Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-025-05735-0

RESEARCH PAPER

Application of predictive modeling tools for the identification of *Ocimum* spp. herbal products

Evelyn J. Abraham[1] · Sarah J. Chamberlain[2,3] · Wilmer H. Perera[4] · R. Teal Jordan[5] · Joshua J. Kellogg[1,5]

Received: 9 February 2024 / Revised: 20 December 2024 / Accepted: 7 January 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2025

Abstract

Species identification of botanical products is a crucial aspect of research and regulatory compliance; however, botanical classification can be difficult, especially for morphologically similar species with overlapping genetic and metabolomic markers, like those in the genus *Ocimum*. Untargeted LC–MS metabolomics coupled with multivariate predictive modeling provides a potential avenue for improving herbal identity investigations, but the current dearth of reference materials for many botanicals limits the applicability of these approaches. This study investigated the potential of using greenhouse-grown authentic *Ocimum* to build predictive models for classifying commercially available *Ocimum* products. We found that three species, *O. tenuiflorum*, *O. gratissimum*, and *O. basilicum*, were chemically distinct based on their untargeted UPLC-MS/MS profiles when grown in controlled settings; combined with an orthogonal high-performance thin-layer chromatography (HPTLC) approach, *O. tenuiflorum* materials revealed two distinct chemotypes which could confound analysis. Three predictive models (partial least squares, LASSO regression, and random forest) were employed to extrapolate these findings to commercially available products; however, the controlled materials were significantly different from external samples, and all three chemometric models were unreliable in classifying external materials. LASSO was the most successful when classifying new greenhouse samples. Overall, this study highlights how growing and processing conditions can influence the complexity of botanical metabolome profiles; further studies are needed to characterize the factors driving herbal products' phytochemistry in conjunction with chemometric predictive modeling.

Keywords Chemometrics · Natural products · Botanical identity · Metabolomics · Predictive modeling

Introduction

In 1994, Congress passed the Dietary Supplement Health and Education Act (DSHEA), which required the US Food and Drug Administration to set clear good manufacturing processes (GMP) regulations for the dietary supplement industry. Currently, 21 CFR 111 sets guidelines for supplement testing and regulations, including meeting specifications for product identity, purity, strength, composition, and limiting contaminants. While each of these categories is crucial for a safe marketplace, understanding and defining proper specifications and testing approaches is quite complex, especially regarding herbal supplements, where the chemical complexity of botanicals makes identification convoluted. Three primary techniques used in the industry for botanical identification include morphology, genetics (both whole genome and PCR-based approaches), and analytical chemistry.

A major complication in herbal identity studies is that closely related species can have very similar morphological and genetic markers but variations in therapeutic properties. Most identification approaches compare a sample to a limited number of reference standards meant to represent the characteristics of the entire species. However, this fails to incorporate the intraspecies variation that results from differences in chemotypes, environmental factors, and processing

✉ Joshua J. Kellogg
jjk6146@psu.edu

1 Intercollege Graduate Degree Program in Plant Biology, Pennsylvania State University, University Park, PA, USA

2 Department of Biology, Pennsylvania State University, University Park, PA, USA

3 PAC Herbarium, Pennsylvania State University, University Park, PA, USA

4 CAMAG Scientific, Inc., Wilmington, NC, USA

5 Department of Veterinary and Biomedical Sciences, Pennsylvania State University, University Park, PA, USA

Published online: 20 January 2025

practices [1]. For example, highbush blueberries harvested at different altitudes have different anthocyanin profiles [2], and green tea leaves have differential metabolomes based on season, geography, and harvest time [3]. In these cases, a single reference standard would not incorporate all potential chemical variations [2]. And since herbal products' therapeutic benefits are a result of their unique metabolite compositions, species labels may not fully incorporate these distinctions.

Recent developments in mass spectrometry provide a potential avenue to improve this limitation. Improvements in MS instrumentation yield analyses with high resolution and mass accuracy, yielding detection and relative abundance determination for thousands of compounds in a plant sample. This shifts the analytical approach, moving from reliance on a single marker compound or small subset of compounds to examining a sample's entire metabolome. These non-targeted metabolite profiles provide a more extensive look into plants' chemical similarities and relationships [4]; these methods, combined with multivariate statistical approaches, have been successful in identifying botanical adulteration [5].

However, translation of these approaches beyond the lab and into commercial workflows faces several hurdles. The first is that introducing more metabolite information does not overcome the lack of chemical diversity that limited reference standards provide; where they exist, many species are represented by one or two reference materials, and those available often lack appropriate voucher specimens and identifying information [1]. Recent efforts by the NIH Office of Dietary Supplements have expanded reference material availability, but it is still infeasible to generate numerous standards for every commercially viable herb [6]. For successful predictive modeling, a study must incorporate a range of reliably identified samples, especially for plants with many cultivars or closely related species.

Metabolomics data requires the application of an appropriate chemometric tool to robustly interpret the data. Linear regression models are among the simplest and most widespread approaches. Perhaps the most common supervised linear model used in herbal product studies to combat the small sample size limitation is partial least squares (PLS). PLS models reduce large datasets to smaller components that encompass the covariance and correlation of features, similar to principal component analysis (PCA). Unlike PCA, dependent variables guide the data reduction in PLS so that the covariance of sample groups is maximized. New samples can then be added to the model to determine their classification based on the similarities in metabolite profiles [7]. The use of PLS-discriminant analysis (PLS-DA) in herbal authentication studies has been thoroughly reviewed [8–10]; however, linear models like PLS tend to overfit data containing more variables than samples (as is the case with untargeted metabolomics) [5, 11, 12]. To mitigate the limitations of small sample numbers is to add a penalty parameter to the linear model, conservatively biasing model parameters towards zero [13]. Least absolute shrinkage and selection operator (LASSO) regression is a form of penalization particularly well suited for producing sparse, simple models: in the present context, it encourages models where only a small number of metabolites are used to predict the outcome [14]. Studies exemplify LASSO as a preferable method for variable selection over other data reduction techniques, like PLS-DA, for reducing overfitting and model complexity [15]. However, Zhu et al. (2017) reported that LASSO models had lower prediction accuracies than similar penalized regression techniques like elastic net and weighted fusion when classifying *Ganoderma* species [16]. Non-linear models, like random forest (RF), are also gaining popularity as herbal product authentication tools. RF is an ensemble approach that combines the predictive classifications of multiple decision trees. At each node of a decision tree, a different metabolite is considered. Samples are separated at each node based on the abundance or presence of the specific metabolite, and each sample moves down the tree's nodes until a final classification is made [17]. RF uses the principle of majority voting; the most common prediction from all decision trees is the final classification [17].

Thus far, few studies have directly evaluated the performance of untargeted metabolomics for herbal product identification studies for their ability to predict and model the identity of external, consumer-available products. There is a lack of reference materials for most herbal species, and acquiring enough samples with reliable identifications is complicated. To this end, this study used a controlled greenhouse experiment to incorporate inter- and intra-species variation into the dataset while ensuring a proper sample size for modeling is met. We grew thirty-one varieties of *Ocimum*, or basil/Tulsi, belonging to three species, *Ocimum gratissimum* L. (OG), *Ocimum tenuiflorum* L. (OT), and *Ocimum basilicum* L. (OB), in a greenhouse. These materials were ordered based on the claimed identity on the seed distributors' websites, and subsequent morphological identification was performed on the seeds. Furthermore, samples were also analyzed using high-performance thin-layer chromatography (HPTLC). HPTLC is one of the most popular analytical techniques in compendial methods and regulatory settings for botanical identity testing but can also be applied as a powerful quantitative tool. Herein, we compared the ability of HPTLC and untargeted metabolite profiling to distinguish the materials based on claimed species and morphological identity, ultimately identifying 4 distinct chemotypes across three species of *Ocimum* that did not necessarily align with the species. However, despite the chemotype designation, the greenhouse-grown materials were not translatable to predict the classification of external materials.

Experimental

Materials

All solvents and chemicals used, unless otherwise noted, were of reagent or spectroscopic grade, as required, and obtained from VWR (Radnor, PA, USA) or Sigma-Aldrich (St. Louis, MO, USA).

Greenhouse growth

Thirty-one varieties of *Ocimum*, belonging to three species, *O. basilicium*, *O. tenuiflorum*, and *O. gratissimum*, were ordered from thirteen sources and stored at 4 °C until grown in a greenhouse under controlled environmental conditions (Table 1). Seeds were sowed in one-inch germination squares with common potting soil. Four seeds were planted per square, and two plants were allowed to grow following germination. Each variety had a total of five one-inch squares for germination. Seedlings were held in a walk-in growth chamber at 24 °C and watered every 3 days or when the top 0.5 inch of soil was dry. Once seedlings reached three internodes, one of the two plants was transferred into a four-inch pot with the same common potting soil mixed with Miracle-Gro water-soluble fertilizer (The Scotts Company, Marysville, OH, USA) and moved to a greenhouse. The greenhouse was maintained at 24 °C with an 8:16 light:dark cycle and watered every other day. In total, five plants of each variety were grown until harvest. The entire growth cycle was repeated for three rounds, resulting in 3 biological

Table 1 Greenhouse-grown *Ocimum* extraction information. [†]Morphological identification provided from seed taxonomic analysis; missing values indicate no reserve seeds available for taxonomic verification (for images, see Table S2)

Sample name	Labeled species	Morphological ID[†]	Variety	Average extraction yield ± SD (%)
BRS_G	*O. gratissimum*		West African Wild Clove	14.49 ± 0.69
CP_G	*O. gratissimum*	*O. gratissimum*	n/a	11.3 ± 2.2
CP_S	*O. tenuiflorum*	*O. tenuiflorum*	Purple Tulsi	11.6 ± 5.1
FS_G	*O. gratissimum*	*O. gratissimum*	n/a	13.9 ± 1.3
FS_S	*O. gratissimum*	*O. basilicum*	n/a	15.3 ± 2.8
JS_G	*O. basilicum*	*O. basilicum*	Genovese	14.3 ± 2.4
JS_H	*O. tenuiflorum*	*O. tenuiflorum*	n/a	11.3 ± 5.0
JS_R	*O. basilicum*	*O. basilicum*	Red Rubin	12.3 ± 6.4
JS_S	*O. basilicum*	*O. basilicum*	Sweet Thai	14.0 ± 2.4
LSC_T	*O. gratissimum*		n/a	15.9 ± 3.1
PGS_A	*O. basilicum*	*O. basilicum*	Amethyst Purple	17.2 ± 1.2
PGS_G	*O. basilicum*	*O. basilicum*	Genovese	12.5 ± 2.0
PGS_K	*O. tenuiflorum*	*O. basilicum*	Kapoor	15.0 ± 1.4
PGS_R	*O. tenuiflorum*	*O. basilicum*	Rama	15.3 ± 1.7
PGS_T	*O. basilicum*	*O. basilicum*	Thai	11.2 ± 1.5
PGS_V	*O. gratissimum*	*O. gratissimum*	Vana	9.3 ± 3.4
SE_K	*O. tenuiflorum*	*O. basilicum*	Kapoor	13.1 ± 2.6
SMS_A	*O. tenuiflorum*	*O. tenuiflorum*	Amrita	14.6 ± 1.4
SMS_G	*O. basilicum*	*O. basilicum*	Genovese	15.75 ± 0.98
SMS_K	*O. tenuiflorum*		Krishna	10.4 ± 6.4
SMS_M	*O. basilicum*	*O. basilicum*	Mriani	12.8 ± 1.8
SMS_R	*O. tenuiflorum*	*O. tenuiflorum*	Rama	15.1 ± 4.3
SMS_S	*O. basilicum*	*O. basilicum*	Sweet lettuce leaf	10.4 ± 3.0
SMS_T	*O. basilicum*	*O. basilicum*	Thai	13.9 ± 4.1
SMS_V	*O. gratissimum*	*O. gratissimum*	Vana	17.4 ± 3.9
SN_S	*O. tenuiflorum*	*O. tenuiflorum*	n/a	11.5 ± 2.1
STS_K	*O. tenuiflorum*	*O. basilicum*	Kapoor	15.4 ± 3.0
TLM_H	*O. tenuiflorum*	*O. tenuiflorum*	Holy green leaf	12.7 ± 3.2
TLS_K	*O. tenuiflorum*	*O. basilicum*	Kapoor	14.9 ± 1.8
TLS_V	*O. gratissimum*	*O. gratissimum*	Vana	10.5 ± 1.4
TS_G	*O. tenuiflorum*	*O. tenuiflorum*	Holy green leaf	16.2 ± 3.0

replicates (separate greenhouse growth times), with five technical replicates (same greenhouse growth time) each.

