Integrating Metal Phenolic Networks-Mediated Separation and Machine Learning-Aided SERS for High-Precision Quantification and Classification of Nanoplastics

*Haoxin Ye[a], Shiyu Jiang[b], Yan Yan[c], Bin Zhao[d], Edward Grant[e], David Kitts[a], Rickey Y. Yada[a], Anubhav Pratap-Singh[a], Alberto Baldelli[a] , Tianxi Yang[a]**

a Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, Vancouver V6T1Z4, Canada
b Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
c Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA
d Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, V6T 1Z4, Canada
e Department of Chemistry, Faculty of Chemistry, The University of British Columbia, Vancouver V6T1Z4, Canada

*Corresponding author: Tianxi Yang (tianxi.yang@ubc.ca)

Keywords: nanoplastics; SERS; metal–phenolic networks; machine learning; accurate classification; high-precision quantification

ABSTRACT

The increasing accumulation of nanoplastics across ecosystems poses a significant threat to both terrestrial and aquatic life. Surface-enhance Raman scattering (SERS) is an emerging technique used for nanoplastic detection. However, the identification and classification of nanoplastics using SERS have challenges regarding sensitivity and accuracy, as nanoplastics are sparsely dispersed in the environment. Metal-phenolic networks (MPNs) have the potential to rapidly concentrate and separate various types and sizes of nanoplastics. SERS combined with machine learning may improve prediction accuracy. Herein, for the first time, we report the integration or MPNs-mediated separation with machine learning-aided SERS methods for the accurate classification and high-precision quantification of nanoplastics which is tailored to include the complete region of characteristic peaks across diverse nanoplastics in contrast to the traditional manual analysis of SERS spectra on a singular characteristic peak. Our customized machine learning system (e.g., outlier detection, classification, qualification) allows for the identification of detectable nanoplastics (accuracy 81.84%), accurate classification (accuracy > 97%) and the sensitive quantification of various types of nanoplastics (PS, PMMA, PE, PLA) down to ultra-low concentrations (0.1 ppm) as well as the accurate classification (accuracy > 92%) of nanoplastics mixtures to sub-ppm level. The effectiveness and novelty of this approach are substantiated by its ability to discern between different nanoplastics mixtures and detect nanoplastics samples in natural water systems.

1. Introduction

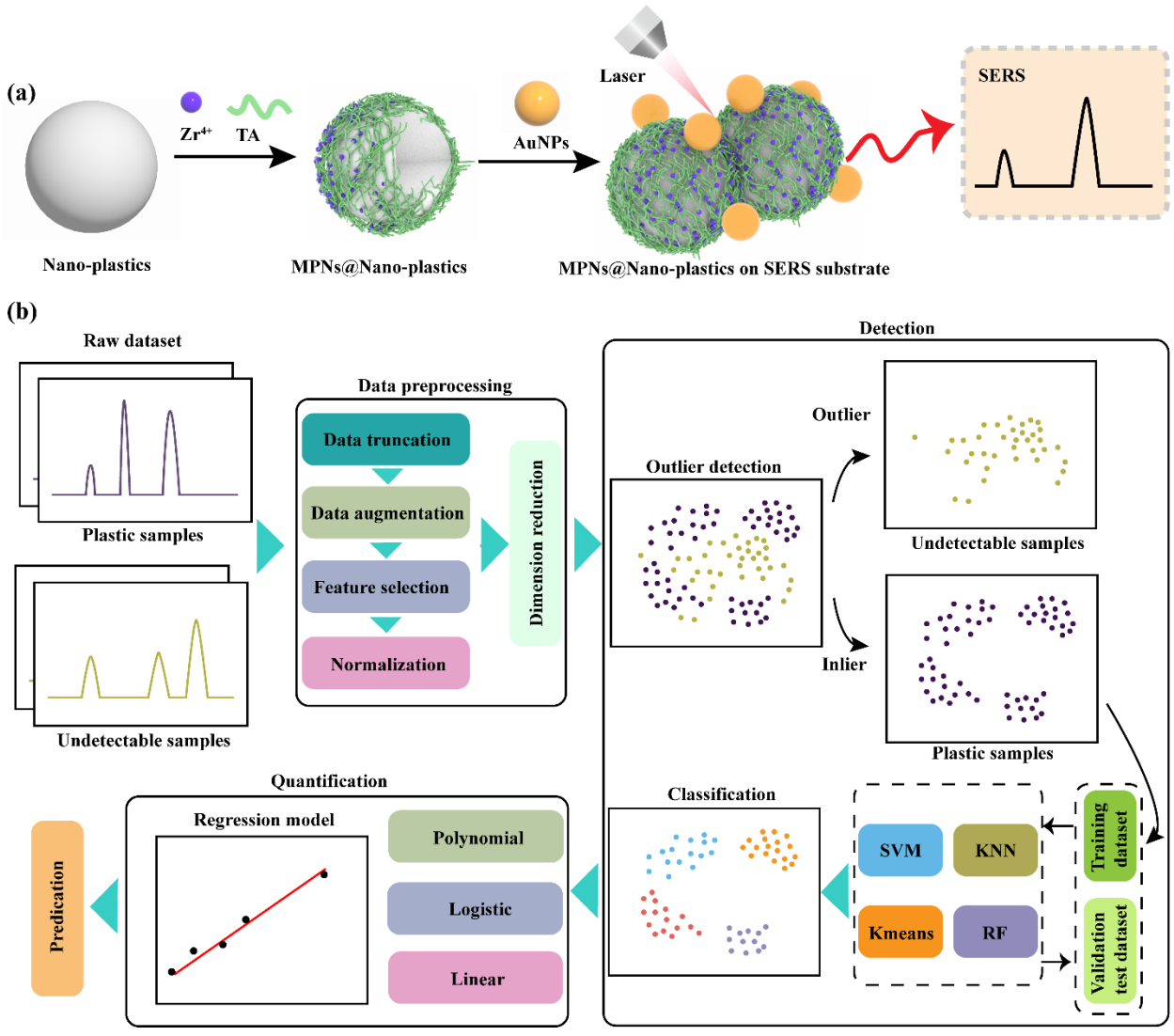
The increasing contamination of plastic in the environment is becoming a global concern, with annual emissions from plastic waste reaching approximate 53 million metric tons by 2030[1]. Large plastic debris degrades into microplastic particles (1 μm–5 mm) which then further fragments into smaller and potentially more hazardous nanoplastics (1–1000 nm)[2,3]. These nanoplastics, remain suspended in water, and contaminate various ecosystems including marine, soil and freshwater ecosystems, underscoring their pervasive nature[4–6]. Given their potential adverse effects on nature, organisms, and notably, humans, nanoplastics have emerged as a significant environmental and health concern. Numerous studies have identified the human health risks associated with nanoplastics, that include disruption of vascular endothelial cadherin junctions, induction of acute inflammation, and perturbation of gut microbiome composition and functionality[7–9].

Separation and enrichment processes are critical for nanoplastic assessment in aquatic samples. Currently methods for nanoplastic enrichment and separation operate in isolation and encounter various challenges. Enrichment techniques, including ultracentrifugation[10], evaporation[11], and vacuum drying[12], lack efficacy or the possibility of sample destruction. Similarly, widely used separation strategies such as asymmetrical flow field-flow fractionation/cross-flow filtration[13], density separation[14], and sieving[15] prove to be laborious and intricate. Metal phenolic networks (MPNs), a unique category of metal-organic materials synthesized from metal ions and phenolic molecules, exhibit a remarkable spectrum of surfaces due to the inherent properties of polyphenols. These networks can rapidly (~5 min) assemble into a diverse array of particles (e.g., organic, inorganic, and biological entities)[16–18] and demonstrate high stability to various aqueous environments (e.g., high-salt, highly acidic, and alkaline conditions)[19,20]. As a result, MPNs have garnered significant interest for applications such as drug delivery systems[21], tissue repair[17], biochemical sensing, and catalysis[22]. Furthermore, MPNs have shown potential in facilitating the collection of polymer particles through aggregation, a process influenced by the concentration ratio of metal ions to phenolic ligands[18,20,23] which can serve as a fast and straightforward method for the enrichment and separation of environmental contaminants, playing a crucial role in enhancing the sensitivity of these techniques. For instance, our previous research employed luminescent MPNs as coating layers for the labeling, separation and detection of various types and sizes of micro- and nanoplastics using a custom-designed portable fluorescence microscope. Although our approach enables sensitive detection of plastic particles, fluorescence imaging systems have limited capacity to identify different plastic types.

Recent innovations for the compositional examination of nanoplastics have employed methods such as inductively coupled plasma mass spectrometry (ICP-MS)[24], attenuated total reflectance Fourier-transform infrared spectroscopy (ATR-FTIR)[25], and infrared photothermal heterodyne imaging (IR-PHI)[26]. However, these methods are time-consuming and require extensive instrumentation at substantial cost of operation Surface-enhanced Raman spectroscopy (SERS), an advanced and ultra-sensitive vibrational spectroscopic technique combining Raman scattering and nanotechnology, is frequently utilized for various substances identification owing to its exceptional detection limits and expeditious characterization capabilities[27–29]. SERS measurements, combined with the use of a portable Raman instrument, allows for rapid on-site detection[30,31]. Manual analysis of a characteristic SERS peak of the targeted analyte is a common approach for SERS identification of the analyte. This approach, however, can potentially introduce errors in classification, especially for targets analysis in a complex matrix. Machine learning combined with the analysis of Raman or SERS data has been used effectively in a variety of application contexts, including, but not limited to, cancer detection, bacterial classification, and identification of contaminants[32–34]. Machine learning provides a robust means of accurately identifying target analytes in a range of complex systems (e.g., *in vivo* and natural environments) through the creation of customized strategies for spectral analysis. For instance, the integration of Raman spectroscopy with partial least squares (PLS) predictive regression models presents a viable strategy for monitoring lactate and pH values in body fluids[35]. Random forest models can be combined with Raman techniques to achieve accurate identification of nanoplastics, even in complex environmental systems such as natural rainwater (accuracy > 97%)[36]. However, this method only allowed the detection of nanoplastics at very high concentrations (approximately 5000 ppm) by using a traditional filtration method for separating nanoplastics. Currently, there is still a huge challenge for the accurate identification and quantification of nanoplastics simultaneously. As such, the integration of MPNs-mediated separation and customized machine learning-aided SERS addresses the latter.

In the present study, we utilized tannic acid (TA) and Zr^{4+} as model reagents to form MPNs (Scheme 1a) and optimized MPN-mediated enrichment and separation of diverse nanoplastic types, including Polystyrene (PS), Polymethyl Methacrylate (PMMA), Polyethylene (PE), Polylactic acid (PLA), to enhance separation efficiency and minimize interference other matrices, prior to SERS detection. Subsequently, a customized machine learning method was developed for the analysis of SERS spectra obtained from nanoplastics. By collecting the whole SERS peak region of nanoplastics, our methods enabled accurate classification and high-precision quantification of diverse nanoplastics (Scheme 1b). Utilizing MPNs-mediated separation process and the customized machine learning-based SERS approach, we achieved a classification accuracy of >97%

for nanoplastics and notably low detection limits in ultrapure water (0.1 ppm), tap water (1 ppm) and lake water (10 ppm) samples. Our study presents a novel strategy that not only allows for rapid and simple enrichment and separation of various types of nanoplastics, but also provides a means for the accurate classification and sensitive quantification of nanoplastics in a complex system.



