Customizable Machine Learning Models for Rapid Microplastic Identification Using Raman Microscopy

Benjamin Lei[a], Justine R. Bissonnette[a], Úna E. Hogan[a], Avery E. Bec[a], Xinyi Feng[a], Rodney D. L. Smith[a,b,c]*

[a]*Department of Chemistry, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1*

[b]*Waterloo Institute for Nanotechnology, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1*

[c]*Waterloo Artificial Intelligence Institute, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1*

Correspondence to:

rodsmith@uwaterloo.ca

ABSTRACT

Raman spectroscopy is commonly used in microplastics identification, but equipment variations yield inconsistent data structures that disrupt development of communal analytical tools. We report a strategy to overcome the issue using a database of high-resolution, full-window Raman spectra. This approach enables customizable analytical tools to be easily created – a feature we demonstprate by creating machine learning classification models using open-source random forest, K-nearest neighbors, and multi-layer perceptron algorithms. These models yield >95% classification accuracy when trained on spectroscopic data with spectroscopic data downgraded to 1, 2, 4 or 8 cm^{-1} spacings in Raman shift. The accuracy can be maintained even in non ideal conditions, such as with spectroscopic sampling rates of 1 kHz and when microplastic particles are outside the focal plane of the laser. This approach enables the creation of classification models that are robust and adaptable to varied spectrometer setups and experimental needs.

INTRODUCTION

The ability to accurately identify microplastics is impeded by a combination of their dimensions, structural diversity, tendency to adsorb contaminants, and presence in natural environments in concentration ranges that span multiple orders of magnitude.[1–7] Global interest in this class of environmental contaminant has rapidly increased in recent years, spurred by reports of microplastics identified in ecosystems around the planet,[8–13] within living tissue,[14–16] and in the human body.[17,18] This has led to significant research and development in environmental processing techniques,[19–21] analytical techniques to detect and quantify particles,[20,22–26] and discussions regarding research data management and accessibility. Many instrumental techniques have found use in the identification of microplastics, most notably including infrared spectroscopy, Raman spectroscopy, ^{1}H NMR spectroscopy and GC/MS. Many of these techniques continue to rely upon labor-intensive analyses that typically include preparation of samples, pre-screening of purified samples, then quantitative analysis of characterization data. The time-intensive and costly nature of the process greatly decreases the rate at which microplastics research can be carried out. An accurate and robust protocol capable of identifying microplastic particles in an automated or semi-automated fashion would streamline global research; such capabilities would facilitate global efforts to establish a quantitative understanding of the true scale and potential impact of the microplastics problem.

The vibrational spectroscopies, Fourier transform infrared (FTIR) and Raman spectroscopy, are the most prominent of techniques being used for microplastics identification. The molecular vibrations of each type of plastic are dependent structure, effectively providing a unique “fingerprint” for each type of plastic. The conventional approach to identification of samples using these spectroscopies is therefore through comparison of spectra for unknown samples with one of known standards. Such standards can be found in databases but, as discussed in a recent review, the selection of publicly available databases is quite limited.[20] Further, spectra within these databases are often variable in characteristics such as the frequency range and spacing of vibrational frequency for datapoints. Manual identification of plastics by direct comparison with known standards is further complicated because over 10,000 chemicals have been documented as additives or substances used in the manufacture of plastics.[27,28] This complexity in the chemistry of pristine plastics is compounded by known, and potentially unknown, environmental degradation processes.[29–31] Most researcher groups have attempted to overcome the challenges by creating internal collections of spectroscopic standards that are optimized to their specific research objectives, some of which have been made publicly accessible.[32,33] The complexity of commercial plastics can be difficult to match even when using such customized databases, as acknowledged in a combined FTIR/Raman spectroscopy analysis.[34] There is a pressing need to explore and develop strategies to accurately and consistently identify plastics using imperfect spectroscopic data. Such strategies should be developed in parallel for both Raman and FTIR spectroscopy as their varied strengths and limitations lead to differences in quantitative capabilities.[35]

The rapid development of machine learning (ML) algorithms provides many accessible tools that are being explored to improve microplastics analysis protocols.[36] An ML classification model based on the random forest (RF) algorithm trained on 306 Raman spectra from the open-source SLOPP and SLOPP-E databases was found to be capable of 89% classification accuracy.[37] This value was boosted to 94% with extensive pre-processing of the spectra to increase compatibility. Inclusion of a principal component analysis (PCA) step was found to offer improvements in the classification accuracy of Raman spectroscopic maps, as compared to assignments based on conventional fingerprinting.[38] A convolution neural network (CNN) model trained on Raman spectra from the diverse open source RRUFF database[39] enabled identification and quantitation of solvated species during real-time measurements in a flow-cell, but struggled to consistently identify microplastics.[40] A K-nearest neighbors (KNN) classification model trained on 906 FTIR spectra from manually labelled environmental microplastics offered 90% classification accuracy with no user intervention.[41] Beyond identification, efforts have also targeted the improvement of spectrum quality – including the automated removal of artefacts, baselines, and contaminants.[42]

Herein, we demonstrate near-quantitative classification accuracy of Raman spectra using three different ML algorithms. We strategically create a database of high-resolution Raman spectra spanning the full frequency range of molecular vibrations to enable the database to adapt to different experimental setups and objectives. We apply an augmentation protocol to convert 108 different Raman spectra into 4520 spectra for training ML algorithms. Classification models trained on this dataset using KNN, RF and multi-layer preceptor (MLP) algorithms are applied to analyze a series of Raman microscopic maps on diverse microplastics. The results demonstrate that this approach can yield near perfect accuracy even with spectroscopic acquisition times as low as 1 ms, and for samples that are outside of the focal plane of the microscope.

EXPERIMENTAL

Materials. A total of 108 plastic samples purchased from diverse sources, including commercial suppliers such as Sigma Aldrich, specialty suppliers such as Cospheric, bulk plastic suppliers contacted through Ali-Baba, commercial plastic goods purchased through McMaster Carr, and consumer goods sourced locally. Raman spectra collected for the database used the plastic samples in their as-received state. Microparticles used for testing the classification models were created from the plastic samples using a metal grater. All spectra are linked to their identity and supplier in a dataset available under CC BY 4.0 license.

Raman Microscopy. Spectroscopic measurements were performed on a Renishaw inVia Reflex Raman microscope. Measurements were performed using a 532 nm laser (Renishaw DPSSL, 50 mW) and a 2400 lines mm^{-1} diffraction grating, or a 633 nm laser (Renishaw HeNe laser, 17 mW) with an 1800 lines mm^{-1} diffraction grating. Samples that were in powder or microparticle form were loaded on glass microscope slides, while bulk plastics were placed directly on the sample stage. Microscopic maps were performed using a single Raman acquisition per pixel, with the timing indicated for each map. The total acquisition time is therefore the number of pixels multiplied by the acquisition time. The Renishaw WiRE 5.3 software package was used to control the instrument and process the data, which included polynomial baseline subtraction and removal of cosmic rays. Processed spectra were saved as ASCII text files and stored in an in-house database using mongoDB 5.0.8.

