Environmental Science & Technology

pubs.acs.org/est

Article

Machine Learning Applications for Chemical Fingerprinting and Environmental Source Tracking Using Non-target Chemical Data

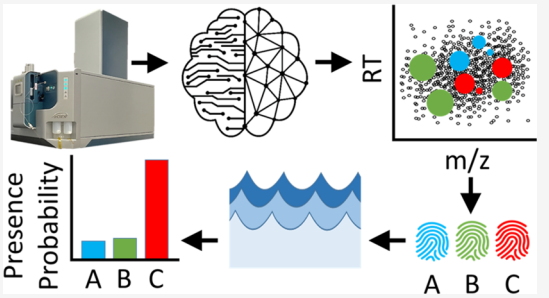
Emmanuel Dávila-Santiago, Cheng Shi, Gouri Mahadwar, Bridgette Medeghini, Logan Insinga, Rebecca Hutchinson, Stephen Good, and Gerrad D. Jones*

Cite This: https://doi.org/10.1021/acs.est.1c06655

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A frequent goal of chemical forensic analyses is to select a panel of diagnostic chemical features—colloquially termed a chemical fingerprint—that can predict the presence of a source in a novel sample. However, most of the developed chemical fingerprinting workflows are qualitative in nature. Herein, we report on a quantitative machine learning workflow. Grab samples (n = 51) were collected from five chemical sources, including agricultural runoff, headwaters, livestock manure, (sub)urban runoff, and municipal wastewater. Support vector classification was used to select the top 10, 25, 50, and 100 chemical features that best discriminate each source from all others. The cross-validation balanced accuracy was 92−100% for all sources (n = 1,000 iterations). When screening for diagnostic features from each source in samples collected from four local creeks, presence probabilities were low for all sources, except for wastewater at two downstream locations in a single creek. Upon closer investigation, a wastewater treatment facility was located ∼3 km upstream of the nearest sample location. In addition, using simulated in silico mixtures, the workflow can distinguish presence and absence of some sources at 10,000-fold dilutions. These results strongly suggest that this workflow can select diagnostic subsets of chemical features that can be used to quantitatively predict the presence/absence of various sources at trace levels in the environment.



KEYWORDS: *chemical forensics, chemical fingerprinting, machine learning, high-resolution mass spectrometry, multivariate analysis, non-target chemical analysis*

INTRODUCTION

Observations of health declines in humans and wildlife (e.g., cancer rates,[1−4] metabolic disorders,[5−7] reproductive abnormalities,[8−12] and die-off events[13]) are increasingly being reported in the literature. Often, the exact causes of these maladies are unknown, but uncharacterized anthropogenic compounds in the environment are hypothesized to drive these observations.[1,14,15] Ecotoxicological research has identified thousands of toxic substances, yet most chemical forensic studies fail to identify the causative agents driving health declines. Those that are successful are laborious, taking years to decades to solve.[13,16−18] A prominent challenge of chemical forensic studies, which seek to identify the source of a chemical/mixture of interest,[19] is that 10−100 s of thousands of chemicals exist in the environment. However, state and federal monitoring programs screen for mere hundreds of compounds (e.g., see ref 20), which ignore ∼>99.9% of all chemical features present (i.e., the chemical space), so the chances of identifying the specific chemicals driving adverse environmental phenomena are small. Therefore, new chemical forensic strategies are needed for routine monitoring that utilize broader swaths of the chemical space.

In recent years, non-target chemical analysis and multivariate computational techniques have been heralded as critical tools in chemical forensic analyses.[21,22] Using high-resolution mass spectrometry (HRMS) instrumentation, non-target analyses seek to collect data on all chemical features that hit an instrument's detector. HRMS instruments can detect thousands of features in environmental samples. With so much data, multivariate tools are necessary for analyzing non-target data sets because unique patterns emerge when considering multiple features simultaneously compared to each feature individually.[23] The fundamental hypothesis underlying recent forensic studies, including this one, is that the holistic chemical composition of a sample is not random and at least some fraction therein contains discriminatory information about a particular source or environmental phenomena. The challenge of forensic analyses is selecting a subset of chemical features

Received: September 30, 2021
Revised: March 1, 2022
Accepted: March 2, 2022

that are useful for diagnostic purposes. Such a diagnostic panel is often referred to as a "chemical fingerprint". In this study, our aim is to select a subset of chemical features that are most predictive of a chemical/pollutant source. We argue that a goal of chemical fingerprinting should be to develop quantitative predictive tools. Although several studies have advanced the chemical forensic literature,[17,22,24−29] this goal remains an open challenge due to limitations of applied multivariate techniques.

Most statistical tools used to develop chemical fingerprints are poorly suited for quantitative or probabilistic predictions. The most popular multivariate tools that have been used in chemical forensic analyses include co-occurrence analysis (e.g., Venn diagrams),[17,24] hierarchical cluster analysis,[17,25,26] principal component analysis,[22,26,27] and multiple machine learning classification tools.[28,29] Venn analyses utilize only presence/absence information, so all features are weighted equally, regardless of peak intensity.[30] Any discriminatory information that may come from differences in concentration or peak areas/intensities is lost. Therefore, sources with very similar chemical compositions yet distinguishable (e.g., via ratios of critical features or other approaches) may not be separable using co-occurrence tools. Clustering also has limitations for chemical fingerprinting and source apportionment. Clustering relies on the entire chemical signature for group assignment and provides no information on which subset of chemical features is important for prediction.[17] Furthermore, clustering can only assign samples to a single group[31] and thus provides no insight into how many pollution sources may be present in a mixed sample. Masiá et al.[27] and Schollée et al.[22] used PCA to demonstrate that sources (e.g., influent/effluent and wastewater/surface water) were strongly differentiated by the non-target chemical gradients present in each sample. Although it is possible to use factor loadings to interpret the important chemical features most correlated with each component, the correlation coefficients between the non-target chemical features and the principal components are often very weak (e.g., $|r| < 0.05$),[22,32] making it difficult to select the subset of chemical features that are most responsible for separating different sources in ordination space. Although ordination has been used in attempts to select targeted features that could be used to predict sources (e.g., see ref 26 for the PFAS example), the purpose of ordination is to summarize the largest gradients present within a data set instead of selecting a subset of diagnostic features. Similar to clustering, ordination may miss source signals that are masked at low concentrations. The best recent example of source tracking and chemical fingerprinting with targeted data utilized various machine learning classification models to predict PFAS sources (AFFF vs non-AFFF) based on the concentration of 10 measured PFAS.[28] In this study, the authors iterated through all possible combinations (n = 1023) of PFAS and compared the classification accuracy of each combination to determine if a particular PFAS subset was disproportionately capable of predicting source membership. Although it is possible to iterate through all possible combinations with data sets that have few chemical features, this quickly becomes impractical with even slightly larger data sets. For example, with just 30 compounds, over 1 billion unique chemical combinations exist. Therefore, manual iteration is impractical for selecting a subset of chemical features that are diagnostic in many target and all large non-target chemical data sets.

Chemical fingerprinting workflows have been called for in the literature;[33−35] however, selecting a subset of diagnostic chemical features (e.g., 10−100) out of the thousands of candidate features has been an obstacle. Herein, we report on a chemical forensic workflow that overcomes this limitation. Based on the non-target chemical data generated from user-defined sources of interest, our goal was to quantify the importance of every non-target chemical feature within a data set based on its ability to discriminate sources. Using the most diagnostic features of each source, our follow-up goal was to use this chemical fingerprint to probabilistically predict the presence or absence of these predefined sources in environmental water samples. This workflow is expected to benefit water quality monitoring programs, chemical forensic studies, and studies seeking to extract discriminatory information from chemical data.

