried sesame leaf powders and processed foods. *PLoS ONE.* (2020) 15:e0243175. doi: 10.1371/journal.pone.0243175
37. Blumenthal M, Busse WR. *The Complete German Commission E Monographs*. Austin, TX: American Botanical Council. (1998).
38. U. S. Pharmacopeia. Herbal Medicines Compendium [Internet]. Available online at: http://hmc.usp.org/ (accessed August 13, 2021).
39. Tyler VE. *Herbs of Choice: The Therapeutic Use of Phytomedicinals*. Pharmaceutical Products Press (1994). p. 209.
40. Kellogg JJ, Paine MF, McCune JS, Oberlies NH, Cech NB. Selection and characterization of botanical natural products for research studies: a NaPDI center recommended approach. *Nat Prod Rep.* (2019) 36:1196–221. doi: 10.1039/C8NP00065D
41. Chandra A, Li Y, Rana J, Persons K, Hyun C, Shen S, et al. Qualitative categorization of supplement grade Ginkgo biloba leaf extracts for authenticity. *J Funct Foods.* (2011) 3:107–14. doi: 10.1016/j.jff.2011.03.004
42. Gafner S, Blumenthal M, Foster S, Cardellina II JH, Khan IA, Upton R. Botanical ingredient adulteration – how some suppliers attempt to fool commonly used analytical techniques. *Acta Hortic.* (2020) 1287:15–24. doi: 10.17660/ActaHortic.2020.1287.3
43. Baume N, Mahler N, Kamber M, Mangin P, Saugy M. Research of stimulants and anabolic steroids in dietary supplements. *Scandanavian J Med Sci Sports.* (2006) 16:41–8. doi: 10.1111/j.1600-0838.2005.00442.x
44. Geyer H, Parr MK, Koehler K, Mareck U, Schänzer W, Thevis M. Nutritional supplements cross-contaminated and faked with doping substances. *J Mass Spectrom.* (2008) 43:892–902. doi: 10.1002/jms.1452
45. Cohen PA, Travis JC, Keizers PHJ, Deuster P, Venhuis BJ. Four experimental stimulants found in sports and weight loss supplements: 2-amino-6-methylheptane (octodrine), 1,4-dimethylamylamine (1,4-DMAA), 1,3-dimethylamylamine (1,3-DMAA) and 1,3-dimethylbutylamine (1,3-DMBA). *Clin Toxicol.* (2018) 56:421–6. doi: 10.1080/15563650.2017.1398328
46. Lv X, Li Y, Tang C, Zhang Y, Zhang J, Fan G. Integration of HPLC-based fingerprint and quantitative analyses for differentiating botanical species and geographical growing origins of *Rhizoma coptidis*. *Pharm Biol.* (2016) 54:3264–71. doi: 10.1080/13880209.2016.1223699
47. Donno D, Beccaro GL, Mellano MG, Bonvegna L, Bounous G. Castanea spp. buds as a phytochemical source for herbal preparations: botanical fingerprint for nutraceutical identification and functional food standardization. *J Sci Food Agric.* (2014) 94:2863–73. doi: 10.1002/jsfa.6627
48. Parveen A, Wang Y-H, Fantoukh O, Alhusban M, Raman V, Ali Z, et al. Development of a chemical fingerprint as a tool to distinguish closely related Tinospora species and quantitation of marker compounds. *J Pharm Biomed Anal.* (2020) 178:112894. doi: 10.1016/j.jpba.2019.112894
49. Official Methods of Analysis of AOAC International. *Withanolide Glycosides and Aglycones of Ashwagandha (Withania somnifera).* 22nd ed. Gaithersburg, MD, USA: AOAC International (2015).