Data Handling. Python (version 3.7.10) was used to process spectroscopic data, train classification models, and apply them to classify Raman spectroscopic maps. Data was accessed from the database using pymongo (version 3.10.1), handled and augmented using numpy (version 1.21.4), and machine learning steps were executed using scikit-learn (version 1.0.2). Samples of code used for the entire process are available under CC BY 4.0 license. Classification metrics were calculated using the number of true positive (*tp*), false positive (*fp*), true negative (*tn*), and false negative (*fn*) predictions. Metrics are defined in equations (1) through (4):

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \tag{1}$$

$$Sensitivity = \frac{tp}{tp+fn} \tag{2}$$

$$Specificity = \frac{tn}{fp+tn} \tag{3}$$

$$Precision = \frac{tp}{tp+fp} \tag{4}$$

RESULTS & DISCUSSION

Database Creation. An in-house Raman spectroscopy database was initiated with 108 unique plastic samples representing 14 different plastic types. Plastic samples were selected to be diverse in nature to capture some of the diversity inherent in both plastic types and plastic additives that are commonly used in the global economy. The sample set consisted of polyethylene (PE; 43), polypropylene (PP; 27), polytetrafluoroethylene (PTFE; 24), nylon (21), poly(methylmethacrylate) (PMMA; 17), polystyrene (PS; 14), polyurethane (PU; 10), silicone (9), polyethylene terephthalate (PET; 6), polyoxymethylene (POM; 4), polyvinyl chloride (PVC; 4), polycarbonate (PC; 3), acrylonitrile butadiene styrene (ABS; 2), and polyester (PEs; 2). The SynchroScan^{TM} feature on a Renishaw inVia Raman microscope was used to acquire spectra between 100 and 4000 cm^{-1} twice per sample: once with a 633 nm laser paired to an 1800 lines mm^{-1} diffraction grating and once with a 532 nm laser paired to a 2400 lines mm^{-1} diffraction grating. This provides continuity throughout the vibrational spectroscopy window with spacings of *ca*. 0.6 and 0.7 cm^{-1} between datapoints, respectively. All individual spectra were manually processed by polynomial baseline subtraction and removal of any cosmic rays. Spectra exhibiting very low quality, such as those with strong fluorescence, were removed from consideration to yield a total of 186 high-resolution Raman spectra for further analysis.

The fingerprinting approach conventionally used to identify plastic samples from their Raman spectra becomes challenging for non-ideal samples, or samples with similar chemical structures. The fingerprint nature of spectra for individual plastics is readily visible in comparisons of the spectra for pure PE, PS, nylon, PTFE, ABS, and PMMA (Figure 1A). The utility of this approach can be readily seen in this optimal situation – samples with similar spectra such as PS and ABS show sufficient differences for accurate identification. Such optimal situations cannot be relied upon in analysis of environmental microplastics, however, because commercial plastics may contain such a diversity of additives[27,28] and experience environmental degradation that may alter their spectra.[29–31] Such additives or chemical modifications disrupt the fingerprints of plastic samples by masking features, introducing new dominant peaks, or significantly degrading the quality of acquired spectra due to fluorescence. These issues are seen upon comparison of a series of polyethylene samples with varied additives (Figure 1B). It has been demonstrated that chemical similarity can also lead to false positives. For example, long alkyl chains, such as sodium dodecyl sulfate used to clean environmental samples or stearates used in the manufacture of laboratory gloves, produce spectra with sufficient similarity to polyethylene that false positives can occur.[43] Such challenges impede identification of samples using conventional fingerprinting approaches.

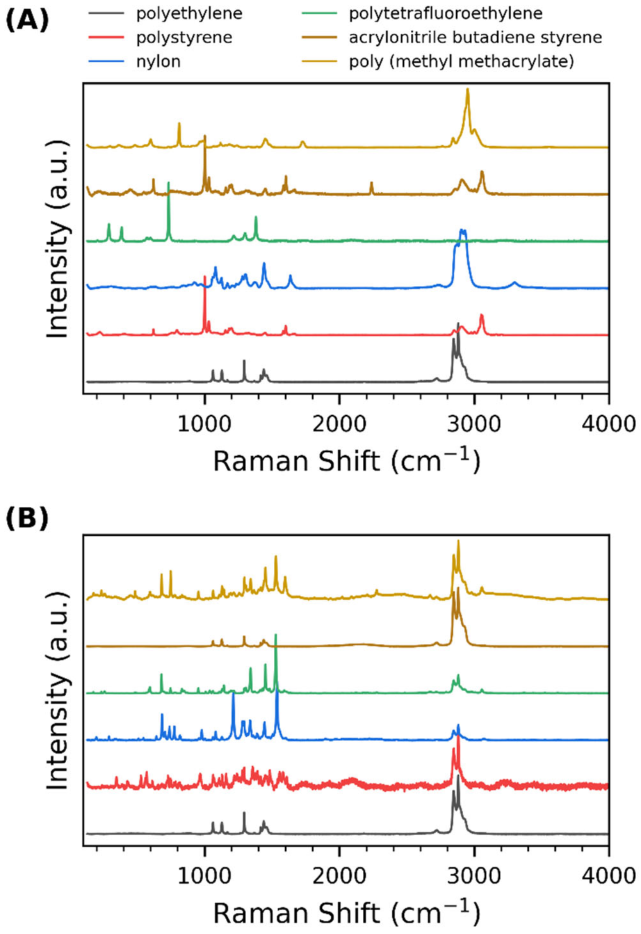

Figure 1. Raman spectra on representative plastics. **(A)** Variation in spectra that enables differentiation of composition by “fingerprinting,” with six different types of plastic. **(B)** Raman spectra on pure polyethylene (bottom curve) and 5 samples containing unknown chemical

additives. Note the red dataset shows the effect of severe fluorescence from an additive. All results shown are baseline subtracted.

Training Dataset Preparation. The rapid identification of unknown samples necessitates low spectroscopic acquisition times, a Raman shift window that is wide enough to differentiate similar samples, and an ability to analyze data with low signal-to-noise ratios (SNR) and possible contaminants. The Raman spectra acquired in this report are clearly incompatible with these needs, but the spectra span the complete vibrational window and provides both high resolution and high signal-to-noise ratios. This provides flexibility to downgrade the data, for example by truncating all spectra to a region of interest or interpolating to a uniform Raman shift grid, such that ML models can be tailored to individual experimental setups. Processing of the Raman spectra to train ML models that are compatible with the experimental demands is accomplished through systematic data truncation, data interpolation, and data augmentation. Data preparation is demonstrated here for optimized use with our Raman microscopic mapping process.

Data preparation begins with truncation of spectra to the desired region and interpolation onto a consistent Raman shift grid. The models trained here are truncated between 720 to 1800 cm^{-1} – a range selected to capture the most prominent of vibrations for plastic samples while maintaining compatibility with the 633 nm laser and 1800 lines mm^{-1} diffraction grating in our instrument. Interpolation of the data ensures that all spectra share uniform range and spacing in Raman shift values, which is important because the intensity at these values serve as the features in the ML models. We tested four sets of models here, with spacings of 1, 2, 4 and 8 cm^{-1}. The 2 cm^{-1} models are used for detailed discussions.

Appending a series of “blank” samples to the dataset enables ML models to handle any spectra that contain no plastics. We accomplish this by introducing artificial spectra, with 20 flat lines and 20 broad inverted parabola. The parabola are randomly centered between 500 and 4000 cm^{-1} with peak intensities randomly set between 0.1 and 0.3.

Spectra that we wish to analyze will contain significant variation in intensity of the desired plastic peaks and signal-to-noise ratios that will be substantially worse than the training data. These features are introduced into the training dataset by an augmentation process that simultaneously increases the size of the training dataset. Each individual spectrum is treated by (i) introducing random noise across the spectrum to generate an SNR randomly between 3 and 15 for the strongest peak, then (ii) normalizing the spectrum such that the maximum intensity is between 0.3 and 1.0. This process is repeated *n* times to introduce randomized variations in SNR and peak intensity for each individual sample within the training dataset. The models trained here use an *n* of 20, effectively converting the 186 unique spectra and 40 baseline spectra into a final training dataset that consists of 4520 spectra.

Classification Model Training. A significant portion of the normalized Raman scattering intensity data exists in regions where no peaks reside and therefore carry no useful information. The efficiency of model training and application can be increased by applying a feature selection protocol to select only the most informative features. We employ the *SelectKBest* algorithm using the *Mutual Information (Classification)* scoring function from *sklearn* to select the most informative features within the dataset. Application of this algorithmic approach to the 2 cm^{-1} spacing dataset highlights the utility of this approach (Figure 2A): reduction of the initial 541 datapoints to the 190 of which are deemed most informative results in consideration of only datapoints that are associated with peaks in one or more of the spectra within the training dataset. The number of discrete datapoints (K_{best}) being considered is a user-selected hyperparameter that is optimized during model training.