METHODS

Source Selection and Site Description. Detailed descriptions of sampling locations, sampling protocols, and analytical protocols are included in the Supporting Information. Briefly, we collected source-specific samples, which were used to generate chemical fingerprints. In addition, we collected water samples from four local riverine systems surrounding Corvallis, Oregon, which were expected to contain a mixture of sources. The source samples included runoff from suburban and urban surfaces [(sub)urban runoff, n = 12], runoff from agricultural fields (agricultural runoff; n = 6), livestock manure (manure; n = 8), municipal wastewater (WWTP, n = 20), and water from lakes and streams with no to minimal human development upstream (headwaters, n = 5). Briefly, (sub)urban samples were collected from curbs, gutters, and parking lots in Corvallis. Agricultural runoff samples were collected from grazed and ungrazed fields owned by Oregon State University. Grazed fields were grazed by alpacas, dairy cows, or horses, and ungrazed fields were used to grow silage. Both (sub)urban and agricultural runoff samples were collected following winter rain events. Livestock manure was collected from dairy, beef, and swine animal facilities operated by Oregon State University. Wastewater was collected by the municipal treatment facility in Philomath, Oregon. This facility consists of a series of stabilization ponds, which were sampled in spring, summer, and winter. Finally, headwater samples were collected from high-elevation lakes and streams located within the Oregon Coast and Oregon Cascade ranges. Detailed descriptions of the sample locations and source types are included in the Supporting Information (Table S1). All water (1 L) and solid samples (100 g) for each source were collected in triplicate.

In addition to source-specific samples, surface water samples were collected from four nearby creek systems (Dixon Creek, Marys River, Oak Creek, and Rickreall Creek). Dixon Creek is urbanized and located entirely within the Corvallis city limits. Marys River, Oak Creek, and Rickreall Creek originate in forested regions and traverse agricultural, rural residential, and (sub)urban landscapes. Therefore, it was plausible that each source we selected could contribute to the chemical composition of these systems. Grab samples were collected from five longitudinal locations that were approximately evenly spaced along the length of each water body. Creek samples were collected during the spring, summer, and winter of 2018/19.

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

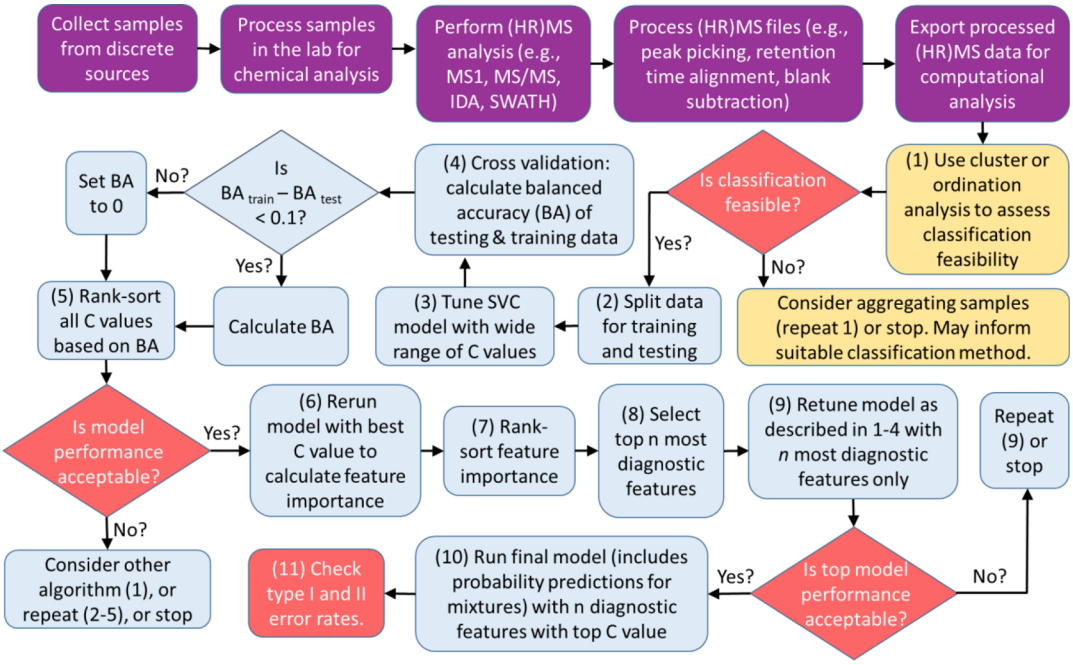


Figure 1. Flowchart illustrating the steps used in this study including sample collection and instrument analysis (purple), classification feasibility (yellow), and the chemical fingerprinting workflow developed herein (blue). Steps developed in this forensic analysis are numbered (1−11). Critical "go/no-go" checks associated with the fingerprinting workflow are in red. Rectangles and diamonds represent process and decision steps, respectively.

Sample Collection and Processing. Non-polar organics were extracted from manure and water samples according to Webster et al.[36] and Jones et al.[37] Briefly, manure samples (~20 g wet weight) and methanol (~15 mL) were added to each of three 50 mL centrifuge tubes for triplicate analysis. Each tube was hand-shaken (1 min) and centrifuged (5 min at 4500 rpm). The supernatant was decanted into a 4 L amber glass bottle. This process was repeated three times total. Each bottle was filled with DI water to dilute the methanol. This entire process was repeated without manure as a negative control. Water samples (both source and creek samples) were collected in a 4 L amber glass bottle. For each sampling trip, a DI blank (1 L) was taken into the field and was handled and processed identically to all samples for a negative control. All samples and controls were immediately transported to the laboratory where they were pressure-filtered (0.7 μm glass microfiber filter, 9 cm, Millipore, Ireland) using ultra-high-purity nitrogen gas. After filtration, samples were split into three 1 L bottles for triplicate analysis and spiked with 10 deuterated internal standards (100 ng each; see Table S2). Non-polar organic compounds were extracted from samples using C18 solid-phase extraction (SPE) cartridges (6 mL, 1000 mg, Restek, Bellefonte, PA, USA). Samples were eluted using 95:5 (v/v) methanol and water (9 mL), dried in a vacuum oven, transferred to LC vials, dried to completion, and resuspended in 1 mL of 90:10 (v/v) water and methanol before instrument analysis.

Non-target Chemical Analysis. HRMS analysis was performed at Oregon State University's Mass Spectrometry Center using Shimadzu Nexera UHPLC coupled to an AB Sciex 5600 triple time-of-flight mass spectrometer. Samples were run in five batches, which corresponded to different sampling efforts. Samples within each batch were randomized. Every 15th sample, instrument negative (90:10 v/v DI/methanol) and positive (100 or 500 ng of mixed standard) controls were run. Non-target methods were adapted from Schollée et al.[22] and Schymanski et al.[38] All samples were analyzed in the positive electrospray ionization (ESI+) mode. Instrument performance was stable across all runs and was assessed by inspecting peak intensities of instrument positive controls (Figure S1). Recoveries of all internal standards were high, but this is potentially confounded by sensitivity limitations of the instrument (see Supporting Information for discussion). Only chemical features with an m/z range from 100 to 1000 were included. Sciex MasterView v1.1, PeakView v2.2, and MarkerView v1.3.1 were used to process all MS1 files (see Supporting Information for detailed descriptions of all software settings). Following peak picking and retention time alignment, feature intensities were imported into Excel and averaged across triplicate samples. A feature was retained in a sample only if it was present in all replicates. Regardless of peak intensity in a sample, all chemical features present in negative controls (both field and instrument) were removed from samples collected on the same date. These two criteria were used to be as conservative as possible in what was considered a feature. Following blank subtraction, replicate averaging, and blank subtraction, 7771 features out of 10,000 were retained across all samples. Detailed descriptions of QA/QC protocols are included in the Supporting Information.