50. Lee J-E, Lee B-J, Chung J-O, Kim H-N, Kim E-H, Jung S, et al. Metabolomic unveiling of a diverse range of green tea (Camellia sinensis) metabolites dependent on geography. *Food Chem.* (2015) 174:452–9. doi: 10.1016/j.foodchem.2014.11.086
51. Lee J-E, Lee B-J, Chung J-O, Hwang J-A, Lee S-J, Lee C-H, et al. Geographical and climatic dependencies of green tea (Camellia sinensis) metabolites: A 1H NMR-based metabolomics study. *J Agric Food Chem.* (2010) 58:10582–9. doi: 10.1021/jf102415m
52. Yuk J, McIntyre KL, Fischer C, Hicks J, Colson KL, Lui E, et al. Distinguishing Ontario ginseng landraces and ginseng species using NMR-based metabolomics. *Anal Bioanal Chem.* (2013) 405:4499–509. doi: 10.1007/s00216-012-6582-6
53. Avula B, Wang Y-H, Isaac G, Yuk J, Wrona M, Yu K, et al. Metabolic profiling of hoodia, chamomile, terminalia species and evaluation of commercial preparations using ultrahigh-performance liquid chromatography quadrupole-time-of-flight mass spectrometry. *Planta Med.* (2017) 83:1297–308. doi: 10.1055/s-0043-109239
54. Kumar D. Nuclear magnetic resonance (NMR) spectroscopy for metabolic profiling of medicinal plants and their products. *Crit Rev Anal Chem.* (2016) 46:400–12. doi: 10.1080/10408347.2015.1106932

55. Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, et al. The future of NMR-based metabolomics. *Curr Opin Biotechnol.* (2017) 43:34–40. doi: 10.1016/j.copbio.2016.08.001
56. Jorge TF, Rodrigues JA, Caldana C, Schmidt R, van Dongen JT, Thomas-Oates J, et al. Mass spectrometry-based plant metabolomics: metabolite responses to abiotic stress. *Mass Spectrom Rev.* (2016) 35:620–49. doi: 10.1002/mas.21449
57. Beale DJ, Pinu FR, Kouremenos KA, Poojary MM, Narayana VK, Boughton BA, et al. Review of recent developments in GC–MS approaches to metabolomics-based research. *Metabolomics.* (2018) 14:152. doi: 10.1007/s11306-018-1449-2
58. Bracewell-Milnes T, Saso S, Abdalla H, Nikolau D, Norman-Taylor J, Johnson M, et al. Metabolomics as a tool to identify biomarkers to predict and improve outcomes in reproductive medicine: a systematic review. *Hum Reprod Update.* (2017) 23:723–36. doi: 10.1093/humupd/dmx023
59. Considine EC, Thomas G, Boulesteix AL, Khashan AS, Kenny LC. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics.* (2017) 14:7. doi: 10.1007/s11306-017-1299-3
60. Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol Biol.* (2002) 48:155–71. doi: 10.1007/978-94-010-0448-0_11
61. Héberger K. Chapter 7 - Chemoinformatics—multivariate mathematical–statistical methods for data evaluation. In: Vékey K, Telekes A, Vertes A, editors. Medical Applications of Mass Spectrometry. Amsterdam: Elsevier (2008). p. 141–69. Available from: https://www.sciencedirect.com/science/article/pii/B9780444519801500094 (accessed August 28, 2021).
