ACS OMEGA

Article

Cite This: ACS Omega 2019, 4, 15636−15644

http://pubs.acs.org/journal/acsodf

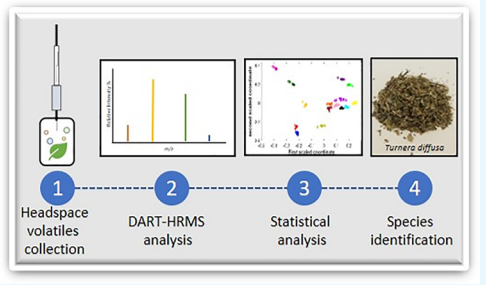
Random Forest Processing of Direct Analysis in Real-Time Mass Spectrometric Data Enables Species Identification of Psychoactive Plants from Their Headspace Chemical Signatures

Meghan Grace Appley, Samira Beyramysoltan, and Rabi Ann Musah*

Department of Chemistry, University at Albany—State University of New York, 1400 Washington Avenue, Albany, New York 12222, United States

Supporting Information

ABSTRACT: The United Nations Office on Drugs and Crime has designated several "legal highs" as "plants of concern" because of the dangers associated with their increasing recreational abuse. Routine identification of these products is hampered by the difficulty in distinguishing them from innocuous plant materials such as foods, herbs, and spices. It is demonstrated here that several of these products have unique but consistent headspace chemical profiles and that multivariate statistical analysis processing of their chemical signatures can be used to accurately identify the species of plants from which the materials are derived. For this study, the headspace volatiles of several species were analyzed by direct analysis in real-time high-resolution mass spectrometry (DART-HRMS). These species include *Althaea officinalis*, *Calea zacatechichi*, *Cannabis indica*, *Cannabis sativa*, *Echinopsis pachanoi*, *Lactuca virosa*, *Leonotis leonurus*, *Mimosa hositlis*, *Mitragyna speciosa*, *Ocimum basilicum*, *Origanum vulgare*, *Piper methysticum*, *Salvia divinorum*, *Turnera diffusa*, and *Voacanga africana*. The results of the DART-HRMS analysis revealed intraspecies similarities and interspecies differences. Exploratory statistical analysis of the data using principal component analysis and global *t*-distributed stochastic neighbor embedding showed clustering of like species and separation of different species. This led to the use of supervised random forest (RF), which resulted in a model with 99% accuracy. A conformal predictor based on the RF classifier was created and proved to be valid for a significance level of 8% with an efficiency of 0.1, an observed fuzziness of 0, and an error rate of 0. The variables used for the statistical analysis processing were ranked in terms of the ability to enable clustering and discrimination between species using principal component analysis−variable importance of projection scores and RF variable importance indices. The variables that ranked the highest were then identified as *m*/*z* values consistent with molecules previously identified in plant material. This technique therefore shows proof-of-concept for the creation of a database for the detection and identification of plant-based legal highs through headspace analysis.



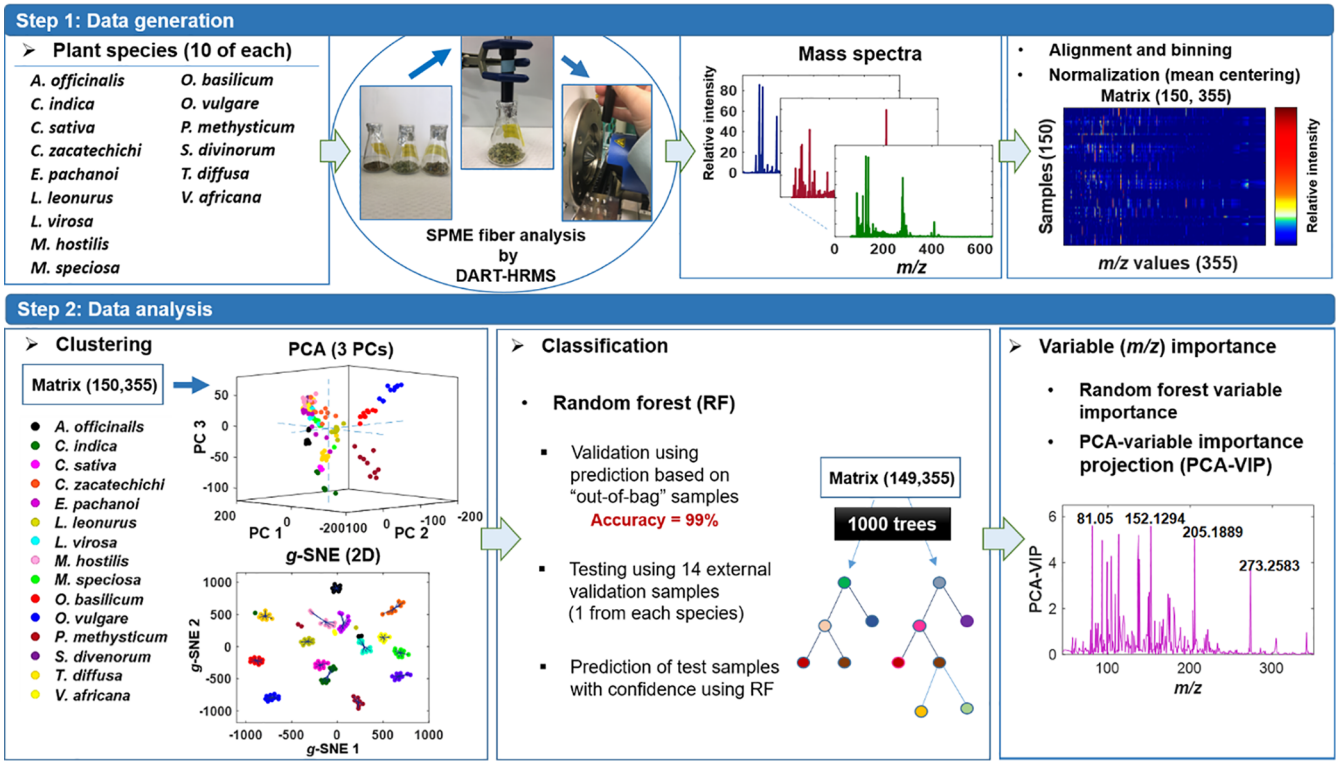
INTRODUCTION

While significant attention has been given in recent years to the surge of the opioid epidemic, the dramatic increase in the abuse of unregulated psychoactive plants remains troublesome. The rising concern is such that the United Nations Office on Drugs and Crime (UNODC) has designated 20 species as plants of concern.[1] These plants are perceived by users to be a more safe and natural alternative to achieving altered states of consciousness than synthetic drugs. Products derived from these materials are readily available through Internet commerce and are difficult to regulate in large part because of the challenge of distinguishing them from innocuous plant materials such as food, spices, and medicinal herbs. Examples of such species include *Salvia divinorum* and *Turnera diffusa*, both endemic to Central and South America, and *Mitragyna speciosa*, native to Southeast Asia.