Scheme 1. Schematic representation of integrating MPNs-mediated separation and machine learning-aided SERS for nanoplastics analysis. a) Self-assembly of MPNs coatings composed of TA and Zr^{4+} onto nanoplastics particles and SERS detection of nanoplastics using gold nanoparticle (AuNPs) substrates. **b)** Outline of the machine learning process: spectral datasets including plastic and undetectable samples are prepared, preprocessed, and analyzed. Data preprocessing includes truncation, augmentation, feature selection, normalization, and dimensionality reduction. The outlier detection is performed using the Isolation Forest algorithm,

and classification is executed on plastic samples (inliers) using four machine learning models (SVM, KNN, K-Means, and RF) to determine nanoplastic types. Quantitative analyses are conducted using polynomial, logistic, and linear regression models.

2. Results and discussion

2.1. SERS characterization of MPNs@nanoplastics

TA and Zr^{4+} were used as model reagents to form MPNs due to their high stability under different pH and biological environments[23]. 500 nm PS particles were selected to characterize MPNs-mediated enrichment and separation of nanoplastics. In contrast to PS nanoplastics without a MPN coating, the MPNs-assembled PS particles interconnected via MPNs (Figure 1a−b) with a clear coating observed by Scanning Electron Microscopy (SEM) imaging (Figure 1c). The MPNs-mediated aggregation and enrichment could facilitate precipitation of nanoplastics following regular centrifugation (Figure 1d)[37,38]. The Dynamic light scattering (DLS) analysis demonstrated a size-dependent aggregation characteristic among nanoplastics particles, with a pronounced increase in aggregation as the particle size decreased. (Figure S2). Nanoplastics with sizes of 866 nm and 478 nm predominantly aggregated into particles measuring between 4-5 μm. Conversely, nanoplastics with a smaller size of 42 nm exhibited a tendency to aggregate into substantially larger particles, exceeding 6 μm in size. The assembly of MPNs was demonstrated through a distinct shift (from 1194 cm^{-1} in TA to 1206 $cm^{-1)}$ in the vibrational peak of the HO-C bond from TA hydroxyl groups , as shown by Fourier-transform infrared spectroscopy (FTIR) analyses[39] (Figure S3).

We further used SERS to characterize MPNs-assembled nanoplastics (MPNs@NPs), utilizing PS nanoplastics as an example, to analyze SERS spectra of individual constituents (Figure S4). TA rendered robust SERS signals compared to weak signals from Zr^{4+}. Characteristic MPNs spectral regions were identified between 722–889, 1227–1256, and 1451–1651 cm^{-1}, which remained after the formation of MPNs-assembled PS particles, indicating that MPNs were assembled into PS particles. The dominant MPNs peak at 1487 cm^{-1} was attributed to TA[40]. PS nanoplastics exhibited a distinct peak at 998 cm^{-1} and this signature was retained after MPNs coating. It was evident that SERS is a viable tool for nanoplastics detection following MPNs-mediated enrichment and separation. It is imperative to optimize the molar concentrations of TA and Zr^{4+} as they affected the MPNs coating properties and separation performance[20,41]. We aimed to minimize MPNs peaks and accentuate the 998 cm^{-1} intrinsic signal from PS nanoplastics as the MPNs coating may overlap with other types of signals from nanoplastics[38]. Figure 1e demonstrates

the relationship between the TA/Zr^{4+} concentration ratio and the resultant SERS intensity at 998 cm^{-1} from PS. When fixing Zr^{4+} concentration, an initial SERS intensity increased, followed by a decline as TA/Zr^{4+} concentration ratio increased. The low concentration of TA may hinder MPNs formation, as the branched structure of TA determines the physicochemical property of MPNs[42], thereby attenuating separation efficiency for nanoplastics. Conversely, excessive TA could result in an overly dense coating layer[18], which might could interfere with SERS signals from nanoplastics and weaken SERS detection of nanoplastics since the ideal interparticle distance for highest SERS enhancement should be within the distance of < 10 nm[43]. Previous workers showed that precise modulation of the TA/Zr^{4+} concentration ratio will lead to a MPNs coating layer thinner than 10 nm[23] and this indicated that the MPNs coating could create the ideal distance between nanoplastics and SERS substrate (AuNPs) to enable the most effective SERS enhancement of nanoplastics. As expected, elevating TA concentrations increased the 1487 cm^{-1} peak from MPNs (as depicted in Figure 1f). To maintain a minimal level of interference while simultaneously ensuring high SERS intensities of PS nanoplastics, MPNs composed of 400 μM (TA) and 10 μM (Zr^{4+}) were selected as the optimal concentration for the formation of MPNs@NPs. We further demonstrated that by the use of optimized MPNs-mediated separation and enrichment SERS approach for the detection of PS particles (10 ppm), the characteristic SERS signals of PS at 1002 cm^{-1} was greatly enhanced compared to the direct SERS detection of PS particles without MPNs-mediated separation (Figure 1g).

This optimized MPNs separation protocol was subsequently used for other types of nanoplastics including 500 nm PS, 500 nm PMMA, 740–4990 nm PE, and 250 nm PLA. The characteristic peaks identified in the SERS spectra for each type of nanoplastic are shown in Figure 1h. PS exhibited a peak at 1002 cm^{-1}, attributed to the ring-breathing modes; PMMA displayed peaks at 811 cm^{-1} (C=O stretching) and 1452 cm^{-1} (C–H bending); PE showed peaks at 1060 cm^{-1} (symmetric C−C stretching), 1135 cm^{-1} (asymmetric C−C stretching), 1297 cm^{-1} (CH_2 twisting), and 1441 cm^{-1} (C−H bending); and PLA was characterized by peaks at 870 cm^{-1} (C−COO vibration) and 1441 cm^{-1} (C−H bending). These observed SERS peaks are in line with data reported by others[32,44,45]. SERS experiments for nanoplastics at various concentrations following MPNs-mediated separation were conducted and their respective highest characteristic peak is labeled as shown in Figure 1i–l. The limit of detection (LOD) of SERS assay was 0.1 ppm (PS), 1 ppm (PMMA), 5 ppm (PE), and 1 ppm (PLA), respectively (Figure S5a−d). The notably lower LOD for PS can be attributed to an intrinsic higher Raman activity, particularly when compared to other materials including PE, PMMA, and PLA. Typically, PS molecules enriched with electron-donating functional groups exhibit prominent Raman cross-sections, which are attributed to extended π−π systems, and enhance molecular polarizability[45]. We also compared the

sensitivity of the direct detection of nanoplastics using SERS with MPNs-mediated separation and SERS detection. The sensitivity of MPNs-mediated separation and SERS detection achieved a 500-fold improvement compared to direct SERS detection of nanoplastics, with LOD at 50 ppm (PS, PMMA, PE) and 100 ppm (PLA) respectively (Figure S6a−d).

The accurate identification of nanoplastics using SERS techniques, especially in a complex system, remains a huge challenge. Different types of nanoplastics may have similar or overlapping characteristic risks. The fingerprinting spectrum for specific nanoplastics are often affected by other substances in certain spectral regions, thereby diluting the unique "fingerprinting" attributes. Even more challenging is the complexity of identifying nanoplastic signals across various concentrations. For each type of plastic analyzed, the most prominent characteristic peak was selected. The distribution of SERS intensity of these peaks was systematically examined for four types of nanoplastics: polystyrene (PS), polymethylmethacrylate (PMMA), polyethylene (PE), and polylactic acid (PLA), each with concentrations exceeding the limit of detection (LOD) (Figure S7). Our findings reveal that, while PS and PE displayed characteristic peaks at 998 cm^{-1} and 1297 cm^{-1}, respectively, their spectra also exhibited intensities at 811 cm^{-1} and 870 cm^{-1}, which are indicative of PMMA and PLA, respectively. Analogously, the spectra of PMMA included signals at non-characteristic peaks (870 cm^{-1}, 998 cm^{-1}, and 1297 cm^{-1}), and the spectra of PLA was complicated by the presence of peaks at 811 cm^{-1} and 1297 cm^{-1}. These spectral interferences might be attributed to contributions from MPNs, AuNPs, or intrinsic weak peaks inherent to the plastics themselves (Figure S4). Due to these interferences and the variations in characteristic peaks, manual differentiation of nanoplastic types remains particularly challenging. Consequently, using a tailored machine learning approach coupled with SERS emerges as a promising avenue for the accurate classification of nanoplastics. Such strategies have demonstrated success in pinpointing target analytes within intricate systems analyzing by spectrometric techniques[46].