62. Clark TN, Houriet J, Vidar WS, Kellogg JJ, Todd DA, Cech NB, et al. Interlaboratory comparison of untargeted mass spectrometry data uncovers underlying causes for variability. *J Nat Prod.* (2021) 84:824–35. doi: 10.1021/acs.jnatprod.0c01376
63. Kellogg JJ, Graf TN, Paine MF, McCune JS, Kvalheim OM, Oberlies NH, et al. Comparison of metabolomics approaches for evaluating the variability of complex botanical preparations: green tea (camellia sinensis) as a case study. *J Nat Prod.* (2017) 80:1457–66. doi: 10.1021/acs.jnatprod.6b01156
64. Liu Y, Finley J, Betz JM, Brown PN. FT-NIR characterization with chemometric analyses to differentiate goldenseal from common adulterants. *Fitoterapia.* (2018) 127:81–8. doi: 10.1016/j.fitote.2018.02.006
65. Hendriks MMWB, Eeuwijk FA van, Jellema RH, Westerhuis JA, Reijmers TH, Hoefsloot HCJ, et al. Data-processing strategies for metabolomics studies. *TrAC Trend Anal Chem.* (2011) 30:1685–98. doi: 10.1016/j.trac.2011.04.019
66. Yi L, Dong N, Yun Y, Deng B, Ren D, Liu S, et al. Chemometric methods in data processing of mass spectrometry-based metabolomics: a review. *Anal Chim Acta.* (2016) 914:17–34. doi: 10.1016/j.aca.2016.02.001
67. Myers OD, Sumner SJ Li S, Barnes S, Du X. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. *Anal Chem.* (2017) 89:8696–703. doi: 10.1021/acs.analchem.7b00947
68. Li Z, Lu Y, Guo Y, Cao H, Wang Q, Shui W. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Anal Chim Acta.* (2018) 1029:50–7. doi: 10.1016/j.aca.2018.05.001
69. Hemmer S, Manier SK, Fischmann S, Westphal F, Wagmann L, Meyer MR. Comparison of three untargeted data processing workflows for evaluating LC-HRMS metabolomics data. *Metabolites.* (2020) 10:378. doi: 10.3390/metabo10090378
70. Klåvus A, Kokla M, Noerman S, Koistinen VM, Tuomainen M, Zarei I, et al. "Notame": workflow for non-targeted LC–MS metabolic profiling. *Metabolites.* (2020) 10:135. doi: 10.3390/metabo10040135
71. Du X, Smirnov A, Pluskal T, Jia W, Sumner S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. In: Li S, editor. Computational Methods and Data Analysis for Metabolomics. New York, NY: Springer US (2020). p. 25–48. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-0716-0239-3_3 (accessed October 18, 2021).
72. Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, et al. Pretreatment of mass spectral profiles: application to proteomic data. *Anal Chem.* (2007) 79:7014–26. doi: 10.1021/ac070946s
73. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics.* (2006) 7:142. doi: 10.1186/1471-2164-7-142
74. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* (2016) 3:160018. doi: 10.1038/sdata.2016.18
75. Jackson E. *PCA With More Than Two Variables*. In: A Use's Guide to Principal Components. John Wiley & Sons, Ltd. (1991). p. 26–62. Available online at: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471725331.ch2 (accessed August 29, 2021).
76. Zhang J, Yu Q, Cheng H, Ge Y, Liu H, Ye X, et al. Metabolomic approach for the authentication of berry fruit juice by liquid chromatography quadrupole time-of-flight mass spectrometry coupled to chemometrics. *J Agric Food Chem.* (2018) 66:8199–208. doi: 10.1021/acs.jafc.8b01682
77. Farag MA, Labib RM, Noleto C, Porzel A, Wessjohann LA. NMR approach for the authentication of 10 cinnamon spice accessions analyzed via chemometric tools. *LWT.* (2018) 90:491–8. doi: 10.1016/j.lwt.2017.12.069
78. Kellogg JJ, Kvalheim OM, Cech NB. Composite score analysis for unsupervised comparison and network visualization of metabolomics data. *Anal Chim Acta.* (2020) 1095:38–47. doi: 10.1016/j.aca.2019.10.029
79. Wallace ED, Todd DA, Harnly JM, Cech NB, Kellogg JJ. Identification of adulteration in botanical samples with untargeted metabolomics. *Anal Bioanal Chem.* (2020) 412:4273–86. doi: 10.1007/s00216-020-02678-6
