analytical chemistry

pubs.acs.org/ac Article

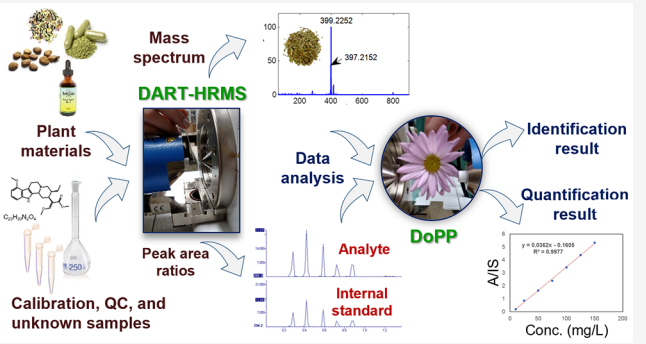
Introducing "DoPP": A Graphical User-Friendly Application for the Rapid Species Identification of Psychoactive Plant Materials and Quantification of Psychoactive Small Molecules Using DART-MS Data

Samira Beyramysoltan, Megan I. Chambers, Amy M. Osborne, Mónica I. Ventura, and Rabi A. Musah*

Cite This: *Anal. Chem.* 2022, 94, 16570−16578 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The widespread abuse of "legal high" psychoactive plants continues to be of global concern because of their negative impacts on public health and safety. In forensic science, a major challenge in controlling these substances is the paucity of methods to rapidly identify them. We report the development of the Database of Psychoactive Plants (DoPP), a new user-friendly tool featuring an architecture for the identification of plant unknowns, and the necessary regression statistics for the development and validation of psychoactive compound quantification. The application relies on the knowledge that terrestrial plants exhibit species-specific chemical signatures that can be revealed by direct analysis in real time—high-resolution mass spectrometry (DART-HRMS). Subsequent automated machine learning processing of libraries of these spectra enables rapid discrimination and species identification. The chemical signature database includes 57 available plant species. The rapid acquisition of mass spectra and the ability to sample the materials in their native form enabled the generation of the vast amounts of spectral replicates required for database construction. For the identification of sample unknowns, a data analysis workflow was developed and implemented using the DoPP tool. It utilizes a hierarchical classification tree that integrates three machine learning methods, namely, random forest, k-nearest neighbors, and support vector machine, all of which were fused using posterior probabilities. The results show accuracies of 98 and 99% for 10-fold cross-validation and external validation, respectively, which make the classification model suitable for identity prediction of real samples.



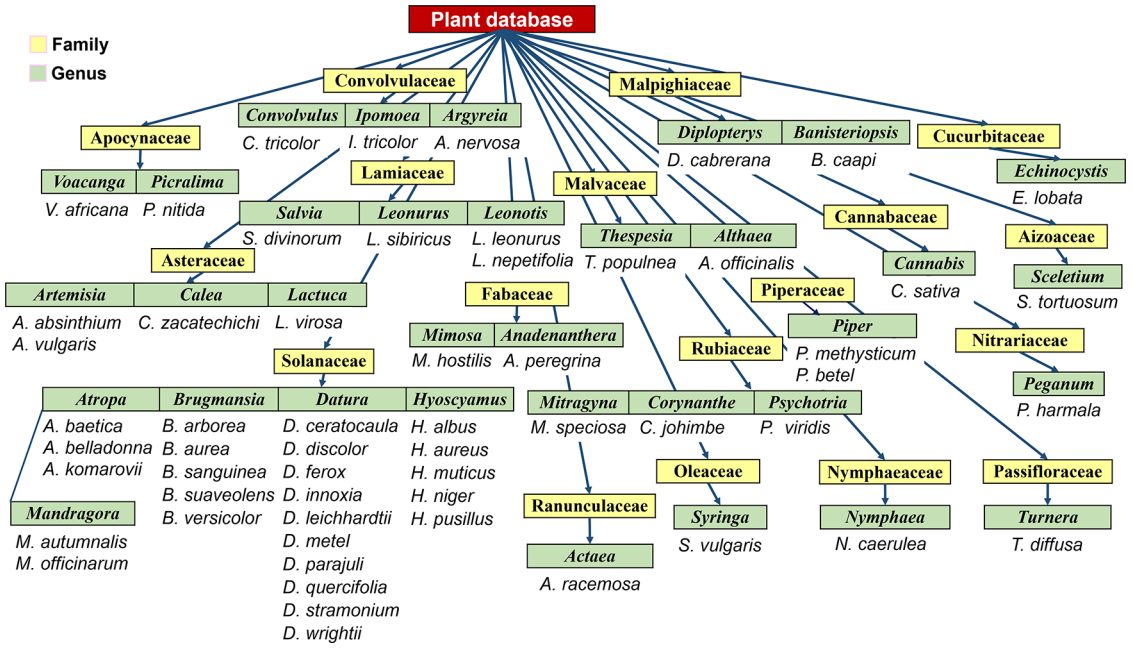
INTRODUCTION

One of the continuing challenges in analytical chemistry is the paucity of efficient approaches for the rapid identification of plant-derived complex matrices. This is of particular relevance in forensics where the ingestion of psychoactive plant materials can cause impairment that leads to the commission of crimes, the improper handling of machinery resulting in workplace accidents, driving under the influence, agitation and disorientation leading to violence, and mental and physical health challenges that can result in death.[1−3] Because of its relevance to possible criminal activity or liability, it is essential that the species identity of the plant material that was ingested be known. Although such determinations are relatively straightforward for the small number of mind-altering plants that have physical characteristics that are readily recognized by visual examination (e.g., observation of cystolithic hairs unique to *Cannabis sativa*), the vast majority of psychoactive plants and the materials derived from them (e.g., crumbled leaves and other aerial parts, seeds, tinctures, extracts, etc.) do not have distinguishing features that enable them to be readily differentiated from innocuous products such as foods and spices. Some psychoactive plants have served as sources of modern-day drugs that continue to be clinically relevant, such as atropine and scopolamine from *Datura* species plants.[4,5] However, the vast majority of known psychoactive plants are typically regarded as dangerous, with no generally accepted clinical use. It is for this reason that the active small-molecule components of many of these plants, when known, have been scheduled. Those shown to have addictive properties and no established medical use have been designated as Schedule I drugs and those that are addictive but have clinical utility are categorized as Schedule II.[6] Examples of the former include ibogaine found in plants in the Apocynaceae family and *N,N*-