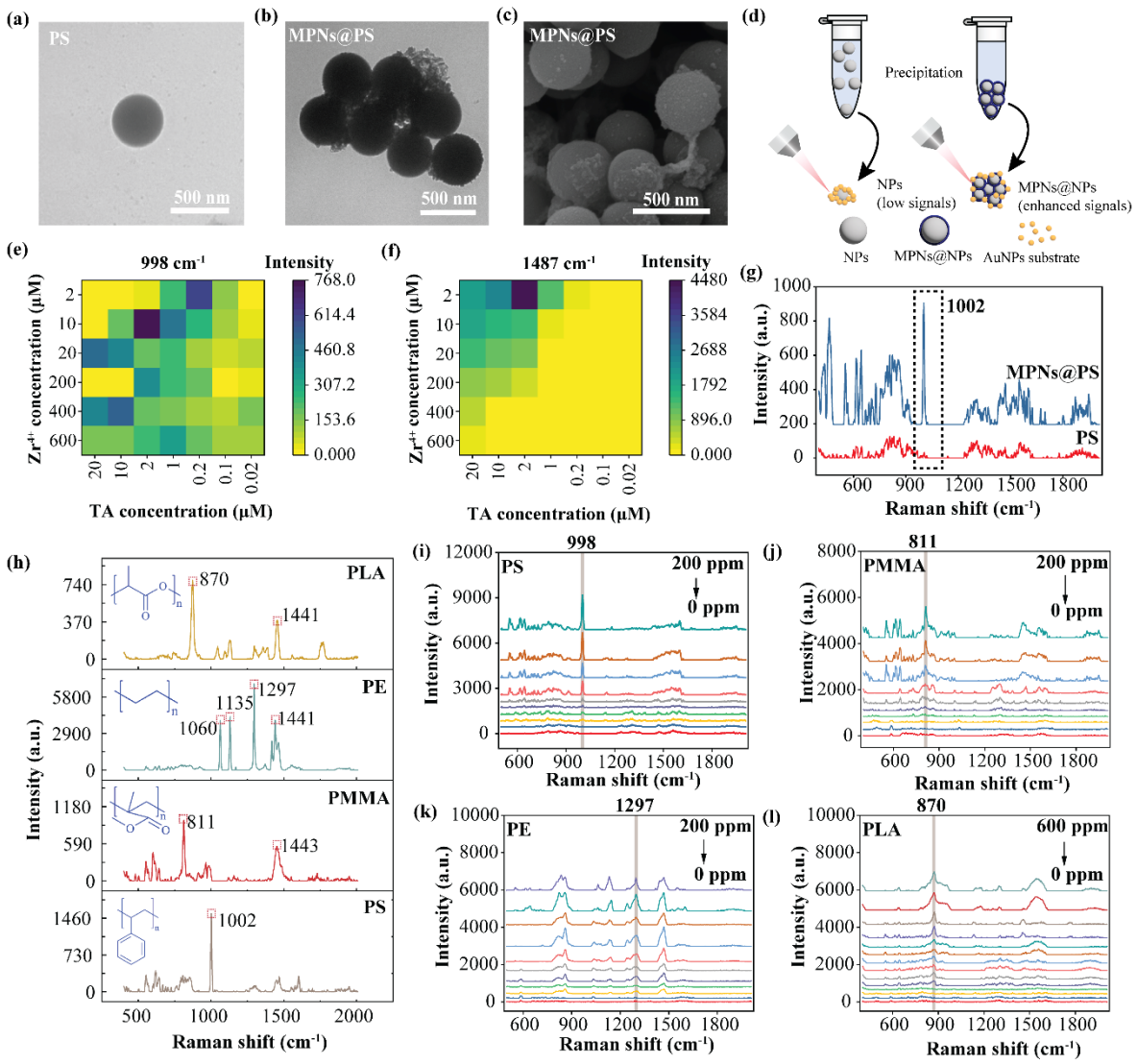


Figure 1. SERS characterization and detection of MPNs@NPs by. a-b Transmission Electron Microscopy (TEM) images of 500 nm PS nanoplastics before (a) and after (b) MPNs-mediated separation. **c** SEM image of 500 nm PS nanoplastics after MPNs-mediated separation. **d** Schemic illustration of MPNs-mediated method for SERS signal enhancement of nanoplastics. **e-f** Peak intensities derived from SERS spectra at wavenumber of 998 (e) and 1487 cm^{-1} (f) for MPNs@PS, presented over diverse TA and Zr^{4+} concentrations (2, 10, 20, 200, 400, 600 μM for TA, 0.02, 0.1, 0.2, 1, 2, 10, 20 μM for TA). **g** SERS spectra of 500 nm PS (10 ppm) with and without MPNs-mediated separation **h** SERS spectra distinguishing between four nanoplastics variants: 500 nm polystyrene (PS), 500 nm poly(methyl methacrylate) (PMMA), 250 nm polyvinyl chloride (PVC), and polyethylene (PE). **i-l** SERS spectral data for MPNs@NPs assorted by plastic type (PS, PMMA, PE, PLA) across a spectrum of concentrations. Concentrations include 0, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 200 ppm for PS, PMMA, and PE, and additionally 300, 400, 500, 600 ppm for

PLA. For panels a-f, the PS nanoplastics concentration is pegged at 10 ppm. In panels a-d, TA and Zr^{4+} concentrations are standardized at 10 and 400 μM, respectively. MPNs@PS: MPNs coated PS nanoplastics. MPNs@NPs: MPNs coated nanoplastics

2.2. Data pre-processing and outlier detection

SERS is known for its ability to detect a diverse range of substances at ultra-low concentrations[47,48]. Identifying unknown samples between detectable plastic (from our experimental settings) and undetectable plastic specimens is necessary for subsequent accurate classification. The types of plastics currently in use are numerous, and it was impossible to purchase all the template plastics due to limited commercial availability, therefore, four representative commercially available plastic types with concentrations exceeding LOD after MPNs-mediated separation were selected as plastic samples, resulting in total number of SERS spectra for each sample: PS, 24, PMMA, 18, PE, 15, PLA, 30. 40 undetectable plastic sample spectra were also created by introducing broad inverted parabolas, sporadically centered between 300 and 2080 cm^{-1}. These spectra were divided into two distinct groups: the first represented Raman-inactive substances or low concentration entities, and the second portrays Raman active substances or those with high concentrations (Supporting Note S1 and Figure S8).

Prior to introducing SERS spectra to machine learning classifiers, essential preprocessing was used to simulate typical environmental conditions. Key preprocessing measures encompassed data truncation, augmentation, feature selection, normalization, and dimensionality reduction. SERS spectra were truncated between 500–2080 cm^{-1}, a range optimized to highlight key vibrational features of plastic samples while excluding potential interference from AuNPs in the 300–500 cm^{-1} band that could possibly occur from the interparticle plasmonic coupling between AuNPs (Figure 2a–b). Data augmentation served dual purposes: expanding the training dataset and incorporating potential interferences (Figure 2c). Random noise was added to non-signal regions by adjusting existing peak intensities at a signal-noise ratio (SNR) randomly exceeding 15, where the signal represents the most pronounced peak intensity in the spectrum (Figure 2d). This procedure was repeated tenfold to include randomized SNR variations within the training dataset, resulting in 1397 spectra in total. To accurately identify peak positions for specific nanoplastics, we accommodated a peak shift tolerance of ± 6 cm^{-1} based on the characteristic peaks from various nanoplastics (Figure 2d). Characteristic peaks were selected from each plastic type instead of considering the entire spectrum for machine learning in order to reduce interference from MPNs and AuNPs (Figure 1h). As such, distinctive peaks were marked as features for machine learning: 998 cm^{-1} for PS, 811 cm^{-1} for PMMA, 1297 cm^{-1} for PE, and 870 cm^{-1} for PLA (Figure 1i–l). A detailed discussion regarding data augmentation and peak finding was provided in Supporting

Note S2 and Figure S9. The efficacy of feature selection was evaluated by cross-validation accuracy which underscored the need for representation of the selected features to the samples[49]. An increase in the number of features selected showed positive correlation with the enhanced accuracy, indicating the importance of selected four features in machine learning statistics (Figure 1e). Subsequent to normalization for these features, principal component analysis (PCA) was utilized for dimensionality reduction. PCA score plots highlighted plastic-specific cluster separations and undetectable sample distributions, showing the potential to differentiate different types of plastic samples as well as plastic samples with undetectable plastic samples, although there were some overlaps (Figure 1f).

To differentiate detectable and undetectable plastic samples, the isolation forest model was employed as an outlier detection method, which is particularly suitable for extensive, high-dimensional datasets[44]. Due to the capabilities of random forests, this technique discriminates between inliers (plastic samples) and outliers (undetectable plastic samples) without the need for a separate outlier model[44]. The isolation forest algorithm yields a measure of normality determined by the path lengths. This score is illustrated through the gradient of background colors on the path length decision boundary plots. Shorter paths are F-associated with reduced normality scores, pinpointing undetectable plastic samples (outliers), whereas longer paths signify detectable plastic samples (inliers) with elevated scores (Figure 1g). The score set to distinguish plastic samples from undetectable plastics stood at 0.53, clearly manifested in the binary decision boundary (Figure S10). Cumulatively, the model proficiently discerned plastics from undetectable samples, achieving an accuracy of 81.84%, a recall of 81.84%, a precision of 82.58%, and an F1 score of 82.09% (Figure S11). It is noteworthy that the model exhibited enhanced predictive capability for inliers (accuracy: 84.08%) as opposed to outliers (accuracy: 76.92%) (Figure 1h). This could possibly be attributed to intrinsic structural resemblances from certain artificial undetectable samples to experimental plastics. Challenges were encountered in the identification of low concentrations of non-degradable nanoplastics (PS, PMMA, PE) while the greater error rates were observed in medium concentrations of degradable nanoplastics (PLA) (Figure S12).

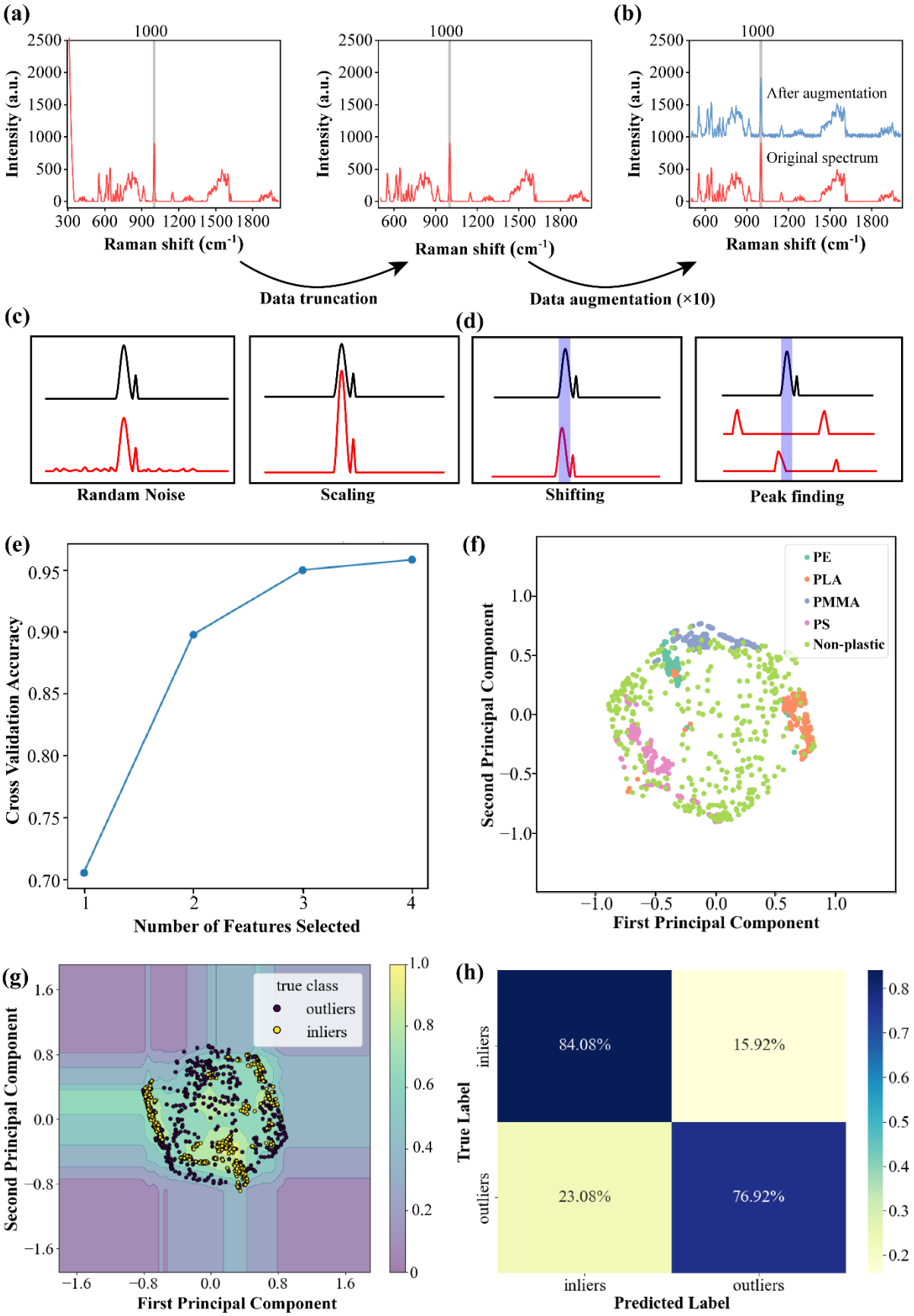


Figure 2. **SERS Spectra Preprocessing and Outlier Detection for plastic and undetectable plastic samples. a-c** Depictions of a representative SERS spectrum (a), and its subsequent transformations following data truncation (b) and augmentation (c). **d** Illustration of the data augmentation procedure, encompassing the introduction of random noise, scaling adjustments, and meticulous peak identification to address peak shifting. **e** The Feature Selection Profile (FSP) delineates cross-validation accuracy corresponding to the incremental addition of features. Notably, each feature's significance was methodically appraised leveraging a Support Vector Classifier (SVC) with a linear kernel, complemented by a 5-fold cross-validation strategy. **f** Principal Component Analysis (PCA) plot contrasting detectable plastic samples against undetectable ones. **g** Path-length decision boundary showing outlier detection results through the Isolation Forest algorithm. **h** Confusion matrix articulating the performance of the Isolation Forest model on discerning between outliers and inliers.