80. Luo D, Liu Y, Wang Y, Zhang X, Huang L, Duan B. Rapid identification of Fritillariae Cirrhosae Bulbus and its adulterants by UPLC-ELSD fingerprint combined with chemometrics methods. *Biochem Syst Ecol.* (2018) 76:46–51. doi: 10.1016/j.bse.2017.12.007
81. Zhou S, Allard P-M, Wolfrum C, Ke C, Tang C, Ye Y, et al. Identification of chemotypes in bitter melon by metabolomics: a plant with potential benefit for management of diabetes in traditional Chinese medicine. *Metabolomics.* (2019) 15:104. doi: 10.1007/s11306-019-1565-7
82. Kesanakurti P, Thirugnanasambandam A, Ragupathy S, Newmaster SG. Genome skimming and NMR chemical fingerprinting provide quality assurance biotechnology to validate Sarsaparilla identity and purity. *Sci Rep.* (2020) 10:19192. doi: 10.1038/s41598-020-76073-7
83. Cebi N, Yilmaz MT, Sagdic O, A. rapid ATR-FTIR spectroscopic method for detection of sibutramine adulteration in tea and coffee based on hierarchical cluster and principal component analyses. *Food Chem.* (2017) 229:517–26. doi: 10.1016/j.foodchem.2017.02.072
84. Kalogiouri NP, Aalizadeh R, Dasenaki ME, Thomaidis NS. Application of high resolution mass spectrometric methods coupled with chemometric techniques in olive oil authenticity studies - a review. *Anal Chim Acta.* (2020) 1134:150–73. doi: 10.1016/j.aca.2020.07.029
85. Torrecilla JS, Cancilla JC, Matute G, Díaz-Rodríguez P, Flores AI. Self-organizing maps based on chaotic parameters to detect adulterations of extra virgin olive oil with inferior edible oils. *J Food Eng.* (2013) 118:400–5. doi: 10.1016/j.jfoodeng.2013.04.029
86. Menezes R, Sessions Z, Muratov E, Scotti L, Scotti M. Secondary Metabolites Extracted from Annonaceae and Chemotaxonomy Study of Terpenoids. *J Braz Chem Soc.* (2021) 32:2061–70. doi: 10.21577/0103-5053.20210097
87. Barbosa Silva Cavalcanti A, Costa Barros RP, Costa VC de O, Sobral da Silva M, Fechine Tavares J, Scotti L, et al. Computer-aided chemotaxonomy and bioprospecting study of diterpenes of the lamiaceae family. *Molecules.* (2019) 24:3908. doi: 10.3390/molecules24213908
88. Cavalcanti É, Scotti M, Scotti L, Emereciano V. Application of Self-Organizing Maps generated from Molecular Descriptors of Flavonoid in the Chemotaxonomy of the Asteraceae Family. In: Proceedings of MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition. Sciforum.net: MDPI (2017). p. 5063. Available online at: http://sciforum.net/conference/mol2net-03/paper/5063 (accessed September 15, 2021).
89. Brereton RG, A. short history of chemometrics: a personal view. *J Chemom.* (2014) 28:749–60. doi: 10.1002/cem.2633
90. Kjeldahl K, Bro R. Some common misunderstandings in chemometrics. *J Chemometrics.* (2010) 24:558–64. doi: 10.1002/cem.1346
91. Marshall DD, Powers R. Beyond the paradigm: combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Prog Nucl Magn Reson Spectrosc.* (2017) 100:1–16. doi: 10.1016/j.pnmrs.2017.01.001

92. Kellogg JJ, Todd DA, Egan JM, Raja HA, Oberlies NH, Kvalheim OM, et al. Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds. *J Nat Prod.* (2016) 79:376–86. doi: 10.1021/acs.jnatprod.5b01014
93. Ismail S, Maulidiani M, Akhtar M, Abas F, Ismail I, Khatib A, et al. Discriminative analysis of different grades of gaharu (Aquilaria malaccensis Lamk) via 1H-NMR-based metabolomics using PLS-DA and random forests classification models. *Molecules.* (2017) 22:1612. doi: 10.3390/molecules22101612
94. Windarsih A, Wijayanti T, Irnawati I, Rohman A. The use of 1H-NMR spectroscopy coupled with chemometrics for authentication of curcuma Xanthorrhiza adulterated with curcuma aeruginosa. *Key Eng Mater.* (2021) 884:320–6. doi: 10.4028/www.scientific.net/KEM.884.320