Once the plants showed their first inflorescence, but before full flowering, the leaves were harvested. All leaves from each plant were cut using sterile scissors at the base and flash frozen in liquid nitrogen before transfer to a − 80 °C freezer for storage. In the final greenhouse replicate, an extra plant was grown of each variety. This plant was morphologically identified as *Ocimum* and used as a voucher specimen. All voucher specimens can be found at the Pennsylvania State University Herbarium, or online from the Mid-Atlantic Herbaria Consortium (record numbers 30760376–30760406).

Morphological identification

We used seed characteristics based on the work of Patel et al. (2015) [18] as well as the formation of mucilage to confirm seed identity [19]. To identify seeds, 3–5 seeds were removed from each seed packet. Seeds were examined under a stereomicroscope to record shape, color, and texture. Seeds were then placed in a Petri dish filled with tap water for a minimum of 5 min and observed for the presence of a mucilaginous coat around the testa. Seeds were identified via the following criteria:

- *Ocimum basilicum*: ellipsoid, black seeds with pitted surface texture, forming a thick, cloudy mucilaginous coat in the presence of water.
- *Ocimum gratissimum*: rounded, brown seeds with pebbly surface texture, not mucilaginous in water.
- *Ocimum tenuiflorum*: ellipsoid, yellow seeds with streaks of black, surface smooth, forming a thin, translucent mucilaginous coat in the presence of water.

Metabolite extraction

The five technical replicates were combined for each growth replicate, resulting in three samples per variety. After the final harvest, samples were lyophilized at the Pennsylvania State University CSL Behring Fermentation Facilities. Lyophilized samples were then ground to a fine powder with a mortar and pestle under liquid nitrogen. Powdered samples were weighed, combined with 80% aqueous methanol with 0.1% formic acid, and shaken at 200 rpm at room temperature for 16–18 h. Following extraction, the solid material was separated via vacuum filtration with 0.2 um filter paper, and the solvent was dried to completion with a Buchi rotoevaporator. Final extracts were stored at room temperature in sealed vials until mass spectrometry preparation. Extract yields can be found in Table 1.

Obtaining consumer-available *Ocimum* products and *Ocimum* reference materials

Reference materials

Eleven reference materials were obtained through online retailers to construct an external validation set. Four samples were ordered from herbal product suppliers and arrived with certificates of analysis (COAs) verifying their species identification (two *O. tenuiflorum* and two *O. gratissimum*). Two samples (one *O. tenuiflorum* and one *O. gratissimum*) were donated from the University of Georgia Medicinal Plant Farm and were morphologically identified with an associated publication [20]. An *O. tenuiflorum* VRBM was obtained from ChromaDex (Torrance, CA, USA), and a "Holy Basil" reference standard was obtained from USP (however, no specific species was listed). Two out-group reference materials were added to the external validation set—one *Salvia rosmarinus* L. (rosemary) standard (USP) and one *Salvia officinalis* L. (sage) leaf reference material (ChromaDex). One *O. basilicum* sample was obtained from an organic farm and morphologically identified on-site prior to shipment. Samples were extracted following the method described above for greenhouse sample extractions, except materials that arrived as a powder were not ground with liquid nitrogen prior to extraction. VRBMs and USP reference standards were only extracted once due to limited sample size. All samples were stored at room temperature in the dark until use, and final extracts were stored at room temperature in sealed vials until mass spectrometry preparation. See Table 2 for a complete list of reference materials and extraction yields.

Consumer products

Seventeen consumer products were ordered online from a variety of retailers. Eleven products were bulk, dried herbs from organic farms, herbal product suppliers (with no associated COAs), or eBay retailers. Two were tinctures or liquid softgels, two were *O. gratissimum* and *O. tenuiflorum* tea blends, and two were culinary products. There were six *O. basilicum* products, six *O. tenuiflorum* products, two *O. gratissimum*, and three "holy basil" products with no species identification. Dried leaves and powdered materials were prepared as described above. Most products were extracted in triplicate, but some only underwent a single extraction due to limited sample quantity. Liquid samples were not further extracted, and their mass spectrometry preparation is described below. All samples were stored at room temperature in the dark until extraction, and the final extracts were stored at room temperature in sealed vials until mass spectrometry preparation. See Table 2 for a complete list of consumer products and extraction yields.

Table 2 Reference material (EVS) and consumer product (C) information

Sample name	Set	Type	Species	Average extraction yield ± SD (%)
B1	C	Dried herb	*O. tenuiflorum*	18.4 ± 6.9
B2	C	Dried herb	*O. tenuiflorum*	10.8 ± 1.3
B3	C	Dried herb	*O. gratissimum*	14.36 ± 0.68
B4	C	Dried herb	*O. tenuiflorum*	26.7 ± 2.4
B5	C	Dried herb	*O. gratissimum*	17.53 ± 0.87
B6	C	Dried herb	*O. tenuiflorum*	14.25 ± 0.91
B7	C	Dried herb	*O. tenuiflorum*	15.0 ± 1.6
B8	C	Dried herb	*O. basilicum*	24.2 ± 2.1
B9	C	Dried herb	*O. basilicum*	21.3 ± 1.3
B10	C	Dried herb	*O. tenuiflorum*	16.19 ± 0.16
B11	C	Dried herb	*O. basilicum*	18.6 ± 3.5
B12	C	Culinary herb	*O. tenuiflorum*	16.97
B13	C	Culinary herb	*O. basilicum*	30.91
B14	C	Tulsi tea, mix	*O. tenuiflorum* and *O. gratissimum*	17.57
B15	C	Holy basil extract gel capsule	*O. tenuiflorum*	38.42
B16	C	Holy basil tincture	*O. tenuiflorum*	n/a
B17	C	Tulsi tea, mix	*O. gratissimum* and *O. tenuiflorum*	22.77
VS1	EVS	Dried herb—COA	*O. gratissimum*	15.1 ± 1.2
VS2	EVS	Dred herb—COA	*O. tenuiflorum*	11.8 ± 4.1
VS3	EVS	Dried herb—publication	*O. tenuiflorum*	16.9 ± 2.6
VS4	EVS	Dried herb—publication	*O. gratissimum*	14.1 ± 2.0
VS5	EVS	Holy basil VRBM	*O. tenuiflorum*	n/a
VS6	EVS	Rosemary USP reference standard	*Non-target*	n/a
VS7	EVS	Sage leaf USP reference material	*Non-target*	n/a
VS8	EVS	Holy basil USP reference material	*O. tenuiflorum* or *O. gratissimum*	n/a
VS9	EVS	Dried herb—morphological ID	*O. basilicum*	21.2 ± 1.4

Ultraperformance liquid chromatography tandem mass spectrometry analysis

All samples were prepared at a concentration of 1 mg/mL in LCMS grade methanol with 1 uM chlorpropamide (Stanta Cruz Biotechnology, Dallas, TX, USA) as an internal standard.

Untargeted metabolomic analyses were performed on a Vanquish Duo UHPLC system connected to a Thermo Orbitrap Fusion Lumos Mass Spectrometer (ThermoFisher Scientific, Waltham, MA). A Waters Acquity UPLC BEH C18 (1.7 µm, 2.1 × 150 mm) column was used with a flow rate of 0.1 mL/min at 55 °C. Solvent A was 0.1% formic acid (v/v) in LC–MS water and solvent B was 0.1% formic acid (v/v) in LC–MS acetonitrile. The mobile phase gradient of solvent B was as follows: 3% for 0.01 min, 45% for 10 min, 75% for 2 min, 100% for 4.5 min, and 3% for 0.2 min. A 2 uL injection was used for all samples [21, 22].

Mass spectrometry was conducted using an electrospray ionization source with a positive ion spray voltage of 3500 V, sheath gas pressure of 25 Arb, auxiliary gas pressure of 5 Arb, ion transfer temperature of 275 °C, and vaporizer temperature of 75 °C. MS^1 data was acquired with an Orbitrap resolution of 120,000, a scan range of 100–1000 Da, and an RF lens of 50% in the profile mode. MS^2 data was collected in a data-dependent manner using an intensity threshold of 2.5e4 and dynamic exclusion (ions excluded after 1 detection for 30 s).

Data processing and preparation

The UPLC-MS/MS data were analyzed and processed using MZmine 3.1 software [23]. Peaks were detected with a noise level of 5.5E5 counts, minimum peak duration of 0.25 min, and 25% tolerance for *m/z* intensity variation. The ADAP algorithm was used to build chromatograms with the following parameters: minimum group size = 5, group intensity threshold = 5.5E5, minimum highest intensity = 5.5E5, and scan-to-scan accuracy = 0.05 Da or 10.0 ppm. Chromatograms were resolved using the ADAP intensity window chromatogram

resolution feature with a signal/noise threshold of 7, minimum feature height of 80, coefficient/area threshold of 110, peak duration range of 0.00–0.10 min, and RT wavelet range of 0.00 to 0.10 min. Isotopes were filtered before integrating all features with the join aligner algorithm with the following parameters: *m/z* tolerance = 0.05 Da or 10.00 ppm, weight for *m/z* = 50, RT tolerance = 0.25 min, weight for RT = 50, and mobility weight = 1.00.

Features not present at an intensity greater than fivefold the average intensity compared to the blank samples were removed. Additionally, all samples not present in at least 3 replicates were removed. Then, the three replicates of each variety were averaged for further data analysis. Raw spectral data was deposited in the MASSive database (ID: MSV000094012, https://doi.org/10.25345/C5V980317). The final dataset contained 1123 features after all processing steps.

Metabolomics data analysis

Data preprocessing was performed in R version 4.1.1. Preprocessing settings were selected based on the greenhouse data and applied to the external validation and consumer product datasets in future steps (see "Supervised predictive models" section). Features were Hellinger transformed (square root) to transform from heteroscedastic to homoscedastic noise and auto-scaled (sample intensity—average feature intensity/feature intensity standard deviation) to limit the dependence of the variation on the mean concentration and improve biological relevance [24]. The R script is available at: https://github.com/kelloggresearchgroup/MetabolomicsAnalysis/

After transformation, a permutational multivariate analysis of variance (PerMANOVA) was performed using the pairwise.adonis() function in the *vegan* package using Euclidean distances and Benjamini & Hochberg *p*-value adjustment to determine if a statistical difference exists between the overall metabolomes of the three species [25].

Unsupervised PCA

An unsupervised principal component analysis (PCA) was performed to visualize potential metabolite patterns between the three species in the greenhouse data and the variation between the greenhouse, external validation set, and consumer products. For all three PCAs, the key components were selected using scree plots generated with the viz_contrib() tool within the *factoextra* package.