Fingerprinting Workflow. All scripts were written in Python using Jupyter notebook (version 6.0.2). Scripts are available at https://github.com/EcoChem-OSU/Chemical-Fingerprinting, and a detailed description of the fingerprinting workflow is presented in the Supporting Information. Briefly, the workflow consists of 11 steps (Figure 1). Step 1: first, the data are screened to assess classification feasibility using ordination or clustering. We performed hierarchical clustering using the Bray−Curtis dissimilarity index as the distance metric and the average linkage method.[39] Euclidean distance has been used for non-target analysis,[17,25] but this method is

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

inappropriate for sparse (i.e., data sets with a disproportionate number of 0 s) or high-dimensional (i.e., data sets with hundreds to thousands of variables) matrices.[39] We predict that as source clusters become more chemically distinct, the feasibility of classification increases. It is important to note that no classification algorithm is best suited for all data sets and that the observed clustering/ordination patterns (e.g., radial vs linear separation) may provide insight into which classification algorithm is best suited for a particular data set. We developed this workflow using support vector classification (SVC) with a linear kernel, but this algorithm could be replaced with other machine learning tools (e.g., random forest and gradient boosting classification) to take advantage of their properties.

Steps 2−5: using the source samples of interest, the SVC model is tuned to best predict the presence or absence (i.e., one-versus-all classification) of each source. The SVC regularization parameter (C) is varied to maximize the balanced accuracy of the testing data set. As the number of sources in a data set grows, the fraction of any one source decreases, leading to an imbalance in the number of presences versus absences for each source. This occurs with one-versus-all classification and is problematic because the overall classification performance can be skewed by that of the dominant class (i.e., absence). Balanced accuracy equally weights the true positive and negative classification rates when assessing performance, regardless of the number of samples represented by each class. The C values are rank-sorted based on the balanced accuracy.

Steps 6−8: using the best C value, the model is rerun to quantify the coefficient weight of each chemical feature. The coefficient weight is a quantitative measure of the predictive power of a feature, and for each source, the top n chemical features are selected as diagnostic chemical fingerprints.

Steps 9−11: the SVC model is retuned as described in steps 2−5 using only the top n chemical features. If sample mixtures are present (e.g., receiving bodies of water), the model also predicts the probability that a source is present in a sample. The final step of the workflow is to assess the validity of the results by inspecting critical "go/no-go" checks, which includes checking the balanced accuracies at each tuning step and checking final type I and II error rates. These critical thresholds are data-dependent and subject to the needs of the study. Although the thresholds may vary, we found two rules of thumb. First, the balanced accuracy of the initial tuning should be >50%. Second, sources should be more distinctive when only the diagnostic features are used, so the balanced accuracy should increase when using the top n diagnostic features compared to when using all features.

Fingerprint Performance Limits. Because samples were collected as close to the source as possible (i.e., "pure" source samples), a presence probability of 0.5 is appropriate for distinguishing a pure source from its associated outgroup; however, this threshold should not be confused with a presence/absence threshold. We surmise that as a chemical source is diluted in a receiving body of water, the presence probability will decrease until it is statistically indistinguishable compared to when the source is absent (i.e., 0% source). At or below this critical presence/absence probability threshold (herein referred to as $p_{0\%}$), the source should be considered absent. Ideally, $p_{0\%}$ could be identified using samples containing actual mixtures of multiple sources, but we were unable to perform this analysis herein. Instead, we created in silico mixtures as a proof of concept. A single source and creek sample were chosen at random and removed from the data set. From these samples, the peak intensities of all features in each sample were multiplied by different ratios (source/creek—0:100, 0.01:99.99, 0.1:99.9, 1:99, 25:75, 50:50, and 100:0) to approximate dilution, and the resulting intensities were summed to create an in silico mixture. This approach assumes a linear response between source proportion and peak intensity, which is not representative of the true instrument response; however, we argue that these simulated mixtures provide general insight into the workflow's lower performance limits. The workflow was used to predict the presence probability of each source in each in silico mixture. At some ratio, we expect the presence probability to be just distinguishable from $p_{0\%}$. This critical point will be considered the fingerprint's lower performance limit.

RESULTS AND DISCUSSION

Clustering. The cophenetic correlation coefficient of the clustering dendrogram was 0.97 with Bray−Curtis dissimilarity and average linkage and was 0.84 when using Euclidean distance with Ward's linkage (Figures S2 and S3), suggesting that Bray−Curtis dissimilarity is a better approach for clustering this data compared to Euclidean distance. Furthermore, obvious chaining was present when using Euclidean distance. Chaining occurs when individual samples are added sequentially to an existing cluster (i.e., no substantial differences in chemical composition) and suggests that Euclidean distance was less able to differentiate sources within the data.[40] Based on the Bray−Curtis dendrogram, the chemical composition of samples collected from the same source was more similar to each other compared to samples collected from different sources. This suggests that the sources are separable based on the chemical composition and that chemical fingerprinting is possible. Nevertheless, clustering is not a prerequisite for fingerprinting as this technique utilizes the entire chemical composition to identify groupings. It could be possible to have substantial overlap in chemical composition (i.e., little to no clustering) but still have a small subset of features that are diagnostic.

Model Tuning and Fingerprint Development. During the initial tuning and retuning, model performance was low (50% balanced accuracy) for all sources when C values were $<\sim 10^{-3}$ and high ($\leq$100% balanced accuracy) when C values were $>\sim 10^{-2}$ (Figures 2a and S4). It is important to note that overfitting was low in all iterations as balanced accuracy scores were only retained if the difference in training and testing accuracy was $\leq$10% (see Supporting Information for detailed descriptions on model tuning). When the data were randomly shuffled, the balanced accuracy was typically 0% but never exceeded 50%, indicating that high balanced accuracies observed during the initial tuning were not generated from random chance alone. Although we did not assess every instance, balanced accuracies of 0% were due to overfitting when the difference in the training and testing accuracies was >10%. Balanced accuracies of 50% occurred when the SVC model classified a particular source as absent for every sample, which is more common using one-versus-all classification schemes given that most samples are absent. Although C values $>10^{-2}$ typically resulted in the maximum observed balanced accuracy (100% correct classification for most sources), choosing any of these C values did not change downstream results, although this was not extensively assessed.

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

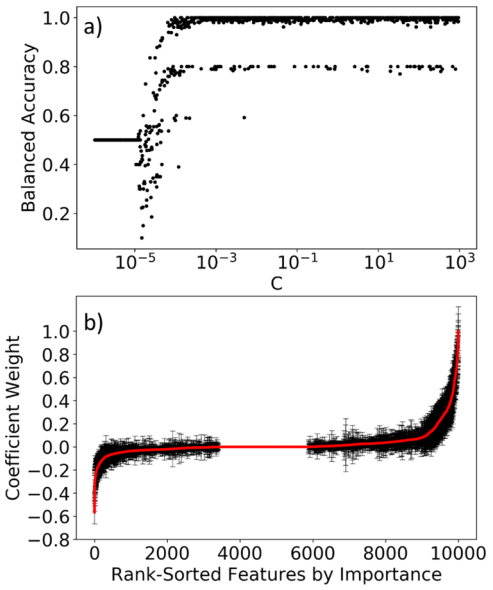


Figure 2. Altering the C value during the initial SVC tuning resulted in various balanced accuracies [WWTP results illustrated; (a)]. Normalized coefficient weights were rank-sorted for all chemical features [WWTP results illustrated; (b)]. Features with positive and negative weights are predictive of presence and absence, respectively. The closer the weight is to ±1, the more predictive power a chemical feature has. Error bars represent standard deviations.