Received: April 12, 2022
Accepted: October 11, 2022
Published: November 17, 2022

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

Scheme 1. Plant Species Represented in the DoPP Platform and the Taxonomical Relationships between Them



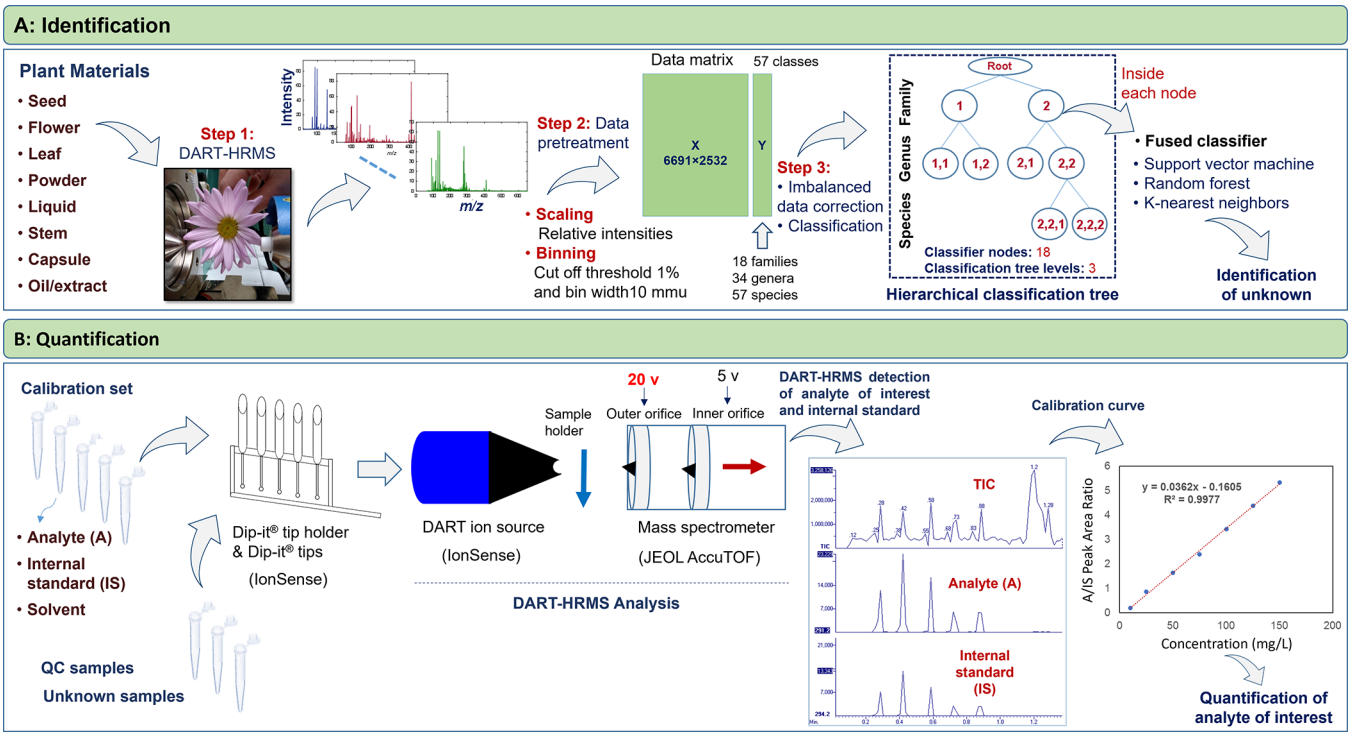
dimethyltryptamine (DMT) found in multiple species such as *Mimosa hostilis (tenuiflora)*, *Diplopterys cabrerana*, and *Psychotria viridis*. Examples of the latter include atropine and scopolamine, which are found in many plants in the *Datura* genus. Ironically, while the purified forms of most of the known addictive small-molecule natural products are scheduled, the plants from which many of them are derived are not. For instance, atropine and scopolamine are Schedule II drugs, but the *Datura* spp. plants that contain them are not. For this reason, the plants are known as "legal highs", because unlike their purified active components, in most countries/U.S. states, they can be possessed and ingested without fear of prosecution.[7,8] The exponential rise in the abuse of these dangerous materials has raised alarm and caused the United Nations Office on Drugs and Crime (UNODC) to declare 20 species as "plants of concern", including *Mitragyna speciosa* and *Salvia divinorum*.[9] An important prerequisite to the legislation of the manufacture, sale, distribution, and ingestion of these substances is the ability to identify them rapidly and definitively. However, a systematic way in which to routinely accomplish this for the ever-increasing range of plant materials and their evolving forms has proven elusive. This is because: (1) the plant materials themselves often do not possess distinguishing features, making them unrecognizable in a forensic context; (2) standard well-established analytical methods (such as GC−MS and LC−MS) that are useful in the identification of purified or semi-purified substances are time-consuming to perform on the whole plant material and/or have not been developed for the analysis of whole plant products; (3) there is generally no statistical reporting of the level of certainty of positive identification of a particular plant drug based on screening it against a bona fide database; and (4) unlike the case for purified compounds for which libraries of spectroscopic and mass spectrometric data are available that can serve to facilitate confirmation of the structures of unknowns, there is no available analogous database with accompanying software to aid in the rapid detection of plant materials. Therefore, there is an urgent need for the development of a rapid analysis approach that circumvents some of the present challenges associated with the identification of dangerous psychoactive plant-derived substances. In addition, since the amounts of the scheduled molecules that are contained within the bulk psychoactive materials will influence sentencing guidelines, an analysis approach that also enables quantification of the active compounds present is highly desirable.[3]

Previous studies have shown that direct analysis in real time−high-resolution mass spectrometry (DART-HRMS), with minimal if any sample preparation required, reveals within a single analysis of the bulk material a range of detected molecules extending across the dielectric constant spectrum.[10−16] Furthermore, it has been shown that when analyzed by DART-HRMS, plants exhibit species-specific chemical signatures that can be utilized to predict the identities of species within a given genus, using advanced statistical analysis tools.[5,17−21] These findings imply the possibility that the application of machine learning tools to a library of DART-HRMS-derived species-specific chemical signatures might provide a mechanism to predict the species identity of plant material unknowns with a statistical level of certainty. Furthermore, this same instrumental technique can be used to quantify the psychoactive molecules present.[22−25] In principle, it could provide a more universal approach for the identification of new psychoactive materials, rather than relying on current conventional methods, which require nuanced method development that is also time- and resource-intensive. Importantly, the analysis can be conducted in less than 1 min per sample.