These drugs are bulk-shipped into the United States in large containers and are often purposefully mislabeled. Because of the difficulty in identifying them, it is impossible for border protection agents to assess the veracity of the species identity listed on the product labels. In principle, a technique that could be exploited for the identification of these materials is headspace analysis. This approach would be successful if the plant materials exhibit headspace volatiles profiles that are consistent for a given plant material but distinct from the headspace of others. The number of studies that have explored this hypothesis is limited. A few reports have shown that a handful of psychoactive materials can be detected and identified through the use of headspace analysis, including cocaine and 3,4-methylenedioxymethamphetamine.[2−4] Additional studies have shown that cannabis can also be detected and identified through the use of headspace analysis by targeting specific compounds.[2,5−7] This technique has also been applied to innocuous plant materials including basil and oregano,[8−10] but the exploration of this approach for the

Received: July 11, 2019
Accepted: August 20, 2019
Published: September 11, 2019

Scheme 1. Steps in the Workflow for the Species Identification of Psychoactive Plants Based on Chemometric Processing of DART-HRMS Data Acquired from Headspace Analysis



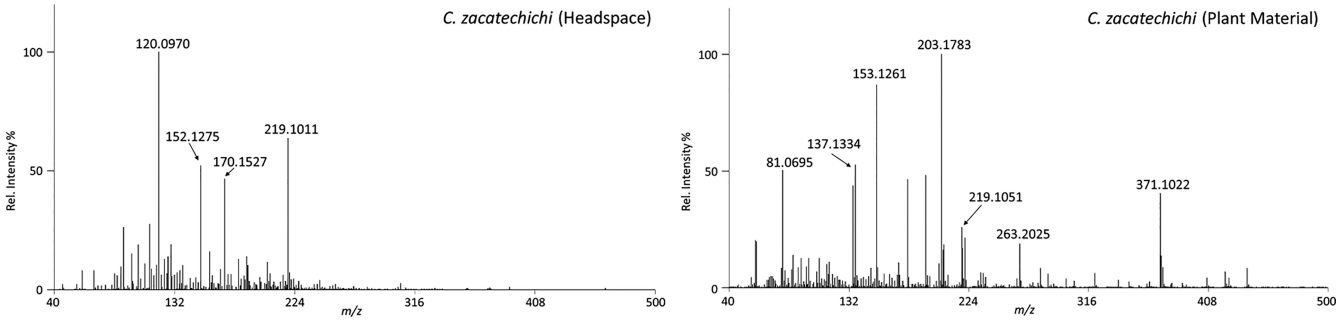


Figure 1. Representative DART mass spectra for the headspace (left panel) and plant material (right panel) analysis of *Calea zacatechichi*.

identification of psychoactive plant-based legal highs has not been reported.

Should it be demonstrated that plant materials exhibit fingerprint profiles that are diagnostic for a given species, it should be possible to create a database of these against which the headspace of unknown materials can be screened to make an identification. The feasibility of creating such a database hinges on being able to generate hundreds of replicates of the requisite data and the development of an appropriate statistical analysis approach for classification. In this regard, the utilization of ambient ionization mass spectral techniques such as direct analysis in real-time high-resolution mass spectrometry (DART-HRMS), shows significant promise in promoting the rapid analysis of samples to generate the data required to create a robust database. For example, previous research shows that DART-HRMS can be used for the identification of different forensically relevant samples including entomological specimens and condom residue evidence, based on the ability to rapidly generate large replicate datasets.[11,12] DART-MS analysis facilitated by the concentration of analytes on solid supports (e.g., sorbents) has previously been reported.[13] Furthermore, the blending of these techniques with headspace collection specifically has been used to detect reaction intermediates induced by plant defense mechanisms in *Mimosa pudica* roots,[14] as well as for the study of the volatile profiles of beers.[15]

Herein, we describe a proof-of-concept for the identification of plant-based legal highs through the use of sorbent-facilitated DART-HRMS analysis and multivariate statistical analysis processing of the generated data.

RESULTS AND DISCUSSION

The overall approach that was devised to accomplish the identification of plant-based materials from headspace analysis is presented in Scheme 1. To assess whether the headspace of psychoactive legal highs exhibits consistent and diagnostic chemical signatures, the headspace volatiles of 11 plant-based legal highs identified by the UNODC as plants of concern, as well as two nonpsychoactive controls (*Ocimum basilicum* and *Origanum vulgare*), were sampled using solid-phase micro-

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

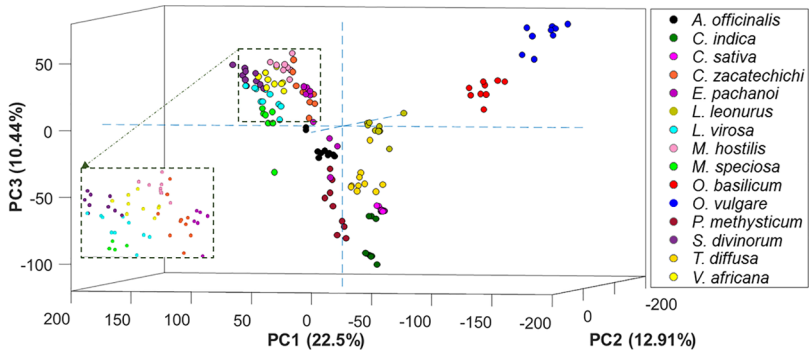


Figure 2. 3-D scores plot featuring principal components (PCs) 1−3 derived from principal component analysis (PCA) of DART-HRMS data generated by analysis of the headspace of each of the indicated species. The score plot displays clear separation for species *O. basilicum*, *O. vulgare*, *L. leonurus*, *C. sativa*, *C. indica*, *T. diffusa*, *P. methysticum*, and *A. officinalis*. The inset, which is enclosed in the smaller rectangle, is expanded for ease of visualization to further illustrate the relationships between the clustered species *M. hostilis*, *M. speciosa*, *S. divinorum*, *L. virosa*, *V. africana*, *C. zacatechichi*, and *E. pachanoi*. The percentage variance accounted for by each of the indicated PCs is shown in parentheses.