Supervised predictive models

General ensemble approach

We designed three predictive models (PLS-DA, LASSO, and RF) to dually classify samples as a given species (e.g., "is *O. basilicum*" or "is not *O. basilicum*"), developing a Monte Carlo training scheme for each model to evaluate multiple test and training set combinations. This resulted in an ensemble approach that allows probabilistic predictions. The model workflow (Fig. 1) was performed 100 times to evaluate the distribution of validation parameters and determine the species predictions within the external validation set and consumer products set based on the probability (out of 100) that a sample is or is not the specified species.

PLS-DA Binary partial least squares—discriminatory analysis (PLS-DA) models were constructed in R with the splsda() function in the *mixOmics* package [26]. The tune.splsda() function was used to optimize the model with the training set with a fivefold cross-validation method to select the ideal number of components and variables. The test, external validation, and consumer sets were predicted using the predict() function, and validation scores and classification rates were generated using the confusionMatrix() function from the *caret* package. The entire process was repeated 100 times. Final predictions of the external validation and consumer products sets were made by evaluating the probability that each sample is "positive" over the 100 predictions and choosing the species with the highest probability out of the three separate binary models. The distribution of validation scores was expressed as the average and standard deviation over 100 runs.

LASSO Binary LASSO regression models were built using the *glmnet* package in R [27]. The key parameter of LASSO models is the regularization parameter lambda (λ), which controls the amount of shrinkage the coefficients receive. Larger λ values result in more zero coefficients, and thus a simpler model. However, a smaller λ yields less bias with more features [14]. Lambda was chosen to minimize tenfold cross validation error on the training set using cv.glmnet(). The lambda that resulted in the lowest mean validation error was used to build the LASSO model with the glmnet() function. The test, external validation, and consumer sets were predicted using the predict() function, and model evaluation and final predictions over 100 runs were conducted as described in the PLS-DA section.

Random forest Random forest ensemble models were constructed using the randomForest() function from the *randomForest* package in R [28]. Each model was constructed from 500 individual trees. A bootstrap approach was used to evaluate model performance on the training set, and the test, external validation, and consumer sets were predicted as described above.

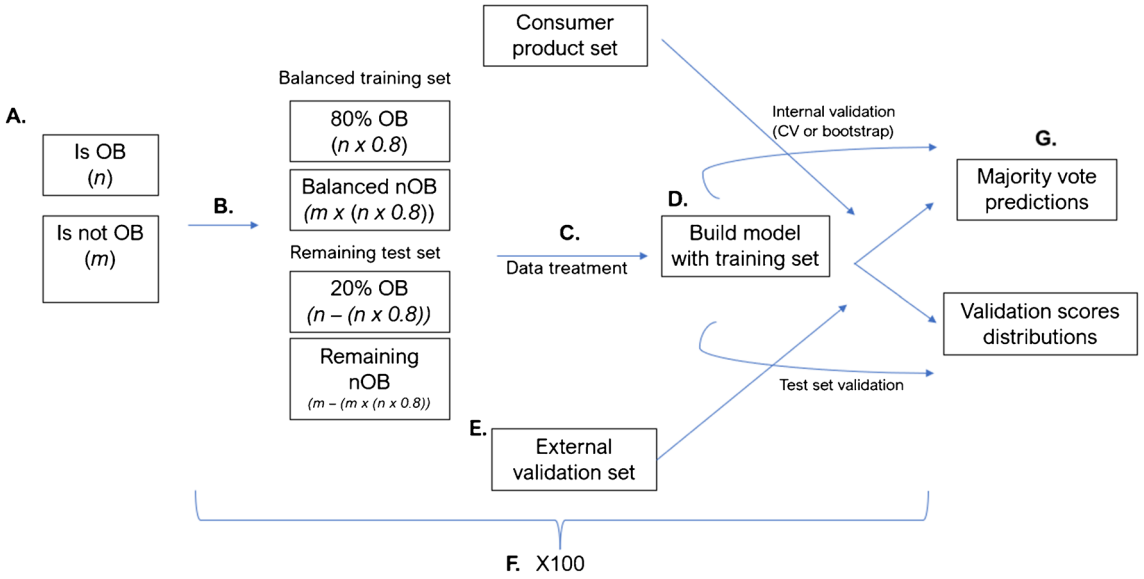


Fig. 1 Monte-Carlo based resampling workflow for supervised model construction. **A** Assign binary variable names, **B** split data into balanced test and training sets, **C** auto-scale all data based on the mean and standard deviations of each variable in the training set, **D** build predictive model with training set, **E** use model to predict species of each new sample in the test, external validation, and consumer product sets, **F** repeat steps B–E 100 times, **G** evaluate validation parameters and predictions over 100 runs

Model comparisons and validation parameters

Model performance was evaluated by directly comparing each model's specificity, sensitivity, balanced accuracy, and correct classification rate (CCR) over 100 runs. Each of these measurements considers the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the model's prediction.

Specificity is the model's ability to determine FNs (e.g., a model's ability to predict that a not-*O. basilicum* sample is not-*O. basilicum*):

$$Specificity = TN/(TN + FP)$$

Sensitivity is the opposite of specificity; it is the model's ability to predict TPs (e.g., correctly predicting *O. basilicum* samples):

$$Sensitivity = TP/(TP + FN)$$

Balanced accuracy is a combination of specificity and sensitivity, with an adjustment to account for the uneven distribution of classes in the predicted dataset.

$$Balanced\ accuracy = (specificity + sensitivity)/2$$

We further evaluated the predictions from each model. The probability of a sample being positive or negative (OB or nOB, for example) was calculated based on the number of each prediction over the 100 runs. Using this information, we calculated the correct classification rate (CCR) of each model:

$$CCR = (TP + TN)/(TP + FP + TN + FN)$$

High-performance thin-layer chromatography

Standards and test solutions

Stock solutions for quercetin, rutin, and chlorogenic acid were prepared at 1 mg/mL in methanol. Chlorogenic acid and quercetin were further diluted at 200 µg/mL while rutin was prepared at 400 µg/mL. The universal HPTLC mix (UHM) was prepared in-house and used as a system suitability test (SST) [29]. Sample extracts were prepared at 10 mg/mL in methanol.

Instrumentation

HPTLC analyses were conducted with an HPTLC system (CAMAG), equipped with a TLC-Visualizer 2, Automatic TLC Sampler 4, Automatic Developing Chamber 2, TLC Scanner 4, TLC Plate Heater III, and a Derivatizer. Analyses were run and data was processed using the software visionCATS version 3.2. HPTLC parameters were in agreement with the United States Pharmacopoeia (USP) general chapter 203 [30].

Application, development, and detection

Two microliters of the reference standards and UHM and 8 µL of the sample were applied onto an HPTLC Si 60 F254 plate (Supelco), then developed with n-butyl acetate:methanol:water:formic acid (7.5:2:1:1, v/v) [31]. The detection was performed under shortwave UV (254 nm), longwave UV (350 nm broad band), and white light in reflection + transmission (RT) prior to derivatization and in longwave UV and white light RT after derivatization with natural product reagent (NP), and anisaldehyde (AS) subsequently to NP (NP + AS).

Reagent preparation and spraying

Natural product reagent was prepared by dissolving 1 g of 2-aminoethyl diphenylborinate in 200 mL of methanol. The plate was heated at 100 ℃ for 3 min, then cooled down to room temperature and sprayed with 3 mL of the solution using a Derivatizer with a green nozzle, set at level 3. Images were recorded after 2 min.

Anisaldehyde in sulfuric acid reagent was prepared as follows: 170 mL of cooled methanol were slowly mixed with 20 mL of acetic acid and 10 mL of sulfuric acid. The mixture was cooled down to room temperature, then 0.5 mL of anisaldehyde (p-methoxy benzaldehyde) was added. Three milliliters of the solution was sprayed onto the plate using a Derivatizer with a blue nozzle at level 4, then the plate was heated at 100 °C for 3 min and the images were recorded.

Results and discussion

Ocimum growth and variation

This study utilized the natural variation within *Ocimum*, or basil/Tulsi, to investigate the application of statistical models in natural product studies. *Ocimum* has been widely cultivated worldwide for centuries. There are over 60 reported species of *Ocimum*, each with a number of cultivars and chemotypes [19]. This extensive taxonomic collection, along with morphological overlap between varieties, results in convoluted *Ocimum* identification. *O. tenuiflorum*, often called *O. sanctum*, is the most common therapeutic species of *Ocimum*. *O. tenuiflorum* cultivars include Krishna and Rama and are typically characterized by strong, spicy scents and fuzzy leaves, although wide morphological variation is reported among *O. tenuiflorum* varieties [32]. *O. gratissimum* also contains valuable therapeutic properties. Vana is a common *O. gratissimum* cultivar, and other names for *O. gratissimum* include African and Clove Tulsi [33]. *O. basilicum*, or sweet basil, possesses high economic value, mainly for its culinary uses. *O. basilicum* is widely cultivated, resulting in numerous cultivars, including Thai basil, Purple basil, Genovese basil, and more. The extensive collection of common names and *Ocimum* varieties often results in misclassification upon harvest and the potential substitution of rare types with cheaper varieties [34]. For example, the *O. basilicum* variety purple basil can easily be confused with *O. tenuiflorum* due to the similarity in leaf colors. This makes *Ocimum* an ideal model system for investigating the potential for predictive models to differentiate between closely related herbal species.

The commercial seeds were labeled as *O. basilicum* (10), *O. tenuiflorum* (13), or *O. gratissimum* (8); however, based on morphological investigations using seed characteristics, our dataset contained 5 *O. gratissimum*, 7 *O. teniflorum*, and 16 *O. basilicum* materials. We observed no species-based trends in biomass, leaf dry weight, or growth patterns (data not shown). We found that morphological identification based on leaf and floral characteristics was quite tedious and challenging, and we posit a potential reason for the misclassification of commercially available seeds. To better characterize the samples, we used seed morphology (Supplemental Table 2) to highlight differences between the three species' seeds. All mislabeled seeds were *O. basilicum* characterized as a holy basil species, highlighting the extent of cultivation and morphological variation within *O. basilicum* which can complicate taxonomic evaluation.

Unsupervised evaluation of sample relationships

Before developing predictive models, we investigated the relationships between the greenhouse-grown *Ocimum* species. Permutational multivariate analysis of variance (PerMANOVA) is a Euclidian-distance-based measurement that evaluates group differences within a multivariate space with multiple permutations which accommodates the covariates and unbalanced datasets common to metabolomics studies [29]. We compared the relationships of the labeled species (the species on the seed packet) and the morphologically ID'd species to investigate the chemical differences with different classifications. Among the greenhouse samples, there was a significant difference between all three labeled species (Table 3). *O. basilicum*–labeled materials differed the most from the other two species, with a *p*-value of 0.0001 for both *O. basilicum* vs *O. tenuiflorum* and *O. gratissimum*. *O. tenuiflorum* vs *O. gratissimum* resulted in a *p*-value of 0.0014, demonstrating that the three species as listed on the seed packets have unique metabolite profiles and are chemically distinct when grown in a controlled setting. PerMANOVA based on the morphological identification show similar results; all three species are significantly different from each other (Table 3).