2.3. Classification of nanoplastic types

Accurate identification of nanoplastics at low concentrations is difficult especially in complex matrices such as natural environmental systems.[50] Machine learning techniques were utilized to categorize different nanoplastic types across various concentrations (PS 0.1–200 ppm, PMMA 1–200 ppm, PE 5–200 ppm, and PLA 1–600 ppm) following MPNs-mediated separation. The dataset was formed from features derived from four characteristic peaks of 957 SERS spectra of nanoplastics. Interrelationships between these features were assessed using Cramér's V coefficients, grounded in the chi-squared statistic[51]. Our findings revealed that most selected peaks were completely independent (Cramér's V = 0), except for peaks 811 and 870, which displayed strong correlation coefficients with peaks 998 and 1297, registering values of 0.39 and 0.22, respectively (Figure 3a). We employed t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction, which outperformed PCA in producing better clustered data[52] (Figure 3d and Figure S13). Kernel Density Estimate (KDE) plots further revealed that PS, PMMA, PE and PLA across different concentrations exhibit multiple peaks in each cluster (Figure S14).

After the pre-data processing, we further evaluated four prevalent models for multi-class classification: RF, K-Means, KNN, and SVM, as illustrated in Figure 3g. In the context of supervised learning classifiers, SVM, KNN and RF are common models for multi-classification. [53,54] Among them, RF models are trained without the necessity for dimensionality reduction, attributed to their inherent capability to proficiently manage high-dimensional data[55]. We also compared the supervised learning with unsupervised learning algorithm (K-Means) to optimize the machine learning model[56]. To ensure a rigorous and unbiased model selection process, we employed a 5-fold cross-validation strategy across the entire dataset. This cross-validation was

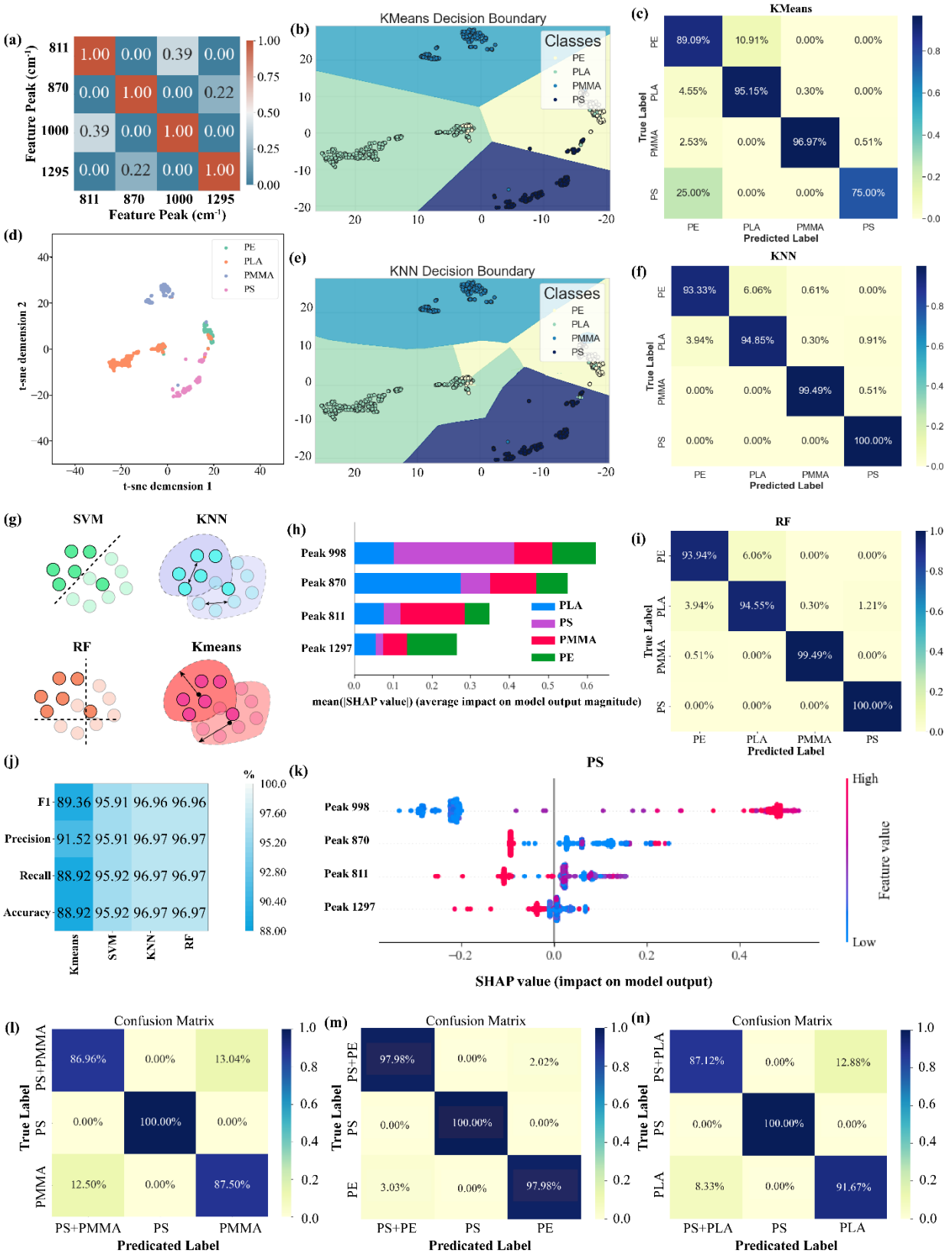
implemented to enable reasonable comparison among the models under evaluation and to optimize the use of the available, albeit limited, number of positive samples in the dataset.

Due to the imbalance in the number of datasets generated, we evaluated model performances using four metrics to find the optimal model: accuracy, precision, recall, and F1 score. Figure 3j presents our analysis that showed minimal difference among these values, indicating that the imbalanced number of datasets had a limited impact on model performance. The supervised RF (96.97%), KNN (96.97%), and SVM (95.92%) models notably outperformed the unsupervised K-Means model (88.92%) according to accuracy, thereby demonstrating enhanced reliability and repeatability for nanoplastic classification. The decision boundary plots showed that the unsupervised K-Means model encountered difficulty in differentiating between PS and PE, as evidenced by the relatively low accuracy of 75.00% and 89.09% observed in the confusion matrix (Figure 3b–c). For supervised models, confusion in differentiating PE from PLA was the primary factor leading to the low accuracy of SVM compared to RF and KNN (Figures 3f, 3i, and S15a). A comparison between SVM and KNN decision boundary plots clearly illustrated this phenomenon (Figures 3e and S15b). Therefore, we demonstrated that RF and KNN are the ideal models for multi-classification of nanoplastics samples.

To elucidate the contributions of selected features to predictive outcomes[57], we computed the average SHapley Additive exPlanations (SHAP) value magnitudes for features in the RF model. As the RF model utilized features without dimensionality reduction, it allowed for a transparent demonstration of the contribution of each feature to nanoplastic prediction (Figure 3h). Features were stratified in terms of importance, from the most critical (top) to the least (bottom). All four selected characteristic peaks notably influenced the classification of the four nanoplastic types. This suggested the intrinsic complexity in distinguishing nanoplastic types following MPNs-mediated separation solely through observing specific characteristic peak from one type. Consequently, machine learning approaches were found to be vital for the accurate classification of nanoplastics across varying concentrations. The dominant contributions were observed as follows: peak 998 cm^{-1} for PS, peak 870 cm^{-1} for PLA, peak 811 cm^{-1} for PMMA, and peak cm^{-1} 1297 for PE. These associations were congruent with our SERS experimental findings (Figure 1h). SHAP summary plots further showed the specific influence of each feature on nanoplastic classification[58]. For all nanoplastics, the most important feature (present in the first row of the SHAP summary) exhibited elevated values predominantly at locations of large SHAP values, signifying its positive role in prediction (Figure 3k and S16). Conversely, high values of non-characteristic peaks were predominantly associated with low SHAP values, indicative of their mitigating impact on the prediction. Certain high values corresponding to non-characteristic peaks,

such as peaks 870 cm^{-1} and 811 cm^{-1}, were found to enhance the likelihood of PS prediction (Figure 3k). This could be attributed to these non-characteristic peaks existing within the characteristic peak region of MPNs (Figure 1d). As PS plastic concentration increased, a concurrent increase in these characteristic peaks from MPNs and PS was observed—a pattern also observed across other plastic categories (Figure 1i).

We further explored the identification capability of the machine learning approach on nanoplastic samples with different species mixtures. We used the t-SNE dimensionality reduction and KNN model for further testing since it can offer the visualization of results and have the highest predication accuracy. PS-PMMA, PS-PE and PS-PLA mixtures across various concentrations with mass ratio of 1:1 were first detected by SERS following MPNs-mediated separation (Figure S17a–c). These plastic groups exceeding LOD were selected as mixture samples and their visualization after t-SNE dimensionality reduction were shown in Figure S17d–f. Multiple random clusters for both single nanoplastics and nanoplastic mixtures as well as some overlapping clusters indicated the difficulty in identifying nanoplastic mixtures compared to the single nanoplastics. The decision boundary determined by KNN were shown in Figure S17g–i, respectively. The overall accuracy achieved 92.87%, 98.11% and 94.63% for PS-PMMA, PS-PE, PS-PLA mixtures, respectively, proving the potential capability for discriminating nanoplastic mixtures through machine learning approaches (Figure 3l–n).