extraction (SPME) fibers, which were subsequently analyzed by DART-HRMS in positive-ion mode. Bulk materials derived from plant parts that have historically been used for their psychoactive effects were analyzed (i.e., *Mimosa hostilis* (ground leaves), *Voacanga africana* (ground root bark), *T. diffusa* (ground leaves), *Piper methysticum* (ground leaves), etc.). The headspace of each sample was concentrated on poly(dimethylsiloxane) (PDMS) SPME fibers for 30 min, and this was followed by the analysis of the adsorbed compounds by DART-HRMS (Scheme 1, step 1). Also performed were direct DART-HRMS analyses of the bulk plant material, the results of which served to enable comparison with the headspace results. A representative example of the mass spectra generated from these DART-HRMS experiments is shown in Figure 1, while the DART-HRMS spectra of the headspace and plant-based legal high material are presented in Figure S1. The spectra of the headspace of cannabis species are presented in Figure S2.

In Figures 1 and S1, the left panels show the spectra of the headspace profiles, while those on the right are of the spectra obtained from direct analysis of the bulk material. The results of these analyses revealed two trends. First, multiple replicates of the material of the same species, even when acquired from different sources, exhibited similar headspace small-molecule profiles, and this was also true for the direct analysis of the bulk plant material. For example, the headspace spectra of *Calea zacatechichi* (Figure 1) all contained the m/z values (± 0.005) of 120.0970, 170.1527, and 219.1011, and the bulk material spectra all contained the m/z values (± 0.005) of 137.1318, 203.1768, and 219.1011. Not only did each of the respective spectra have similar m/z values, but they also had similar mass spectral patterns that were unique to that species. The trends seen for *C. zacatechichi* were also observed for the other species analyzed in this study (Figures S1 and S2). This indicated that the replicates representing different samples of the same species showed diagnostic intraspecies similarities and differentiating interspecies distinctions. Second, while there was some duplication of compounds between the plant material and the headspace constituents, the spectra of the two were markedly different. For example, both the mass spectra of the headspace and plant material of *C. zacatechichi* (Figure 1) contain the high-resolution m/z value 219.1011 (± 0.005), which has been identified as representing protonated euparone $[(C_{12}H_{10}O_4) + H^+]$ based on its fragmentation pattern.[16] Similar findings were observed with the other plant species analyzed (Figure S1). This observation was anticipated, since by and large, the headspace signatures would be composed of the subset of compounds contained within the plant materials that are volatilized under ambient conditions. Interestingly, it was also observed that the headspace profiles of two different strains of cannabis could be distinguished visually (Figure S2). This aligns with the previously reported observations.[5,7]

In the spectra of each species (both plant material and headspace), several of the observed high-resolution masses could be correlated to formulas that were consistent with compounds well known to be present in the plant. For example, m/z 137.1330 (± 0.005) is consistent with terpene compounds known to be present in *T. diffusa*,[17] *S. divinorum*,[18] and other plant species.[19] The m/z value 149.0966 (± 0.005) found in *O. basilicum* corresponds to $C_{10}H_{13}O$, which is consistent with the presence of protonated estragole, which has previously been shown to be present in the plant material.[20]

The observations of consistent intraspecies similarities and interspecies differences set the stage for the successful development of a database and a corresponding statistical analysis model that could serve as a screening device against which the chemical fingerprints of unknowns could be compared for species identification purposes. To study the possibility of utilizing plant material headspace for differentiation between species, a mass data matrix that aligned plant-derived DART-HRMS spectra according to common m/z values was created and subjected to statistical analysis processing methods. Thus, as indicated in Scheme 1 (step 1), the mass spectral data from 15 species (in replicates of 10 each, resulting in a total of 150 spectra) were first binned and normalized, yielding a 150 × 355 data matrix (355 represents the total number of m/z values). Then, as indicated in Scheme 1 step 2, principal component analysis (PCA) and global t-distributed stochastic neighbor embedding (g-SNE), as unsupervised methods, were applied to explore and visualize the structure inherent in the data and to reveal clustering of species within a lower-dimensional space. With PCA, the data were resolved to scores and loadings. Figure 2 illustrates the three-dimensional (3-D) score plot (along principal components (PCs) 1−3). These three PCs explained ∼46% of the variance of the data. Each point in the plot corresponds to a sample, and the distances between points reveal the relative level of similarity and dissimilarity between samples. For ease of visualization, each species is represented by a color. From the plot, a clear separation of the species *O. basilicum*, *O.*

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

vulgare, *Leonotis leonurus*, *Cannabis sativa*, *Cannabis indica*, *T. diffusa*, *P. methysticum*, and *Althaea officinalis* is readily apparent. The rectangular panel embedded within the plot is a magnification of the upper-left quadrant and shows that the species *M. hostilis*, *M. speciosa*, *S. divinorum*, *Lactuca virosa*, *V. africana*, *C. zacatechichi*, and *Echinopsis pachanoi* are clustered together.

The plot displayed in Figure 3 shows the results of the application of the *g*-SNE technique in two dimensions. Similar

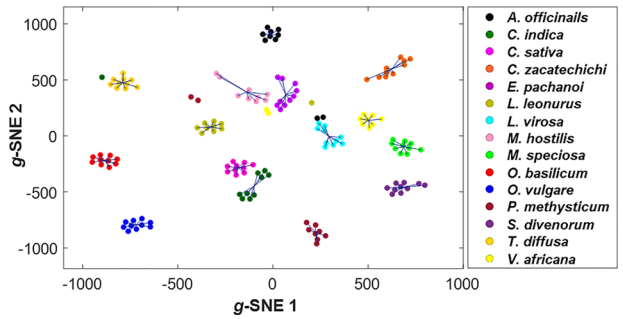


Figure 3. Clustering results observed from the application of global *t*-distributed stochastic neighbor embedding (*g*-SNE) to DART-HRMS data generated from plant headspace analysis. This 2-D rendering shows points that appear in clusters that are color-coded to species. The clustering is based on the relative similarities of the data points that correspond closely to the true labels and illustrates a clear separation of species.

to a PCA score plot, the points define the positions of observations based on the relative *g*-SNE similarities, and each species is defined by a color. The plot illustrates the clustering of the samples of each species and shows a clear separation between them that corresponds closely to the true labels. Of note is the fact that the local similarity relations between species are comparable with the PCA results. However, one sample belonging to the *C. indica* class was observed to be an extreme outlier and was thus removed prior to further analysis.