There were distinct clusters of the materials labeled and morphologically ID'd as *O. basilicum* in the PCA scores

Table 3 PerMANOVA results comparing the untargeted metabolite profiles of species and sample source. $P = 0.05$

Greenhouse-only morphological ID	**OB vs OG**	**OB vs OT**	**OG vs OT**
Adjusted *P*-value	0.0008	0.0003	0.0018
Greenhouse-only labeled species	**OB vs OG**	**OB vs OT**	**OG vs OT**
Adjusted *P*-value	0.0001	0.0001	0.0014
All samples	**GH vs consumer**	**GH vs EVS**	**Consumer vs EVS**
Adjusted *P*-value	0.0001	0.0001	0.5333

plots (Fig. 2). *O. basilicum*, commonly called sweet basil, differs from holy basil in its chemical properties, bioactive potential, and intended cultivation as a culinary, not therapeutic, herb [35]. While there is a significant difference between all three species based on claimed ID and morphological ID, there was an unexpected overlap seen in the PCA scores plots (Fig. 2). Figure 2a shows that 5 *O. basilicum* materials are more chemically similar to the *O. gratissimum* materials, and Fig. 2b highlights that these same materials are all labeled as *O. tenuiflorum* by the seed distributors. Four of the five samples were labeled as "Kapoor" varieties, which is taxonomically ambiguous; some sources list Kapoor as a variety of *O. tenuiflorum* [36], while others classify it as *O. africanum* [37, 38]. This confusion extends to commercial seed labels, as most "Kapoor" seeds were labeled as *O. tenuiflorum* when ordered in 2021, but their label has since changed to *O. africanum*. Upon further investigation via HPTLC, we confirmed that these materials belong to a unique cultivar separate from *O. tenuiflorum* and *O. gratissimum* (Fig. 2e). Possible explanations for this overlap include potential gene flow, local adaptation, or hybridization. However, this is beyond the focus of this study, and therefore further studies investigating evolutionary history and hybridization could clarify the classification of these materials. However, the current study highlights the benefits of orthogonal chemical approaches—i.e., untargeted LC–MS metabolomics and HPTLC with multivariate modeling—that are capable of identifying hidden patterns between materials that morphological evaluation (whether by leaf or seed) alone may be overlooked.

Together, these results indicate that while conclusive taxonomic classification is difficult, there are three chemotypes among the greenhouse materials based on the untargeted metabolomics information, and a potential fourth chemical group based on the HPTLC fingerprints. Since we observed such distinct chemical groupings, we further investigated if chemotype groupings are more reliable for classifying new materials than species identification. While straying from traditional identification categories, chemotype classifications can be informative for consumers and herbal supplement manufacturers since chemical diversity drives differences in therapeutic effects.

Regardless of the classification approach, an essential question surrounding botanical authentication is to know if "reference materials" (i.e., the greenhouse-grown samples) are a reliable predictive model for external samples. Thus, we evaluated the chemical variation between the greenhouse and commercial samples, including reference materials and consumer products (Table 3). PerMANOVA results revealed that the greenhouse samples (all three species combined) were significantly different from both the external validation samples ($p = 0.0001$) and the consumer products ($p = 0.0001$). However, the consumer and external validation profiles did not significantly differ ($p = 0.53330$). Similarly, consumer and validation set samples were positioned close together on the PCA scores plot (Fig. 3a). The three consumer products that were separated from the remaining samples (in the upper right quadrant) were the two tea blends—B14 and B17—and a product marketed for culinary use—B13. While it is difficult to conclude why these products are so unique, we can hypothesize the chemical profile was altered as a result of processing, harvest, or formulation conditions. These results are supported with HPTLC fingerprints, where the consumer product fingerprints are much more complicated than the greenhouse-grown materials (Fig. 3b).

These findings agreed with the original hypotheses: greenhouse samples had highly controlled growth and processing conditions, which reduced sources of variation typical of herbal product manufacturing. When introducing external conditions, like temperature, moisture, drying conditions, and storage, *Ocimum* chemical profiles experience large variations [39, 40]. Since the conditions under which the ordered samples were grown, processed, packaged, and shipped were unknown (as is typical in a regulatory testing environment), we could not introduce all potentially relevant variables into our dataset. This is representative of industry conditions in which having such a varied collection of samples for any species would be quite time- and cost-intensive.

Predictive model construction and evaluation

The PerMANOVA and PCA results clearly demonstrated that the greenhouse samples are chemically distinct from the commercial samples. Adding in a supervision component could potentially leverage the known metadata to improve the predictive capabilities of the metabolomics datasets; thus, we evaluated if the greenhouse-produced

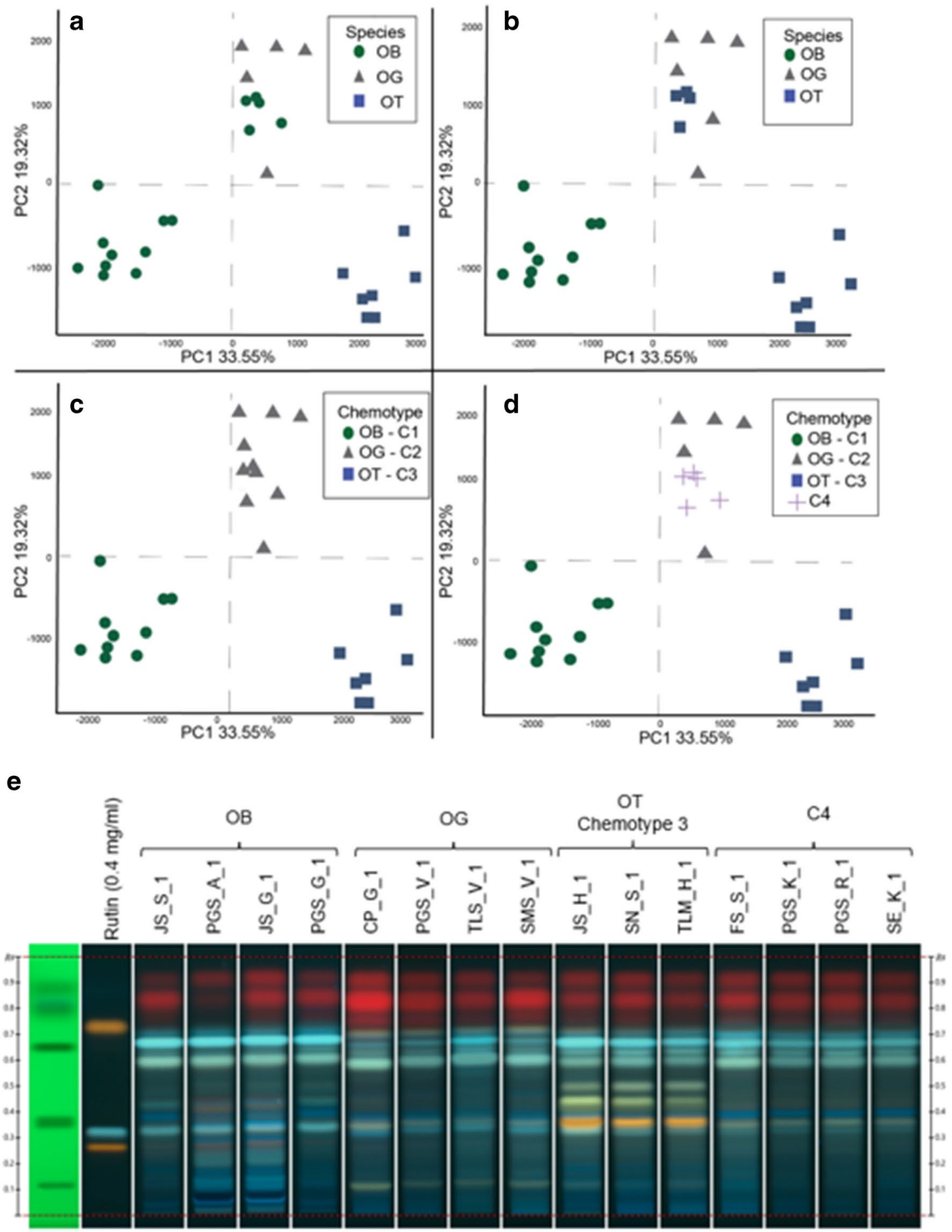


Fig. 2 Principal component analysis (PCA) scores plots of the untargeted metabolomics profiling of *Ocimum* samples from greenhouse-produced samples. PCA scores plot of greenhouse samples coded by **a** morphological ID, **b** labeled species, **c** LC–MS-based chemotype (3 groups), **d** HPTLC-based chemotype, as demonstrated by HPTLC fingerprints (**e**)

"reference materials" could predict the identity of commercial samples when employing a supervised machine learning approach. To this end, we chose three models to evaluate predictive capabilities: LASSO, PLS-DA, and RF. LASSO and PLS-DA are both linear models that are quite common in the herbal product literature. RF is increasingly

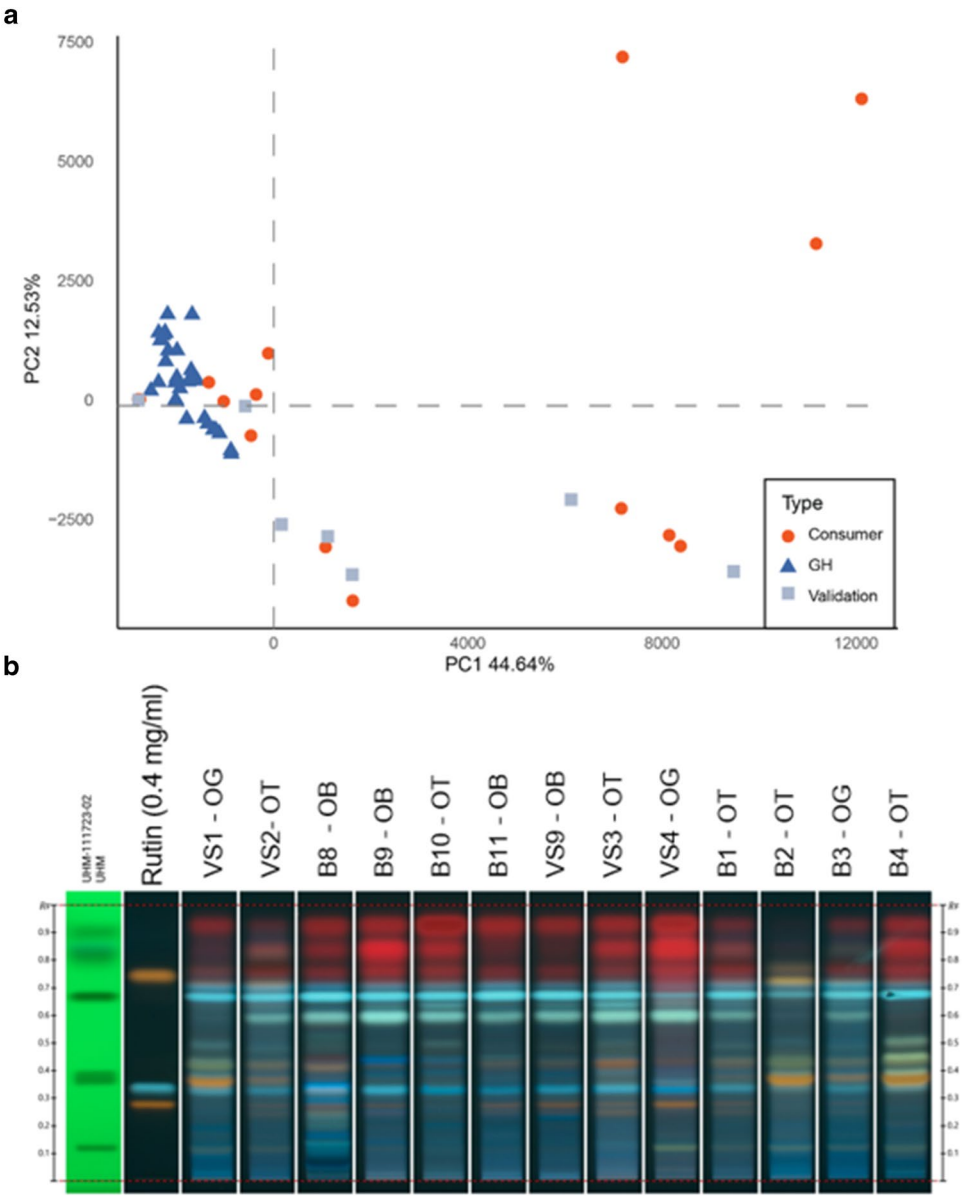


Fig. 3 **a** PCA scores plot profiling *Ocimum* samples from greenhouse-produced and external products. **b** HPTLC fingerprints of consumer products are more complex than greenhouse fingerprints (Fig. 2e)

popular, especially considering the often non-linear relationship of metabolites. We constructed binary models (e.g., each sample was classified as "is *O. basilicum*" or "is not *O. basilicum*") for all evaluations. We had the additional goal of determining if classification based on chemotype or species is more reliable when building predictive models. All data shown is based on the seed morphological identity of the materials, not the provided commercial label identification.