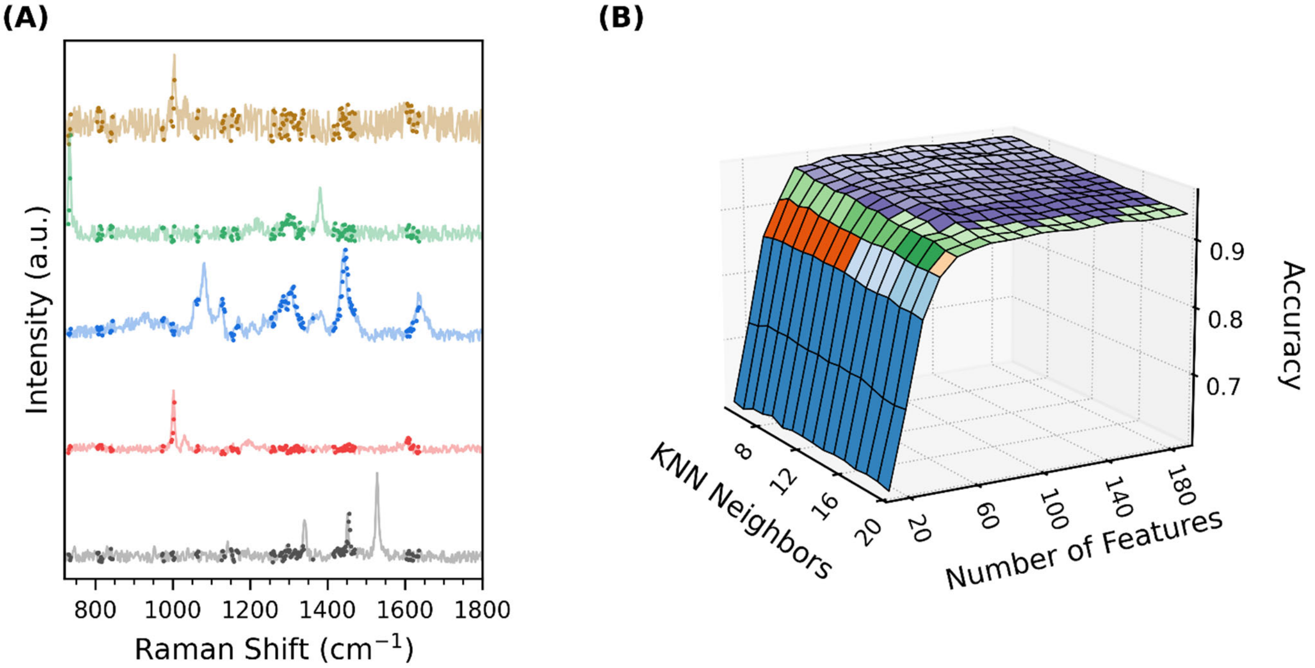

Figure 2. Sample selection of data points and algorithm hyperparameters for training machine learning models. (**A**) The Raman scattering intensity at individual wavenumbers serve as features in the models. The training and application of the model is expedited by using only a

subset of features. The mutual information algorithm in *sci-kit learn* selects the most informative points (solid circles) from the full spectrum (faded lines). (**B**) Model hyperparameters are optimized by monitoring classification accuracy in a grid search spanning the number of KNN neighbors (K_N) and the number of features selected (K_{best}). Color scheme is to enhance visibility of the performance gradient.

The KNN, RF and multi-layer perceptor (MLP) algorithms were optimized for classification of plastic samples. Hyperparameters for each algorithm were optimized by systematically training and testing ML models using a stratified shuffle-split process with 5 splits. A uniform grid was used for all hyperparameters, which included the K_{best} parameter introduced above and the quantile cut-off parameter for all models, alongside algorithm-specific hyperparameters (Table 1). The optimization process was repeated for datasets using Raman shift spacings of 1, 2, 4 and 8 cm^{-1}. A sample optimization grid showing classification accuracy against K_{best} and K_N for KNN classification with the 2 cm^{-1} spacing data demonstrates the process (Figure 2B) and the optimized values (Table 1); all other optimization grids are available (Figures S1-S3). Final classification models were trained for each combination of classification algorithm and Raman shift spacing using the hyperparameters that yielded the highest classification accuracy.

Table 1. Hyperparameters for ML models and their optimized values for 2 cm^{-1} dataset.

Model	Parameter	Minimum	Maximum	Step Size	Optimized
All	K_{best}	10	200	10	120 (KNN) 190 (MLP & RF)
	Quantile cut-off	2.5%	7.5%	2.5%	7.5% (KNN) 2.5% (MLP & RF)
KNN	K_N (neighbors)	5	21	1	5
RF	*N* (estimators)	10	200	10	110
	D (depth)	10	200	10	190
MLP	Neurons	10	1000	20	30

Strengths and weaknesses of models based on each classification algorithm can be identified through inspection of confusion matrices and classification metrics. Each of the three classification models correctly identifies the majority of plastics (ie, true positive classifications) with over 90% consistency (Figure S4, Tables 2, S1, and S2). Errors appear across all models to induce a decrease in both precision and sensitivity of the models, with the dominant error being the assignment of ABS to PS (Figure S4). This error arises due to the truncated spectroscopic window and the similarity in chemistry of these two plastics – the nitrile vibrational peak that is the key differentiator between these two plastics and is located at a higher frequency than the window maximum. Resolution of these two plastics would require expansion of the window to *ca*. 2200 cm^{-1}, but this would negatively impact identification of all other plastics by truncating the information-rich fingerprint region of the spectrum. Low sensitivity also appears in several models due to assignment of PVC as PC, POM as PTFE, and POM as a "blank" (Figure S4; Tables 2, S1 and S2). It is expected that these errors will be minimized as the size of the spectroscopic database grows. Beyond these significant errors, the classification models show unique combinations of strengths and weaknesses that are visible as varied patterns of errors in the confusion matrices.

Table 2. Classification metrics obtained while training a model using the KNN algorithm.

Plastic	Accuracy	Precision	Sensitivity	Specificity
ABS	0.994	0.929	0.325	1.000
blank	0.987	0.972	0.956	0.994
nylon	0.995	0.998	0.952	1.000
PMMA	0.999	0.994	0.988	1.000
PC	0.997	0.841	0.967	0.998
PEs	0.999	1.000	0.925	1.000
PE	0.993	0.981	0.981	0.996
PET	0.994	0.832	0.992	0.995
POM	0.988	0.656	0.738	0.993
PP	0.991	0.940	0.989	0.991
PS	0.994	0.912	0.996	0.994
PTFE	0.990	0.976	0.933	0.997
PU	0.998	0.971	0.990	0.999
PVC	0.996	0.942	0.812	0.999
silicone	0.997	0.977	0.950	0.999

A vote-based strategy was introduced as a fourth approach to classification of spectra. Spectra for unknown samples were first classified using each of the KNN, MLP and RF models. The results were subsequently tallied and compared to each other. Classification as a plastic type was committed to a spectrum only if two or three models agreed on the classification; the spectrum was otherwise assigned as a “blank.” This fourth approach explores whether the three unique algorithms may have complementary properties that can improve accuracy and consistency when classifying unknown spectra.