Reported here for the first time is the accomplishment of two main aims: (1) the development of a DART-HRMS chemical signatures database of available psychoactive plants; and (2) the development of a user-friendly and intuitive data analysis tool for the rapid identification of unknown materials and quantification of the psychoactive compounds contained within them [termed Database of Psychoactive Plants (DoPP)]. The application allows users to simply import the

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

Scheme 2. Overview of the Data Analysis Workflow for Psychoactive Plant Materials



DART-HRMS data of the unknown into the platform, which then reveals species identity with a statistical level of certainty. It can also be used for the quantification of the psychoactive components present. The performance of the application is demonstrated using commercial psychoactive plant samples, and the quantification feature is illustrated for the determination of DMT concentrations in the plant material.

MATERIALS AND METHODS

Materials. Plant materials representing 18 families, 34 genera, and 57 species, including various plant parts (e.g., seeds, flowers, roots, leaves, bark, roots, and stems) and processed products such as resins, powders, extracts, and capsules from different vendors, were analyzed. Detailed information on the analyzed plants, including order, family, genus, and species, as well as the material type and vendor, is presented in Table S1. Scheme 1 illustrates taxonomical relationships between families, genera, and species of the represented plants, with the families and genera highlighted in yellow and light green boxes, respectively.

Instrumentation. A DART-SVP ion source (IonSense Inc., Saugus, MA, USA) coupled with a JEOL AccuTOF high-resolution time-of-flight mass spectrometer (JEOL USA, Peabody, MA, USA) operating in positive-ion mode was used to collect spectra in the range m/z 40−1100 (as indicated in Scheme 2A-Step 1). Mass spectrometer settings were as follows: gas heater temperature, 350 °C; orifice 1, 20 V; orifice 2, 5 V; ring lens, 5 V; peak voltage, 400 or 600 V; grid voltage, 50 V; and ion source helium flow rate, 2.0 L/min. For the DART-HRMS analysis of seeds and bark, samples were divided into smaller segments using a razor blade and each of the segments was suspended via tweezers directly within the path of the DART gas stream in the open-air space between the ion source and mass spectrometer inlet.

Liquids, powders, resins, extracts, crushed leaves, and the pulverized content of the interiors of gelatin-based capsules were each sampled three times by suspending the closed end of a melting point capillary tube into the material and presenting the coated surface into the DART gas stream. For the seeds and bark, each of the generated DART mass spectra represented the average of the spectra of the segments, while for the liquid, powder, resin, extract, ground leaves, and capsule samples, each spectrum was composed of an average of three spectra. With each set of analyses for each product, polyethylene glycol 600, which served as a mass calibrant, was analyzed. TSSPro3 software (Schrader Software Solutions, Grosse Pointe, MI, USA) was used for processing the mass spectra for background subtraction, mass calibration, and peak centroiding.

Ten samples of *M. speciosa* (aka kratom) and five samples of *Datura* species were analyzed by independent laboratories using the same experimental parameters. Kratom leaves were sampled at IonSense Inc. (Saugus, MA, USA) using an instrument similar to that operated in our laboratory. *Datura* species were analyzed at the Emerging Technology and Entrepreneurship Complex (ETEC) at the University at Albany using a DART SVP Ion Source coupled to a JEOL JMS-T100LP AccuTOF LC-plus 4G mass spectrometer. It was found that for this instrument, increasing the detector voltage to 2200 V and adjusting the sampling interval to 0.25 ns were critical to obtaining mass spectra that could be screened against the database for external validation purposes. It should also be noted that when the gas temperature and/or orifice 1 voltage are altered, the data collected can deviate enough from that of the spectra within the database to lead to false positives or negatives. There are two reasons for this: (1) the relative abundance of the peaks changes as a function of temperature. The spectra at lower temperatures are dominated by peaks from more volatile compounds and at higher temperatures,

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

higher boiling compound peaks are more prominent; and (2) increases in the orifice 1 voltage (and to a much lesser extent increases in temperature), shift the analysis from one that is conducted under soft ionization conditions (i.e., 20 V), where there is minimal fragmentation, to one where there is collision-induced dissociation. This can lead to spectra that will appear quite different from those that populate the database because the spectra will be dominated by fragment peaks that appear at the expense of the protonated precursor peaks from which they are derived. Therefore, it is essential that the instrument parameters are well replicated.