A common hindrance to metabolomics-based predictive modeling is building a representative training set with a small sample size. The most common solution is randomly selecting a subset of data to serve as a training set; while this is adequate for large datasets with balanced groups, it becomes complicated with small, unbalanced groups [41]. Consider, for example, using 80% of the overall data as the training set for the *O. gratissimum* model. However, less than 20% of the samples were *O. gratissimum*, and therefore it is possible that no *O. gratissimum* samples would be selected to build its own model! Thus, we used a Monte Carlo–based ensemble approach which randomly generated a training set that is balanced based on the number of samples in the "positive" group with the remaining samples placed in the test set (Fig. 1). The model was built and optimized as described in the experimental section, and the process was repeated for 100 iterations. Notably, this ensemble approach allowed the final prediction of new samples to be probability-based; the final prediction was based on the probability the sample belonged to each class over the 100 generated models. We evaluated model performance by comparing the distribution (average and standard deviation) of selectivity, specificity,

balanced accuracy, and CCR over 100 runs (Table 4, standard deviations can be found in Supplemental Table 1).

Predictive models classify new greenhouse samples with varying success

We first evaluated three supervised models' ability to classify new greenhouse samples based on species and chemotype; LASSO (Fig. 4a), PLS-DA (Fig. 4b), and RF (Fig. 4c). Overall, the models based on species had higher accuracies and correct classification rates (CCRs) than the models based on chemotype. The lowest accuracy (0.62) based on species resulted from the *O. gratissumum* binary LASSO model, whereas the lowest accuracy based on chemotype was 0.28 with the C3 LASSO model. We originally hypothesized that the chemotype classifications would be more reliable than morphological species for classifying new materials. However, these results indicate that even though there are interspecies chemical overlaps, species classifications provide more accurate predictions for new materials. These chemical relationships could be further investigated in future studies by expanding the sample size of all three species.

Within the species classifications, there are *O. basilicum* materials with high chemical similarity to *O. gratissium* which could have confounded the results. Indeed, the *O. gratissimum* models resulted in lower accuracies than the *O. tenuiflorum* and *O. basilicum* models. *O. tenuiflorum* and *O. basilicum* models resulted in high balanced accuracies (above 0.85), meaning all three models were mostly able to correctly predict the species identity of unseen samples. The CCR highlights that the LASSO models resulted in the most correct classifications for all three species—*O. basilicum* (100%), *O. tenuiflorum* (100%), and *O. gratissimum* (95%) (Table 4). It is of note that the CCR was often higher than the average balanced accuracy over the 100 runs, suggesting that a probability-based prediction scheme could be useful for optimizing prediction results. Taken together, these results indicate that a LASSO predictive model can successfully predict the species, but not chemotype, of new materials that were grown in the same controlled setting.

Developed models are not applicable to external samples

Before evaluating consumer products with the developed models, we assessed each model's ability to predict the species identity of validated external reference materials. This external validation set included certified reference materials from USP and Chromadex, as well as dried herbs from well-established sources with accompanying certificates of analyses (COAs) or publications to confirm species identity. We also included sage (*Salvia officinalis* L.) and rosemary (*Salvia rosmarinus* L.) materials to determine if the model can identify non-*Ocimum* samples. One note, only one *O. basilicum* sample was included in the external validation set due to a lack of available reference materials.

Overall, none of the three models performed well in classifying the reference materials (Table 4, Fig. 5 (row 2)). The LASSO model for *O. basilicum* had the highest balanced accuracy at 71%. Both the PLS-DA and RF had low balanced accuracies overall, and there was no indication that the models built with greenhouse materials could successfully predict the species of external materials (Table 5). When considering specificity and sensitivity, we observed inverse relationships between *O. basilicum* and *O. gratissimum*/*O. tenuiflorum*; as *O. basilicum's* sensitivity increased, the other two species' sensitivity decreased. So, when a model determined a new material was *O. basilicum*, its ability to reliably recognize the other species decreased. Similarly, when a model correctly predicted if a new material was not *O. basilcum*, it could not recognize if a material was not the other two species. However,

Table 4 Average validation scores across 100 runs of each supervised model (LASSO, PLS-DA, and RF) in predicting the greenhouse test sets based on species and chemotype

Species									
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.91	1.00	0.37	0.91	1.00	0.82	0.96	1.00	0.99
Specificity	0.97	0.96	0.87	0.94	0.99	0.56	0.92	0.71	0.50
Accuracy	0.94	0.98	0.62	0.92	0.99	0.69	0.94	0.86	0.75
CCR	1.00	1.00	0.95	1.00	1.00	0.63	1.00	0.76	0.58
Chemotype									
	LASSO			PLS-DA			RF		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
Sensitivity	0.54	0.13	0.05	0.83	0.80	1.00	0.83	0.87	1.00
Specificity	0.92	0.73	0.52	0.93	0.42	0.93	0.87	0.73	0.78
Accuracy	0.73	0.43	0.28	0.88	0.61	0.97	0.90	0.80	0.88
CCR	0.87	0.42	0.64	1.00	0.55	1.00	1.00	1.00	0.79

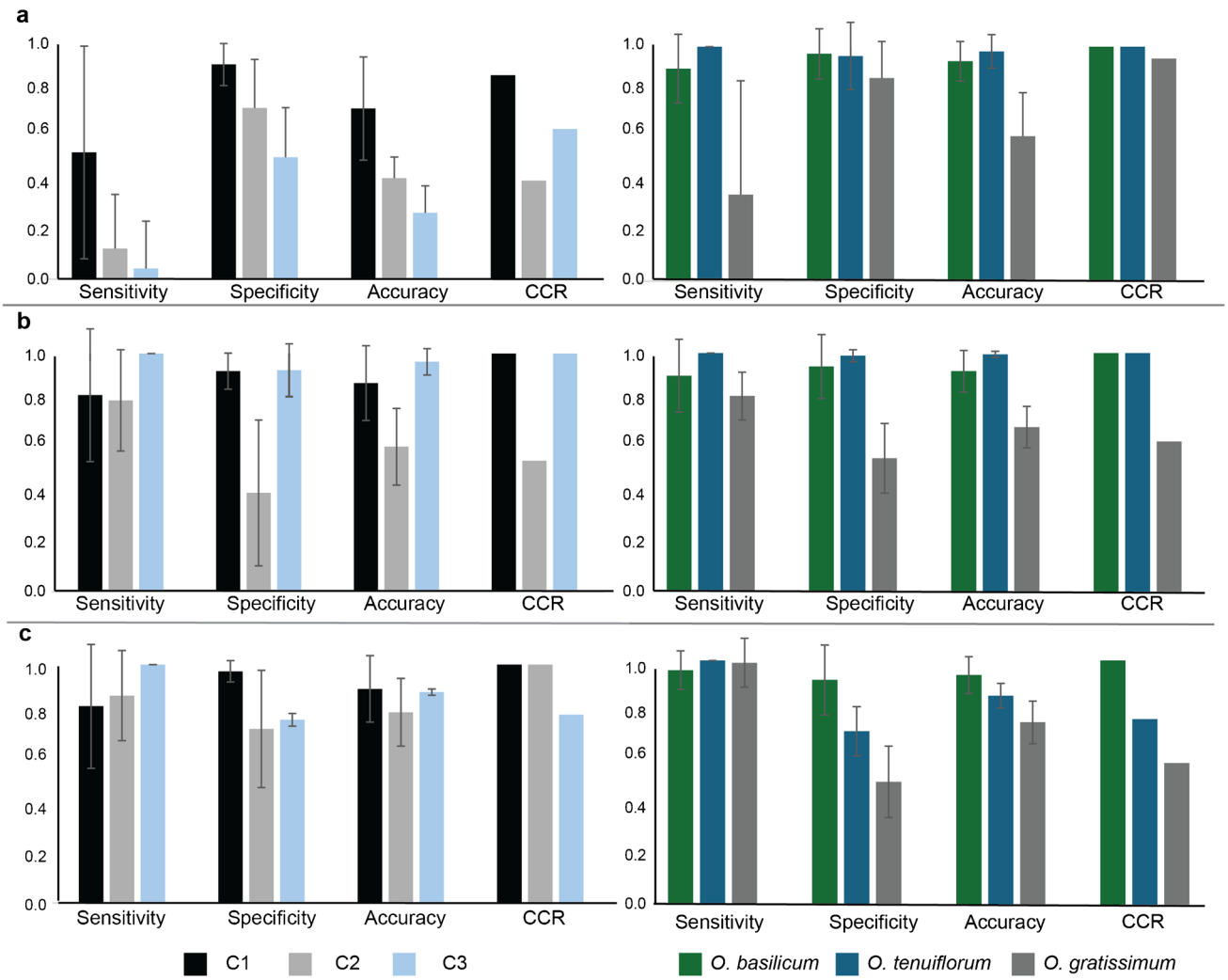


Fig. 4 Validation scores bar plots for **a** LASSO, **b** PLS-DA, **c** RF. Validation scores are represented as the average and standard deviation over 100 runs of each model predicting the greenhouse test set based on chemotype (left) and species (right)

irrespective of the models' accuracies, when the three species' models were combined for overall probability-based predictions, it was rare that the final prediction was correct (Table 6). The LASSO model, which was the most reliable for predicting new greenhouse materials, resulted in a final assignment that over half of the materials are not one of the three *Ocimum* species at all (Table 6).

These results are not all together unexpected; a botanical's chemical profile is greatly impacted by environmental, harvesting, and processing conditions. The untargeted metabolomics PCA scores plots combined with HPTLC fingerprints confirmed that the greenhouse materials were not chemically representative of the external materials, and introducing a supervised aspect via multivariate predictive modeling did not improve this relationship.

Consumer product identity predictions

While no model built on the greenhouse samples was able to adequately distinguish external reference samples, they were evaluated for their abilities to classify consumer-available products. It should be noted that these samples were not verified materials, so their actual species identity has not been confirmed. Thus, reporting sensitivity, specificity, and balanced accuracy was our best estimate of model performance.