Raman Microscopic Mapping. The accuracy of classification models was tested on Raman microscopic maps of blended microplastics. Spectroscopic maps of an area containing a mix of PS, PMMA and PE serve as the first test case (Figure 3). The map consists of 4131 spectra acquired over a 400 x 250 μm area in 5 μm steps in each direction. Individual spectra were acquired for 100 ms at full laser power and were processed and classified in an automated process. All spectra were prepared for classification by (i) asymmetric least squares baseline subtraction,[44] (ii) spectrum normalization, and (iii) truncation of the data to the datapoints with Raman shifts closest to those used as features in the respective ML models. Normalization of spectra by the maximum of the baseline-subtracted data is straightforward in map pixels where plastics exist, but the exaggeration of noise caused by applying this approach to “blank” pixels invariably leads to assignments of random plastics in these pixels. Screening of strategies to identify such “blank” spectra such that a secondary normalization protocol can be applied led us to institute an approximated SNR check on all spectra. The SNR check compared the maximum of a baseline-subtracted spectrum, which provides a measure of signal intensity, to the minimum, which provides an approximation of noise. Any spectrum with an SNR ratio below 1.5 was designated as a likely “blank” pixel and divided by the global maximum intensity of the spectroscopic map. The SNR cutoff value was empirically chosen based on its ability to yield consistent, high-quality classifications with each individual algorithm. This value also lies significantly below the SNR threshold of 3 that is conventionally considered for analytical purposes, meaning that it should not introduce excessive bias into the outcomes.

Spectroscopic maps created by applying the KNN (Figure 3A), MLP (Figure 3B) and RF (Figure 3C) models to the data show general agreement in assignments. All major particles that reside within the focal plane of the laser are successfully identified – this includes PMMA spheres, a PE shard, and PS fragments. The maps deviate from one another in regions where low-quality spectra exist, however, such as at the edges of particles and where particles are outside the focal plane of the laser. Such errors include the incorrect assignment of a PP particle where one doesn’t exist in the KNN map, PVC particles in the MLP map, and a scattering of pixels in isolated regions to a variety of plastics for all three algorithms. The combination of all three classification maps through the voting protocol removed most of these errors from the spectroscopic maps (Figure 3D). This automated classification process was repeated with different blends of microplastic particles to test the limits of this approach. Good performance was obtained for most samples (Figures S5-S6), but the classic limitations of Raman spectroscopy were observed. For example, the 2cm^{-1} KNN model successfully identified PS in mixtures of PS and PET at 1000 ms and 100 ms spectroscopic acquisition times, but the fluorescence originating from PET led to its incorrect assignment as nylon at 1000 ms acquisition times and to a “blank” at 100 ms (Figure S7). Microparticles of PS were also observed to be somewhat problematic in that spectra were occasionally misidentified as the related ABS plastic (Figure S8) – as noted above, the distinguishing feature between these two plastics is a nitrile vibration that lies outside of the window used by the classification models. Each of the individual classification models is therefore capable of accurate identification of microplastics, and the combination of multiple ML algorithms can be a useful strategy in minimizing errors in challenging situations.

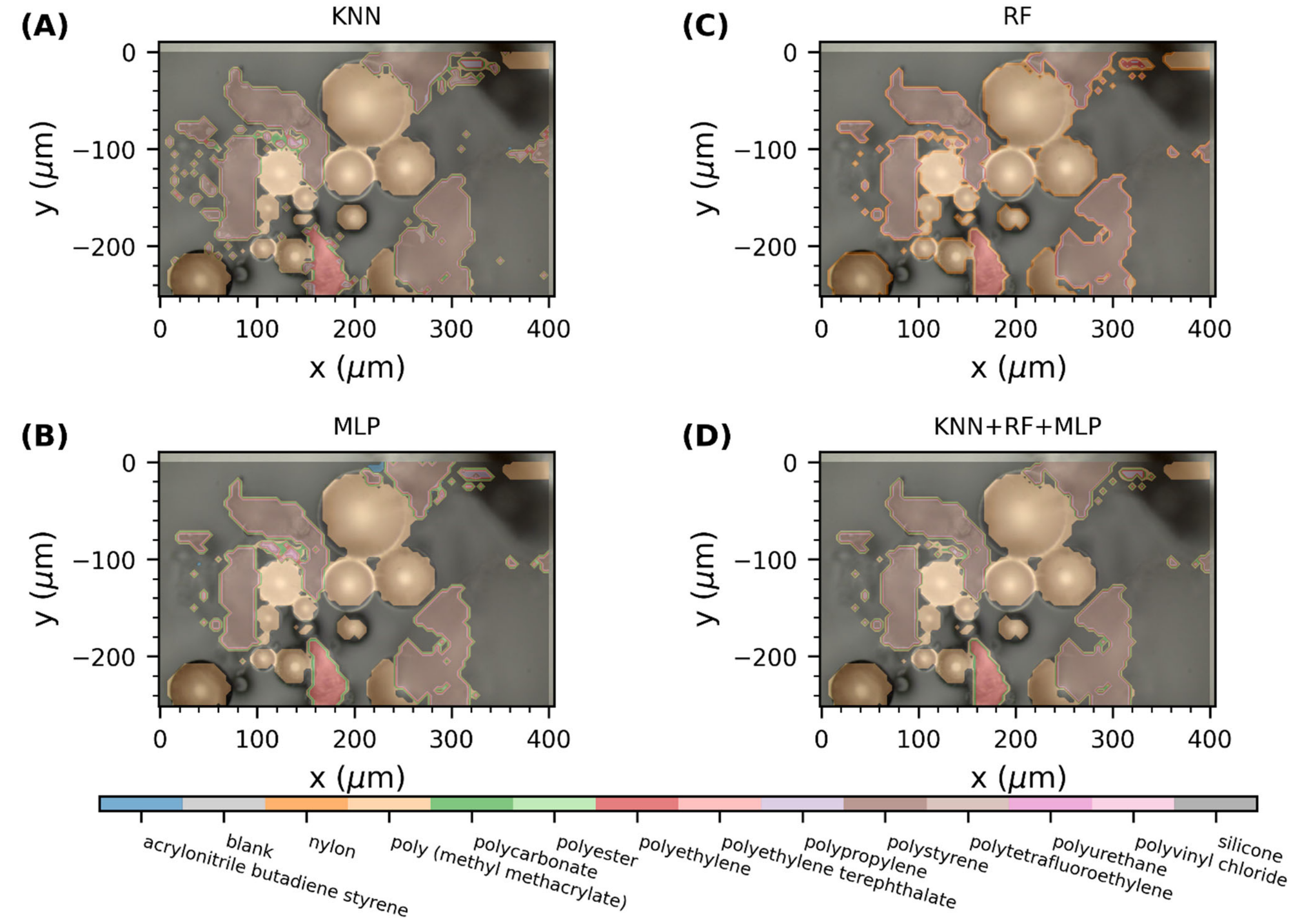

Figure 3. Comparison of optimized model performance in identifying microplastics in Raman microscopic mapping experiments. The **(A)** KNN, **(B)** MLP, and **(C)** RF models all show reasonable accuracy, but each exhibits unique inaccuracies. **(D)** Combination of the three models in a voting algorithm overcomes the inaccuracies inherent in each individual model. Maps consist of 4131 spectra acquired with 100 ms spectroscopic acquisition times with 5 μm x 5 μm grid size. Images show white light image with assignments overlaid with 50% transparency.

High Sampling Frequency. The dilute concentrations and physical size of microparticles in environmental samples tends to greatly increase the experimental times necessary to make confident classifications of particles. The ability to make confident assignments using rapid spectroscopic measurements would help to decrease time commitments, thereby increasing the rate of quality data acquisition. The ability of classification models to assign spectra acquired at high sampling rates, and therefore with very low SNR ratios, was tested by mapping the same area of maps with varied spectroscopic acquisition times. A spectroscopic acquisition of 1000 ms yields high SNR ratios that yield perfect assignments from all three models (Figures 4 and S9). When particles are within the focal plane of the laser, all models continue to perform very well as the spectroscopic acquisition time decreases through 10 and 1 ms. There is an increase, however, in the number of incorrectly assigned pixels as the SNR ratio decreases. The RF model yields the worst performance in these high acquisition rate tests (Figure 4A-C), while the KNN (Figure S9A-9C) and MLP models (Figure S9D-9F) continue to perform remarkably well even for 1 ms spectroscopic acquisition times. In line with the limitations of Raman spectroscopy, classification accuracy begins to suffer when particles are outside of the laser focal plane. Under optimal conditions, the ML models trained here can accurately assignment spectra even when acquired as fast as 1000 spectra per second; sampling at 10-100 spectra per second is feasible even under non-ideal conditions.