Multivariate Data Analysis. The devised psychoactive plant material identification workflow, which was based on the machine learning processing of a database of the species-specific chemical fingerprints of psychoactive plants, is described here. The sample identification aspect of this workflow is comprised of mass spectral data pre-processing, application of advanced statistical analysis, and identification of plant material unknowns. To develop the approach, the processed DART mass spectra (6691 spectra overall), which were collected from plant materials representing 18 families, 34 genera, and 57 species, were imported as text files into Python 3.7 software (Python Software Foundation, DE, USA) in the form of two-column tables of m/z values and their corresponding relative intensities. As indicated in Scheme 2A—Step 2, the spectra were aligned in a matrix with an optimal bin width [10 millimass units (mmu)] and a relative abundance cutoff threshold of 1%. Due to the variability of sample numbers and availability, there was a significant disparity between the numbers of samples of each species. This imbalance was addressed using the support vector machine-synthetic minority oversampling technique (SVM-SMOTE),[26,27] which served to increase the number of samples in minor classes through the generation of "synthetic data". The synthetic data were randomly created along the lines adjoining each minority class support vector with several of its nearest neighbors. Since the species share taxonomical relationships (as shown in Scheme 1), a supervised top-down hierarchical classification tree[5,28] was designed to simplify the complex 57 flat classification problem into 18 multiclasses (as illustrated in Scheme 2A—Step 3). The classification tree had 18 classification nodes organized within 3 levels of discrimination (family, genus, and species) and ended at 57 leaf nodes representing the individual species. Thus, samples were first categorized into families at the first level of discrimination and subsequently discriminated by genus and then to the corresponding species at the second and third levels, respectively. To increase the performance of the classification model,[29−31] the results of three machine learning methods were fused using posterior probabilities. Therefore, within the classification node of each tree, random forest (RF), k-nearest neighbors (KNN), and support vector machine (SVM) were trained, and each trained model assigned a probability value to each class label for the samples in each classification node. Prediction of the sample label is based on the average of the probabilities resulting from the application of the SVM, KNN, and RF models. For the assignment of samples to each class in each node, a probability threshold was computed for each class using the prediction results of 100 × randomly selected test set (30% of data) and the precision-recall (sensitivity) curves.

RESULTS AND DISCUSSION

To develop a classification model for rapid identification of psychoactive plant-derived materials, hierarchical classification tree-based supervised methods were used. The overall approach, including data acquisition and statistical analysis, is summarized in Scheme 2A. Assessment of mass spectra in both positive- and negative-ion modes revealed that much more chemical information (i.e., many more peaks) was contained in positive-ion mode spectra. Given that the greater the number of peaks, the more refined a prediction model that can be built, we chose to use the spectra generated in positive-ion mode. Representative spectra (average of 10 DART-HRMS analysis replicates) for all 57 species are presented in Figure S1 for one of the forms of the material. As an example, spectra of diverse forms of *Artemisia absinthium* are shown in Figure S2. The figure displays the spectra of dried herb powders and seeds, as well as a processed form of the materials (an *A. absinthium* tincture). From the figure, similarities and differences between the spectra are noted. For example, some peaks are common to multiple sample forms (such as m/z 231.125). On the other hand, the seed was observed to exhibit the greatest number of peaks. The spectra of the different forms of each species were compared to remove the variables related to the plant matrix and not related to the species identity. As indicated in Scheme 2A—Step 2, the collected spectra were aligned along common m/z values using a relative abundance threshold cutoff of 1% and binned (with a bin width of 10 mmu). The bin width and relative abundance threshold cutoff values were determined by iterative evaluation of the goodness of the classification model as a function of changes in bin width and relative abundance threshold cutoffs. The resulting matrix with dimensions of 6691 × 2532 was subjected to the application of SVM-SMOTE to handle the class imbalances. Species discrimination was then achieved by adopting hierarchical classification tree-based supervised methods using scikit-learn[32] and its interfaces.[33] The spectra of 30% of the samples were randomly selected to serve as external validators for the testing of the trained models, and the hierarchical classification tree was trained against a fused classifier comprised of SVM, RF, and KNN methods (Scheme 2A−Step 3). The trained model was then validated using 10-fold cross-validation and external validation, yielding prediction accuracies of 98 and 99%, respectively. Figure S3 illustrates the corresponding normalized confusion matrix for the external validation of the fused classifier. The x- and y-axes display the predicted and expected values, respectively. The color gradient extends from blue to white, with blue and white representing a 0 and 100% prediction rate for identification, respectively. The diagonal values in the matrix correspond to true positive rates, and the off-diagonal entries represent false negative and false positive rates. As illustrated in Figure S3, with the exception of the three species *A. nervosa* (Sp 32), *S. divinorum* (Sp 42), and *S. tortuosum* (Sp 52), for which the true positive rates fell between 70 and 90%, all other species were predicted with ≥90% accuracy.

To facilitate the utilization of the fused classifier model as a tool for the screening and identification of psychoactive plant material unknowns, an intuitive and user-friendly graphical interface named Database of Psychoactive Plants (DoPP) was designed and developed as a stand-alone application in Windows (using the programming language Python). It is composed of three parts termed "Identification", "Quantifica-

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

Figure 1. Illustration of the application of DoPP for the identification of a plant sample (*M. speciosa*) analyzed by DART-HRMS. As shown in Panel A, when the mass spectrum of the solid material is imported, the interface reveals the mass data table containing *m/z* values and the corresponding relative intensities, and the mass spectrum of the query sample. The results present (1) the family, genus, and the species of the query sample, along with the posterior probabilities from the fused classifier in the three levels of the hierarchical classification tree; (2) the identity and structure of any known psychoactive components; and (3) a bar plot showing the probabilities associated with the identification of the family, genus, and species by the embedded classifiers (i.e., SVM, RF, K-NN, and a fused classifier comprised of all three) in the hierarchical classification tree. Three other bar plots (Panels B–D) display the probabilities for identification of the family, genus, and species levels acquired using the fused classifier.

tion," and "Psychoactive Plant Directory", which are accessible via tabs (Figure 1).

The "Identification" tab displays the species identity prediction that DoPP assigns to the DART mass spectrum of the material that is screened. The "Quantification" tab enables the performance of the tasks required for the quantification of psychoactive small molecules detected in the analyzed sample, including computation of the calibration curve and method validation using quality control (QC) samples. The necessary steps for the successful accomplishment of quantification of small molecules within complex matrices by DART-HRMS have been described previously and are summarized in Scheme 2B.[22−25] Calibration and QC samples containing the analyte and internal standard analyzed by DART-HRMS enable validated method development and determination of concentrations using ratios of the peak areas of the analyte and internal standard. In this regard, the univariate statistical modeling capabilities of DoPP play an important role in the development of the quantification method. The application employs the *statsmodels* module[34] of Python as well as ANOVA to calculate and validate the regression. The validation criteria utilized are based on those defined by the U.S. Food and Drug Administration (FDA) Bioanalytical Method Validation Guidance for Industry.[35] The interface highlights the results to reveal when: (1) the measured non-zero calibrators differ from the nominal (theoretical) concentrations by greater than 15%, and (2) the calculated concentration of the lower limit of quantification (LLOQ) calibrator deviates by greater than 20% of its nominal concentration.