The results of these exploratory analyses unmasked the hidden discrimination structure between species. Subsequent application of the supervised random forest (RF) technique (using the "RandomForest" package in R) (Scheme 1, step 2, center panel) was performed on the 149 × 355 data matrix to define the discrimination model for the classification of plant species using DART-HRMS data and class labels. The RF method hyperparameters, the minimum number of nodes and the number of variables (m/z) randomly sampled as candidates at each split, were optimized based on a random search of their values within a range. The minimum number of nodes was explored within the range of 1−5, and the number of sampled variables was set to between 20 and 350 variables. Cross-validation (10-fold) of the created RF classifiers was repeated 100 times to find the optimum parameter values that enabled the building of an accurate model. The optimum values were observed to coincide with 1 node for the minimum number of nodes and 55 randomly sampled variables for each split. The RF technique set with these optimized parameters was then performed with different numbers of trees, and in this case, a forest with 1000 trees was found to provide a model with an improved error rate in prediction. The RF algorithm categorizes approximately a third of the dataset as "out-of-the-bag" (OOB) samples (for validation purposes) and performs training with the remaining two-thirds. Thus, the votes for the OOB samples are aggregates of only those decision trees that were not included in the training set. The OOB samples were used to calculate error rates and variable importance values. Figure S3 illustrates the estimated error rate for the OOB classifier on the training set for the grown trees in the RF model. The error converged to a plateau at a value of 0.007 after growing around 382 trees. Table S1 shows the performance results of the discrimination model for each species (i.e., classification precision, sensitivity, and specificity), and it displayed an accuracy of 99% for the OOB sample predictions. The sensitivity and specificity illustrate the true positive and true negative rates for species identification, respectively. The results show that a single sample of *C. indica* was incorrectly predicted to be *E. pachanoi* but that all other observations were identified correctly. This indicates that DART-HRMS analysis of plant-derived headspace in combination with the RF model is a satisfactory approach for identifying plant species.

One of the important properties of RF is the added possibility of computing a "proximity matrix" as a descriptive measure. The proximity matrix quantifies the similarity between samples and is calculated in those instances when two samples are placed in the same terminal node. The results of the application of multidimensional scaling to this distance matrix (1-proximity) to obtain the two principal coordinate components are shown in Figure S4 (with each species assigned a color). Like points were observed to cluster correctly, but the plot also revealed the close similarities between *O. vulgare*, *O. basilicum*, *L. leonurus*, *T. diffusa*, and *P. methysticum* on the one hand and between *E. pachanoi*, *V. africana*, *L. virosa*, *A. officinalis*, and *M. hostilis* on the other. In comparing these results with those obtained by PCA and *g*-SNE, it was deduced that the three methods provide complementary information in presenting the similarities between species, as is described below.

To determine the accuracy of the method for predicting the identity of unknowns, 14 samples were analyzed blindly by DART-HRMS. Their mass spectrometric data were then aligned and binned with the training samples. A conformal predictor based on the RF classifier was created to determine the prediction with an assigned confidence level for each test sample. All training samples were considered as members of the bag of calibration samples, and an off-line experiment using the leave-one-out (LOO) approach was applied. The conformity measure and *p*-values (from eq 1, see Experimental Section) were then calculated for LOO sample prediction. Of the 149 LOO samples, 15 were assigned to multiple classes (at the ε = 8% significance level), but all of the other samples were assigned a single label, which was correct in each case. Thus, the designed conformal predictor proved to be valid for a significance level of 8% with an efficiency of 0.1, an observed fuzziness of 0, and an error rate of 0. The efficiency is the number of multiple predictions over all tested samples, and the observed fuzziness is defined as the sum of all *p*-values for the incorrect class labels. A predictor makes an error when the predicted region does not contain the true label, and the error rate refers to the number of observations predicted incorrectly.

Table 1 presents the performance outcomes for the prediction of the identities of these unknowns, as well as the prediction credibility and confidence level using the RF model. The results show that the true class labels fall within the correct prediction region (with a significance level of 8%) for all unknown samples. The confidence level for the unknown

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

Table 1. Prediction Results for the Indicated 14 Test Samples Representing Each Species[a]

species	prediction	credibility[b]	confidence level[c]
A. officinalis	true	0.09	1
C. indica	true	0.27	1
C. sativa	true	0.09	0.91
C. zacatechichi	true	0.9	1
E. pachanoi	true	0.09	0.91
L. leonurus	true	0.09	1
L. virosa	true	1	1
M. hostilis	true	0.09	0.91
M. speciosa	true	0.45	1
O. basilicum	true	0.09	1
O. vulgare	true	0.18	0.91
P. methysticum	true	0.45	1
S. divinorum	true	0.36	1
T. diffusa	true	0.55	1

[a]The credibility and confidence levels are reported for each. [b]Credibility corresponds to the highest computed *p*-value. [c]Confidence level refers to 1 minus the second-highest *p*-value.

samples representing *M. hostilis*, *O. vulgare*, *E. pachanoi*, and *C. sativa* samples indicates that the *p*-value for some other class(es) should be 0.09. The calculated *p*-value for each species is displayed for each sample in Table S2. The table illustrates that the four aforementioned samples can each be classified as members of two species.

Aiming to rank the variables in terms of their ability to facilitate clustering and discrimination between species, the importance of the variables was quantified using principal component analysis−variable importance of projection (PCA−VIP) scores and RF variable importance indices (Scheme 1, step 2, last panel). The importance of the primary variables identified by PCA as contributing to the maximum variance are defined in eq 3 (see Experimental Section). The average relative importance of the variables (m/z values) in the bootstrap analysis PCA−VIP for the three principal components (which accounted for ∼46% of the variance of the data) is illustrated in Figure 4a, in which the 30 most important m/z values with PCA−VIPs are labeled. These include monoterpenoids (β-myrcene, camphene, β-pinene, β-phellandrene, γ-terpinene, and α-pinene at m/z 137.1096 in *O. basilicum*, *C. zacatechichi*, *C. zacatechichi*, *T. diffusa*, *T. diffusa*, and *L. leonurus*, respectively), sesquiterpenoids (α-curcumene at m/z 203.1789 in *C. zacatechichi* and *T. diffusa*; *trans*-α-bergamotene, caryophyllene, and β-sesquiphellandrene at m/z 205.1889 in *O. basilicum*, *T. diffusa*, and *T. diffusa*, respectively), and estragole (at m/z 149.0895 in *O. basilicum*). In addition, the permutation-based importance of predictive variables in the 10 repeats of the RF modeling was applied to show which m/z values were useful for discrimination between plant species. All variables (m/z values) were considered for all of the trees in all 10 RF classifiers. Each variable's importance is the average of the importance values derived from the classifiers. The bar plot in Figure 4b displays the rankings for the 30 most important variables computed by this method.