Accuracy for Out-of-Focus Microparticles. Acquisition of optimal Raman spectra requires that the focal plane of the laser to be well positioned relative to surface of the object being analyzed. This poses a challenge for 2-dimensional mapping or attempts to rapidly target individual microplastics because the particles often move out of focus. The classification models trained here accurately classify Raman maps acquired on odd-shaped particles that are out of focus, such as the polypropylene particle in Figure 5. This capability does begin to break down as spectrum quality further decreases at millisecond spectroscopic acquisition times.

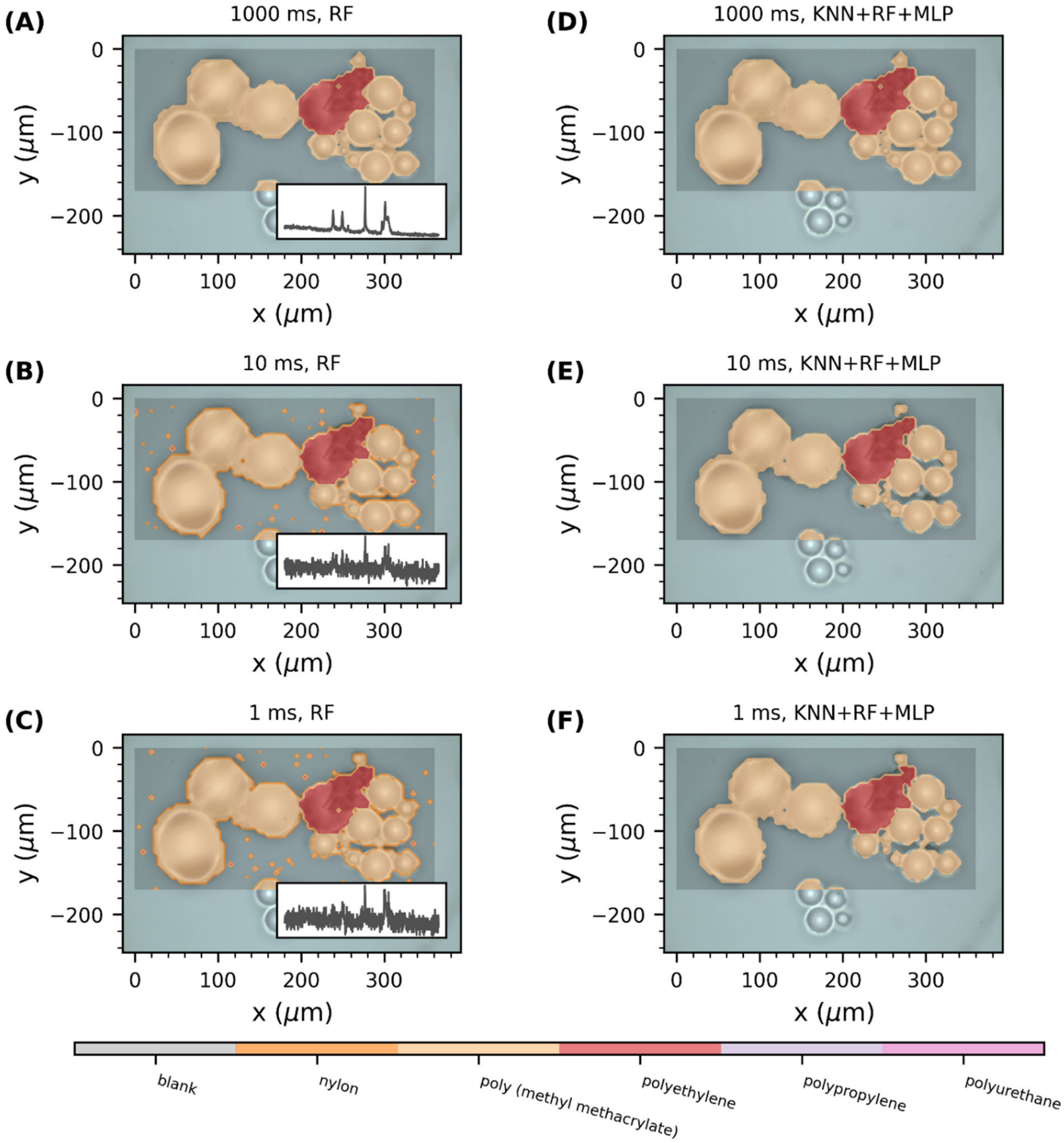

Figure 4. Comparison of classifications as a function of spectroscopic acquisition time. Classifications made using optimized RF classification model with (**A**) 1000 ms, (**B**) 10 ms, and (**C**) 1 ms spectroscopic acquisition times. Classifications made using the combination of all three models with (**D**) 1000 ms, (**E**) 10 ms, and (**F**) 1 ms spectroscopic acquisition times. Images show white light image with assignments overlaid with 50% transparency. Insets show sample spectrum to demonstrate approximate SNR in spectra. Data from the RF and MLP models are available Figure S9. Each map contains 2555 spectra, acquired in 5 μm steps in x and y dimensions.

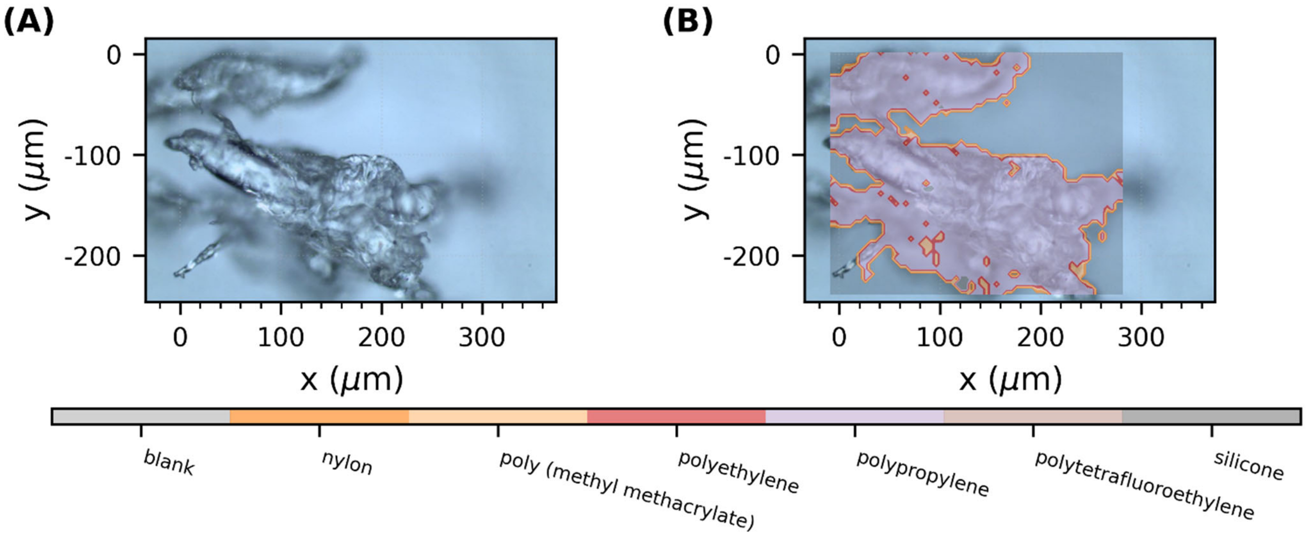

Figure 5. Influence of microscope focus on the accuracy of classification models. (**A**) White light image of an irregularly shaped polypropylene particle. (**B**) Spectroscopic map assigned using a KNN model with 2 cm^{-1} spacing. Spectroscopic acquisition time was 1 second for individual pixels. Raman map contains 2891 spectra, acquired in 5 μm steps in x and y dimensions.