Using the "Psychoactive Plant Directory" tab, the user can access a repository of the mass spectra of different forms of the species in the database (e.g., from different areas of the plant, such as the aerial parts, roots, seeds, etc.), or processed forms such as extracts, in order to make comparisons and visualize the chemical structure(s) of the psychoactive component(s), among other features. Details for the plant species, such as molecule(s) of interest with their respective monoisotopic masses, chemical formulas, and structures, can be found in Table S2. The "Psychoactive Plant Directory" tab also serves as a resource for information about the plant species represented within DoPP. Clicking this tab opens the window shown in Figure S4A, where a list of each of the species that fall under the "Sample Information" section can be found. If, for example, a search of *Lactuca virosa* is performed within this tab, mass spectra of different analyzed forms of this species appear in the "Display Data" section of the tab (Figure S4B). Also, a link to the Wikipedia page that describes the species and the structures of its known psychoactive components appears under the "Psychoactive Compound" tab. As DoPP contains DART mass spectra of powder, leaf, flower, resin, seed, and tincture forms of *L. virosa*, representative mass spectra of each can be viewed via the "Display Data" section, where the mass spectra of each form are shown in Figure S4B.

Approach for the Identification of Sample Unknowns. In order to illustrate the utilization of DoPP for the identification of plant material, the interrogation of

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

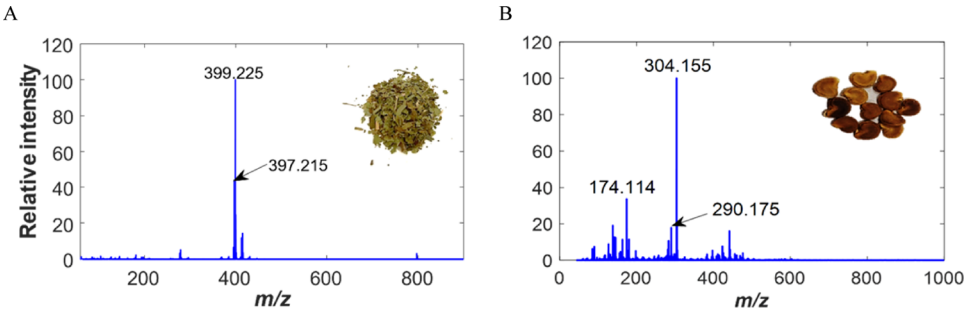


Figure 2. Representative 20 V soft ionization DART mass spectra of (A) *M. speciosa*, aka kratom, and (B) *D. innoxia*. The base peak at nominal *m/z* 399 in the kratom mass spectrum (A) corresponds to the protonated form of its psychoactive component mitragynine. Prominent peaks in the *D. innoxia* spectrum (B) correspond to the protonated forms of atropine (*m/z* 290) and scopolamine (*m/z* 304).

materials comprised of *M. speciosa* commonly known as kratom (leaf), *Datura innoxia* (seed), *Datura wrigthii* (seed), *Ricinus communis* in castor oil form, *Salvia miltiorrhiza* in tablet form, and a plastic bag are described here. Kratom has been identified by the UNODC as a plant of concern because of its increased recreational use, the potential to cause dependence, its various adverse health effects, and because it has been implicated in drug overdose deaths.[36] Its major psychoactive component is mitragynine, which has been shown to act on various opioid receptors including the mu, delta, and kappa receptors.[36] *Datura* species are legal highs containing atropine and scopolamine, which are controlled substances in many countries. For this study, kratom and *Datura* species were also analyzed by independent laboratories using experimental parameters identical to those described earlier (see Methods section). This enabled determination of the utility of DoPP using data generated from a different instrument and acquired by different analysts. Figure 2 displays the DART mass spectra of commercially available kratom (comprised of crumbled leaves (Figure 2A)) and *D. innoxia* seeds (Figure 2B). The mass spectra of other samples are shown in Figure S5. As indicated in Figure 2A, the base peak at nominal *m/z* 399 in the kratom mass spectrum corresponds to the protonated form of mitragynine ($[C_{23}H_{30}N_2O_4 + H]^+$; *measured:* 399.2278). Interestingly, despite the complexity of the kratom raw material, the spectrum is relatively simple and is dominated by the mitragynine peak. Prominent peaks in Figure 2B (*D. innoxia* seed) correspond to the protonated form of atropine ($[C_{17}H_{23}NO_3 + H]^+$; *measured:* 290.1751) and scopolamine ($[C_{17}H_{21}NO_4 + H]^+$; *measured:* 304.1543) with the scopolamine peak being the more dominant. Figure S6 illustrates the similarities and differences in correlation coefficient for ten kratom, five *D. innoxia*, and five *D. wrightii* samples that were analyzed independently in each of the two different laboratories. The brightest shade of yellow represents the highest correlation and the darkest shade of blue represents the lowest. To compare the interlaboratory spectra for reproducibility, the interspectral correlation scores for the spectra were computed. Then, the correlations for each spectrum were averaged. The average scores for the datasets from each laboratory for each species were examined to reveal whether they fell within the normal distribution.[37] Using the average scores of the correlation metrics along with the paired *t* test, it was found that the spectra of the three species from the two independent laboratories were statistically the same at the 95% confidence level. It should be noted that the mass resolving power and mass accuracy can vary between different mass analyzers and that different types of mass analyzers may influence not only reproducibility but also DoPP results. Future studies will be devoted to the assessment of the scope and limitations of DoPP as a function of differences in mass analyzer type.

In conducting classification in real-world scenarios, a classifier not only must correctly group unknown samples into the classes that are defined in the model but must also correctly reject: (1) samples that represent novel classes against which the model was not trained; and (2) other anonymous data such as background or poor quality data. *R. communis* and *S. miltiorrhiza*, which are species not represented in the database, were used to investigate how the classifier would handle data from a species that should not be recognizable. Also, the plastic bag and a poor quality mass spectrum of *D. wrightii* material (by virtue of its not having been properly processed for background correction) were screened against the database to test how the model would treat data that should not be recognized and poor quality data, respectively. Screening of the spectra using DoPP resulted in correct identification of kratom in all tested cases, a result featured in the "Identification" tab section. The prediction outcomes for all of the other samples are presented in Figures S7−S14.