While in most cases the balanced accuracies were slightly higher with the consumer products than the validation set predictions, they were still too low for reliable identification in an industry setting. For example, the highest balanced accuracy for consumer product predictions is 71% (RF, *O. tenuiflorum* model), accompanied by

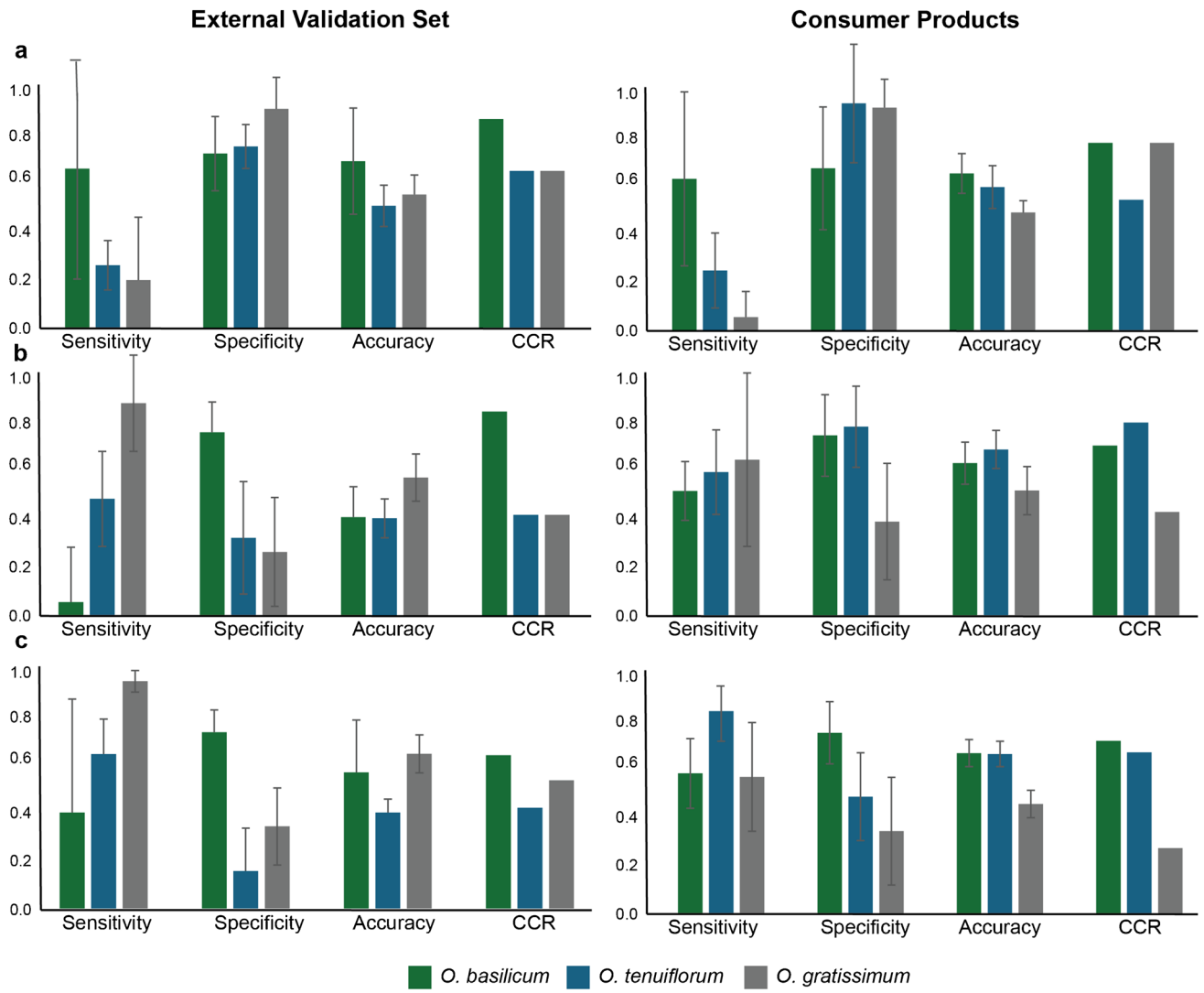


Fig. 5 Validation scores bar plots for **a** LASSO, **b** PLS-DA, **c** RF. Validation scores are represented as the average and standard deviation over 100 runs of each model predicting the external validation set (left) and consumer product set (right) based on species

Table 5 Average validation scores across 100 runs of each supervised model (LASSO, PLS-DA, and RF) in predicting the greenhouse test sets based on species and chemotype

External validation									
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.68	0.27	0.21	0.06	0.51	0.93	0.42	0.68	0.99
Specificity	0.74	0.74	0.93	0.80	0.34	0.28	0.77	0.16	0.34
Accuracy	0.71	0.52	0.57	0.43	0.43	0.60	0.56	0.42	0.68
CCR	0.89	0.67	0.67	0.22	0.44	0.44	0.67	0.44	0.56
Consumer products									
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.62	0.24	0.01	0.49	0.57	0.62	0.62	0.89	0.60
Specificity	0.66	0.92	0.90	0.71	0.74	0.37	0.79	0.52	0.36
Accuracy	0.64	0.58	0.48	0.60	0.66	0.49	0.71	0.70	0.48
CCR	0.76	0.53	0.76	0.35	0.76	0.41	0.76	0.71	0.29

Table 6 Probabilities of each external validation material belonging to each species over 100 runs of each model

LASSO									
Actual sample ID:	OG	OT	OT	OG	OT	Non-target	Non-target	OT or OG	OB
OB	0.01	0.23	0.79	0.02	0.08	0.37	0.46	0.09	0.68
OG	0.2	0.00	0.00	0.42	0.20	0.00	0.20	0.00	0.00
OT	0.16	0.06	0.06	0.05	0.92	0.00	0.07	0.89	0.00
Predicted ID:	None	None	OT	None	OT	None	None	OT	OB
PLS-DA									
Actual sample ID:	OG	OT	OT	OG	OT	Non-target	Non-target	OT or OG	OB
OB	1.00	0.97	0.68	0.99	0.83	0.93	0.04	0.96	0.94
OG	0.89	0.82	0.66	0.89	0.89	0.77	0.73	1.00	0.46
OT	1.00	0.06	0.36	0.49	1.00	0.49	0.62	1.00	0.32
Predicted ID:	All	OB	OB	OB	OT	OB	OG	OG or OT	OB
RF									
Actual sample ID:	OG	OT	OT	OG	OT	Non-target	Non-target	OT or OG	OB
OB	0.00	0.21	0.61	0.00	0.00	0.02	1.00	0.00	0.42
OG	0.98	0.67	1.00	1.00	0.88	0.99	0.02	1.00	0.29
OT	1.00	0.82	0.88	0.90	1.00	0.82	0.00	1.00	0.46
Predicted ID:	OG	OG	OT	OG	OG	OG	OB	OG or OT	OG

a 76% CCR. If approximately 30% of the time the model incorrectly identifies the product, this would translate to a 30% chance of mislabeled products reaching the market or unnecessary supply chain investigations.

Taken together, these results cast a doubtful light on the usefulness of controlled-environment samples serving as the foundation of machine learning and predictive modeling in herbal product identification studies. Overall, the models developed based on controlled greenhouse samples could not be applied to these external samples. These results were congruent with the unsupervised analysis, in which the external samples were separated from the greenhouse samples in the PCA scores plot (Fig. 2d). Recently, Harnly and Upton reported similar findings when using chemometrics models to investigate variation between *Actaea racemosa* L. reference materials and commercial standards [42]. These authors investigated multiple pretreatment approaches and model manipulation techniques to ultimately report that statistical modeling results in generally high sensitivity and low specificity when assessing the species of new materials, even when looking at a conserved subset of compounds. In our current study, samples in the external validation set were reference standards; they all originated from different sources, locations, and growing conditions that can influence their chemical profiles. These studies together suggest that external factors play a major role in metabolite composition and highlight the crucial need for the inclusion of more data from environmental, processing, and other external variables into predictive models before integrating such approaches into identification or authentication applications.

Conclusions

Multivariate modeling is a primary method for analyzing untargeted metabolomics data and is well-suited to discover patterns within complex chemical profiles. Chemometric approaches are instrumental in identifying and authenticating botanical materials due to their independence from a small subset of specific chemical markers, as well as their ability to incorporate numerous validated samples with inter- and intraspecies variation. As machine learning and predictive modeling become more popular across scientific disciplines, the limitations and applicability of new models in the herbal product space must be critically evaluated.

This study demonstrated that using chemometric models built on untargeted metabolomics data from controlled greenhouse samples could aid in classifying new *O. tenuiflorum* and *O. basilicum* greenhouse samples. Since the greenhouse samples all had the same environmental and processing conditions, this loosely resembled the conditions that a raw material grower would face. Thus, it is conceivable that using a LASSO model would be acceptable as an internal material verification system to confirm whether the chemical composition of the plants grown/processed under similar conditions conforms as expected or if there is unusual metabolite variation between batches.

Ultimately, chemometric modeling of controlled reference materials was not able to accurately predict the identity of commercially available products, as it was evident that there were significant differences in chemical

composition between the greenhouse and commercial samples. This suggests that environmental variation is a large driver of *Ocimum* diversity, potentially greater than taxonomic classifications; this is borne out by other recent botanical studies. While this has been extensively reported upon in agricultural and crop science studies, the relationship between genetics, metabolomics, and environmental interactions in herbal products has not been thoroughly investigated [43]. Including more variation in the original models, such as different post-harvest drying conditions or water and nutrient levels, may improve their ability to predict external samples. Attempting to encompass all potential sources of variation would result in a large and complex study requiring extensive data collection stages; however, smaller studies could potentially identify specific factors driving the shifts in chemistry and yield more successful predictive models. Therefore, more studies are needed to determine which of these myriad factors have the greatest impact on chemical diversity.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-025-05735-0.

Acknowledgements We would like to thank Dr. Justin Silverman for his guidance in statistical model construction and editorial review. We would also like to thank the Pennsylvania State University Herbarium for their support in *Ocimum* identification and voucher specimen preparation, and the Huck Institutes' Metabolomics Core Facility (RRID:SCR_023864) for use of the ThermoFisher Orbitrap Fusion Lumos LC-MS. Finally, we would like to thank Scott DiLoreto and the Pennsylvania State University Greenhouses for their guidance and assistance with growing *Ocimum*.

Author contribution Conceptualization: Evelyn Abraham, Joshua Kellogg. Methodology: Evelyn Abraham, Joshua Kellogg. Validation: Evelyn Abraham, R. Teal Jordan, Wilmer Perera, Sarah Chamerlain. Formal analysis: Evelyn Abraham, R. Teal Jordan, Wilmer Perera, Joshua Kellogg. Investigation: Evelyn Abraham, R. Teal Jordan, Wilmer Perera, Sarah Chamerlain. Resources: Evelyn Abraham, Wilmer Perera, Joshua Kellogg. Data curation: Evelyn Abraham, Joshua Kellogg. Writing—original draft: Evelyn Abraham, Joshua Kellogg. Writing—review and editing: Evelyn Abraham, Sarah Chamberlain, Wilmer Perera, R. Teal Jordan, Joshua Kellogg. Visualization: Evelyn Abraham, R. Teal Jordan, Wilmer Perera. Supervision: Joshua Kellogg. Project administration: Wilmer Perera, Joshua Kellogg. Funding acquisition: Evelyn Abraham, Joshua Kellogg.

Funding This work was funded by the National Institute of Health, National Center for Complementary and Integrative Health grant number 1F31AT01213-01A1 and supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04956 and Accession #7006496. Additional funding was provided by the Anne C. Chatham Fellowship in Medicinal Botany from the Garden Club of America.

Declarations

Competing interests Evelyn Abraham receives a salary from Eurofins Food and Feed Testing, where she is a marketing content writer, but her work with Eurofins is independent and unrelated to research at Pennsylvania State University. Eurofins is not a collaborator or financial supporter of Josh Kellogg's research. All other authors have no potential conflicts of interest related to the current manuscript.