Application to External Data. The ML models and deployment strategy developed here maintain their performance when tested against real-world spectra acquired from open databases. A study comparing Raman to FTIR spectroscopy by Cabernard *et al.* included publication of 1,199 Raman spectra from 237 unique plastic samples.[45] These spectra are labeled and have data spacing of 0.8 cm^{-1}. Removal of spectra for samples that were not present in our training data, such as polybutadiene, petroleum jelly and silk, left 324 spectra that are compatible with our models. The SLoPP and SLoPP-E databases contain a total of 261 labeled spectra with data spacing of *ca*. 3 cm^{-1}, of which 187 spectra are compatible with the models trained here.[33] Performance metrics were calculated following application of the combined approach using the 2 cm^{-1} (Cabernard et al. database) or 4 cm^{-1} (SLoPP and SLoPP-E databases) ML models here (Tables S3-S4). Metrics from each database are comparable with each other and to those obtained during model training (Table 2), with high classification accuracy on all plastics. The fundamental issues associated with ABS and POM continue to yield the largest issues; PEs exhibits problems arising from false negatives for both databases, which is likely associated with the small number of Pes spectra integrated into the training data. Comparisons of classifications by each model to sample identity are available.

Outlook. The strategy behind the Raman spectroscopy database created here provides significant experimental flexibility to ensure that it can adapt to evolution of the rapidly expanding study of environmental microplastics. The data structure, where all spectra are high resolution and span the full molecular vibration frequency range, enables ML models to be tailored to any variation of instrumentation. The classification models created by downgrading Raman shift spacings to 1, 2, 4 and 8 cm^{-1} show that the KNN, MLP and RF classification algorithms can provide excellent accuracy for each of these options (Figures 2, S1-S3). Models can therefore be optimized for a low-cost, low-resolution system suitable for portability and field research, all the way up to research grade instrument capable of high spatial resolution mapping. The approach can be tailored for development of flow cells and real-time analysis of microplastics,[46–48] analytical mapping of surfaces or filtered samples,[21] or to support the development of algorithms and techniques to automate the counting and measurement of microparticles.

Problems associated with unpredictable background signals and potential contributions from fluorescence and contaminants have led to development of strategies to improve the quality of spectra prior to use in fingerprinting or ML analysis.[42] The training and application approaches used here successfully remove much of these concerns. The models themselves are trained using baseline-corrected spectra with a range of artificially installed noise such that noisy real-world data can be analyzed with no smoothing. The datapoints are ranked in terms of information content such that the most useful datapoints can be selected and used; this minimizes worries about contaminants but also prevents their analysis. Finally, an automated asymmetric least squares baseline subtraction algorithm was found to be applicable to a wide range of spectra without a need to tweak parameters, which greatly simplifies data preparation. The ML classification models created here are both simple to apply and amenable to automated analysis.

Each of the KNN, MLP and RF algorithms yield high accuracy classification even for spectra with low signal-to-noise ratios. Quality Raman spectra can often be acquired on plastics with spectroscopic acquisition times on the seconds timescale. Such high-quality spectra (e.g., Figure 4A) are easily analyzed and all models here are capable of >96% classification accuracy even for complex mixtures of plastics (Figure 3). This accuracy is maintained for most plastics even when increasing the sampling frequency by three orders of magnitude

(Figure 4), or when samples are outside of the focal plane of the laser (Figure 5). We note that weakly scattering samples such as PTFE are challenging to classify from low quality spectra, such as those acquired using low spectroscopic acquisition times or when samples are outside of the focal plane of the laser. The fluorescence from PET and the structural similarity between PS and ABS also pose classic problems for Raman analysis. The ability to analyze such challenging samples with spectroscopic acquisition times on the order of 100 ms nonetheless shows the remarkable robustness of the classification strategy used here.

The strengths of the ML models created here are remarkable considering the moderate sized spectroscopic database used to generate them. Future expansion of the size and diversity of the plastics in the database will expand the capabilities and reliability of classification models created with it. Microplastics collected from the environment are often contaminated by biofilms, inorganic materials, and organic molecules such as pesticides; determining the impact of such contaminants on the accuracy of the classification models marks the next step in which development must proceed.

CONCLUSIONS

A database containing 108 high resolution Raman spectra was shown to be effective at creating machine learning classification models for microplastic analysis. Random forest, K-nearest neighbors, and multi-layer perceptron algorithms in the scikit-learn python package were shown to be capable of producing classification models that yield >95% classification accuracy. Plastic samples that produce strong peaks in Raman spectra could be accurately assigned even at spectroscopic sampling rates of 1 kHz. This feature may make the models suitable for real-time monitoring applications – especially considering that similar performance can be obtained when downgrading the spectroscopic data to Raman spacings up to 8 cm^{-1}. More challenging plastic samples required sampling rates to be decreased towards 1-100 Hz to maintain high accuracy. These results demonstrate the value of creating databases using only consistent, complete, high-quality spectroscopic data. It is anticipated that this database will continue to grow in both size and diversity, providing the community with a valuable tool to advance microplastics research.

AUTHOR INFORMATION

Corresponding Author. rodsmith@uwaterloo.ca

Present Address. University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1

ASSOCIATED CONTENT

Data Availability. Sample computer code and data that support the findings of this study are available under Creative Commons Attribution 4.0 International License in Borealis: The Canadian Dataverse Repository at https://doi.org/10.5683/SP3/LSN0R0.

Supporting Information. Hyperparameter optimization grids, confusion matrices, tables containing ML classification metrics, additional Raman microscopic maps.

ACKNOWLEDGEMENTS

We thank Environment and Climate Change Canada (ECCC) and the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding through the *Plastics for a Cleaner Future* program. The Canada Foundation for Innovation and Ontario Ministry of Research are thanked for support in the purchase of instrumentation through the *John R. Evans Leaders Fund* and *Small Infrastructure Funds*, respectively.

REFERENCES

(1) Amobonye, A.; Bhagwat, P.; Raveendran, S.; Singh, S.; Pillai, S. Environmental Impacts of Microplastics and Nanoplastics: A Current Overview. *Front. Microbiol.* **2021**, *12*, 768297. https://doi.org/10.3389/fmicb.2021.768297.

(2) Akdogan, Z.; Guven, B. Microplastics in the Environment: A Critical Review of Current Understanding and Identification of Future Research Needs. *Environ. Pollut.* **2019**, *254*, 113011. https://doi.org/10.1016/j.envpol.2019.113011.

(3) Malankowska, M.; Echaide-Gorriz, C.; Coronas, J. Microplastics in Marine Environment: A Review on Sources, Classification, and Potential Remediation by Membrane Technology. *Environ. Sci. Water Res. Technol.* **2021**, *7* (2), 243–258. https://doi.org/10.1039/D0EW00802H.

(4) Zhang, Q.; Xu, E. G.; Li, J.; Chen, Q.; Ma, L.; Zeng, E. Y.; Shi, H. A Review of Microplastics in Table Salt, Drinking Water, and Air: Direct Human Exposure. *Environ. Sci. Technol.* **2020**, *54* (7), 3740–3751. https://doi.org/10.1021/acs.est.9b04535.

(5) Du, J.; Xu, S.; Zhou, Q.; Li, H.; Fu, L.; Tang, J.; Wang, Y.; Peng, X.; Xu, Y.; Du, X. A Review of Microplastics in the Aquatic Environmental: Distribution, Transport, Ecotoxicology, and Toxicological Mechanisms. *Environ. Sci. Pollut. Res.* **2020**, *27* (11), 11494–11505. https://doi.org/10.1007/s11356-020-08104-9.