In comparison to the PCA−VIP results, it is noteworthy that 40% of the m/z values detected by PCA−VIP aligned with those that emerged by RF modeling. For visualization of this correspondence, Figure 5 illustrates the 3-D loading plot created using the first three PCs, along with the marked loadings for the important m/z values detected by PCA and RF analyses. The solid navy points in the figure show the loadings for 355 variables, while the magenta stars and red circles are markers for m/z values and loadings that were derived from PCA and RF, respectively. This rendering makes apparent that both methods furnish similar results and that about 40% of the m/z values that emerged in RF analysis as important were also essential in explaining the maximum variance of the data. Table

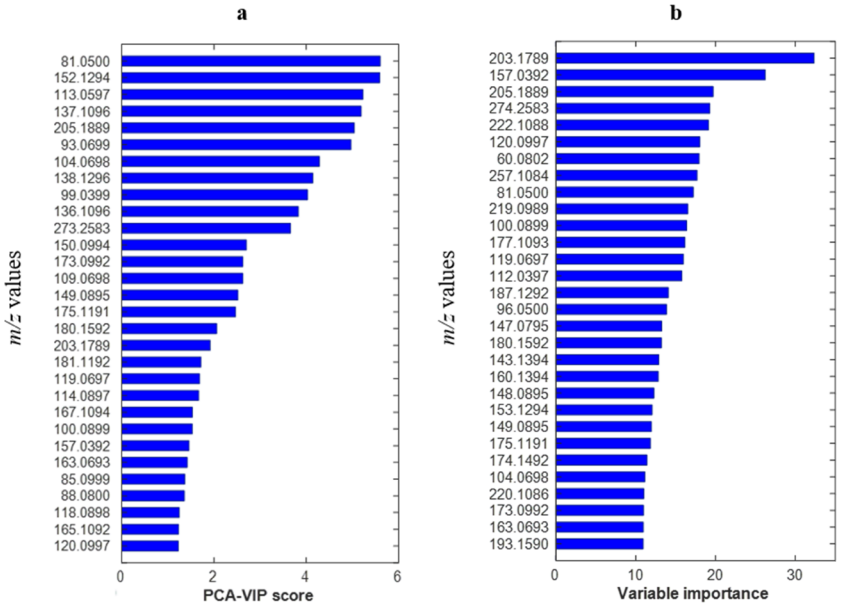


Figure 4. Values (30 m/z) observed to be most important for enabling clustering and species discrimination, calculated using PCA and RF modeling of DART-HRMS-derived data from the analysis of plant headspace. (a) Variables (m/z values) of importance in discrimination, revealed through bootstrap PCA−VIP analysis based on the three principal components, which explained ∼46% of the variance of the data, and their corresponding average scores. (b) The m/z values important for discrimination were extracted using permutation-based importance of predictive variables in RF. In both panels, the m/z values are listed in the order of decreasing PCA−VIP scores and variable importance RF values.

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

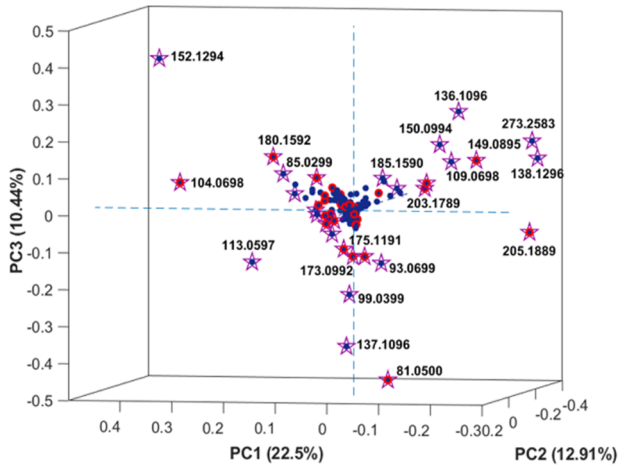


Figure 5. Equivalent semantic relationships between PCA–VIP and RF variable importance methods from within the set of important predictors (m/z values), rendered as a 3-D loading plot. The navy points display the loadings of 355 m/z values. The loadings of the m/z values representing the top-ranking variables obtained from the RF and PCA–VIP analyses are indicated with red circles and magenta stars, respectively. The observed overlap of circles and stars illustrates alignment in the predictions of the two methods regarding the m/z variables that were the most important contributors to the ability to differentiate between species.

S3a–f reports the average relative intensities for m/z values that were ranked by both methods to be important.

From the point of view of the local variable importance for clustering of species based on PCA score and loading plots (Figures 2 and 5, respectively), m/z values 152.1294, 104.0698, 180.1592, and 85.0299 were important in the clustering of *M. hostilis*, *M. speciosa*, *S. divinorum*, *L. virosa*, *V. africana*, *C. zacatechichi*, and *E. pachanoi*. The m/z values 81.0500, 137.1096, 99.0399, 93.0699, 173.0992, and 175.1191 were important for the detection of the similarities between *L. leonurus*, *C. sativa*, *C. indica*, *T. diffusa*, *P. methysticum*, and *A. officinalis*, respectively.

Table S4a–c lists information on the characteristics of the important variables and shows the 20 most important discriminating features for each species that were revealed by the RF approach and which represent the mean of the importance of each variable in the samples belonging to each species. These m/z values illustrate the features that were significant for enabling the discrimination of a specific species from all other species. However, it should be noted that these variables do not necessarily match with those indicated in Figure 4 and Table S3 that enabled the creation of the classification model. This is because there were two types of investigations accomplished using the RF results. One was differentiation of a given species from the 14 others that were the subject of the investigation. The m/z values that enabled the accomplishment of this were described as being of "local" importance and are listed in Table S4. The second enabled discrimination between all species simultaneously such that the discrimination between species could be readily visualized through the clustering observed in Figure S4 (two-dimensional (2-D) plot of the proximity matrix analysis). The m/z values associated with this type of classification are described here as "global" and appear in Figure 4. As the two types of exploration accomplish different tasks, the variables that are most heavily weighted in achieving the two types of classification are not necessarily the same.