374

Figure 3. Classification of Nanoplastics through Machine Learning Models. a Cramér's V Correlation Coefficient heatmap detailing relationships amongst categorical features (Peaks at 811, 870, 998, and 1297 cm^{-1}). **b-c** Decision boundary plot from K-means clustering (b), accompanied by a corresponding confusion matrix (c) elucidating the accuracy of nanoplastic classification. **d** t-SNE plot after dimensionality reduction, capturing groupings within nanoplastics. **e-f** Decision boundaries established by KNN (e), alongside a confusion matrix (f) demarcating the accuracy of KNN model in nanoplastic classification. **h** Bar graph portraying the average magnitude of SHAP values. **i** Confusion matrix of the accuracy achieved in classifying nanoplastics utilizing the RF. **j** Model prediction assessment for an array of metrics, including accuracy, recall, precision, and the F1 score. **k** SHAP summary plot, spotlighting the consequentiality of individual features on the prediction of PS nanoplastics. Analyses pertinent to other nanoplastic categories are presented in Figure S16. **l-n** Confusion matrices of classification accuracy for various nanoplastic mixtures: PS-PMMA (l), PS-PE (m), and PS-PLA (n).

2.4. Quantification analysis

SERS facilitates quantification via the examination of characteristic peak intensities. The intensity of specific peaks in a SERS spectrum scales with the analyte concentration, thus enabling accurate and sensitive measurements[59]. To discern the quantitative relationship between nanoplastic concentrations and SERS intensity at characteristic peaks, we applied three prominent regression models: logistic, polynomial, and linear, targeting high-precision quantification. The concentration gradients of nanoplastics exceeding LOD were used, which were elucidated by SERS experiments in Section 2.1.

Taking 500 nm PS nanoplastics as a representative, the polynomial function (R^2 = 0.9745) outperformed the logistic (R^2 = 0.9674) and linear models (R^2 = 0.9128), as evident from Figure 4a–c. Notably, as the PS concentration rose, the curve flattened. This observation aligns with previous studies on quantifying PS nanoplastics using AuNPs as SERS substrates[12]. High concentrations of PS MPNs@NPs might cause the AuNPs to be shielded or buried, which reduces their effectiveness in enhancing the Raman signal given the optimal SERS enhancement occurs within a distance of < 10 nm proximity to the metal surface[60]. The polynomial regression model was shown to have the optimal relationship. To streamline data analysis and predictions, we linearized the fitting curves for all nanoplastic samples by adjusting the X-axis (concentrations). Tests on PS nanoplastics of varied sizes (50 nm, 500 nm, 1 μm) reinforced the method's versatility across nanoplastic dimensions (Figure 4d–f). It is important to note that the quantitative fitting curve is size-dependent, emphasizing the necessity of pre-determining size using techniques such as DLS. Subsequent quantitative assessments were conducted on PE, PLA, and PMMA. The

results, depicted in Figure S18, revealed the superiority of the polynomial model across nanoplastic types, consistent with the PS nanoplastic findings. We also converted the fitting curve to its linear form, as showcased in Figures 4g–i, underscoring the applicability of our MPNs-mediated separation in tandem with SERS for quantifying various types and sizes of nanoplastics.

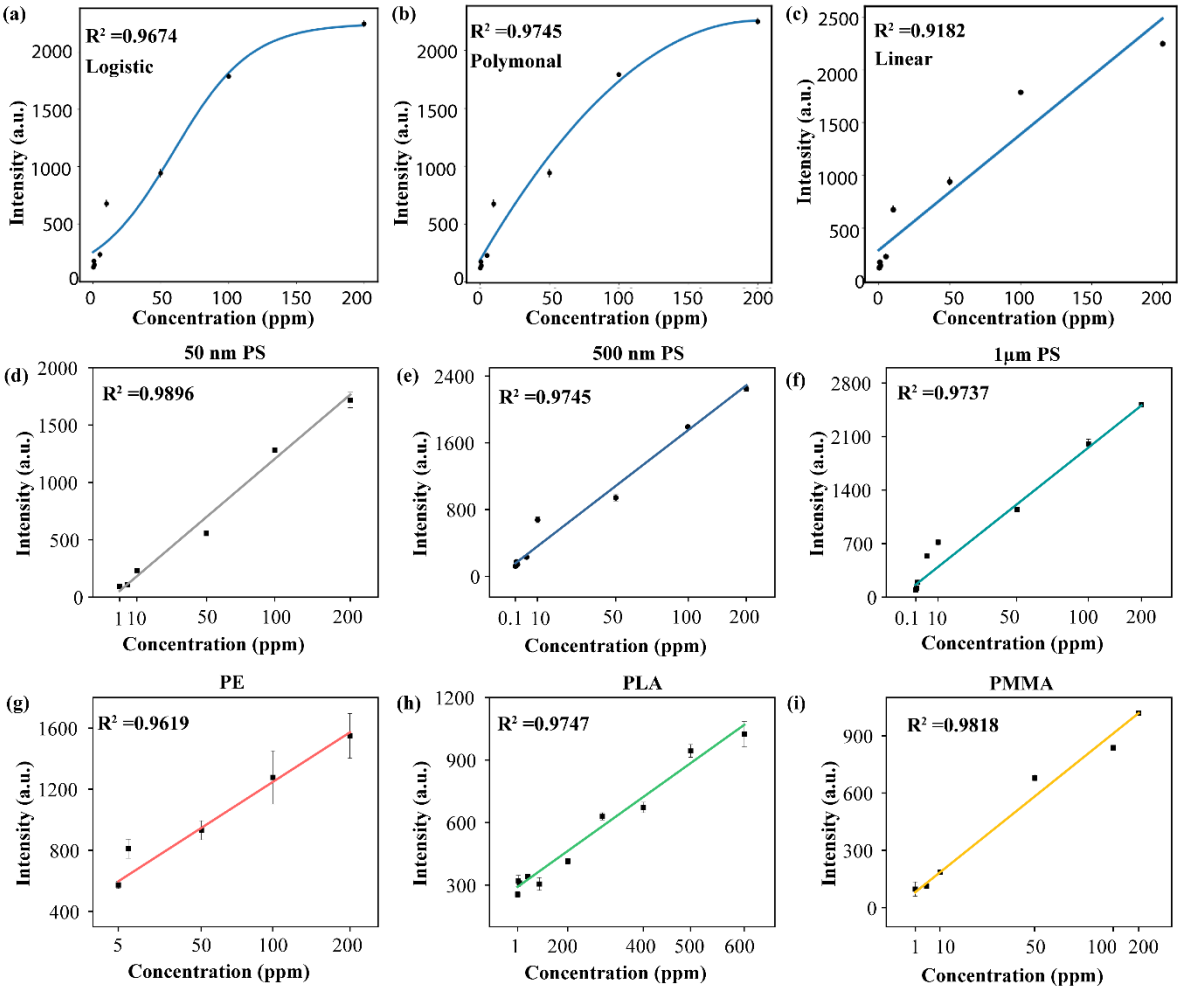


Figure 4. Quantitative Analysis of Nanoplastics using Regression Models. a-c Exploration of the relationship between 500 nm PS nanoplastic concentration and its characteristic peak intensity at 998 cm^{-1} via distinct regression models: logistic (a), polynomial (b), and linear (c). Complementary findings for other nanoplastic types are delineated in Supplementary Figure S18. **d-f** Linearized curves following X-axis adjustments via the polynomial model across different PS nanoplastic sizes: 50 nm (d), 500 nm (e), and 1 μm (f). $p < 0.000001$ by ANOVA analysis for all groups. **g-i** Linearized curves post X-axis adjustments for various nanoplastics: PS (g), PE (h), PLA (i), and PMMA (j).

2.5. Detection of nanoplastics in complex systems

The accurate detection of nanoplastics in real-world systems remains an imposing challenge. The difficulties for accurate detection arise from the interference of environmental impurities for nanoplastics, obscuring their isolation and subsequent analysis. As a result, a large proportion of existing research focuses on the detection of template nanoplastics in ultrapure water instead of quantifying them in real world systems[61–63]. To demonstrate the applicability of our method for nanoplastic detection in natural environmental water conditions, we introduced four types of nanoplastics (PS, PMMA, PE, PLA) into tap and lake water. These nanoplastics were then detected and quantified using SERS following MPNs-mediated separation, providing external validation for our established machine learning protocols.

In evaluating potential interference for nanoplastic classification and quantification, all tap and lake water samples were performed with MPNs-mediated separation and SERS detection. Figure 5a illustrates the SERS spectra control of different water after introducing MPNs-mediated separation. SERS spectra of tap water showed the similar SERS pattern compared to ultrapure water. This could be attributed to the fact that primary constituents in tap water—metal electrolytes such as Na^+, K^+, Ca^{2+}, and Mg^{2+},[64] cannot be separated through the MPNs protocol and are insensitive to Raman scattering due to a lack of pronounced molecular vibrational patterns. In contrast, the lake water spectrum revealed three broad peak regions: 600−900, 1200−1575, 1780−1960 cm^{-1}, which might result from a combination of signals from MPNs and intrinsic lake water substances. The composition of wetland water is more complex due to both its high electrolyte content and biological entities[65]. Biological entities may interact with MPNs[66], leading to a SERS spectrum reflecting interference from MPNs and lake water. After spiking 0−200 ppm of nanoplastics in both tap and lake water, we used MPNs-mediated separation and machine learning-aided SERS for the detection of various types of nanoplastics. Our approach also showed great potential for identifying multiple nanoplastics as validated by the observable peak characteristic from these nanoplastics (Figure 5b-c). However, the increased LOD for nanoplastics were observed in these real-world waters compared to ultrapure water: 1 ppm for PS, 10 ppm for PMMA, 10 ppm for PE, and 1 ppm for PLA in tap water, and 10 ppm for PS, PMMA, PE, and PLA in lake water (see Figure S19). One potential reason could be metal ions present in these water systems could compete with Zr^{4+} for adsorption sites on TA, due to the capacity of TA to form MPNs with a range of metal ions[18]. For further analysis, nanoplastic groups surpassing the LOD were chosen and classified using a trained KNN model. The classification accuracy for nanoplastics reached 90.63% in tap water and 95.02% in lake water, underscoring the robustness of machine learning model in analyzing potential nanoplastic samples across varied environmental contexts. However, as highlighted in the confusion matrix (Figure 5d–e), PLA classification posed challenges in both waters, with accuracies of 84.51% in tap water and 74.24% in lake water,

458 respectively. PLA, a known biodegradable material, may be biodegraded by microorganisms such
459 as fungi or bacteria[67] thus complicating the identification of PLA nanoplastics at low
460 concentrations, particularly in lake water.