References

1. Applequist WL, Miller JS. Selection and authentication of botanical materials for the development of analytical methods. Anal Bioanal Chem. 2013;405(13):4419–28. https://doi.org/10.1007/s00216-012-6595-1.
2. Spinardi A, Cola G, Gardana CS, Mignani I. Variation of anthocyanin content and profile throughout fruit development and ripening of highbush blueberry cultivars grown at two different altitudes. Front Plant Sci. 2019;10:1045. https://doi.org/10.3389/fpls.2019.01045.
3. Wang H, Cao X, Yuan Z, Guo G. Untargeted metabolomics coupled with chemometrics approach for Xinyang Maojian green tea with cultivar, elevation and processing variations. Food Chem. 2021;352:129359. https://doi.org/10.1016/j.foodchem.2021.129359.
4. Perez De Souza L, Alseekh S, Naake T, Fernie A. Mass spectrometry-based untargeted plant metabolomics. CP Plant Biol. 2019;4(4):e20100. https://doi.org/10.1002/cppb.20100.
5. Abraham EJ, Kellogg JJ. Chemometric-guided approaches for profiling and authenticating botanical materials. Front Nutr. 2021;8:780228.
6. Hosbas Coskun S, Wise SA, Kuszak AJ. The importance of reference materials and method validation for advancing research on the health effects of dietary supplements and other natural products. Front Nutr. 2021;8:786261. https://doi.org/10.3389/fnut.2021.786261.
7. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58(2):109–30. https://doi.org/10.1016/S0169-7439(01)00155-1.
8. Walkowiak A, Ledziński Ł, Zapadka M, Kupcewicz B. Detection of adulterants in dietary supplements with Ginkgo biloba extract by attenuated total reflectance Fourier transform infrared spectroscopy and multivariate methods PLS-DA and PCA. Spectrochim Acta Part A Mol Biomol Spectrosc. 2019;208:222–8. https://doi.org/10.1016/j.saa.2018.10.008.
9. Sarkar R, Chatterjee N, Shaikh N, et al. Authentication of Tinospora cordifolia derived herbal supplements using high resolution mass spectrometry-based metabolomics approach – a pilot study. Ind Crops Prod. 2023;200:116835. https://doi.org/10.1016/j.indcrop.2023.116835.
10. Barbosa S, Saurina J, Puignou L, Núñez O. Classification and authentication of paprika by UHPLC-HRMS fingerprinting and multivariate calibration methods (PCA and PLS-DA). Foods. 2020;9(4):486. https://doi.org/10.3390/foods9040486.
11. Gad HA, El-Ahmady SH, Abou-Shoer MI, Al-Azizi MM. Application of chemometrics in authentication of herbal medicines: a review. Phytochem Anal. 2013;24(1):1–24. https://doi.org/10.1002/pca.2378.
12. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411–21.
13. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol). 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
14. Ranstam J, Cook JA. LASSO regression. Br J Surg. 2018;105(10):1348–1348. https://doi.org/10.1002/bjs.10895.
15. Bujak R, Daghir-Wojtkowiak E, Kaliszan R, Markuszewski MJ. PLS-based and regularization-based methods for the selection of

relevant variables in non-targeted metabolomics data. Front Mol Biosci. 2016;3. https://doi.org/10.3389/fmolb.2016.00035.
16. Zhu Y, Tan TL, Cheang WK. Penalized logistic regression for classification and feature selection with its application to detection of two official species of Ganoderma. Chemom Intell Lab Syst. 2017;171:55–64. https://doi.org/10.1016/j.chemolab.2017.09.019.
17. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
18. Patel RP, Singh R, Saikia SK, et al. Phenotypic characterization and stability analysis for biomass and essential oil yields of fifteen genotypes of five Ocimum species. Ind Crops Prod. 2015;77:21–9. https://doi.org/10.1016/j.indcrop.2015.08.043.
19. Zahran EM, Abdelmohsen UR, Khalil HE, et al. Diversity, phytochemical and medicinal potential of the genus Ocimum L. (Lamiaceae). Phytochem Rev. 2020;19(4):907–53. https://doi.org/10.1007/s11101-020-09690-9.
20. Fuller NJ, Pegg RB, Affolter J, Berle D. Variation in growth and development, and essential oil yield between two Ocimum species (O. tenuiflorum and O. gratissimum) grown in Georgia. Horts. 2018;53(9):1275–82. https://doi.org/10.21273/HORTSCI13156-18.
21. Kellogg JJ, Todd DA, Egan JM, et al. Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds. J Nat Prod. 2016;79(2):376–86. https://doi.org/10.1021/acs.jnatprod.5b01014.
22. Kellogg JJ, Alonso MN, Jordan RT, et al. A methoxylated flavone from *Artemisia afra* kills *Mycobacterium tuberculosis*; preprint. Microbiology. 2023. https://doi.org/10.1101/2023.10.11.561885.
23. Schmid R, Heuckeroth S, Korf A, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol. 2023;41(4):447–9. https://doi.org/10.1038/s41587-023-01690-2.
24. Van Den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, Van Der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7(1):142. https://doi.org/10.1186/1471-2164-7-142.
25. Oksanen J, Simpson GL, Blanchet FG, et al. vegan: Community Ecology Package. R package version 2.6-4. 2022. https://CRAN.R-project.org/package=vegan
26. 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(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
27. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Soft. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01.
28. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2/3:18–22.
29. Do TKT, Schmid M, Phanse M, et al. Development of the first universal mixture for use in system suitability tests for high-performance thin layer chromatography. J Chromatogr A. 2021;1638:461830. https://doi.org/10.1016/j.chroma.2020.461830.
30. U. S. Pharmacopeia. 203 High-performance thin-layer chromatography procedure for identification of articles of botanical origin. First Supplement to USP 38–NF 33. 2017:7044-7045.
31. Perera WH, Frommenwiler DA, Sharaf MHM, Reich E. An improved high-performance thin-layer chromatographic method to unambiguously assess Ginkgo biloba leaf finished products. JPC-J Planar Chromat. 2021;34(6):559–60. https://doi.org/10.1007/s00764-021-00146-0.
32. Malav P, Pandey A, Bhatt KC, Gopala Krishnan S, Bisht IS. Morphological variability in holy basil (Ocimum tenuiflorum L.) from India. Genet Resour Crop Evol. 2015;62(8):1245–56. https://doi.org/10.1007/s10722-015-0227-5.
33. Kumar N, editor. Biotechnological approaches for medicinal and aromatic plants: conservation, genetic improvement and utilization. Singapore: Springer Singapore; 2018. https://doi.org/10.1007/978-981-13-0535-1.
34. Bhamra SK, Heinrich M, Johnson MRD, Howard C, Slater A. The cultural and commercial value of Tulsi (Ocimum tenuiflorum L.): multidisciplinary approaches focusing on species authentication. Plants. 2022;11(22):3160. https://doi.org/10.3390/plants11223160.
35. Novak J, Blüthner W-D (Eds) Medicinal, aromatic and stimulant plants; Handbook of plant breeding. Cham: Springer International Publishing. 2020;12. https://doi.org/10.1007/978-3-030-38792-1.
36. Parveen A, Wang Y-H, Fantoukh O, 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. https://doi.org/10.1016/j.jpba.2019.112894.
37. Srivastava S, Lal RK, Maurya R, et al. Chemical diversity of essential oil among basil genotypes (Ocimum viride Willd.) across the years. Ind Crop Prod. 2021;173:114153. https://doi.org/10.1016/j.indcrop.2021.114153.
38. Maurya S, Sangwan NS. Profiling of essential oil constituents in Ocimum species. Proc Natl Acad Sci India Sect B Biol Sci. 2020;90(3):577–83. https://doi.org/10.1007/s40011-019-01123-8.
39. Ekren S, Sönmez Ç, Özçakal E, et al. The effect of different irrigation water levels on yield and quality characteristics of purple basil (Ocimum basilicum L.). Agric Water Manag. 2012;109:155–61. https://doi.org/10.1016/j.agwat.2012.03.004.
40. Abozeed A, El Shafey R, Osman Y. Effect of location and environmental conditions on growth, yields and chemical constituents of sweet basil (*Ocimum basilicum* L.). J Agric Chem Biotechnol. 2015;6(1):1–13. https://doi.org/10.21608/jacb.2015.43961.
41. Beleites C, Salzer R. Assessing and improving the stability of chemometric models in small sample size situations. Anal Bioanal Chem. 2008;390(5):1261–71. https://doi.org/10.1007/s00216-007-1818-6.
42. Harnly J, Upton R. Variation in botanical reference materials: similarity of Actaea Racemosa analyzed by flow injection mass spectrometry. J AOAC Int. 2023:qsad137. https://doi.org/10.1093/jaoacint/qsad137.
43. Napier JD, Heckman RW, Juenger TE. Gene-by-environment interactions in plants: molecular mechanisms, environmental drivers, and adaptive plasticity. Plant Cell. 2023;35(1):109–24. https://doi.org/10.1093/plcell/koac322.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Supplemental Information for

Application of Predictive Modeling Tools for the Identification of *Ocimum* spp. Herbal Products

Abraham, Evelyn J., Chamberlain, Sarah J., Perera, Wilmer H., Jordan, R Teal, Kellogg, Joshua J.

Supplemental Table 1: Validation scores' standard deviations across 100 runs of each supervised model (LASSO, PLS-DA, and RF) in predicting the greenhouse test sets, external validation set, and consumer product set.

	Greenhouse Species								
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.15	0.00	0.49	0.15	0.00	0.39	0.079	0.00	0.310
Specificity	0.11	0.14	0.16	0.13	0.025	0.24	0.14	0.10	0.15
Accuracy	0.084	0.16	0.19	0.087	0.013	0.13	0.075	0.050	0.089

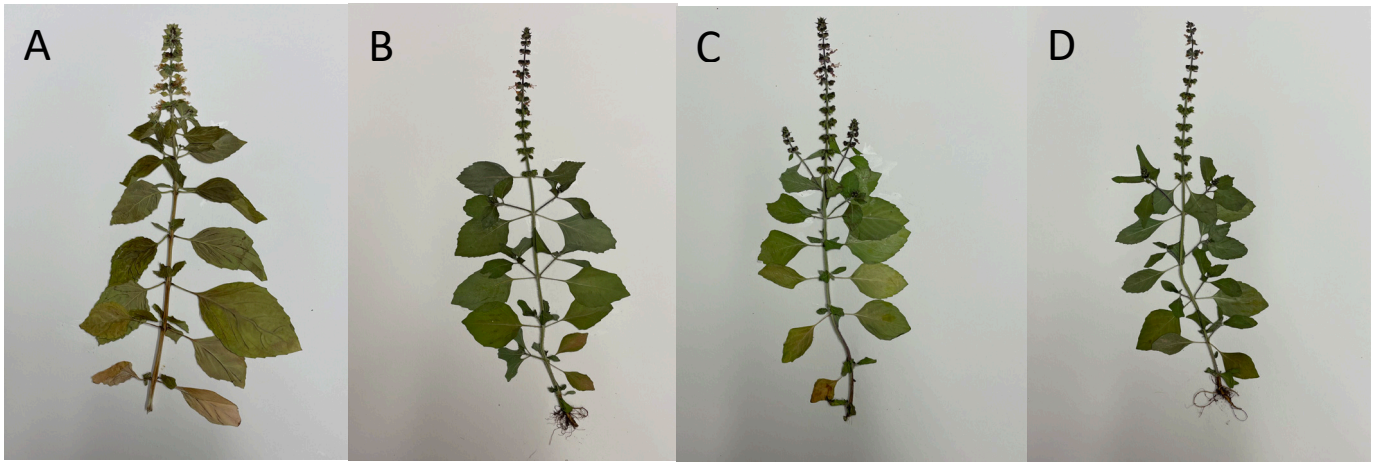
	Greenhouse Chemotype								
	LASSO			PLS-DA			RF		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
Sensitivity	0.45	0.23	0.20	0.28	0.21	0.00	0.26	0.19	0.00
Specificity	0.090	0.21	0.21	0.076	0.31	0.11	0.045	0.25	0.027
Accuracy	0.22	0.090	0.12	0.16	0.11	0.055	0.14	0.14	0.013