95. Barbosa S, Campmajó G, Saurina J, Puignou L, Núñez O. Determination of phenolic compounds in paprika by ultrahigh performance liquid chromatography–tandem mass spectrometry: application to product designation of origin authentication by chemometrics. *J Agric Food Chem.* (2020) 68:591–602. doi: 10.1021/acs.jafc.9b06054
96. Rácz A, Gere A, Bajusz D, Héberger K. Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition? *RSC Adv.* (2018) 8:10–21. doi: 10.1039/C7RA08901E
97. Martín-Torres S, Jiménez-Carvelo AM, González-Casado A, Cuadros-Rodríguez L. Differentiation of avocados according to their botanical variety using liquid chromatographic fingerprinting and multivariate classification tree. *J Sci Food Agric.* (2019) 99:4932–41. doi: 10.1002/jsfa.9725
98. Little JG, Marsman DS, Baker TR, Mahony C. In silico approach to safety of botanical dietary supplement ingredients utilizing constituent-level characterization. *Food Chem Toxicol.* (2017) 107:418–29. doi: 10.1016/j.fct.2017.07.017
99. Breiman L. Random Forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
100. Vigneau E, Courcoux P, Symoneaux R, Guérin L, Villière A. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. *Food Qual Prefer.* (2018) 68:135–45. doi: 10.1016/j.foodqual.2018.02.008
101. Deklerck V, Finch K, Gasson P, Bulcke JV den, Acker JV, Beeckman H, et al. Comparison of species classification models of mass spectrometry data: kernel discriminant analysis vs. random forest a case study of Afrormosia (Pericopsis elata (Harms) Meeuwen). *Rapid Communications in Mass Spectrometry.* (2017) 31:1582–8. doi: 10.1002/rcm.7939
102. Hou L, Liu Y, Wei A. Geographical variations in the fatty acids of Zanthoxylum seed oils: a chemometric classification based on the random forest algorithm. *Ind Crops Prod.* (2019) 134:146–53. doi: 10.1016/j.indcrop.2019.03.070
103. Deng X. Predictive geographical authentication of green tea with protected designation of origin using a random forest model. *Food Control.* (2020) 107:106807. doi: 10.1016/j.foodcont.2019.106807
104. Martín-Torres S, Jiménez-Carvelo AM, González-Casado A, Cuadros-Rodríguez L. Authentication of the geographical origin and the botanical variety of avocados using liquid chromatography fingerprinting and deep learning methods. *Chemometr Intell Lab Syst.* (2020) 199:103960. doi: 10.1016/j.chemolab.2020.103960
105. Wu X-M, Zuo Z-T, Zhang Q-Z, Wang Y-Z. Classification of Paris species according to botanical and geographical origins based on spectroscopic, chromatographic, conventional chemometric analysis and data fusion strategy. *Microchem J.* (2018) 143:367–78. doi: 10.1016/j.microc.2018.08.035
106. Pan H, Yao C, Yao S, Yang W, Wu W, Guo D, et al. metabolomics strategy for authentication of plant medicines with multiple botanical origins, a case study of Uncariae Rammulus Cum Uncis. *J Sep Sci.* (2020) 43:1043–50. doi: 10.1002/jssc.201901064
107. Wang Y, Zuo Z-T, Huang H-Y, Wang Y-Z. Original plant traceability of Dendrobium species using multi-spectroscopy fusion and mathematical models. *R Soc Open Sci.* (2019) 6:190399. doi: 10.1098/rsos.190399
108. Elansary HO, Mahmoud EA. Basil cultivar identification using chemotyping still favored over genotyping using core barcodes and possible resources of antioxidants. *Null.* (2015) 27:82–7. doi: 10.1080/10412905.2014.982874
109. Goodacre R. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *J Exp Bot.* (2004) 56:245–54. doi: 10.1093/jxb/eri043
110. Gil M, Reynes C, Cazals G, Enjalbal C, Sabatier R, Saucier C. Discrimination of rosé wines using shotgun metabolomics with a genetic algorithm and MS ion intensity ratios. *Sci Rep.* (2020) 10:1170. doi: 10.1038/s41598-020-58193-2