461 Our subsequent analysis involved quantifying nanoplastics in tap and lake water systems at
462 varying concentrations. Utilizing a trained polynomial regression model, we determined the
463 nanoplastic concentrations and computed relative recovery ratios across concentrations exceeding
464 LOD (as shown in Figure 5f–i). In tap water, PS, PMMA, and PE exhibited robust recovery ratios
465 (80-120%). However, PLA nanoplastics at lower concentrations presented diminished ratios,
466 specifically 36.97±2.42% at 1 ppm and 35.38±0.23% at 5 ppm. Lake water displayed analogous
467 trends, with PLA recording 47.84±5.50% at 10 ppm and 40.15±5.07% at 50 ppm, likely
468 attributable to the effect of microbial activity. In lake water, both PS and PMMA nanoplastics
469 exhibited recovery ratios comparable to those observed in tap water. It is noteworthy that PE
470 nanoplastic groups in lake water yielded consistently low recovery ratios. This could be attributed
471 to weak CH_2 twisting vibration from PE affected by complex constituents in lake water
472 constituents, which diminished peak intensity at 1297 cm^{-1}. Such observations indicate potential
473 difficulties in accurately quantifying certain nanoplastics with inherently low Raman activity,
474 especially in intricate environmental systems. This necessitates the re-establishment of quantitative
475 analysis for PE in lake water instead of direct analysis using the quantitative method established
476 in ultrapure water. Similarly, the polynomial regression model performed best among other models
477 in analyzing PE in lake water, suggesting its potential applicability for quantifying nanoplastics
478 across diverse environmental systems (Figure S20) that further indicates the potential for
479 developing unique fitting curves for each nanoplastic to optimize detection across various
480 environments. The utilization of MPNs-mediated separation, combined with machine learning-
481 aided-SERS techniques, has proven successful for both classification and quantification of
482 nanoplastics in complex systems. Such a combination promises rapid on-site detection, accurate
483 classification, and high-precision quantification of nanoplastics in environmental systems.

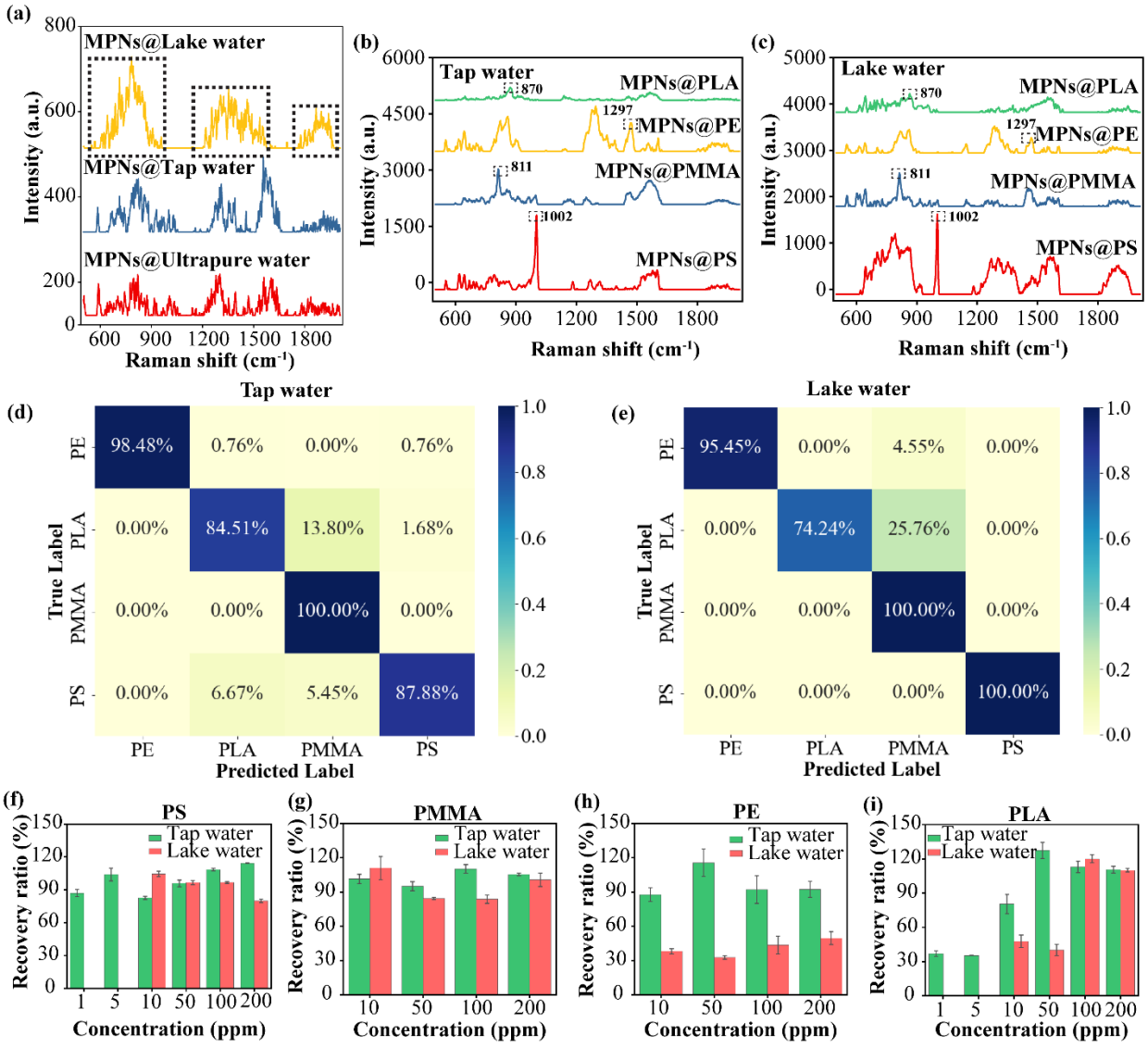


Figure 5. Classification and quantification of nanoplastics in environmental samples. a SERS spectra from ultrapure, tap, and lake water after MPNs mediated separation. **b-c** SERS spectra of various nanoplastics at 100 ppm spiked in tap (b) and lake (c) water after MPNs-mediated separation **d-e** Confusion matrices showcasing the classification accuracy of different nanoplastic types (PS, PMMA, PE, PLA) in tap water (b) and lake water (c) using the established KNN model. **f-j** Recovery rates of nanoplastics at concentrations exceeding detection limits in tap and lake water: PS (d), PMMA (e), PE (f), and PLA (g).

2.6. Comprehensive analysis of our method

To highlight the merits of our assay, we conducted a comparative assessment of this work for

nanoplastic analysis against recently established Raman/SERS techniques, considering factors such as SERS substrate, pretreatment, LOD for PS nanoplastics, and operation time (Table S1). Earlier investigations focused predominantly on enhancing the LOD via utilizing innovative SERS substrates. However, they confronted challenges: the protracted synthesis process for SERS substrates[68] and the exorbitant cost of their commercial counterparts[69]. Moreover, many of these studies employed drying-related methodologies as a pre-treatment for nanoplastic samples, which sought to refine the LOD[68,70]. This also prolonged the overall operation time to exceed 3 hours. Removing the pre-treatment procedure decreased the processing time to approximately one hour, however, it adversely affected the LOD, increasing it to ppm for PS nanoplastics[71]. In contrast, our methodology leveraged AuNPs as the SERS substrate. Employing minimal volumes of AuNPs solution and sample (each at 1 μL) promises substantial cost savings, especially when applied to large-scale applications. The streamlined MPNs-mediated separation techniques reduce the total operational duration to approximately 30 minutes but achieve low LOD of 0.1 ppm for 500 nm PS, which performs better than most current research. This efficiency, combined with the use of a portable Raman instrument and a mini centrifuge, enables effective on-site detection.

Furthermore, machine learning augments the classification and quantification accuracy of nanoplastics following MPNs separation and SERS detection. Traditional methodologies typically relied on manual analysis of Raman spectra, focusing primarily on a singular characteristic peak[45]. Such an approach potentially introduces judgement errors. Our machine learning approach, by considering the complete region of characteristic peaks across diverse nanoplastics, has been validated to accurately differentiate among PS, PMMA, PE, and PLA. We can achieve this differentiation even among nanoplastic mixtures and nanoplastics in environmental systems such as tap and lake water. This methodology is scalable, able to incorporate additional nanoplastics as they emerge in the market. More importantly, the universal labeling capacity of MPNs allows the technique to extend its detection capabilities beyond nanoplastics. It shows great potential to detect a myriad of entities, encompassing biological, inorganic, and organic particles, improving the current challenges of LOD and classification in multifaceted environmental or biological samples[23]. It is essential to recognize our method's potential for shaping sustainable practices, facilitating pollution mitigation, and ultimately fostering a more harmonious coexistence with our environment.

3. Conclusion

Nanoplastic contamination within environmental ecosystems has emerged as a significant concern. Conventional methodologies for nanoplastic detection have grappled with the challenge of classifying nanoplastics at suboptimal concentrations and extended processing times from complex pretreatment requisites. MPNs show their potential as a potent tool, offering rapid and proficient pretreatment avenues for nanoplastic enrichment and separation. Integrating MPN-mediated separation with SERS has enabled us to realize detection limits of 0.1 ppm for PS, 1 ppm for PMMA, 5 ppm for PE, and 1 ppm for PLA. Customized machine learning methodologies were utilized to facilitate accurate classification and quantification of nanoplastics following preprocessing of SERS spectra. The Isolation Forest model, with an accuracy of 81.84%, adeptly identifying outliers, distinguishing between detectable and undetectable plastic samples. For classification, the KNN model, with an accuracy of 96.97%, stood out, while polynomial regression was identified as the most efficacious model for quantification analysis. External validation for nano-plastics spiked in tap and lake water systems demonstrated robust performance. The application of machine learning models to nanoplastic identification post-MPN-mediated separation heralds an innovative approach, enhancing the precision and efficiency of environmental plastic pollution surveillance. While this investigation was limited by the number of commercially available nanoplastics with specific types of sizes, its potential applications are expansive, encompassing a myriad of nanoplastic variants and diverse particulates, including organic, inorganic, and biological entities, within complex systems.

4. Materials and Methods

4.1. Chemical and materials

Polystyrene (PS) particles of varying sizes (1 μm, 500 nm, and 50 nm) and polymethyl methacrylate particles (PMMA, 500 nm) were purchased from Phosphorex (Hopkinton, MA, USA). Polyethylene (PE) particles (740–4990 nm) were purchased from Cospheric (Santa Barbara, CA, USA), while Polylactide (PLA) particles (250 nm) were obtained from CD Bioparticles (Shirley, NY, USA). Tannic acid (ACS reagent ≥99%), zirconyl chloride octahydrate ($ZrOCl_2 \cdot 8H_2O$, 98%), and Tween 80 were acquired from VWR (Edmonton, Canada). Gold nanoparticles (AuNPs, 50 nm ± 4 nm) were sourced from nanoComposix (San Diego, CA, USA). Tap water samples from the Food Nutrition and Health building at The University of British Columbia (UBC) while lake water samples were collected from Nitobe Memorial Garden at UBC were collected.

4.2. Preparation of MPNs@NPs

PS, PMMA, and PLA solutions were directly used after dilution. However, PE solid particles were suspended in a Tween 80 solution as per the following: 100 mL of ultrapure water was boiled for 5 minutes, to which 0.1 g of Tween 80 was added, producing a 0.1% Tween 80 solution. Subsequently, 0.1 g of PE solid plastic particles were introduced and subjected to a 30 minute of ultrasonication pulse in an ice bath using a UP 200ST Hielscher ultrasonicator (Teltow, Brandenburg, Germany). The PE solution was prepared following centrifugation at 10,000 rpm for 10 minutes.