Identification Tab. When the DART mass spectrum of an unknown material is first imported into DoPP, the window that appears in the "Identification" tab is illustrated in Figure 1A. It displays the mass spectral data table and plot, showing *m/z* values and their corresponding relative intensities. On clicking the "Compute Identification" tab, the material is first screened for outlier detection using principal component analysis (PCA) and Hotelling's T^2 statistic, and if it is identified as an outlier, the result will be listed as "Not Detected" in the "Identification Result" section. If it is deemed not to be an outliner, then in the "Display Result" section, a bar plot that reveals the prediction probabilities resulting from classification based on SVM, RF, KNN, and the fused classifiers for identification of the family, genus, and species of the analyzed material is shown (Figure 1A). Three other bar plots (Figure 1B−D) display the identification results for the family, genus, and species levels of the classification tree for the fused classifier. In the "Identification Result" section, the maximum probability computed by the fused classifier for family, genus, and species levels along with their corresponding class labels are shown by DoPP. When the computed probability is lower than the probability threshold for assigning a class label at each level, the background color of the cells changes to pink, indicating that these levels are not assigned. Additional information provided within this tab includes the

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

Table 1. Analysis of the DMT Calibration Model for Goodness of Fit and Precision[a]

A

Regression Statistics: Run 1						
No. of calibration stds.	R-squared	Adjusted R-squared	Std. error	F-statistic	Probability	Confidence level
7	0.998	0.998	0.0596	2820	0	95
Summary						
Index	Coefficient	Std. error	*t*-statistic	P > *t*	Lower limit of 95% CI*	Upper limit of 95% CI*
Intercept	-0.1264	0.078	-1.627	0.165	-0.326	0.073
Slope	0.0457	0.001	53.105	0	0.044	0.048
Analysis of Variance (ANOVA)						
Index	DoF**	Sum of square	Mean square	F-statistic	P > F	
Model	1	33.6284	33.6284	2820.094	0	
Residual	5	0.0596	0.0119			

*CI refers to Confidence Interval; **DoF refers to Degree of Freedom.

B

Low Concentration Level: Statistics						
	Run 1		Run 2		Run 3	
Nominal conc. (mg/L)	Calculated	Percent prediction error	Calculated	Percent prediction error	Calculated	Percent prediction error
30	35.6964	-18.988	33.1561	-10.5205	28.0505	6.4982
30	31.8875	-6.2917	29.2488	2.504	30.712	-2.3732
30	34.0762	-13.5872	34.7232	-15.7441	31.2857	-4.2856
30	33.3406	-11.1352	31.0299	-3.433	30.5872	-1.9575
30	36.1502	-20.5008	28.8825	3.725	27.4513	8.4957
30	35.0123	-16.7077	31.9503	-6.501	29.5288	1.5706
30	30.2974	-0.9913	27.7178	7.6073	27.5536	8.1545
30	27.8138	7.2873	26.1277	12.9076	27.8761	7.0797
30	29.3476	2.1746	26.5624	11.4588	36.7165	-22.3884
30	26.0144	13.2852	23.3186	22.2713	26.4515	11.8284

Cells highlighted in pink show the percent prediction error for concentrations that deviated from the nominal concentrations by greater than ±15%

C

Conc. Level	Parameter	Between-run	Run 1: Within-run	Run 2: Within-run	Run 3: Within-run
Low	Mean calculated conc.	30.2856	31.9636	29.2717	29.6213
	MPE**	-0.9519	-6.5455	2.4275	1.2622
	CV*	11.452	11.6192	11.6696	9.9249
Medium	Mean calculated conc.	75.4304	76.6309	77.6177	72.0427
	MPE**	5.712	4.2113	2.9779	9.9466
	CV*	11.871	13.3137	14.6718	5.552
High	Mean calculated conc.	118.7104	115.5147	122.5506	118.0658
	MPE**	8.6843	11.1425	5.7303	9.1802
	CV*	10.1564	7.8623	9.7194	12.626
LLOQ	Mean calculated conc.	10.7006	9.4847	10.8007	11.8165
	MPE**	-7.0062	5.1528	-8.0065	-18.1648
	CV*	15.4692	9.6496	16.815	9.5642

*CV refers to Coefficient of Variation; **MPE refers to Mean Percentage Error

[a](A) Analysis results of the fitting model for Run 1 for the DMT quantification, showing the linear equation fit with respect to: (1) goodness of fit; (2) parameters for interpretation of the coefficients; and (3) ANOVA; (B) Validation results for the calibration curve using low concentration QC samples; (C) mean, percent prediction error, and CV are reported for between-run and within-run variations.

name(s) and structure(s) of the dominant psychoactive component(s), as well as molecular formula(s). Figure 1 illustrates the results of analyses performed at an independent laboratory (IonSense Inc.) for the identification of an *M. speciosa* (kratom) sample. Figure 1B shows that the probability for the assignment of the plant material to the Rubiaceae family is the highest of all the 18 families represented in the database. The material is further classified as being derived from a *Mitragyna* genus plant, and finally, as the *M. speciosa* species. These are all correct assignments. The prediction results for *D. innoxia* (Figures S7 and S9) and *D. wrigthii* (Figures S8 and S10) were similarly accurate for data collected in our lab and at ETEC. The screening results for *R. communis* and the plastic bag are shown in Figures S11 and S12, respectively. Both are reported as outliers, which is the expected and desired result, as the model should reject both on the grounds that they should not be recognizable. Although *S. miltiorrhiza* (Figure S13) and the poor-quality *D. wrigthii* spectrum (Figure S14) were not rejected in the outlier detection step, they were not assigned to any of the species in the database, as illustrated in the figures. The "not-assigned" status of these samples is visually apparent from the pink background color, which signifies that the observed probability of 0.31 is lower than the threshold of 0.45 that was set for the assignment of an *R. communis* sample to the Rubiaceae family and that the observed probability of 0.26 is lower than the threshold of 0.45 for the assignment of a *D. wrigthii* spectrum to the Asteraceae family. Thus, the results reveal that DoPP was successful not only in determining the identities of species contained within its database, but also in rejecting the samples that represent novel classes or poor matches with entities in the database. They further show that the hierarchical classification tree underlying the fused classifier is a well-fitted model for the identification of psychoactive plant species using the DART-HRMS data. In addition, DoPP provides a useful tool for interrogation of a DART-HRMS database of psychoactive plant species.