111. Cavill R, Keun HC, Holmes E, Lindon JC, Nicholson JK, Ebbels TMD. Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics.* (2009) 25:112–8. doi: 10.1093/bioinformatics/btn586
112. Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: applications of deep learning in metabolomics. *Comput Struct Biotechnol J.* (2020) 18:2818–25. doi: 10.1016/j.csbj.2020.09.033
113. Mendez KM, Broadhurst DI, Reinke SN. The application of artificial neural networks in metabolomics: a historical perspective. *Metabolomics.* (2019) 15:142. doi: 10.1007/s11306-019-1608-0
114. Binetti G, Del Coco L, Ragone R, Zelasco S, Perri E, Montemurro C, et al. Cultivar classification of Apulian olive oils: Use of artificial neural networks for comparing NMR, NIR and merceological data. *Food Chem.* (2017) 219:131–8. doi: 10.1016/j.foodchem.2016.09.041
115. Cajka T, Hajslova J, Pudil F, Riddellova K. Traceability of honey origin based on volatiles pattern processing by artificial neural networks. *J Chromatography A.* (2009) 1216:1458–62. doi: 10.1016/j.chroma.2008.12.066
116. Tušek AJ, Jurina T, Benković M, Valinger D, Belščak-Cvitanović A, Kljusurić JG. Application of multivariate regression and artificial neural network modelling for prediction of physical and chemical properties of medicinal plants aqueous extracts. *J Appl Res Med Aromat Plants.* (2020) 16:100229. doi: 10.1016/j.jarmap.2019.100229
117. Gika HG, Theodoridis GA, Plumb RS, Wilson ID. Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. *J Pharm Biomed Anal.* (2014) 87:12–25. doi: 10.1016/j.jpba.2013.06.032
118. Spiteri M, Dubin E, Cotton J, Poirel M, Corman B, Jamin E, et al. Data fusion between high resolution 1H-NMR and mass spectrometry: a synergetic approach to honey botanical origin characterization. *Anal Bioanal Chem.* (2016) 408:4389–401. doi: 10.1007/s00216-016-9538-4
119. Karioti A, Giocaliere E, Guccione C, Pieraccini G, Gallo E, Vannacci A, et al. Combined HPLC-DAD–MS, HPLC–MSn and NMR spectroscopy for quality control of plant extracts: the case of a commercial blend sold as dietary supplement. *J Pharm Biomed Anal.* (2014) 88:7–15. doi: 10.1016/j.jpba.2013.07.040
120. Deconinck E, De Leersnijder C, Custers D, Courselle P, De Beer JO. A strategy for the identification of plants in illegal pharmaceutical preparations and food supplements using chromatographic fingerprints. *Anal Bioanal Chem.* (2013) 405:2341–52. doi: 10.1007/s00216-012-6649-4
121. Deconinck E, Vanhamme M, Bothy JL, Courselle P. A strategy based on fingerprinting and chemometrics for the detection of regulated plants in plant food supplements from the Belgian market: two case studies. *J Pharm Biomed Anal.* (2019) 166:189–96. doi: 10.1016/j.jpba.2019.01.015
122. Zhou Y, Zuo Z, Xu F, Wang Y. Origin identification of Panax notoginseng by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochim Acta Part A.* (2020) 226:117619. doi: 10.1016/j.saa.2019.117619
123. Calderon AI. Editorial: Combination of mass spectrometry and omics/chemometrics approaches to unravel bioactives in natural products mixtures. *Combinator Chem High Throughput Screen.* (2017) 20:278. doi: 10.2174/1386207320041708111113805
124. Roberts GK, Gardner D, Foster PM, Howard PC, Lui E, Walker L, et al. Finding the bad actor: challenges in identifying toxic constituents in botanical dietary supplements. *Food Chem Toxicol.* (2019) 124:431–8. doi: 10.1016/j.fct.2018.12.026
125. Patras A, Brunton NP, Downey G, Rawson A, Warriner K, Gernigon G. Application of principal component and hierarchical cluster analysis to classify fruits and vegetables commonly consumed in Ireland based on in vitro antioxidant activity. *J Food Composit Anal.* (2011) 24:250–6. doi: 10.1016/j.jfca.2010.09.012