(6) Padervand, M.; Lichtfouse, E.; Robert, D.; Wang, C. Removal of Microplastics from the Environment. A Review. *Environ. Chem. Lett.* **2020**, *18* (3), 807–828. https://doi.org/10.1007/s10311-020-00983-1.

(7) Kazmiruk, T. N.; Kazmiruk, V. D.; Bendell, L. I. Abundance and Distribution of Microplastics within Surface Sediments of a Key Shellfish Growing Region of Canada. *PLOS ONE* **2018**, *13* (5), e0196005. https://doi.org/10.1371/journal.pone.0196005.

(8) Oliveri Conti, G.; Ferrante, M.; Banni, M.; Favara, C.; Nicolosi, I.; Cristaldi, A.; Fiore, M.; Zuccarello, P. Micro- and Nano-Plastics in Edible Fruit and Vegetables. The First Diet Risks Assessment for the General Population. *Environ. Res.* **2020**, *187*, 109677. https://doi.org/10.1016/j.envres.2020.109677.

(9) Shen, M.; Zeng, Z.; Wen, X.; Ren, X.; Zeng, G.; Zhang, Y.; Xiao, R. Presence of Microplastics in Drinking Water from Freshwater Sources: The Investigation in Changsha, China. *Environ. Sci. Pollut. Res.* **2021**, *28* (31), 42313–42324. https://doi.org/10.1007/s11356-021-13769-x.

(10) *Microplastic in the Environment: Pattern and Process*; Bank, M. S., Ed.; Environmental Contamination Remediation and Management; Springer International Publishing: Cham, 2022. https://doi.org/10.1007/978-3-030-78627-4.

(11) Hamilton, B. M.; Jantunen, L.; Bergmann, M.; Vorkamp, K.; Aherne, J.; Magnusson, K.; Herzke, D.; Granberg, M.; Hallanger, I. G.; Gomiero, A.; Peeken, I. Monitoring Microplastics in the Atmosphere and Cryosphere in the Circumpolar North: A Case for

Multi-Compartment Monitoring. *Arct. Sci.* **2022**, AS-2021-0054. https://doi.org/10.1139/AS-2021-0054.

(12) Welsh, B.; Aherne, J.; Paterson, A. M.; Yao, H.; McConnell, C. Atmospheric Deposition of Anthropogenic Particles and Microplastics in South-Central Ontario, Canada. *Sci. Total Environ.* **2022**, *835*, 155426. https://doi.org/10.1016/j.scitotenv.2022.155426.

(13) Roblin, B.; Ryan, M.; Vreugdenhil, A.; Aherne, J. Ambient Atmospheric Deposition of Anthropogenic Microfibers and Microplastics on the Western Periphery of Europe (Ireland). *Environ. Sci. Technol.* **2020**, *54* (18), 11100–11108. https://doi.org/10.1021/acs.est.0c04000.

(14) Amato-Lourenço, L. F.; Carvalho-Oliveira, R.; Júnior, G. R.; dos Santos Galvão, L.; Ando, R. A.; Mauad, T. Presence of Airborne Microplastics in Human Lung Tissue. *J. Hazard. Mater.* **2021**, *416*, 126124. https://doi.org/10.1016/j.jhazmat.2021.126124.

(15) Deng, Y.; Zhang, Y.; Lemos, B.; Ren, H. Tissue Accumulation of Microplastics in Mice and Biomarker Responses Suggest Widespread Health Risks of Exposure. *Sci. Rep.* **2017**, *7* (1), 46687. https://doi.org/10.1038/srep46687.

(16) Atamanalp, M.; Köktürk, M.; Uçar, A.; Duyar, H. A.; Özdemir, S.; Parlak, V.; Esenbuğa, N.; Alak, G. Microplastics in Tissues (Brain, Gill, Muscle and Gastrointestinal) of Mullus Barbatus and Alosa Immaculata. *Arch. Environ. Contam. Toxicol.* **2021**, *81* (3), 460–469. https://doi.org/10.1007/s00244-021-00885-5.

(17) Ragusa, A.; Svelato, A.; Santacroce, C.; Catalano, P.; Notarstefano, V.; Carnevali, O.; Papa, F.; Rongioletti, M. C. A.; Baiocco, F.; Draghi, S.; D'Amore, E.; Rinaldo, D.; Matta, M.; Giorgini, E. Plasticenta: First Evidence of Microplastics in Human Placenta. *Environ. Int.* **2021**, *146*, 106274. https://doi.org/10.1016/j.envint.2020.106274.

(18) Leslie, H. A.; van Velzen, M. J. M.; Brandsma, S. H.; Vethaak, A. D.; Garcia-Vallejo, J. J.; Lamoree, M. H. Discovery and Quantification of Plastic Particle Pollution in Human Blood. *Environ. Int.* **2022**, *163*, 107199. https://doi.org/10.1016/j.envint.2022.107199.

(19) Pfeiffer, F.; Fischer, E. K. Various Digestion Protocols Within Microplastic Sample Processing—Evaluating the Resistance of Different Synthetic Polymers and the Efficiency of Biogenic Organic Matter Destruction. *Front. Environ. Sci.* **2020**, *8*, 572424. https://doi.org/10.3389/fenvs.2020.572424.

(20) Cowger, W.; Gray, A.; Christiansen, S. H.; DeFrond, H.; Deshpande, A. D.; Hemabessiere, L.; Lee, E.; Mill, L.; Munno, K.; Ossmann, B. E.; Pittroff, M.; Rochman,

C.; Sarau, G.; Tarby, S.; Primpke, S. Critical Review of Processing and Classification Techniques for Images and Spectra in Microplastic Research. *Appl. Spectrosc.* **2020**, *74* (9), 989–1010. https://doi.org/10.1177/0003702820929064.

(21) Faltynkova, A.; Johnsen, G.; Wagner, M. Hyperspectral Imaging as an Emerging Tool to Analyze Microplastics: A Systematic Review and Recommendations for Future Development. *Microplastics Nanoplastics* **2021**, *1* (1), 13. https://doi.org/10.1186/s43591-021-00014-y.

(22) Woo, H.; Seo, K.; Choi, Y.; Kim, J.; Tanaka, M.; Lee, K.; Choi, J. Methods of Analyzing Microsized Plastics in the Environment. *Appl. Sci.* **2021**, *11* (22), 10640. https://doi.org/10.3390/app112210640.

(23) Kwon, J.-H.; Kim, J.-W.; Pham, T. D.; Tarafdar, A.; Hong, S.; Chun, S.-H.; Lee, S.-H.; Kang, D.-Y.; Kim, J.-Y.; Kim, S.-B.; Jung, J. Microplastics in Food: A Review on Analytical Methods and Challenges. *Int. J. Environ. Res. Public. Health* **2020**, *17* (18), 6710. https://doi.org/10.3390/ijerph17186710.

(24) Ivleva, N. P. Chemical Analysis of Microplastics and Nanoplastics: Challenges, Advanced Methods, and Perspectives. *Chem. Rev.* **2021**, *121* (19), 11886–11936. https://doi.org/10.1021/acs.chemrev.1c00178.

(25) Cheng, Y.-L.; Zhang, R.; Tisinger, L.; Cali, S.; Yu, Z.; Chen, H. Y.; Li, A. Characterization of Microplastics in Sediment Using Stereomicroscopy and Laser Direct Infrared (LDIR) Spectroscopy. *Gondwana Res.* **2021**. https://doi.org/10.1016/j.gr.2021.10.002.

(26) Primpke, S.; Lorenz, C.; Rascher-Friesenhausen, R.; Gerdts, G. An Automated Approach for Microplastics Analysis Using Focal Plane Array (FPA) FTIR Microscopy and Image Analysis. *Anal. Methods* **2017**, *9* (9), 1499–1511. https://doi.org/10.1039/C6AY02476A.