In DoPP, the approach that was developed for the differentiation of plants is based only on a probabilistic model and species-specific ions as an alternative means to distinguish between species that were not considered. However, using species specific ions can provide another source of information that may be helpful in reducing the false positive rate. Plans are underway to assess the extent to which the inclusion of this consideration could further enhance the utility of the application, particularly as it relates to the

https://doi.org/10.1021/acs.analchem.2c01614
Anal. Chem. 2022, 94, 16570−16578

development of a peak-matching algorithm for unknown sample pre-screening.

Quantification Tab. The data utilized to illustrate the small-molecule quantification application in DoPP have been reported previously[23] for the development of a validated method, consistent with FDA guidelines, for the quantification of DMT. This molecule is a Schedule I psychoactive natural product found in numerous plant species. For its quantification by DART-HRMS in mock ayahuasca brews, the structurally related synthetic compound *N*,*N*-diethyltryptamine (DET) was used as the internal standard. The calibration and QC data were based on the peak area ratios of DMT [$(C_{12}H_{16}N_2 + H)^+$ at *m*/*z* 189.1386] to DET [$(C_{14}H_{20}N_2 + H)^+$ at *m*/*z* 217.1699].

The calibration and QC data were collected over a 3-day period using one calibration curve and two sets of QC standards that were freshly made each day. The calibration data were acquired each day using six standard solutions with DMT concentrations ranging from 10 to 150 mg/L. QC data collected each day included standards at four concentration levels (high, medium, low, and LLOQ). Figure S15 illustrates the display within the "Quantification" tab for the regression and validation of the QC results. The calibration curve and prediction error plots are shown for data collected on the first day of the experiment (termed Run 1), with the results for the LLOQ and the low, medium, and high QC concentrations also being presented. Analysis of the fitting model for the goodness of fit and precision is illustrated in Table 1A and shows the observation of a linear equation to which the experimental data fit well.

For Run 1, Table 1A lists the following: (1) the goodness of the fit (in the "Regression Statistics" section) and reports the standard error of the regression and R-squared; (2) the parameters for interpretation of the regression coefficients (the slope and the intercept) in the "Summary" section; and (3) the ANOVA results table (in the "Analysis of Variance (ANOVA)" section). The calibration curve, prediction error plot, and regression analysis results are illustrated in Figures S16 and S17 for Run 2 and Run 3, respectively. The outcomes for the validation of the calibration curve using QC samples at the low concentration level are shown in Tables 1B, and S3 illustrates the validation results for the medium, high, and LLOQ concentration levels. The tables contain the calculated concentrations and errors for the QC samples in Runs 1−3. Entries highlighted in the pink show the prediction error for the calibrators, which deviated from the concentration of the standard by greater than ±15% or deviated from the concentration of the LLOQ calibrator by over ±20%. The mean calculated concentration, mean percentage error (MPE), and coefficient of variation (CV) are reported in Table 1C for between-run and within-run analyses. Per the identification and quantification results, DoPP is a platform that is well-suited not only for species identification but also for the quantification of detected psychoactive components.

CONCLUSIONS

Comprised of a graphical user interface coupled with a comprehensive database of high-resolution DART mass spectra of psychoactive plant materials, DoPP enables their rapid species identification through screening of their DART mass spectra. In total, 18 families, 34 genera, and 57 species are represented, including multiple species designated by the UNODC as "plants of concern" due to their increased recreational use and their potential to cause addiction and negative health impacts. For the identification of plant material unknowns, DoPP employs a trained hierarchical classification tree constructed from the fusion of SVM, RF, and KNN models. This trained fused model provides discrimination with accuracies of 98 and 99% for 10-fold cross-validation and external validation assessments, respectively. DoPP is a platform that is well-suited not only for species identification but also for the quantification of detected psychoactive components. The quantification feature of DoPP contains the essential statistics measures for the computation and validation of calibration curves. The results show the successful application of DoPP for the identification of unknown psychoactive plant materials and the quantification of their psychoactive components. These features, among several others, enable facile interrogation and identification of plant materials without prior knowledge of botany, in the absence of distinguishing plant morphological features (such as is the case when the plant materials have undergone processing such as grinding or extraction), or the need for extensive sample pre-treatment prior to analysis. DoPP will be compiled as a stand-alone desktop application for windows and mac platforms so that the user will not need to set up any specific software. It also will allow the user to submit their own entries to the host library. Following pre-processing and confirmation of the data, the spectra will be added to the database and will be used to update the trained model.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c01614.

> Representative DART-HR mass spectra of the 57 plant species; representative DART-HR mass spectra of single species represented by different matrices; image of the confusion matrix showing the external validation results for the hierarchical classification tree; illustration of the "Psychoactive Plant Directory" tab of DoPP; DART-HR mass spectra of test plant materials; image of pairwise inter-spectral similarities for DART-HR mass spectra of plant material; identification results for test materials; results of calibration curve analysis for Runs 1, 2, and 3 for DMT quantification; table containing information on the analyzed plant materials; table containing the molecule(s) of interest contained within each of the represented plant species; validation results for the DMT calibration curve using QC samples at high, medium, and LLOQ concentrations (PDF)

AUTHOR INFORMATION

Corresponding Author

Rabi A. Musah − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States;* orcid.org/0000-0002-3135-4130; Email: rmusah@albany.edu

Authors

Samira Beyramysoltan − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States*

Megan I. Chambers − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States*
Amy M. Osborne − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States*
Mónica I. Ventura − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c01614

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The funding support of the National Institute of Justice (NIJ), Office of Justice Programs, U.S. Department of Justice (DOJ) [awards 2015-DN-BX-K057 and 2019-BU-DX-0026] as well as the UAlbany Initiatives for Women Endowment Award received by SB is gratefully acknowledged. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the DOJ. Thanks are extended to the National Science Foundation for support (Award #1429329). Thanks are also extended to Dr. Robert Cody (JEOL USA Inc.) and Ms. Allix Coon (UAlbany) for technical assistance, as well as to William Fatigante at IonSense Inc. for DART-HRMS analysis of kratom samples.