The results of this study reveal that the headspace volatiles of the legal high plant materials analyzed in this study exhibit consistent and unique chemical profiles, the constituents of which can be concentrated using solid-phase microextraction fibers. The results are highly accurate despite the SPME-facilitated volatiles collection having been performed at ambient (as opposed to elevated) temperature and the data variability inherent in the manual DART-MS analysis process. The mass spectra observed were remarkably consistent for samples of the same class. Their chemical signatures, rapidly acquired by DART-HRMS analysis, can then be subjected to multivariate statistical analysis using a conformal predictor based on a random forest model, to predict the species identifies of plant material unknowns at a significance level of 8%, an efficiency of 0.1, an observed fuzziness of 0, and an error rate of 0. This is important, in that it shows proof-of-concept for the creation of a headspace chemical profile database, which can be used to rapidly screen headspace mass spectra of unknowns, to identify plant-based legal highs.

EXPERIMENTAL SECTION

Plant Material. Dried samples of *A. officinalis* leaves, *C. zacatechichi* leaves, *L. virosa* leaves, *L. leonurus* flowering material, and *V. africana* root bark were purchased from World Seed Supply (Mastic Beach, NY). Dried *M. hostilis* root bark was purchased from Mr. Botanicals (MrBotanicals.com). Dried *P. methysticum* root powder and *T. diffusa* leaves were purchased from Bouncing Bear Botanicals (Lawrence, KS). Dried *M. speciosa* leaves were purchased from Kratom King (Reno, NV). Dried *O. basilicum* leaves and *O. vulgare* leaves were purchased from Hannaford Bros. Co. (Scarborough, ME). A fresh *E. pachanoi* plant was purchased from World Seed Supply (Mastic Beach, NY) and then cut and dried. A fresh *S. divinorum* plant was purchased from Undergroundroots.net (La Conner, WA) and then cut and dried. Cannabis samples (i.e., *C. sativa* and *C. indica*) were analyzed at the U.S. Fish and Wildlife Forensics Laboratory (Ashland, OR).

Solid-Phase Microextraction Fibers. Divinylbenzene/carboxen/poly(dimethylsiloxane)-coated 24 ga 50/30 μm solid-phase microextraction fibers and solid-phase microextraction fiber holders for use with manual sampling were purchased from Supelco Inc. (Bellefonte, PA). Fibers were conditioned for 30 min at 250 °C under a stream of helium gas before each headspace sampling.

Headspace Sampling. Roughly 10 g of each plant species was placed in separate 25 mL Erlenmeyer flasks. The mouth of the flask was covered with aluminum foil. A conditioned solid-phase microextraction fiber was then exposed to the headspace of the sample for 30 min at room temperature (Figure 6). This concentration step was performed under ambient conditions (rather than at elevated temperature) to detect volatile components that are more likely to be observed under the ambient conditions present in the vessels containing the samples or within the general vicinity of the samples (in the field). Each of the plant samples was analyzed in replicates of 10. Spectra of *C. sativa* and *C. indica* headspace were acquired by transferring the samples to a 20 mL scintillation vial and placing it uncapped between the ion source and the mass spectrometer inlet.

DART-HRMS Analysis. Exposed SPME fibers were analyzed using a direct analysis in real-time (DART)-SVP

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636–15644

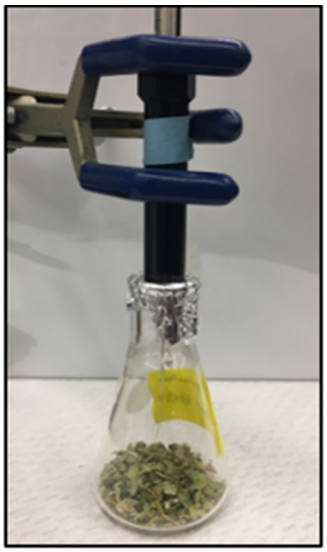


Figure 6. Headspace volatile collection using an SPME fiber.

ion source (IonSense, Saugus, MA) interfaced with a JEOL AccuTOF mass spectrometer (JEOL USA, Peabody, MA). Each fiber, while extended from the holder assembly, was manually "waved" back and forth in the DART gas stream until there was no longer an MS signal that was registered (which signified that the content of the fiber had been fully desorbed and which took ∼1 min) (Figure 7). The fibers were analyzed

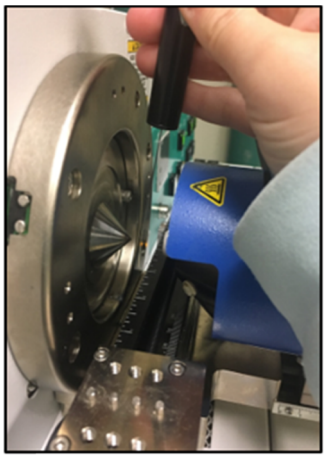


Figure 7. SPME fiber introduction to the DART gas stream.

in positive-ion mode with the gas heater temperature in the DART software set to 250 °C, over a mass range of m/z 40−800. The DART ion source helium flow rate was 2.0 L/min. The mass spectrometer settings were as follows: the orifice 1 voltage was 20 V, the orifice 2 voltage was 5 V to minimize fragmentation, and the peak voltage was 400 V to allow for the detection of ions over m/z 40. The mass spectrometer has a resolving power of 6000 full width at half maximum. Poly(ethylene glycol) (PEG 600) was used to calibrate the mass spectra following the analysis of each individual fiber. Plant material for each species was also analyzed directly using the same DART parameters as the SPME fibers for comparison.

Spectral Analysis. Calibration, background subtraction, and peak centroiding were conducted using TSSPro3 software (Schrader Analytical Laboratories, Detroit, MI). Mass spectral analysis was performed using Mass Mountaineer (Mass-spec-software.com, RBC Software, Portsmouth, NH). The DART mass spectrum of a conditioned SPME fiber that was not exposed to the headspace of any samples was used as a blank for the SPME samples.

Statistical Analysis. To model discrimination between plant species and to discover which features (m/z values) are most important for distinguishing between them, multivariate statistical analysis methods were applied to the DART-HRMS data acquired from the analysis of plant samples. The workflow outlined in Scheme 1 illustrates the approach.

In step 1, SPME fiber-facilitated DART-HRMS was used to generate a mass spectrum for each sample, with the analysis performed using multiple species and 10 replicates. In all, the mass spectra of 150 samples representing 15 different species were imported into MATLAB 9.3.0, R2017b Software (The MathWorks, Inc., Natick, MA), in text format (composed of m/z values and their corresponding intensities) for further analysis in MATLAB and R 3.5.1 (http://cran.r-project.org/). A data matrix with the dimensions 150 × 355 was created from binning of mass spectra, with the optimal bin width and the relative abundance threshold being ±10 mmu and 0.2%, respectively. In step 2, the data matrix was subjected to descriptive and predictive methods to reveal information on species in terms of discriminative markers. This step consisted of three parts: exploration, classification, and determination of variable (m/z) importance, detailed below.