After C selection, each chemical feature was rank-sorted based on its coefficient weight. For all sources, the rank-sorted weights exhibited an S-shaped curve but to various degrees (Figures 2b and S5). Similar importance diagrams have been reported in the literature.[41] The vast majority of chemical features had little predictive power (i.e., normalized coefficient weights near 0), and only a small proportion had coefficient weight magnitudes >0.5. Compared to when all chemical features were used for tuning, the SVC model performance remained equally high or higher when only the top 10, 25, 50, and 100 most diagnostic features were used (Figure S4). During the final iterations, cross-validation model performance was high for all sources. Average balanced accuracies were ≥92% for all sources (averaged across n = 1000 iterations), regardless of how many chemical features were used (Table S3). The chemical fingerprints consisting of the top 10 features are graphically represented in Figure S6. The cross-validation balanced accuracies were 100% when only 10 chemical features were used to predict manure, (sub)urban, and WWTP samples. One hundred percent correct classification was achieved for agricultural runoff and headwater when ≥25 and 100 diagnostic features were used, respectively. These findings are noteworthy because they indicate that few features are needed for chemical fingerprinting. If a relatively small number of chemical features are needed for a diagnostic fingerprint, it becomes possible to use this fingerprinting approach with multiple reaction monitoring methods with more traditional, and widely available, triple quad instrumentation. If these non-target chemical fingerprints could be screened for using instruments typically used in targeted chemical analyses, it would make fingerprinting approaches more readily available to practitioners. Nevertheless, HRMS instruments may be required to develop the fingerprints initially, so fingerprinting advocates with HRMS instrumentation may have to work with monitoring laboratories to make these techniques available for broad-scale use.

Workflow Performance Limits. Until this point in the workflow, all classification predictions have been made using a probabilistic threshold of 0.5 to distinguish a "pure" source from its outgroup; however, this threshold is inappropriate for environmental samples. As the concentration of the source decreases via dilution, the intensity of the signal, and thus the probability of detection, also decreases. Using in silico mixtures, the presence probability when the source percentage was 0% (i.e., $p_{0\%}$) ranged from 0.03 to 0.1 for all sources except headwaters, and when the source percentage was 100%, the presence probability (i.e., $p_{100\%}$) ranged from 0.78−0.99 (Figures 3a and S7). For headwaters, $p_{0\%}$ ranged from 0.17 to 0.22, depending on how many diagnostic chemical features were used, and $p_{100\%}$ ranged from 0.56 to 0.64 (Figure S7). We argue that the difference between $p_{100\%}$ and $p_{0\%}$ represents some indication of fingerprint quality. As the range between these values (i.e., $p_{100\%}-p_{0\%}$) narrows, it becomes increasingly difficult to distinguish presence from absence. Therefore, this suggests that the best fingerprint is WWTP, followed by manure, (sub)urban runoff, agricultural runoff, and headwater.

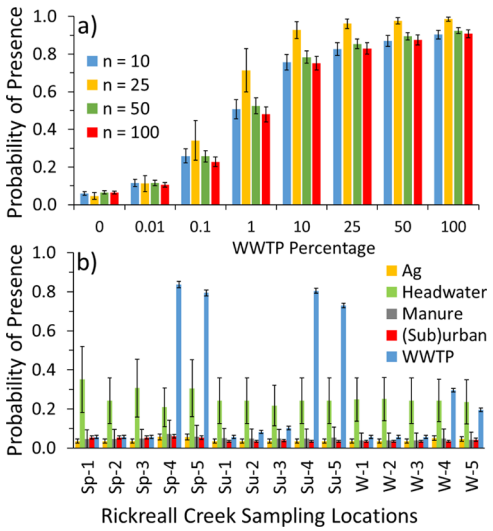


Figure 3. Presence probability distributions were created for in silico mixtures of a single source and creek sample using n = 10−100 diagnostic chemical features [WWTP illustrated; (a)]. The presence probability of each source was calculated for each sample using each source fingerprint [n = 25 diagnostic chemical features; data for Rickreall Creek illustrated; (b)]. Abbreviations include spring (Sp), summer (Su), and winter (W). Sample numbers represent longitudinal locations on the creek. All probabilities were made during the final step of the workflow. Error bars represent standard deviations.

In addition to probability between range $p_{0\%}$ and $p_{100\%}$, the source/creek ratio needed to exceed $p_{0\%}$ could also be an indicator of fingerprint quality. For example, the presence probability for wastewater when the source fraction was 0.01% ($p \approx 0.11$ for n = 10−100 diagnostic features) was nearly two times greater than the critical presence/absence threshold probability ($p_{0\%}$ = 0.06; Figure 3a), suggesting that the machine learning model could potentially detect the presence of this source following a 10,000-fold dilution (i.e., 1 mL in 10 L; see also manure; Figure S7). It is important to emphasize that this should be tested against real mixtures instead of in

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

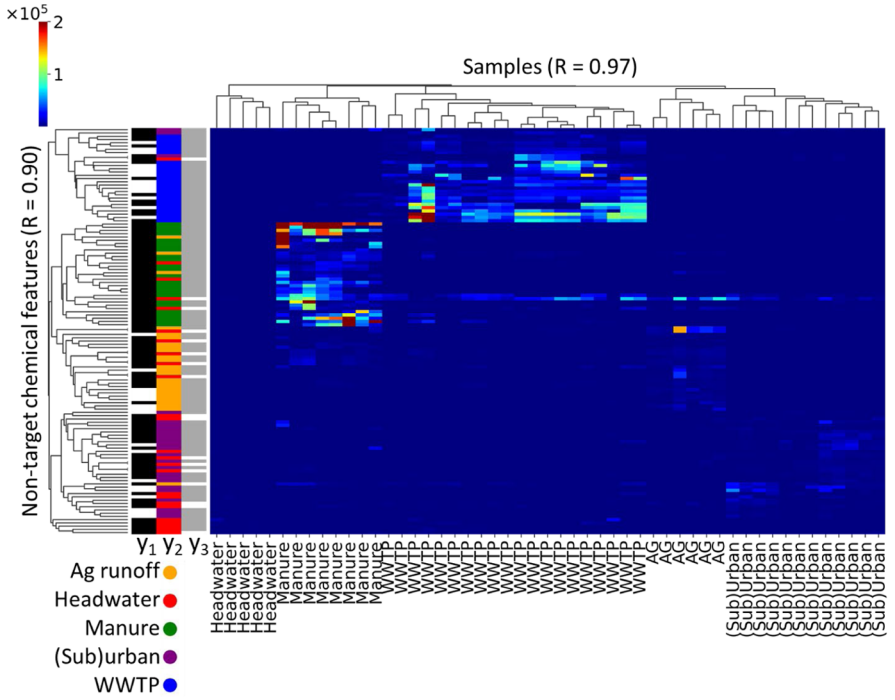


Figure 4. Hierarchical clustering heatmap (Bray−Curtis distance, average linkage) using only diagnostic chemical features (n = 25) for each source. The x and y clustering axes represent sample types and diagnostic chemical features, respectively. The z axis (color scale) represents feature intensity. The cophenetic correlation coefficient (R) for each cluster analysis is included adjacent to the axis title. The y axis color bars indicate whether a feature was unique for a single source (black) or present across multiple sources (white; y_1), which source the feature was diagnostic of (see legend inset; y_2), and whether the coefficient weight was positive (gray) or negative (white; y_3).

silico mixtures because this approach ignores the real behavior of chemicals during analytical detection. For agricultural and (sub)urban runoff, there was virtually no difference in presence probability between 0 and 0.01% source fractions (Figure S7). Only when the source percentage increased to 0.1% did the presence probability increase, suggesting that the SVC model can readily predict the presence or absence of these sources when peak intensities of chemical features are decreased by a factor of 1000. For headwaters, the presence probability was virtually identical to $p_{0\%}$ until the source percentage was 1%, suggesting that with this data set, headwater samples are only distinguishable with a dilution factor less than 100. In practice, we recommend using a suitable mean comparison test (e.g., t-test and Mann−Whitney U-test) to determine which source fraction is statistically different from $p_{0\%}$. Because these were modeled data, the statistical power is inflated with 1000 iterations. Using a t-test, $p_{0.1\%}$ and $p_{0\%}$ are statistically different (t = 3.80, df = 1998, and $p < 0.001$) for headwaters; however, assuming that the same mean and standard deviation were achieved but with only 10 samples, statistical difference would not be observed until a source fraction of 10%. In this instance, we have an example of statistical significance that does not equate to environmental significance. Therefore, with this data, it is hard to estimate an environmentally relevant presence/absence threshold for headwaters. Again, we emphasize that critical thresholds should be tested against real mixtures instead of in silico mixtures.