126. Britton ER, Kellogg JJ, Kvalheim OM, Cech NB. Biochemometrics to identify synergists and additives from botanical medicines: a case study with hydrastis canadensis (Goldenseal). *J Nat Prod.* (2018) 81:484–93. doi: 10.1021/acs.jnatprod.7b00654
127. Kvalheim OM, Chan H, yan Benzie IFF, Szeto Y tong, Tzang AH chung, Mok DK wah, et al. Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products. *Chemometr Intell Lab Syst.* (2011) 107:98–105. doi: 10.1016/j.chemolab.2011.02.002
128. Alvarez-Zapata R, Sánchez-Medina A, Chan-Bacab M, García-Sosa K, Escalante-Erosa F, García-Rodríguez RV, et al. Chemometrics-enhanced high performance liquid chromatography-ultraviolet detection of bioactive metabolites from phytochemically unknown plants. *J Chromatography A.* (2015) 1422:213–21. doi: 10.1016/j.chroma.2015.10.026
129. Chagas-Paula D, Zhang T, Da Costa F, Edrada-Ebel R. A metabolomic approach to target compounds from the asteraceae family for dual COX and LOX inhibition. *Metabolites.* (2015) 5:404–30. doi: 10.3390/metabo5030404
130. Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J Chemom.* (2015) 29:528–36. doi: 10.1002/cem.2736
131. Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometr Intell Lab Syst.* (2009) 95:35–48. doi: 10.1016/j.chemolab.2008.08.004
132. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal Chem.* (2009) 81:2581–90. doi: 10.1021/ac802514y
133. Kim M-O, Lee SU, Yuk HJ, Jang H-J, Lee J-W, Kwon E-B, et al. Metabolomics approach to identify the active substances influencing the antidiabetic activity of Lagerstroemia species. *J Funct Foods.* (2020) 64:103684. doi: 10.1016/j.jff.2019.103684
134. Kiran KR, Swathy PS, Paul B, Shama Prasada K, Radhakrishna Rao M, Joshi MB, et al. Untargeted metabolomics and DNA barcoding for discrimination of Phyllanthus species. *J Ethnopharmacol.* (2021) 273:113928. doi: 10.1016/j.jep.2021.113928
135. Li Z, Du Y, Yuan Y, Zhang X, Wang Z, Tian X. Integrated quality evaluation strategy for multi-species resourced herb medicine of Qinjiao by metabolomics analysis and genetic comparation. *Chin Med.* (2020) 15:16. doi: 10.1186/s13020-020-0292-3
136. Zeliou K, Koui E-M, Papaioannou C, Koulakiotis NS, Iatrou G, Tsarbopoulos A, et al. Metabolomic fingerprinting and genetic discrimination of four Hypericum taxa from Greece. *Phytochemistry.* (2020) 174:112290. doi: 10.1016/j.phytochem.2020.112290
137. Bielecka M, Pencakowski B, Stafiniak M, Jakubowski K, Rahimmalek M, Gharibi S, et al. Metabolomics and DNA-based authentication of two traditional asian medicinal and aromatic species of salvia subg. *Perovskia Cells.* (2021) 10:112. doi: 10.3390/cells10010112
138. Brunáková K, Bálintová M, Henzelyová J, Kolarčik V, Kimáková A, Petijová L, et al. Phytochemical profiling of several Hypericum species identified using genetic markers. *Phytochemistry.* (2021) 187:112742. doi: 10.1016/j.phytochem.2021.112742
139. Simmler C, Anderson JR, Gauthier L, Lankin DC, McAlpine JB, Chen S-N, et al. Metabolite profiling and classification of DNA-authenticated licorice botanicals. *J Nat Prod.* (2015) 78:2007–22. doi: 10.1021/acs.jnatprod.5b00342
140. Handy SM, Pawar RS, Ottesen AR, Ramachandran P, Sagi S, Zhang N, et al. HPLC-UV, metabarcoding and genome skims of botanical dietary supplements: a case study in echinacea. *Planta Med.* (2021) 87:314–24. doi: 10.1055/a-1336-1685
141. Li S-Z, Zeng S-L, Wu Y, Zheng G-D, Chu C, Yin Q, et al. Cultivar differentiation of Citri Reticulatae Pericarpium by a combination of hierarchical three-step filtering metabolomics analysis, DNA barcoding and electronic nose. *Anal Chim Acta.* (2019) 1056:62–9. doi: 10.1016/j.aca.2019.01.004
142. Sultana S, Khan MA, Ahmad M, Bano A, Zafar M, Shinwari ZK. Authentication of herbal medicine neem (azadirachta indica ajuss) By using taxonomic and pharmacognostic techniques Pakistan. *J Botany.* (2011) 43:141–50.