Exploration. An extended form of *t*-distributed stochastic neighbor embedding, termed "*g*-SNE", was used to visualize the data structure in a 2-D scatter plot. This neighbor-embedding technique preserves the pairwise similarities of probable neighbors by minimizing the divergence of similarity distributions between neighboring data points and embedding the points in a lower-dimensional space. The dataset was subjected to principal component analysis (PCA) to explore its similarity structure and to reveal the m/z values which were the primary indicators of similarities and dissimilarities between like and unlike groups, respectively.

Classification. The random forest (RF) technique proposed by Breiman was investigated as a plant species discrimination model.[21] Random forest is a classifier which aggregates a large number of "trees" to reduce overfitting and preserve reliable predictions. Every tree in the forest is "grown" on an independently drawn bootstrap replica of the data matrix and assigned a vote for each class (i.e., the estimated probability of the observation originating from the given class) at each input sample. The samples not included in the replica for a given tree are considered to be "out-of-bag" (OOB) for that tree. The overall accuracy and the performance characteristics of the model are computed based on the predictions of OOB observations. For the prediction of new samples, a conformity measure was used to yield a confidence level prediction based on a random forest classifier.[22] Conformal prediction provides the opportunity to have output region predictions (i.e., a set of predicted labels) with a guaranteed error rate based on the calculated *p*-value. The conformity score for a given observation i in the bag (i.e., the calibration set in the conformal prediction context) for a specific class k (designated as α_i^k) is the proportion of votes of all of the trees for a given class k. The result is a matrix of conformity scores with one row per observation and one column per class.

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

$$p(\alpha_{m+1}^{k}) = \frac{\text{count}\{i \in \{1, ..., m\} | y_i = k \text{ and } \alpha_i^k \leq \alpha_{m+1}^k\} \text{ and } \{i \in \{1, ..., m\} | y_i \neq k \text{ and } \max(\alpha_i^k) \leq \alpha_{m+1}^k\}}{\text{count}\{i \in \{1, ..., m\} | y_i = k\} + 1}, \text{ where } k \in \{1, ..., \text{nc}\} \quad (1)$$

The parameters m and nc indicate the number of samples in the bag and classes, respectively. The resulting scores were then used to calculate the p-value for the labeling of an unknown sample representing a given species, according to eq 1. To calculate the p-value for observation "$m + 1$" for a specific class k (represented as $p(\alpha_{m+1}^k)$), the conformity score of the observation for class k (α_{m+1}^k) was computed and compared with the observations' in-bag scores for class k with the following conditions: the scores of observations belong to class k (α_i^k, $i \in 1,...,m | y_i = k$), and the maximum conformity measure of the observations does not belong to class k (max (α_i^k), $i \in 1,...,m | y_i \neq k$). In the case of single-label predictions, the confidence of the prediction is one minus the second-largest p-value, and the credibility is the largest p-value.

Variable Importance. PCA and RF results were explored to deduce the relative importance of the various m/z values in enabling the clustering of and discrimination between plant species. This was accomplished by generating variable importance of projection (VIP) scores, as proposed by Ginsburg et al.[23] VIPs enable the consideration of the structure of the reduced dimensional PCA space and the class labels according to eqs 2 and 3, where **T**, **P**, y, and b (in eq 2) are the scores, loadings, class labels, and regression coefficients between class labels and scores, respectively. Equation 2 represents the decomposition of the mass data matrix into scores (**T**) and loadings (**P**) matrices, and regression between scores (**T**) and class labels (y). Equation 3 displays the computation equation for VIP scores. The terms npc, m, and nv define the number of principal components, samples, and m/z values, respectively.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \quad y = \mathbf{T}b^{\mathrm{T}} \quad (2)$$

$$\text{VIP}_j = \sqrt{m \frac{\sum_{i=1}^{\text{npc}} b_i^2 t_i^{\mathrm{T}} t_i \left(\frac{p_{ij}}{\|p_i\|}\right)}{\sum_{i=1}^{\text{npc}} b_i^2 t_i^{\mathrm{T}} t_i}}, \text{ where } i \in \{1, ..., \text{npc}\} \text{ and } j \in \{1, ..., \text{nv}\} \quad (3)$$

PCA−VIP scores were calculated by randomized bootstrapping (1000 repetitions), with 80% of the samples used to create a PCA model in each repeat. Determination of the m/z values that were most important in enabling discrimination between sample types was accomplished by defining an importance measure (permutation-based variable importance) that was embedded in the OOB observations in the RF model. The score of a given variable was computed as the average decrease in model accuracy of the OOB samples when the values of the corresponding variable were randomly permuted across the OOB observations. Therefore, for each variable in every tree grown, the difference in the percentage of two votes for the correct class of the OOB observations was measured: a vote for the untouched OOB data and another vote for the variable permuted OOB data. The average of this measure for all of the trees in the ensemble represented the importance score for each variable (i.e., m/z value).[24]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.9b02145.

Representative mass spectra; out-of-bag error for random forest; visualization of proximity matrix for random forest; performance results and species-specific variables for species discrimination; p-values from the RF conformal predictor; average relative intensities for m/z values (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rmusah@albany.edu. Tel: 518-437-3740.

ORCID

Rabi Ann Musah: 0000-0002-3135-4130

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The financial support of the U.S. National Institute of Justice to M.G.F. (grant 2018-R2-CX-0012) and R.A.M. (grants 2015-DN-BX-K057 and 2018-R2-CX-0012), as well as the U.S. National Science Foundation to R.A.M. (grant 1429329), is gratefully acknowledged. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. The authors extend their thanks to Dr. Robert Cody for the analysis of the cannabis samples.

■ REFERENCES

(1) Hammond, B.; Crean, C.; Levissianos, S.; Mermerci, D.; Tun Nay, S.; Otani, T.; Park, M.; Pazos, D.; Piñeros, K.; Umapornsakula, A.; Wong, Y. L.; Chawla, S. In *The Challenges of New Psychoactive Substances*, UNODC Global SMART Programme, 2013.