All plastic solutions (PS, PMMA, PE, PLA) were diluted to concentrations of: 0–200 ppm for PS, PMMA, and PE, with additional concentrations of 300–600 ppm for PLA. For MPNs@NPs preparation, 20 μL each of TA (0.5 mM) and $ZrOCl_2 \cdot 8H_2O$ (20 mM) were added into 960 μL of the aqueous nanoplastics suspension resulting in final concentrations of 10 μM for TA and 400 μM for Zr^{4+}, respectively. After vortexing for 60 seconds, the blend underwent centrifugation at 7500 rpm for 10 minutes using a mini centrifuge (VWR, Edmonton, Canada). The supernatant was discarded, and 1 μL of ultrapure water was introduced to the residue. Gentle agitation using a pipette tip yielded the MPNs@NPs suspension. An identical protocol was applied to detect nanoplastics spiked with tap and lake water to simulate nanoplastic detection in natural water systems.

4.3. Acquisition of SERS spectra

AuNPs served as the SERS substrate for spectral acquisition of samples. The original AuNPs solution was diluted 1:1 volume ratio. A 1 μL aliquot of the AuNPs solution was drop-cast onto the matte side of aluminum foil, followed by amalgamating an equivalent volume of each sample with the AuNPs solution. After air-drying at ambient conditions for 10 minutes, the edge of the coffee ring was tested by WP 785 ER Raman Spectrometer (Morrisville, NC, USA) in order to obtain a consistent signal. The systematic assay procedure was detailed in Supporting Note S3. SERS spectra were recorded by using a 785-nm diode laser and the ENLIGHTEN™ Spectroscopy Software. The incident laser power was 450 mW, the integration time was 60 s, and the spectral domain spanned 300 to 2008 cm^{-1}. Data processing incorporated boxcar smoothing (1 pixel) and Polynomial fitting (Polyfit) for baseline correction.

4.4. Characterization of MPNs@NPs.

DLS and zeta potential assessments were conducted using a Litesizer 500 (Anton Paar, Graz, Austria). For TEM imaging, unstained samples were examined under a Hitachi H7600 TEM

(Tokyo, Japan) at 80 kV. SEM imaging was performed on a Zeiss Crossbeam XB350 (Dublin, CA, USA). Prior to SEM observation, samples were sputter-coated with a 4 nm Pt layer and images were taken at an accelerating voltage of 5.0 kV. UV-vis absorption was measured by the Shimadzu UV-1800 UV–Vis Spectrophotometer (Kyoto, Japan) at a wavelength range from 300 to 900 nm.

4.5. SERS spectra preprocessing.

The SERS spectra, initially spanning a range of 300−2008 cm^{-1}, were truncated to 500−2008 cm^{-1}. Through an augmentation procedure, the dataset was amplified tenfold. From these spectra, the intensities of four diagnostic peaks at 811, 870, 998, and 1297 cm^{-1} were extracted, constituting a 4D vector representation. Each nanoplastic type served as the label, with the respective intensities of the characteristic peaks functioning as the training set.

A feature selection strategy, anchored on Mutual Information (MI), refined the dataset's feature set to ensure their maximal relevance[72]. The significance of each feature was systematically obtained using a Support Vector Classifier (SVC) with a linear kernel, in conjunction with 5-fold cross-validation. This evaluation yielded a key visualization tool, the Feature Selection Profile (FSP). This tool enabled the discernment of the most informationally dense features, which were incorporated to ensure machine learning endeavors. Peak intensities, which displayed variations across different nanoplastic spectra, were normalized to a standard scale spanning 0 to 1. This normalization rendered the intensities inter-comparable and bolstered the convergence efficiency during machine learning training sessions.

For dimensionality reduction, both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) methodologies were adopted to project the high-dimensional data into a 2D plane. Data was visualized in the PCA domain for five distinct nanoplastic categories, namely PE, PLA, PMMA, PS, and undetectable plastic samples. These illustrations were delineated based on the two principal components, highlighting the primary data variance axes. Before embarking on classification training, the t-SNE technique was employed for dimensionality compression of the plastic dataset (PE, PLA, PMMA, PS). Owing to its innate non-linearity, t-SNE adeptly retains local data structures, rendering it optimal for classification tasks[73]. Using parameters set at a perplexity of 70 and a learning rate of 10, we constructed two-dimensional visualization that captures the primary deviations in the dataset. Kernel Density Estimate (KDE) plots supplemented these scatter diagrams, offering insights into data distribution and density in this compacted dimension.

4.6. Machine learning approaches

Outlier Detection: We deployed the Isolation Forest algorithm for recognizing plastic datasets following PCA dimensionality reduction. Post-label transformation into binary classes, plastic samples were mapped as inliers and undetectable plastic as outliers. We trained the model using a set of hyperparameters, exploring the number of estimators in the range [10, 50, 100, 200, 300] and a contamination fraction of [0.35], which corresponds to the number of outliers in entire dataset. To hone in on optimal hyperparameters, a grid search combined with 5-fold cross-validation based on accuracy was performed. The best-fit model was trained on the full dataset and saved for future usage. Our analysis used two distinct decision boundary visualization plots: path length and binary boundary plots. Lastly, we derived a normalized decision threshold from decision function scores of the model to aid in distinguishing inliers from outliers.

Nanoplastic Classification: For categorizing various nanoplastics, we employed the KNN, SVM, RF, and KMeans algorithms. The RF algorithm catered to classification without dimensionality reduction, while t-SNE dimensionality reduction was adopted for the rest. The parameters of models were tuned by combining 5-fold cross-validation with grid search, ensuring varied dataset splits. For RF, parameters under scrutiny were number of trees [100], tree maximum depth [1, 10, 20], samples for node split [2, 5, 10], and samples at leaf node [1, 2, 4]. For the KNN classifier, we evaluated the number of neighbors spanning [10-50] and the weight function ['uniform', 'distance']. SVM evaluations focused on the regularization parameter, C, with values [0.1, 1, 10] and a linear kernel to maximize efficiency. After deducing the best parameters, the classifiers were trained on the complete dataset, storing the models for further analysis. For KMeans, data was organized into four clusters, with repeated initialization of 10 times to reinforce clustering reliability. To safeguard reproducibility and consistency, all models were subjected to a fixed random seed. Visualization of decision boundaries over the dimensionality reduced components for SVM, KNN, and KMeans highlighted distinct regions corresponding to each nanoplastic type, emphasizing the adeptness of models in differentiating PE, PLA, PMMA, and PS plastic samples.

Nanoplastic Quantification: To reveal the relationship between SERS characteristic peak intensity and nanoplastic concentrations, we deployed three regression models: eq. 1 (Logistic Model), eq. 2 (Polynomial Model), and eq. 3 (Linear Model). For each concentration, we derived both the mean and standard deviation of intensity. Model fit was appraised using the coefficient of determination (R^2). To streamline data analysis and prediction, the optimal regression model was linearized by adjusting the X-axis, representing concentrations.

$$f(x) = \frac{a}{1 + e^{-c(x-b)}} \quad (1)$$

$$f(x) = ax^2 + bx + c \quad (2)$$

$$f(x) = ax + b \quad (3)$$

4.7. Machine learning evaluation analysis

The relationships between distinct categorical features were elucidated employing Cramér's V, a measure derived from the chi-squared statistic[51]. An exhaustive computation of Cramér's V was executed for every pair of categorical variables, culminating in constructing a symmetric matrix. This matrix illuminated the interrelationships among all dataset features. A confusion matrix was rendered to offer a graphical depiction of the performance of machine learning models in differentiating the nanoplastic groups. Evaluation of the performance of predictions incorporated a suite of metrics, encompassing accuracy, recall, precision, and the F1 score, delineated in eq. 4–7.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (4)$$

$$recall = \frac{tp}{tp + fn} \quad (5)$$

$$precision = \frac{tp}{tp + fp} \quad (6)$$

$$F1 = \frac{2tp}{2tp + fp + fn} \quad (7)$$

where tp, fp, tn and fn are the number of true positive, false positive, true negative, and false negative predictions, respectively.

To shed light on the significance and impact of features on the RF model predictions, SHapley Additive exPlanations (SHAP) values were determined. A comprehensive SHAP summary plot elucidated the average magnitude and direction of feature contributions specific to each nanoplastic. Given our multi-class classification paradigm, distinct SHAP summary plots were curated for every class, offering nuanced insights into the feature impacts pertinent to each nanoplastic type. For the external validation, all nanoplastics spiked into both tap and lake water samples were classified and quantified utilizing the established machine learning models to determine classification accuracy and the recovery rate. The external validation was detailed in Supporting Note S4.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website. Characterization, raw spectra, outlier detection and classification models evaluation, PCA and t-SNE dimensionality reduction visualization, detection limit determination, creation of undetectable samples, data augmentation, assay procedure, external validation and comprehensive evaluation of this method (PDF).

Source code for can be found at https://github.com/Haoxin01/ML_Nanoplastics_MPNs_SERS

AUTHOR INFORMATION

Corresponding Author

Tianxi Yang - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada.

Authors

Haoxin Ye - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada.

Shiyu Jiang - Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

Dr. Yan Yan - Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA

Dr. Bin Zhao - Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

Dr. Edward Grant - Department of Chemistry, Faculty of Chemistry, University of British Columbia, Vancouver V6T1Z4, Canada

Dr. David Kitts - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada

Dr. Rickey Y. Yada - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada

Dr. Anubhav Pratap-Singh - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada

Dr. Alberto Baldelli - Food, Nutrition and Health, Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver BC, V6T 1Z4 Canada

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the UBC Faculty of Land and Food Systems/Start Up Funds (AWD-020249 UBCLANDF 2022), Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants Program (RGPIN-2023-04100, RYY RGPIN 04598) and NSERC Discovery Grants Program-Discovery Launch Supplement (DGECR-2023-00386). We acknowledge the Canada Foundation for Innovation and John R. Evans Leaders Fund (CFI-JELF #44768), and the British Columbia Knowledge Development Fund (BCKDF). We also thank D. Horne from the UBC Bioimaging Facility (RRID: SCR_021304) for completing SEM and TEM.