■ REFERENCES

(1) Dasgupta, A. Drugs of Abuse: An overview. In *Alcohol, Drugs, Genes and the Clinical Laboratory*, Dasgupta, A., Ed.; Academic Press, 2017; pp 23−51.
(2) Cunningham, N. *Emerging Med. Australas.* **2008**, *20*, 167−174.
(3) Lo Faro, A. F.; Di Trana, A.; La Maida, N.; Tagliabracci, A.; Giorgetti, R.; Busardò, F. P. *J. Pharm. Biomed. Anal.* **2020**, *179*, No. 112945.
(4) Gurib-Fakim, A. *Mol. Asp. Med.* **2006**, *27*, 1−93.
(5) Beyramysoltan, S.; Abdul-Rahman, N. H.; Musah, R. A. *Talanta* **2019**, *204*, 739.
(6) In *Drugs of Abuse: A DEA Resource Guide*, 2020 ed.; U.S. Department of Justice: Drug Enforcement Administration, 2020.
(7) Arunotayanun, W.; Gibbons, S. *Nat. Prod. Rep.* **2012**, *29*, 1304−1316.
(8) Caffrey, C. R.; Lank, P. M. *Open Access Emerging Med.* **2018**, *10*, 9−23.
(9) 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, DOI: 10.1080/00029157.2013.826172.
(10) Coon, A. M.; Beyramysoltan, S.; Musah, R. A. *Talanta* **2019**, *194*, 563−575.
(11) Hayes, J. M.; Abdul-Rahman, N.-H.; Gerdes, M. J.; Musah, R. A. *Anal. Chem.* **2021**, *93*, 15306−15314.
(12) Fowble, K. L.; Musah, R. A. *Methods Mol. Biol.* **2018**, *1810*, 217−225.
(13) Puype, F.; Ackerman, L. K.; Samsonek, J. *Chemosphere* **2019**, *232*, 481−488.
(14) Cody, R. B.; Dane, A. J., Chapter 2: Direct Analysis in Real Time (DART®). In *Ambient Ionization Mass Spectrometry*; The Royal Society of Chemistry, 2015; pp 23−57.
(15) Sisco, E.; Forbes, T. P. *Forensic Chem.* **2021**, *22*, No. 100294.
(16) Beyramysoltan, S.; Giffen, J. E.; Rosati, J. Y.; Musah, R. A. *Anal. Chem.* **2018**, *90*, 9206−9217.
(17) Lesiak, A. D.; Cody, R. B.; Dane, A. J.; Musah, R. A. *Anal. Chem.* **2015**, *87*, 8748−8757.
(18) Lesiak, A. D.; Cody, R. B.; Ubukata, M.; Musah, R. A. *Forensic Sci. Int.* **2016**, *260*, 66−73.
(19) Appley, M. G.; Beyramysoltan, S.; Musah, R. A. *ACS Omega* **2019**, *4*, 15636−15644.
(20) Angelis, E. D.; Pilolli, R.; Bejjani, A.; Guagnano, R.; Garino, C.; Arlorio, M.; Monaci, L. *Foods* **2021**, *10*, 1238.
(21) Wong, M. Y.-M.; So, P.-K.; Yao, Z.-P. *J. Chromatogr., B* **2016**, *1026*, 2−14.
(22) Longo, C. M.; Musah, R. A. *J. Forensic Sci.* **2020**, *65*, 61−66.
(23) Chambers, M. I.; Appley, M. G.; Longo, C. M.; Musah, R. A. *ACS Omega* **2020**, *5*, 28547−28554.
(24) Appley, M. G.; Chambers, M. I.; Musah, R. A. *Drug Test. Anal.* **2021**, *14*, 604.
(25) Chambers, M. I.; Osborne, A. M.; Musah, R. A. *Rapid Commun. Mass Spectrom.* **2019**, *33*, 1915.
(26) Akbani, R.; Kwek, S.; Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets. In *Machine Learning: ECML*; Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; Vol. *2004*, pp 39−50.
(27) Lemaître, G.; Nogueira, F.; Aridas, C. K. *J. Mach. Learn. Res.* **2017**, *18*, 559−563.
(28) Beyramysoltan, S.; Ventura, M. I.; Rosati, J. Y.; Giffen-Lemieux, J. E.; Musah, R. A. *Anal. Chem.* **2020**, *92*, 5439−5446.
(29) Kim, H.-C.; Ghahramani, Z. Bayesian Classifier Combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Neil, D. L.; Mark, G., Eds.; PMLR: Proceedings of Machine Learning Research, 2012; Vol. *22*, pp 619−627.
(30) Ruta, D.; Gabrys, B. *Comput. Inf. Syst.* **2000**, *7*, 1−10.
(31) Deklerck, V.; Finch, K.; Gasson, P.; Van den Bulcke, J.; Van Acker, J.; Beeckman, H.; Espinoza, E. *Rapid Commun. Mass Spectrom.* **2017**, *31*, 1582−1588.
(32) Pedregosa, F.; Varoquaux, G. E.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(33) Sklearn-Hierarchical-Classification: Version = ″1.3.2″. GitHub Repository. https://github.com/globality-corp/sklearn-hierarchical-classification.
(34) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, 2010.
(35) In *Bioanalytical Method Validation Guidance for Industry, 2018 ed.*; Food and Drug Administration, HHS, 2018. https://www.federalregister.gov.
(36) White, C. M. *Am. J. Health Syst. Pharm.* **2018**, *75*, 261−267.
(37) Sheen, D. A.; Rocha, W. F. C.; Lippa, K. A.; Bearden, D. W. *Chemom. Intell. Lab. Syst.* **2017**, *162*, 10−20.