143. Joshi VC, Avula B, Khan IA. Authentication of Stephania tetrandra S. Moore (Fang Ji) and differentiation of its common adulterants using microscopy and HPLC analysis. *J Nat Med.* (2007) 62:117–21. doi: 10.1007/s11418-007-0200-5
144. Soares S, Grazina L, Costa J, Amaral JS, Oliveira MBPP. Mafra I. Botanical authentication of lavender (Lavandula spp) honey by a novel DNA-barcoding approach coupled to high resolution melting analysis. *Food Control.* (2018) 86:367–73. doi: 10.1016/j.foodcont.2017.11.046
145. Kakouri E, Revelou P-K, Kanakis C, Daferera D, Pappas CS, Tarantilis PA. Authentication of the botanical and geographical origin and detection of adulteration of olive oil using gas chromatography, infrared and raman spectroscopy techniques: a review. *Foods.* (2021) 10:1565. doi: 10.3390/foods10071565
146. Harnly J, Chen P, Sun J, Huang H, Colson K, Yuk J, et al. Comparison of flow injection MS, NMR, and DNA sequencing: methods for identification and authentication of black cohosh (Actaea racemosa). *Planta Med.* (2015) 82:250–62. doi: 10.1055/s-0035-1558113
147. Lim DK, Long NP, Mo C, Dong Z, Cui L, Kim G, et al. Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice. *Food Res Int.* (2017) 100:814–21. doi: 10.1016/j.foodres.2017.08.006
148. Anagbogu CF, Zhou J, Olasupo FO, Baba Nitsa M, Beckles DM. Lipidomic and metabolomic profiles of Coffea canephora L. beans cultivated in Southwestern Nigeria. *PLoS ONE.* (2021) 16:e0234758. doi: 10.1371/journal.pone.0234758
149. Zhao N, Cheng M, Lv W, Wu Y, Liu D, Zhang X. Peptides as potential biomarkers for authentication of mountain-cultivated ginseng and cultivated ginseng of different ages using UPLC-HRMS. *J Agric Food Chem.* (2020) 68:2263–75. doi: 10.1021/acs.jafc.9b05568
150. Zhang S, Lai X, Li B, Wu C, Wang S, Chen X, et al. Application of differential proteomic analysis to authenticate ophiocordyceps sinensis. *Curr Microbiol.* (2015) 72:337–43. doi: 10.1007/s00284-015-0950-3
151. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752
152. Ibrahim D. An overview of soft computing. *Procedia Comput Sci.* (2016) 102:34–8. doi: 10.1016/j.procs.2016.09.366
153. Liu Y, Zhang HH, Wu Y. Hard or soft classification? Large-margin unified machines. *J Am Stat Assoc.* (2011) 106:166–77. doi: 10.1198/jasa.2011.tm10319
154. Rivera-Pérez A, Romero-González R, Garrido Frenich A. Feasibility of applying untargeted metabolomics with GC-Orbitrap-HRMS and chemometrics for authentication of black pepper (Piper nigrum L) and identification of geographical and processing markers. *J Agric Food Chem.* (2021) 69:5547–58. doi: 10.1021/acs.jafc.1c01515

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.