Of all the chemical features, only 0.2% were unique to headwaters whereas 37, 20, 4.1, and 2.8% were unique to WWTP, manure, (sub)urban, and agricultural samples, respectively (Figure S8). Furthermore, unlike all other sources, the most diagnostic chemical features for headwaters had negative coefficient weights (Figure 4), which are predictive of a source's absence. Of the top 100 most diagnostic chemical features, only 22 are positive for headwaters, while 84, 99, 100, and 100 are positive for agricultural and (sub)urban runoff, manure, and WWTP samples (Tables S4−S8). In other words, most diagnostic features for headwaters are absent in headwater samples and only present in other sources (Figure 4). These observations suggest that the SVC algorithm selects diagnostic features that are unique and/or disproportionately abundant for one source compared to another. The lack of diagnostic features that are abundant for headwater samples could explain why this is the lowest-quality fingerprint. Although SVC appeared to cue in on intense features, other machine learning algorithms may detect other diagnostic patterns within the data.

As the source percentage increased, the presence probability increased monotonically (Figures 3a and S7). This finding suggests that the presence probability could be used to quantify the source apportionment in a sample; however; this is fraught with uncertainty. Several factors are likely to influence peak intensity of a sample (e.g., matrix effects, instrument performance, and differential in situ attenuation/transformation), which will likely affect the probability estimate. Therefore, we encourage further fingerprint development and testing, particularly the fate and transport of chemical fingerprints, before utilizing this workflow to quantify source apportionment. In addition, we detected no general trend that increasing the number of diagnostic chemical features resulted in increased predictive performance (Figures 3a and S7). This is beneficial from a practical perspective: with fewer chemical fingerprints, this fingerprinting workflow could be translated for MS/MS instruments. Finally, without performing actual mixing experiments of different matrices, it is impossible to assess the lower performance limits of the SVC

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

model because this approach assumes a linear response between instrument response and non-target chemical concentration. Nevertheless, these results indicate that the algorithm is capable of predicting source presence even when chemical features are present at trace levels. Therefore, HRMS instrumentation is likely the limiting factor for being able to detect the presence of different sources instead of the computation limits of the algorithm. Thus, sample processing optimization may be needed to take full advantage of this fingerprinting workflow and other forensic tools.

In Situ Screening. In creek samples, the presence probability for most sources was relatively low except for one notable exception: the WWTP presence probability was >0.80 in spring, >0.73 in summer, and >0.20 in winter in Rickreall Creek at sample sites 4 and 5 (Figures 3b and S9). For $n = 25$ diagnostic features, $p_{0\%}$ (i.e., the preliminary probability threshold for presence) was 0.12, providing strong evidence that WWTP discharge is present. Upon closer inspection, the WWTP of Dallas, OR, is located between sites 3 and 4, which was not known a priori. The lower presence probability in winter is likely due to dilution from the winter rainy season. Interestingly, effluent from the Dallas WWTP was not included in the sampling, which suggests that this fingerprint can be generally applied to detect WWTP effluent in the environment, despite differences in chemical composition across facilities. In addition, the presence probability for wastewater was 0.17 and 0.12 at sites 4 and 5, respectively, in the Marys River during winter (Figure S9). The Philomath WWTP discharges treated effluent between sites 3 and 4, but the facility is only permitted to discharge during the winter rainy season when flows in the Marys River are higher; therefore, the absence of the WWTP fingerprint during the spring and summer sampling events is expected.

Although each system had an urban center within the sampling area, (sub)urban runoff was not detected above the critical presence/absence threshold probability in any creek, even Dixon Creek, which is entirely within the Corvallis city limits. Winter samples were not collected during rain events; therefore, it is possible that (sub)urban runoff had passed through the system prior to sampling. Higher winter base flows are maintained by shallow groundwater, which rises to the surface in winter months. Although it is expected that some portion of shallow groundwater is derived from (sub)urban runoff, our chemical analysis was constrained to only non-polar organics that were extractable using SPE cartridges. Therefore, many of the chemical features we targeted may have been lost due to sorption processes during transport. It is possible that broader portions of the chemical space are needed to capture chemical signatures that are expected to transport overland or through the subsurface. Finally, headwater signatures were moderately low throughout all creek samples, but as previously described, it is difficult to assess whether or not this is above an environmentally relevant presence/absence threshold. Although it is expected that headwater signatures should be present, we are unable to make a meaningful interpretation of the headwater signature, given the limitations of the workflow discussed previously.

Practical Considerations. Our data indicate that this fingerprinting workflow is capable of selecting a subset of chemical features that can distinguish chemical sources; however, there are a variety of considerations that may influence the workflow's success. As the chemical composition of a source becomes increasingly unique, it becomes easier to identify a hyperplane that maximally discriminates sources, regardless of sample size. As the chemical composition becomes increasingly similar, the hyperplane that separates sources becomes less distinct, and it may be necessary to increase the sample size, use non-linear kernels, and/or use other classification algorithms to better distinguish sources. Although decreasing sample size does not necessarily decrease separability, low sample sizes reduce the heterogeneity of the chemical composition of a source. In this study, we were able to detect the presence of a WWTP on Rickreall Creek even though this facility was not sampled to develop the diagnostic signature. This result is not too surprising given that both facilities are located in the Willamette Valley and serve communities of similar demographics; however, it is unlikely that this fingerprint is universally applicable to detect the presence of any municipal wastewater in receiving bodies of water because we have not captured the variability of chemical composition of this source. Depending on the goals of the fingerprint, it may or may not be important to capture the chemical variability of a specific source.

Although it would be advantageous to develop a universal chemical fingerprint that is invariant with time, we hypothesize that societal and ecological systems are continuously evolving; therefore, we predict that the chemical composition of any source is naturally dynamic. Even if a fingerprint is developed from a geographically, temporally, socially, and demographically diverse sampling campaign, the chemical composition may change as societies adopt new technologies and change how we interact with the environment. Therefore, if chemical forensic technologies are to become widely adopted, it is likely that routine updating of diagnostic fingerprints will be needed for at least some sources. Future research should explore the spatial and temporal limits for which a fingerprint is applicable.

Because this is the first data set analyzed using this approach, it is difficult to extract general rules of thumb about the workflow's performance. Nevertheless, we noticed three potential differences between successful and unsuccessful fingerprinting attempts. First, the fingerprinting workflow was ultimately successful when a majority of the most diagnostic chemical features were positive (i.e., predictive of source presence). For agricultural runoff, manure, (sub)urban runoff, and WWTP samples, 100, 87, 100, and 100% of the top 100 most diagnostic chemical features were positive, respectively, whereas for headwater samples, only 22% were positive.

Second, the most diagnostic features typically had large peak intensities that were largely unique to a source. This could potentially explain why the performance of the headwater fingerprint was low, given that there were few unique features present and most peak intensities were low (Figures 4 and S8). In some instances, two sources may have high chemical similarity with few, if any, unique features. In such instances, SVC may not be the most appropriate tool and other algorithms may need to be applied to this workflow. Similar to Kibbey et al.,[28,29] our recommendation for future studies is to incorporate multiple algorithms into the workflow. These algorithms will undoubtedly select other features that are diagnostic, which could help overcome the limitations associated with any individual algorithm.