(27) Hahladakis, J. N.; Velis, C. A.; Weber, R.; Iacovidou, E.; Purnell, P. An Overview of Chemical Additives Present in Plastics: Migration, Release, Fate and Environmental Impact during Their Use, Disposal and Recycling. *J. Hazard. Mater.* **2018**, *344*, 179–199. https://doi.org/10.1016/j.jhazmat.2017.10.014.

(28) Wiesinger, H.; Wang, Z.; Hellweg, S. Deep Dive into Plastic Monomers, Additives, and Processing Aids. *Environ. Sci. Technol.* **2021**, *55* (13), 9339–9351. https://doi.org/10.1021/acs.est.1c00976.

(29) Zhang, K.; Hamidian, A. H.; Tubić, A.; Zhang, Y.; Fang, J. K. H.; Wu, C.; Lam, P. K. S. Understanding Plastic Degradation and Microplastic Formation in the Environment: A Review. *Environ. Pollut.* **2021**, *274*, 116554. https://doi.org/10.1016/j.envpol.2021.116554.

(30) Chamas, A.; Moon, H.; Zheng, J.; Qiu, Y.; Tabassum, T.; Jang, J. H.; Abu-Omar, M.; Scott, S. L.; Suh, S. Degradation Rates of Plastics in the Environment. *ACS Sustain. Chem. Eng.* **2020**, *8* (9), 3494–3511. https://doi.org/10.1021/acssuschemeng.9b06635.

(31) *Freshwater Microplastics: Emerging Environmental Contaminants?*; Wagner, M., Lambert, S., Eds.; The Handbook of Environmental Chemistry; Springer International Publishing: Cham, 2018; Vol. 58. https://doi.org/10.1007/978-3-319-61615-5.

(32) Cowger, W.; Steinmetz, Z.; Gray, A.; Munno, K.; Lynch, J.; Hapich, H.; Primpke, S.; De Frond, H.; Rochman, C.; Herodotou, O. Microplastic Spectral Classification Needs an Open Source Community: Open Specy to the Rescue! *Anal. Chem.* **2021**, *93* (21), 7543–7548. https://doi.org/10.1021/acs.analchem.1c00123.

(33) Munno, K.; De Frond, H.; O'Donnell, B.; Rochman, C. M. Increasing the Accessibility for Characterizing Microplastics: Introducing New Application-Based and Spectral Libraries of Plastic Particles (SLoPP and SLoPP-E). *Anal. Chem.* **2020**, *92* (3), 2443–2451. https://doi.org/10.1021/acs.analchem.9b03626.

(34) Vinay Kumar, B. N.; Löschel, L. A.; Imhof, H. K.; Löder, M. G. J.; Laforsch, C. Analysis of Microplastics of a Broad Size Range in Commercially Important Mussels by Combining FTIR and Raman Spectroscopy Approaches. *Environ. Pollut.* **2021**, *269*, 116147. https://doi.org/10.1016/j.envpol.2020.116147.

(35) Dong, M.; She, Z.; Xiong, X.; Ouyang, G.; Luo, Z. Automated Analysis of Microplastics Based on Vibrational Spectroscopy: Are We Measuring the Same Metrics? *Anal. Bioanal. Chem.* **2022**, *414* (11), 3359–3372. https://doi.org/10.1007/s00216-022-03951-6.

(36) Weisser, J.; Pohl, T.; Heinzinger, M.; Ivleva, N. P.; Hofmann, T.; Glas, K. The Identification of Microplastics Based on Vibrational Spectroscopy Data – A Critical Review of Data Analysis Routines. *TrAC Trends Anal. Chem.* **2022**, *148*, 116535. https://doi.org/10.1016/j.trac.2022.116535.

(37) Ramanna, S.; Morozovskii, D.; Swanson, S.; Bruneau, J. Machine Learning of Polymer Types from the Spectral Signature of Raman Spectroscopy Microplastics Data. arXiv January 14, 2022.

(38) Fang, C.; Luo, Y.; Zhang, X.; Zhang, H.; Nolan, A.; Naidu, R. Identification and Visualisation of Microplastics via PCA to Decode Raman Spectrum Matrix towards Imaging. *Chemosphere* **2022**, *286*, 131736. https://doi.org/10.1016/j.chemosphere.2021.131736.

(39) Lafuente, B.; Downs, R. T.; Yang, H.; Stone, N. The Power of Databases: The RRUFF Project. In *Highlights in Mineralogical Crystallography*; De Gruyter: Berlin, Germany, 2015; pp 1–30.

(40) Post, C.; Brülisauer, S.; Waldschläger, K.; Hug, W.; Grüneis, L.; Heyden, N.; Schmor, S.; Förderer, A.; Reid, R.; Reid, M.; Bhartia, R.; Nguyen, Q.; Schüttrumpf, H.; Amann, F. Application of Laser-Induced, Deep UV Raman Spectroscopy and Artificial Intelligence in Real-Time Environmental Monitoring—Solutions and First Results. *Sensors* **2021**, *21* (11), 3911. https://doi.org/10.3390/s21113911.

(41) Kedzierski, M.; Falcou-Préfol, M.; Kerros, M. E.; Henry, M.; Pedrotti, M. L.; Bruzaud, S. A Machine Learning Algorithm for High Throughput Identification of FTIR Spectra: Application on Microplastics Collected in the Mediterranean Sea. *Chemosphere* **2019**, *234*, 242–251. https://doi.org/10.1016/j.chemosphere.2019.05.113.

(42) Brandt, J.; Mattsson, K.; Hassellöv, M. Deep Learning for Reconstructing Low-Quality FTIR and Raman Spectra─A Case Study in Microplastic Analyses. *Anal. Chem.* **2021**, *93* (49), 16360–16368. https://doi.org/10.1021/acs.analchem.1c02618.

(43) Witzig, C. S.; Földi, C.; Wörle, K.; Habermehl, P.; Pittroff, M.; Müller, Y. K.; Lauschke, T.; Fiener, P.; Dierkes, G.; Freier, K. P.; Zumbülte, N. When Good Intentions Go Bad—False Positive Microplastic Detection Caused by Disposable Gloves. *Env. Sci Technol* **2020**, 9.

(44) He, S.; Zhang, W.; Liu, L.; Huang, Y.; He, J.; Xie, W.; Wu, P.; Du, C. Baseline Correction for Raman Spectra Using an Improved Asymmetric Least Squares Method. *Anal Methods* **2014**, *6* (12), 4402–4407. https://doi.org/10.1039/C4AY00068D.

(45) Cabernard, L.; Roscher, L.; Lorenz, C.; Gerdts, G.; Primpke, S. Comparison of Raman and Fourier Transform Infrared Spectroscopy for the Quantification of Microplastics in

the Aquatic Environment. *Environ. Sci. Technol.* **2018**, *52* (22), 13279–13288. https://doi.org/10.1021/acs.est.8b03438.

(46) Colson, B. C.; Michel, A. P. M. Flow-Through Quantification of Microplastics Using Impedance Spectroscopy. *ACS Sens.* **2021**, *6* (1), 238–244. https://doi.org/10.1021/acssensors.0c02223.

(47) Kaile, N.; Lindivat, M.; Elio, J.; Thuestad, G.; Crowley, Q. G.; Hoell, I. A. Preliminary Results From Detection of Microplastics in Liquid Samples Using Flow Cytometry. *Front. Mar. Sci.* **2020**, *7*, 552688. https://doi.org/10.3389/fmars.2020.552688.

(48) Bianco, V.; Memmolo, P.; Carcagnì, P.; Merola, F.; Paturzo, M.; Distante, C.; Ferraro, P. Microplastic Identification via Holographic Imaging and Machine Learning. *Adv. Intell. Syst.* **2020**, *2* (2), 1900153. https://doi.org/10.1002/aisy.201900153.

FOR TABLE OF CONTENTS ONLY