(2) Lai, H.; Corbin, I.; Almirall, J. R. Headspace sampling and detection of cocaine, MDMA, and marijuana via volatile markers in the presence of potential interferences by solid phase microextraction-ion mobility spectrometry (SPME-IMS). *Anal. Bioanal. Chem.* **2008**, *392*, 105−113.

(3) Gura, S.; Guerra-Diaz, P.; Lai, H.; Almirall, J. R. Enhancement in sample collection for the detection of MDMA using a novel planar SPME (PSPME) device coupled to ion mobility spectrometry (IMS). *Drug Test. Anal.* **2009**, *1*, 355−362.

(4) Viana, M.; Postigo, C.; Querol, X.; Alastuey, A.; López de Alda, M. J.; Barceló, D.; Artíñano, B.; López-Mahia, P.; García Gacio, D.; Cots, N. Cocaine and other illicit drugs in airborne particulates in urban environments: A reflection of social conduct and population size. *Environ. Pollut.* **2011**, *159*, 1241−1247.

(5) Arnoldi, S.; Roda, G.; Casagni, E.; Dell'Acqua, L.; Cas, M. D.; Fare, F.; Rusconi, C.; Visconti, G. L.; Gambaro, V. Characterization of the volatile components of cannabis preparations by solid-phase microextraction coupled to headspace-gas chromatography with mass detector (SPME-HSGC/MS). *J. Chromatogr. Sep. Tech.* **2017**, *8*, No. 350.

(6) Pellati, F.; Brighenti, V.; Sperlea, J.; Marchetti, L.; Bertelli, D.; Benvenuti, S. New methods for the comprehensive analysis of

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644

bioactive compounds in *Cannabis sativa* L. (hemp). *Molecules* **2018**, *23*, No. 2639.
(7) Stenerson, K. K.; Halpenny, M. R. Analysis of terpenes in cannabis using headspace solid-phase microextraction and GC-MS. *LCGC North Am.* **2017**, *35*, 28.
(8) Díaz-Maroto, M. C.; Pérez-Coello, M. S.; Cabezudo, M. D. Headspace solid-phase microextraction analysis of volatile components of spices. *Chromatographia* **2002**, *55*, 723−728.
(9) Gao, B.; Qin, F.; Ding, T.; Chen, Y.; Lu, W.; Yu, L. L. Differentiating organically and conventionally grown oregano using ultraperformance liquid chromatography mass spectrometry (UPLC-MS), headspace gas chromatography with flame ionization detection (headspace-GC-FID), and flow injection mass spectrum (FIMS) fingerprints combined with multivariate data analysis. *J. Agric. Food Chem.* **2014**, *62*, 8075−8084.
(10) Asadollahi-Baboli, M.; Aghakhani, A. Headspace adsorptive microextraction analysis of oregano fragrance using polyaniline-nylon-6 nanocomposite, GC-MS, and multivariate curve resolution. *Int. J. Food Prop.* **2015**, *18*, 1613−1623.
(11) Beyramysoltan, S.; Giffen, J. E.; Rosati, J. Y.; Musah, R. A. Direct analysis in real time-mass spectrometry and kohonen artificial neural networks for species identification of larva, pupa and adult life stages of carrion insects. *Anal. Chem.* **2018**, *90*, 9206−9217.
(12) Coon, A. M.; Beyramysoltan, S.; Musah, R. A. A chemometric strategy for forensic analysis of condom residues: Identification and marker profiling of condom brands from direct analysis in real time-high resolution mass spectrometric chemical signatures. *Talanta* **2019**, *194*, 563−575.
(13) Gómez-Ríos, G. A.; Pawliszyn, J. Solid phase microextraction (SPME)-transmission mode (TM) pushes down detection limits in direct analysis in real time (DART). *Chem. Commun.* **2014**, *50*, 12937−12940.
(14) Musah, R. A.; Lesiak, A. D.; Maron, M. J.; Cody, R. B.; Edwards, D.; Fowble, K. L.; Dane, A. J.; Long, M. C. Mechanosensitivity below ground: Touch-sensitive smell-producing roots in the shy plant *Mimosa pudica*. *Plant Physiol.* **2016**, *170*, 1075−1089.
(15) Cajka, T.; Riddellova, K.; Tomaniova, M.; Hajslova, J. Recognition of beer brand based on multivariate analysis of volatile fingerprint. *J. Chromatogr. A* **2010**, *1217*, 4195−4203.
(16) Elsohly, M. A.; Knapp, J. E.; Slatkin, D. J.; Schiff, P. L.; Doorenbos, N. J.; Quimby, M. W. Euparone, a new benzofuran from *Ruscus aculeatus* L. *J. Pharm. Sci. A.* **1974**, *63*, 1623−1624.
(17) Szewczyk, K.; Zidorn, C. Ethnobotany, phytochemistry, and bioactivity of the genus Turnera (Passifloraceae) with a focus on damiana—*Turnera diffusa*. *J. Ethnopharmacol.* **2014**, *152*, 424−443.
(18) Casselman, I.; Nock, C. J.; Wohlmuth, H.; Weatherby, R. P.; Heinrich, M. From local to global—Fifty years of research on *Salvia divinorum*. *J. Ethnopharmacol.* **2014**, *151*, 768−783.
(19) Mykhailenko, O.; Kovalyov, V.; Goryacha, O.; Ivanauskas, L.; Georgiyants, V. Biologically active compounds and pharmacological activities of species of the genus Crocus: A review. *Phytochemistry* **2019**, *162*, 56−89.
(20) Bucar, F.; Muller, G.; Weissenbacher, D. Investigations of basil cultivars on estragole content. *Spec. Publ. - R. Soc. Chem.* **2001**, *269*, 277−279.
(21) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(22) Devetyarov, D.; Nouretdinov, I. In *Prediction with Confidence Based on a Random Forest Classifier*, International Conference on Artificial Intelligence Applications and Innovations, Larnaca, Cyrus, 2010; pp 37−44.
(23) Ginsburg, S.; Tiwari, P.; Kurhanewicz, J.; Madabhushi, A. Variable Ranking with PCA: Finding Multiparametric MR Imaging Markers for Prostate Cancer Diagnosis and Grading. In *Prostate Cancer Imaging. Image Analysis and Image-Guided Interventions*; Madabhushi, A., Dowling, J., Huisman, H., Barratt, D., Eds.; Springer: Berlin; 2011.
(24) Breiman, L.; Cutler, A. Random Forests, https://www.stat.berkeley.edu/breiman/RandomForests/cc_home.htm.

DOI: 10.1021/acsomega.9b02145
ACS Omega 2019, 4, 15636−15644