Third, the successful application of this method depends on appropriate source selection and matching the data to the right classification algorithm. We developed chemical fingerprints for two point sources (WWTP and manure), two non-point sources [agricultural and (sub)urban runoff], and one

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

landscape source (headwater streams/lakes). Interestingly, the point sources generated fingerprints with the lowest critical presence/absence thresholds, which increased as the sources became increasingly distributed. It is reasonable to assume that point sources have increasingly modular matrices (densely connected subsets of chemical features that do not overlap with other sources) while distributed sources have increasingly nested matrices (overlapping subsets of chemical features). Nestedness is a measure of structures in an ecological system,[42] and the structure of chemical matrices could strongly influence a classification algorithm's performance. WWTP and manure samples had the most unique or unnested chemical composition (Figures 4 and S8) and also the highest-quality chemical fingerprint. Furthermore, the top 25 diagnostic features are unnested (i.e., little overlap in chemical composition; Figure 4). This unnested structure likely makes it easier to discriminate sources with a linear hyperplane, which might explain why classification improved after the non-diagnostic chemical features were removed. As the chemical composition of a source becomes increasingly nested, we predict that the discriminating performance of linear algorithms will decrease. Until non-linear algorithms are incorporated into this workflow, our recommendation is that practitioners should select distinct sources that are expected to have largely unique chemical features. More characterization of non-target data sets and their structure is needed to understand what makes a classification algorithm successful.

The chemicals in this analysis were restricted to non-polar organics and ESI + ionization. Although this segment of the chemical space generates rich data sets with thousands of features,[17,22,25] it is important to acknowledge that our processing protocol biases our results. Chemical bias is unavoidable, and although this is not inherently problematic, researchers will have to be pragmatic in selecting the portions of the chemical space that are best suited for a particular question. As mentioned, non-polar compounds, such as pharmaceuticals, may be retained in soils in overland flow or in subsurface discharge;[43,44] thus, including polar compounds that are more mobile in the environment may be more appropriate when tracking pollution sources in groundwater. Conversely, when tracking changes in the pollution signatures in sediments, non-polar features may be more appropriate.

It is unknown which portion of the chemical space is "best" suited for forensic analysis. Including a broader portion of the chemical space (e.g., using data from both ESI± modes) may be advantageous for selecting the most diagnostic features; however, increasing the chemical space may be met with diminishing returns as large data sets contain redundant information.[45,46] Redundant features are potentially problematic because each feature adds little novel information about the source. Within our data set, we find evidence of redundancy in the cluster analysis (y-axis, Figure 4). Chemical features with increasingly similar peak intensity patterns have smaller Bray Curtis distances and thus cluster together. Without further investigation, we do not know yet why they are similar, but we predict that chemical features with high degrees of similarity also have similar fate and transport characteristics. If all features have the same behavior in the environment, the fingerprint could be lost as a result of a single process (e.g., sorption), leading to false negatives during environmental screening. Therefore, we argue that a robust chemical fingerprint should consist of non-collinear features. Information theory could be applied to select features with low redundancy to maximize the information content contained within a chemical fingerprint.[47] Nevertheless, we recognize that some chemical clustering is expected, given that features are from the same source, and we also recognize that in some situations, having diagnostic features with similar fate and transport characteristics could be advantageous (e.g., non-sorptive features that transport well on the subsurface). Therefore, it is important to have a strong understanding of both the needs of a diagnostic fingerprint and the limitations of the chemical processing method used to develop the fingerprint. Furthermore, work could be carried out to determine which portions of the chemical space are most useful for different applications.

CONCLUSIONS

Multivariate statistics and machine learning tool algorithms are becoming increasingly popular for source tracking.[17,25,26,48,49] Nevertheless, this workflow is unique because it is capable of selecting a subset of features (<1%) that are most diagnostic of a source. For manure and WWTP, the workflow could develop a highly effective fingerprint with just ∼0.1% of the chemical features. Kibbey et al.[29] identified diagnostic subsets of PFAS by testing all possible combinations of chemicals (n = 1023 combinations of 10 PFAS components); however, this is not possible with non-target data sets with thousands of chemical features. For example, with the 7771 non-target chemical features that were considered in this study, our data set contains 2.0×10^{2339} possible feature combinations. Therefore, rank-sorting features based on their coefficient weights is an improvement over recent attempts to select diagnostic subsets of features.

Although this workflow could be used to prioritize features for structure elucidation, source apportionment, or source tracking,[17,18,25] which were part of our original motivation, we argue that the value of this workflow goes beyond these applications. Receiving bodies of water are "data loggers" that collect chemical information from all corners of a watershed. This information is recorded as tens of thousands of dissolved organic molecules, which represent the sum of all processes and activities occurring across a landscape. Therefore, it could be possible to simultaneously quantify many processes occurring within a watershed simply by screening a single water sample for the diagnostic chemical signatures of each process. Additionally, by quantifying transformations of diagnostic fingerprints, it may be possible to develop a smart tracer that could not only detect the presence of a pollution source within a watershed but also its relative location upstream from a sampling point. Finally, thousands of processes occur within ecosystems, yet we surmise that most cannot be quantified with off-the-shelf sensors. With this workflow, scientists can begin to quantify these processes, and we expect that the collective results of this workflow will provide insights into the natural world that have hitherto been unreachable.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c06655.

Detailed descriptions of sampling sites, field and laboratory processing steps, instrument analyses, com-

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

putational scripts, and additional results in the form of tables and figures (PDF)

AUTHOR INFORMATION

Corresponding Author

Gerrad D. Jones − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States;* orcid.org/0000-0002-1529-9506; Email: gerrad.jones@oregonstate.edu

Authors

Emmanuel Dávila-Santiago − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Cheng Shi − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Gouri Mahadwar − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Bridgette Medeghini − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Logan Insinga − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Rebecca Hutchinson − *School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon 97331-5501, United States; Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, Oregon 97331-3803, United States*

Stephen Good − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331-3906, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c06655

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge that Oregon State University in Corvallis, Oregon, is located within the traditional homelands of the Marys River or the Ampinefu Band of Kalapuya. Following the Willamette Valley Treaty of 1855, Kalapuya people were forcibly removed to reservations in Western Oregon. Today, living descendants of these people are a part of the confederated cTribes of Grand Ronde Community of Oregon (grandronde.org) and the confederated tribes of the Siletz Indians (ctsi.nsn.us). We acknowledge Anna Beran, Gary Black, Corey De La Cruz, Kevin Fear, Derek Godwin, Jeffrey Jenkins, Claudia Maier, Jeff Morre, Lewis Semprini, Quirine Van Swaay, Jane Vinesky, Liping Yang, and three anonymous reviewers for their assistance. Funding was provided by the National Science Foundation [NSF, CBET 1949013, EAR 1836768 (partial), and IIS 2046678 (partial)], Oregon State University-Agricultural Research Service (#8986A), and by the Strategic Environmental Research and Development Program (SERDP ER20-1375, partial). The authors declare no financial conflict of interest.

REFERENCES

(1) Knower, K. C.; To, S. Q.; Leung, Y.-K.; Ho, S.-M.; Clyne, C. D. Endocrine Disruption of the Epigenome: A Breast Cancer Link. *Endocr. Relat. Cancer* **2014**, *21*, T33−T55.

(2) Park, S.-J.; Kufareva, I.; Abagyan, R. Improved Docking, Screening and Selectivity Prediction for Small Molecule Nuclear Receptor Modulators Using Conformational Ensembles. *J. Comput. Aided Mol. Des.* **2010**, *24*, 459−471.