	External validation - Species								
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.47	0.10	0.27	0.24	0.21	0.21	0.50	0.15	0.047
Specificity	0.16	0.093	0.13	0.13	0.25	0.24	0.10	0.19	0.17
Accuracy	0.23	0.088	0.08	0.13	0.084	0.10	0.23	0.058	0.083

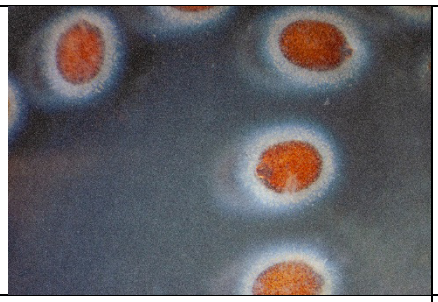
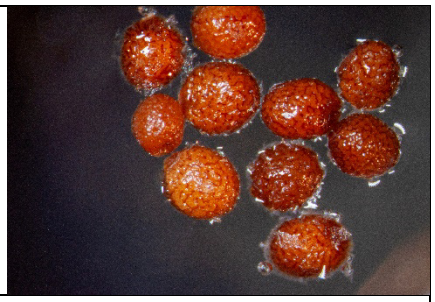
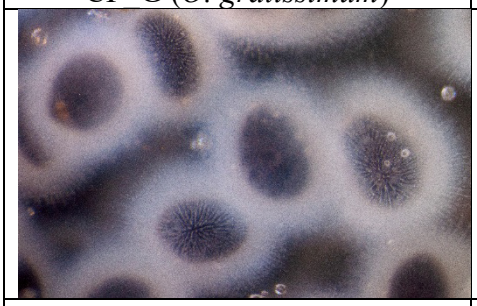
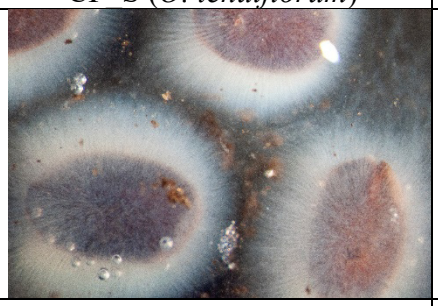
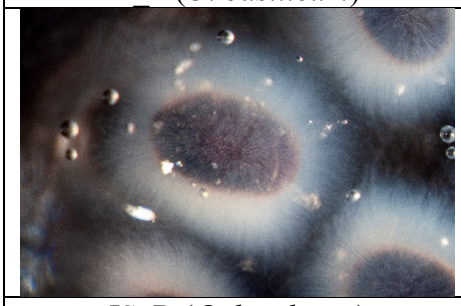
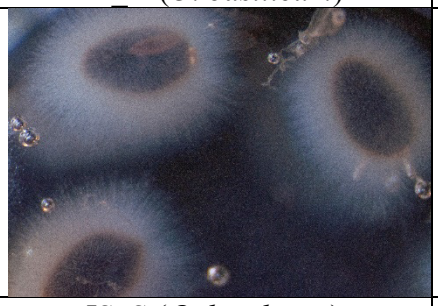
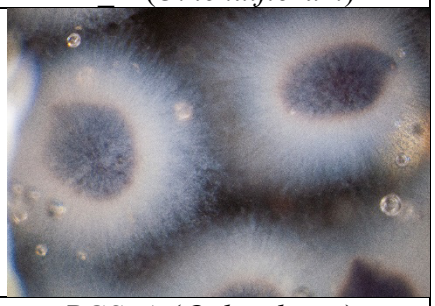
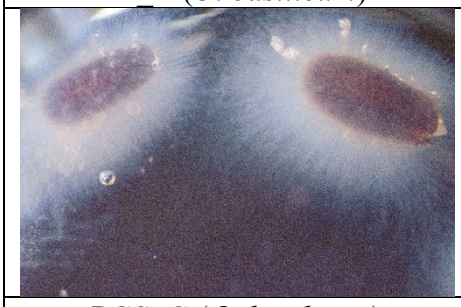
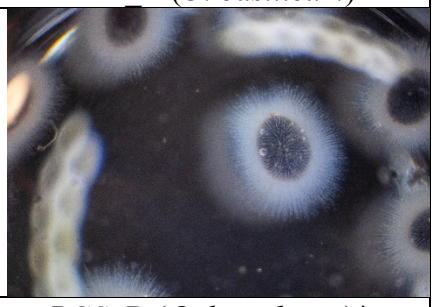
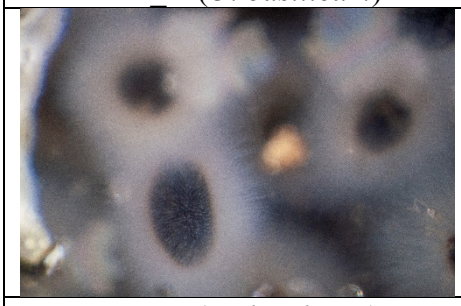
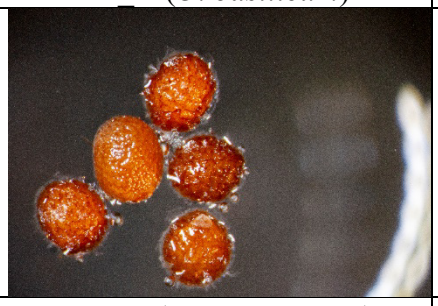
	Consumer Products - Species								
	LASSO			PLS-DA			RF		
	OB	OT	OG	OB	OT	OG	OB	OT	OG
Sensitivity	0.35	0.15	0.10	0.11	0.16	0.34	0.15	0.13	0.24
Specificity	0.25	0.24	0.11	0.16	0.16	0.23	0.14	0.19	0.24

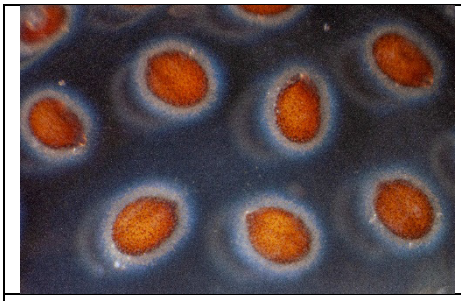
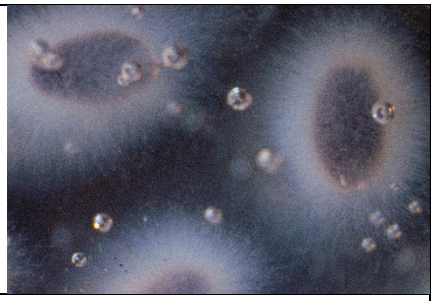
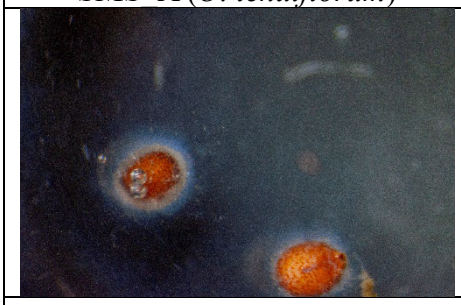
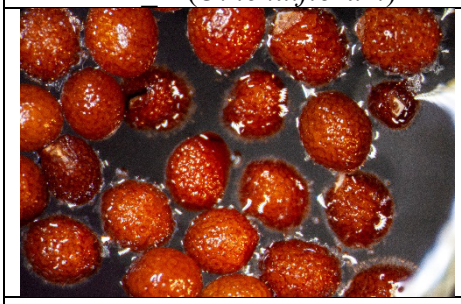
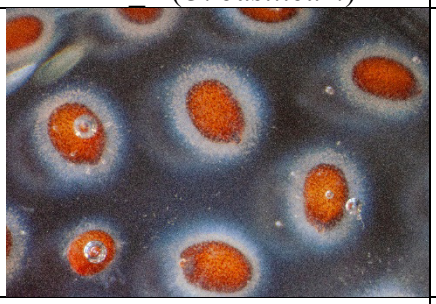
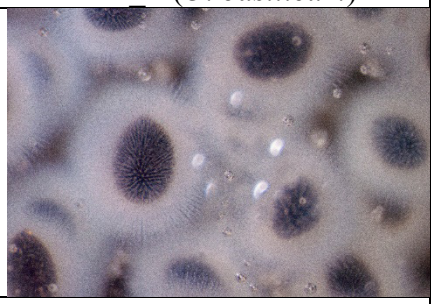
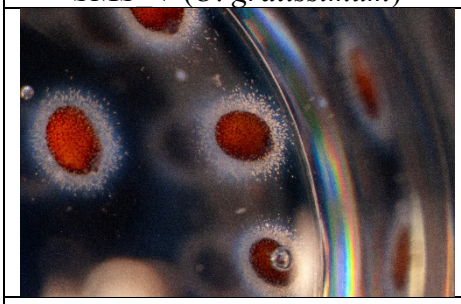
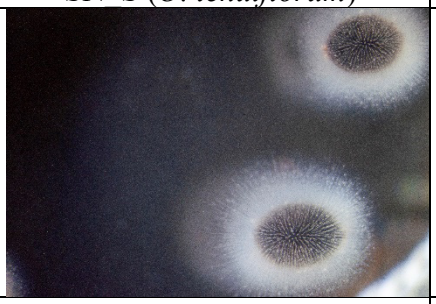
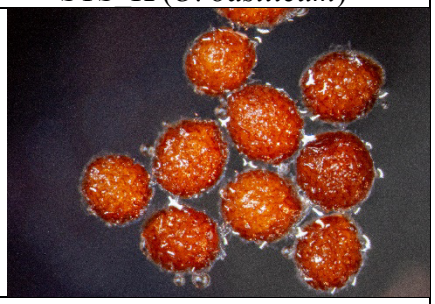
Accuracy	0.080	0.048	0.048	0.082	0.075	0.095	0.059	0.056	0.060

Supplemental Figure 1: Representative Voucher Specimens. A) *O. basilicum* (sweet basil) B) *O. gratissimum* (Vana) C) *O. tenuiflorum* (Kapoor) D) *O. tenuiflorum* (Rama)



Supplemental Table 2: Seed morphological determination, based upon microscopy images

		
CP_G (*O. gratissimum*)	CP_S (*O. tenuiflorum*)	FS_G (*O. gratissimum*)
		
FS_S (*O. basilicum*)*	JS_G (*O. basilicum*)	JS_H (*O. tenuiflorum*)
		
JS_R (*O. basilicum*)	JS_S (*O. basilicum*)	PGS_A (*O. basilicum*)
		
PGS_G (*O. basilicum*)	PGS_K (*O. basilicum*)*	PGS_R (*O. bascilicum*)*
		
PGS_T (*O. basilicum*)	PGS_V (*O. gratissimum*)	SE_K (*O. bascilicum*)*

		
SMS_A (*O. tenuiflorum*)	SMS_G (*O. basilicum*)	SMS_M (*O. basilicum*)
		
SMS_R (*O. tenuiflorum*)	SMS_S (*O. basilicum*)	SMS_T (*O. basilicum*)
		
SMS_V (*O. gratissimum*)	SN_S (*O. tenuiflorum*)	STS_K (*O. basilicum*)*
		
TLM_H (*O. tenuiflorum*)	TLS_K (*O. basilicum*)*	TLS_V (*O. gratissimum*)
		
TS_G (*O. tenuiflorum*)		