(3) Hotchkiss, A. K.; Rider, C. V.; Blystone, C. R.; Wilson, V. S.; Hartig, P. C.; Ankley, G. T.; Foster, P. M.; Gray, C. L.; Gray, L. E. Fifteen Years after "Wingspread"—Environmental Endocrine Disrupters and Human and Wildlife Health: Where We Are Today and Where We Need to Go. *Toxicol. Sci.* **2008**, *105*, 235−259.

(4) Schug, T. T.; Abagyan, R.; Blumberg, B.; Collins, T. J.; Crews, D.; DeFur, P. L.; Dickerson, S. M.; Edwards, T. M.; Gore, A. C.; Guillette, L. J.; Hayes, T.; Heindel, J. J.; Moores, A.; Patisaul, H. B.; Tal, T. L.; Thayer, K. A.; Vandenberg, L. N.; Warner, J. C.; Watson, C. S.; vom Saal, F. S.; Zoeller, R. T.; O'Brien, K. P.; Myers, J. P. Designing Endocrine Disruption out of the next Generation of Chemicals. *Green Chem.* **2013**, *15*, 181−198.

(5) Roberts, J.; Bain, P. A.; Kumar, A.; Hepplewhite, C.; Ellis, D. J.; Christy, A. G.; Beavis, S. G. Tracking Multiple Modes of Endocrine Activity in Australia's Largest Inland Sewage Treatment Plant and Effluent- Receiving Environment Using a Panel of in Vitro Bioassays. *Environ. Toxicol. Chem.* **2015**, *34*, 2271−2281.

(6) Bain, P. A.; Williams, M.; Kumar, A. Assessment of Multiple Hormonal Activities in Wastewater at Different Stages of Treatment. *Environ. Toxicol. Chem.* **2014**, *33*, 2297−2307.

(7) McRobb, F. M.; Sahagún, V.; Kufareva, I.; Abagyan, R. In Silico Analysis of the Conservation of Human Toxicity and Endocrine Disruption Targets in Aquatic Species. *Environ. Sci. Technol.* **2014**, *48*, 1964−1972.

(8) Blazer, V. S.; Iwanowicz, L. R.; Henderson, H.; Mazik, P. M.; Jenkins, J. A.; Alvarez, D. A.; Young, J. A. Reproductive Endocrine Disruption in Smallmouth Bass (Micropterus Dolomieu) in the Potomac River Basin: Spatial and Temporal Comparisons of Biological Effects. *Environ. Monit. Assess.* **2012**, *184*, 4309−4334.

(9) Fent, K. Progestins as Endocrine Disrupters in Aquatic Ecosystems: Concentrations, Effects and Risk Assessment. *Environ. Int.* **2015**, *84*, 115−130.

(10) Kicman, A. T. Pharmacology of Anabolic Steroids. *Br. J. Pharmacol.* **2008**, *154*, 502−521.

(11) Säfholm, M.; Ribbenstedt, A.; Fick, J.; Berg, C.; Safholm, M.; Ribbenstedt, A.; Fick, J.; Berg, C. Risks of Hormonally Active Pharmaceuticals to Amphibians: A Growing Concern Regarding Progestagens. *Philos. Trans. R. Soc., B* **2014**, *369*, 20130577.

(12) Zeilinger, J.; Steger-Hartmann, T.; Maser, E.; Goller, S.; Vonk, R.; Länge, R. Effects of Synthetic Gestagens on Fish Reproduction. *Environ. Toxicol. Chem.* **2009**, *28*, 2663.

(13) Scholz, N. L.; Myers, M. S.; McCarthy, S. G.; Labenia, J. S.; McIntyre, J. K.; Ylitalo, G. M.; Rhodes, L. D.; Laetz, C. A.; Stehr, C. M.; French, B. L.; McMillan, B.; Wilson, D.; Reed, L.; Lynch, K. D.; Damm, S.; Davis, J. W.; Collier, T. K. Recurrent Die-Offs of Adult Coho Salmon Returning to Spawn in Puget Sound Lowland Urban Streams. *PLoS One* **2011**, *6*, No. e28013.

(14) Bruner-Tran, K. L.; Gnecco, J.; Ding, T.; Glore, D. R.; Pensabene, V.; Osteen, K. G. Exposure to the Environmental Endocrine Disruptor TCDD and Human Reproductive Dysfunction: Translating Lessons from Murine Models. *Reprod. Toxicol.* **2017**, *68*, 59−71.

(15) Skakkebaek, N. E.; Rajpert-De Meyts, E.; Buck Louis, G. M.; Toppari, J.; Andersson, A.-M.; Eisenberg, M. L.; Jensen, T. K.; Jørgensen, N.; Swan, S. H.; Sapra, K. J.; Ziebe, S.; Priskorn, L.; Juul, A. Male Reproductive Disorders and Fertility Trends: Influences of Environment and Genetic Susceptibility. *Physiol. Rev.* **2016**, *96*, 55−97.

(16) Spromberg, J. A.; Baldwin, D. H.; Damm, S. E.; Mcintyre, J. K.; Huff, M.; Sloan, C. A.; Anulacion, B. F.; Davis, J. W.; Scholz, N. L. Coho Salmon Spawner Mortality in Western US Urban Watersheds:

Bioinfiltration Prevents Lethal Storm Water Impacts. *J. Appl. Ecol.* **2016**, *53*, 398−407.
(17) Peter, K. T.; Tian, Z.; Wu, C.; Lin, P.; White, S.; Du, B.; McIntyre, J. K.; Scholz, N. L.; Kolodziej, E. P. Using High-Resolution Mass Spectrometry to Identify Organic Contaminants Linked to Urban Stormwater Mortality Syndrome in Coho Salmon. *Environ. Sci. Technol.* **2018**, *52*, 10317−10327.
(18) Tian, Z.; Zhao, H.; Peter, K. T.; Gonzalez, M.; Wetzel, J.; Wu, C.; Hu, X.; Prat, J.; Mudrock, E.; Hettinger, R.; Cortina, A. E.; Biswas, R. G.; Kock, F. V. C.; Soong, R.; Jenne, A.; Du, B.; Hou, F.; He, H.; Lundeen, R.; Gilbreath, A.; Sutton, R.; Scholz, N. L.; Davis, J. W.; Dodd, M. C.; Simpson, A.; Mcintyre, J. K.; Kolodziej, E. P. A Ubiquitous Tire Rubber−Derived Chemical Induces Acute Mortality in Coho Salmon. *Science* **2021**, *371*, 185−189.
(19) Fraga, C. G.; Acosta, G. A. P.; Crenshaw, M. D.; Wallace, K.; Mong, G. M.; Colburn, H. A. Impurity Profiling to Match a Nerve Agent to Its Precursor Source for Chemical Forensics Applications. *Anal. Chem.* **2011**, *83*, 9564−9572.
(20) Pillsbury, L.; Goodwin, K.; Brown, D. *Statewide Water Quality Toxics Assessment Report*; State Of Oregon Department of Environmental Quality, 2015. DEQ15-LAB-0065-TR.
(21) Johnson, G. W.; Ehrlich, R. State of the Art Report on Multivariate Chemometric Methods in Environmental Forensics*. *Environ. Forensics* **2002**, *3*, 59−79.
(22) Schollée, J. E.; Schymanski, E. L.; Avak, S. E.; Loos, M.; Hollender, J. Prioritizing Unknown Transformation Products from Biologically-Treated Wastewater Using High-Resolution Mass Spectrometry, Multivariate Statistics, and Metabolic Logic. *Anal. Chem.* **2015**, *87*, 12121−12129.
(23) Purschke, K.; Vosough, M.; Leonhardt, J.; Weber, M.; Schmidt, T. C. Evaluation of Nontarget Long-Term LC−HRMS Time Series Data Using Multivariate Statistical Approaches. *Anal. Chem.* **2020**, *92*, 12273−12281.
(24) Du, B.; Lofton, J. M.; Peter, K. T.; Gipe, A. D.; James, C. A.; McIntyre, J. K.; Scholz, N. L.; Baker, J. E.; Kolodziej, E. P. Development of Suspect and Non-Target Screening Methods for Detection of Organic Contaminants in Highway Runoff and Fish Tissue with High-Resolution Time-of-Flight Mass Spectrometry. *Environ. Sci.: Processes Impacts* **2017**, *19*, 1185−1196.
(25) Carpenter, C. M. G.; Wong, L. Y. J.; Johnson, C. A.; Helbling, D. E. Fall Creek Monitoring Station: Highly Resolved Temporal Sampling to Prioritize the Identification of Nontarget Micropollutants in a Small Stream. *Environ. Sci. Technol.* **2019**, *53*, 77−87.
(26) Zhang, X.; Lohmann, R.; Dassuncao, C.; Hu, X. C.; Weber, A. K.; Vecitis, C. D.; Sunderland, E. M. Source Attribution of Poly- and Perfluoroalkyl Substances (PFASs) in Surface Waters from Rhode Island and the New York Metropolitan Area. *Environ. Sci. Technol. Lett.* **2016**, *3*, 316−321.
(27) Masiá, A.; Campo, J.; Blasco, C.; Picó, Y. Ultra-High Performance Liquid Chromatography−Quadrupole Time-of-Flight Mass Spectrometry to Identify Contaminants in Water: An Insight on Environmental Forensics. *J. Chromatogr. A* **2014**, *1345*, 86−97.
(28) Kibbey, T. C. G.; Jabrzemski, R.; O'Carroll, D. M. Supervised Machine Learning for Source Allocation of Per- and Polyfluoroalkyl Substances (PFAS) in Environmental Samples. *Chemosphere* **2020**, *252*, 126593.
(29) Kibbey, T. C. G.; Jabrzemski, R.; O'Carroll, D. M. Source Allocation of Per- and Polyfluoroalkyl Substances (PFAS) with Supervised Machine Learning: Classification Performance and the Role of Feature Selection in an Expanded Dataset. *Chemosphere* **2021**, *275*, 130124.
(30) Shade, A.; Handelsman, J. Beyond the Venn Diagram: The Hunt for a Core Microbiome. *Environ. Microbiol.* **2012**, *14*, 4−12.
(31) Callao, M. P.; Ruisánchez, I. An Overview of Multivariate Qualitative Methods for Food Fraud Detection. *Food Control* **2018**, *86*, 283−293.
(32) Schollée, J. E.; Bourgin, M.; von Gunten, U.; McArdell, C. S.; Hollender, J. Non-Target Screening to Trace Ozonation Transformation Products in a Wastewater Treatment Train Including Different Post-Treatments. *Water Res.* **2018**, *142*, 267−278.
(33) Ccanccapa-Cartagena, A.; Pico, Y.; Ortiz, X.; Reiner, E. J. Suspect, Non-Target and Target Screening of Emerging Pollutants Using Data Independent Acquisition: Assessment of a Mediterranean River Basin. *Sci. Total Environ.* **2019**, *687*, 355−368.
(34) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **2017**, *51*, 11505−11512.
(35) Altenburger, R.; Brack, W.; Burgess, R. M.; Busch, W.; Escher, B. I.; Focks, A.; Mark Hewitt, L.; Jacobsen, B. N.; de Alda, M. L.; Backhaus, T.; Ginebreda, A.; Hilscherová, K.; Hollender, J.; Hollert, H.; Neale, P. A.; Schulze, T.; Schymanski, E. L.; Teodorovic, I.; Tindall, A. J.; de Aragão Umbuzeiro, G.; Vrana, B.; Zonja, B.; Krauss, M. Future Water Quality Monitoring: Improving the Balance between Exposure and Toxicity Assessments of Real-World Pollutant Mixtures. *Environ. Sci. Eur.* **2019**, *31*, 12.
(36) Webster, J. P.; Kover, S. C.; Bryson, R. J.; Harter, T.; Mansell, D. S.; Sedlak, D. L.; Kolodziej, E. P. Occurrence of Trenbolone Acetate Metabolites in Simulated Confined Animal Feeding Operation (CAFO) Runoff. *Environ. Sci. Technol.* **2012**, *46*, 3803−3810.
(37) Jones, G. D.; Benchetler, P. V.; Tate, K. W.; Kolodziej, E. P. Mass Balance Approaches to Characterizing the Leaching Potential of Trenbolone Acetate Metabolites in Agro-Ecosystems. *Environ. Sci. Technol.* **2014**, *48*, 3715−3723.
(38) Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibáñez, M.; Portolés, T.; de Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipaničev, D.; Rostkowski, P.; Hollender, J. Non-Target Screening with High-Resolution Mass Spectrometry: Critical Review Using a Collaborative Trial on Water Analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 6237−6255.
(39) Ricotta, C.; Podani, J. On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning. *Ecol. Complex.* **2017**, *31*, 201−205.
(40) McCune, B.; Grace, J.; Urban, D. *Analysis of Ecological Communities*; MjM Software Design, 2002.
(41) Eide, I.; Neverdal, G.; Thorvaldsen, B.; Arneberg, R.; Grung, B.; Kvalheim, O. M. Toxicological Evaluation of Complex Mixtures: Fingerprinting and Multivariate Analysis. *Environ. Toxicol. Pharmacol.* **2004**, *18*, 127−133.
(42) Fortuna, M. A.; Stouffer, D. B.; Olesen, J. M.; Jordano, P.; Mouillot, D.; Krasnov, B. R.; Poulin, R.; Bascompte, J. Nestedness versus Modularity in Ecological Networks: Two Sides of the Same Coin? *J. Anim. Ecol.* **2010**, *79*, 811−817.
(43) Huang, W.; Peng, P. a.; Yu, Z.; Fu, J. Effects of Organic Matter Heterogeneity on Sorption and Desorption of Organic Contaminants by Soils and Sediments. *Appl. Geochem.* **2003**, *18*, 955−972.
(44) Matamoros, V.; Bayona, J. M. Elimination of Pharmaceuticals and Personal Care Products in Subsurface Flow Constructed Wetlands. *Environ. Sci. Technol.* **2006**, *40*, 5811.
(45) Holm, L.; Sander, C. Removing Near-Neighbour Redundancy from Large Protein Sequence Collections. *Bioinformatics* **1998**, *14*, 423−429.
(46) Peng, H.; Long, F.; Ding, C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226−1238.
(47) Solorio-Fernández, S.; Martínez-Trinidad, J. F.; Carrasco-Ochoa, J. A. A Supervised Filter Feature Selection Method for Mixed Data Based on Spectral Feature Selection and Information-Theory Redundancy Analysis. *Pattern Recogn. Lett.* **2020**, *138*, 321.
(48) Charbonnet, J. A.; Rodowa, A. E.; Joseph, N. T.; Guelfo, J. L.; Field, J. A.; Jones, G. D.; Higgins, C. P.; Helbling, D. E.; Houtz, E. F. Environmental Source Tracking of Per- And Polyfluoroalkyl

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Substances within a Forensic Context: Current and Future Techniques. *Environ. Sci. Technol.* **2021**, *55*, 7237−7245.
(49) Hug, C.; Sievers, M.; Ottermanns, R.; Hollert, H.; Brack, W.; Krauss, M. Linking Mutagenic Activity to Micropollutant Concentrations in Wastewater Samples by Partial Least Square Regression and Subsequent Identification of Variables. *Chemosphere* **2015**, *138*, 176−182.

https://doi.org/10.1021/acs.est.1c06655
Environ. Sci. Technol. XXXX, XXX, XXX−XXX