References

1. Borrelle, S. B. *et al.* Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. *Science* **369**, 1515–1518 (2020).
2. Gigault, J. *et al.* Nanoplastics are neither microplastics nor engineered nanoparticles. *Nat. Nanotechnol.* **16**, 501–507 (2021).
3. Urso, M., Ussia, M., Novotný, F. & Pumera, M. Trapping and detecting nanoplastics by MXene-derived oxide microrobots. *Nat Commun* **13**, 3573 (2022).
4. Bergmann, M. *et al.* Plastic pollution in the Arctic. *Nat Rev Earth Environ* **3**, 323–337 (2022).
5. Mitrano, D. M. *et al.* Synthesis of metal-doped nanoplastics and their utility to investigate fate and behaviour in complex environmental systems. *Nat. Nanotechnol.* **14**, 362–368 (2019).
6. Sun, X.-D. *et al.* Differentially charged nanoplastics demonstrate distinct accumulation in Arabidopsis thaliana. *Nat. Nanotechnol.* **15**, 755–760 (2020).
7. Wei, W. *et al.* Anionic nanoplastic exposure induces endothelial leakiness. *Nat Commun* **13**, 4757 (2022).
8. Teles, M., Balasch, J. C., Oliveira, M., Sardans, J. & Peñuelas, J. Insights into nanoplastics effects on human health. *Science Bulletin* **65**, 1966–1969 (2020).
9. Wang, M. *et al.* Oligomer nanoparticle release from polylactic acid plastics catalysed by gut enzymes triggers acute inflammation. *Nat. Nanotechnol.* **18**, 403–411 (2023).
10. Cai, H., Chen, M., Du, F., Matthews, S. & Shi, H. Separation and enrichment of nanoplastics in environmental water samples via ultracentrifugation. *Water Research* **203**, 117509 (2021).
11. Vega-Herrera, A. *et al.* Polymers of micro(nano) plastic in household tap water of the Barcelona Metropolitan Area. *Water Research* **220**, 118645 (2022).
12. Chaisrikhwun, B., Ekgasit, S. & Pienpinijtham, P. Size-independent quantification of nanoplastics in various aqueous media using surfaced-enhanced Raman scattering. *Journal of Hazardous Materials* **442**, 130046 (2023).
13. Mintenig, S. M., Bäuerlein, P. S., Koelmans, A. A., Dekker, S. C. & Wezel, A. P. van. Closing the gap between small and smaller: towards a framework to analyse nano- and microplastics in aqueous environmental samples. *Environ. Sci.: Nano* **5**, 1640–1649 (2018).
14. Grbic, J. *et al.* Magnetic Extraction of Microplastics from Environmental Samples. *Environ. Sci. Technol. Lett.* **6**, 68–72 (2019).
15. Bergmann, M. *et al.* High Quantities of Microplastic in Arctic Deep-Sea Sediments from the HAUSGARTEN Observatory. *Environ. Sci. Technol.* **51**, 11000–11010 (2017).
16. Ejima, H., Richardson, J. J. & Caruso, F. Metal-phenolic networks as a versatile platform to engineer nanomaterials and biointerfaces. *Nano Today* **12**, 136–148 (2017).
17. Li, Y. *et al.* Recent Advances in the Development and Antimicrobial Applications of Metal–

761 Phenolic Networks. *Advanced Science* **9**, 2202684 (2022).
762 18. Guo, J. *et al.* Engineering Multifunctional Capsules through the Assembly of Metal–Phenolic
763 Networks. *Angewandte Chemie International Edition* **53**, 5546–5551 (2014).
764 19. Zhang, X. *et al.* Thin film composite nanofiltration membrane with nanocluster structure
765 mediated by graphene oxide/metal-polyphenol nanonetwork scaffold interlayer. *Journal of*
766 *Membrane Science* **669**, 121330 (2023).
767 20. Chen, J. *et al.* Metal–Phenolic Networks as Tunable Buffering Systems. *Chem. Mater.* **33**,
768 2557–2566 (2021).
769 21. Ping, Y. *et al.* pH-Responsive Capsules Engineered from Metal–Phenolic Networks for
770 Anticancer Drug Delivery. *Small* **11**, 2032–2036 (2015).
771 22. Yun, G. *et al.* Synthesis of Metal Nanoparticles in Metal-Phenolic Networks: Catalytic and
772 Antimicrobial Applications of Coated Textiles. *Advanced Healthcare Materials* **7**, 1700934
773 (2018).
774 23. Lin, Z. *et al.* Luminescent Metal-Phenolic Networks for Multicolor Particle Labeling.
775 *Angewandte Chemie International Edition* **60**, 24968–24975 (2021).
776 24. Jiménez-Lamana, J. *et al.* A Novel Strategy for the Detection and Quantification of
777 Nanoplastics by Single Particle Inductively Coupled Plasma Mass Spectrometry (ICP-MS).
778 *Anal. Chem.* **92**, 11664–11672 (2020).
779 25. Hernandez, L. M. *et al.* Plastic Teabags Release Billions of Microparticles and Nanoparticles
780 into Tea. *Environ. Sci. Technol.* **53**, 12300–12310 (2019).
781 26. Kniazev, K. *et al.* Using Infrared Photothermal Heterodyne Imaging to Characterize Micro-
782 and Nanoplastics in Complex Environmental Matrices. *Environ. Sci. Technol.* **55**, 15891–
783 15899 (2021).
784 27. Haldavnekar, R., Venkatakrishnan, K. & Tan, B. Non plasmonic semiconductor quantum
785 SERS probe as a pathway for in vitro cancer detection. *Nat Commun* **9**, 3065 (2018).
786 28. Zhang, Y. *et al.* General approach to surface-accessible plasmonic Pickering emulsions for
787 SERS sensing and interfacial catalysis. *Nat Commun* **14**, 1392 (2023).
788 29. Feng, E., Zheng, T., He, X., Chen, J. & Tian, Y. A novel ternary heterostructure with dramatic
789 SERS activity for evaluation of PD-L1 expression at the single-cell level. *Science Advances* **4**,
790 eaau3494 (2018).
791 30. He, X., Fan, C., Luo, Y., Xu, T. & Zhang, X. Flexible microfluidic nanoplasmonic sensors for
792 refreshable and portable recognition of sweat biochemical fingerprint. *npj Flex Electron* **6**, 1–
793 10 (2022).
794 31. Zhu, W. *et al.* Rapid and low-cost quantitative detection of creatinine in human urine with a
795 portable Raman spectrometer. *Biosensors and Bioelectronics* **154**, 112067 (2020).
796 32. Lei, B. *et al.* Customizable Machine-Learning Models for Rapid Microplastic Identification

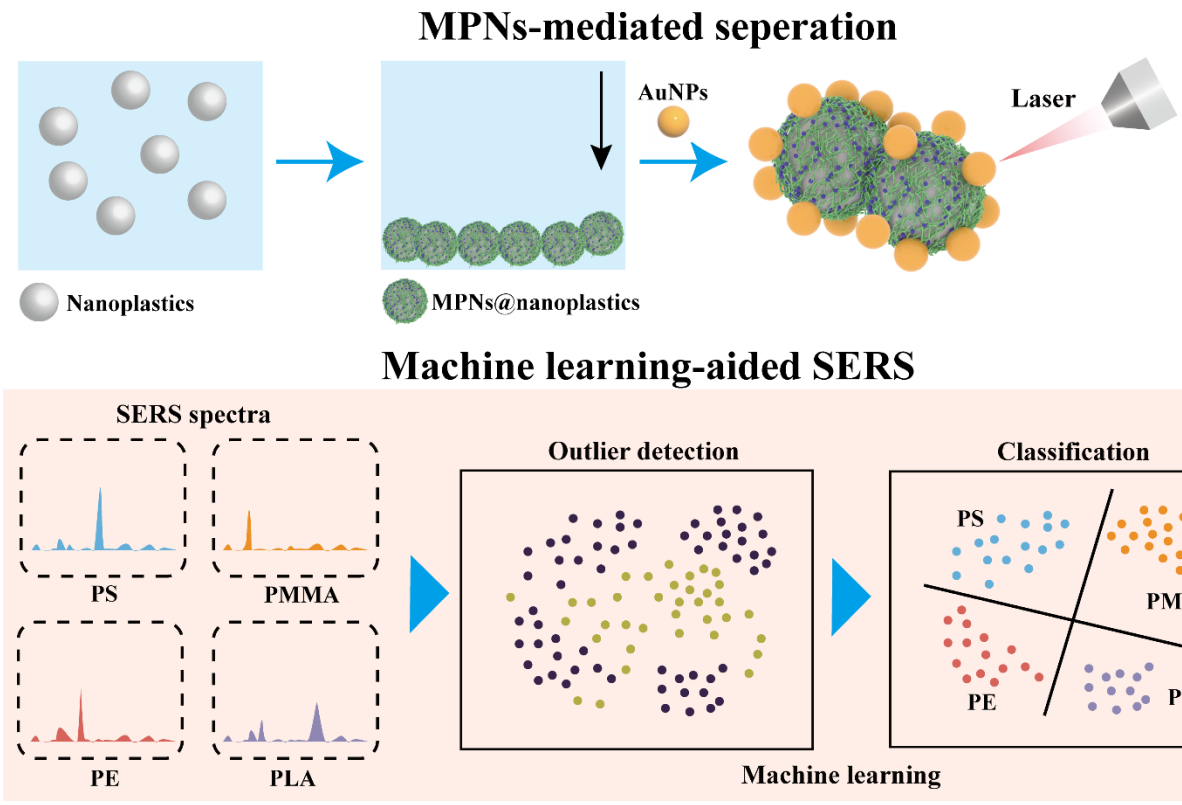
Using Raman Microscopy. *Anal. Chem.* **94**, 17011–17019 (2022).
33. Ho, C.-S. *et al.* Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat Commun* **10**, 4927 (2019).
34. Ralbovsky, N. M. & Lednev, I. K. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chem. Soc. Rev.* **49**, 7428–7453 (2020).
35. Olaetxea, I. *et al.* Machine Learning-Assisted Raman Spectroscopy for pH and Lactate Sensing in Body Fluids. *Anal. Chem.* **92**, 13888–13895 (2020).
36. Xie, L. *et al.* Automatic Identification of Individual Nanoplastics by Raman Spectroscopy Based on Machine Learning. *Environ. Sci. Technol.* **57**, 18203–18214 (2023).
37. Ye, H. *et al.* Rapid Detection of Micro/Nanoplastics Via Integration of Luminescent Metal Phenolic Networks Labeling and Quantitative Fluorescence Imaging in A Portable Device. Preprint at https://doi.org/10.26434/chemrxiv-2023-jnbm1 (2023).
38. Ye, H. *et al.* Quantitative and rapid detection of nanoplastics labeled by luminescent metal phenolic networks using surface-enhanced Raman scattering. *Journal of Hazardous Materials* **470**, 134194 (2024).
39. Liang, H. *et al.* Engineering Multifunctional Films Based on Metal-Phenolic Networks for Rational pH-Responsive Delivery and Cell Imaging. *ACS Biomater. Sci. Eng.* **2**, 317–325 (2016).
40. Chen, C. *et al.* Dynamic Metal–Phenolic Coordination Complexes for Versatile Surface Nanopatterning. *J. Am. Chem. Soc.* **145**, 7974–7982 (2023).
41. Mazaheri, O. *et al.* Assembly of Metal–Phenolic Networks on Water-Soluble Substrates in Nonaqueous Media. *Advanced Functional Materials* **32**, 2111942 (2022).
42. Rahim, Md. A. *et al.* Rust-Mediated Continuous Assembly of Metal–Phenolic Networks. *Advanced Materials* **29**, 1606717 (2017).
43. Gullace, S. *et al.* Universal Fabrication of Highly Efficient Plasmonic Thin-Films for Label-Free SERS Detection. *Small* **17**, 2100755 (2021).
44. Qin, C.-C. *et al.* Melt electrospinning of poly(lactic acid) and polycaprolactone microfibers by using a hand-operated Wimshurst generator. *Nanoscale* **7**, 16611–16615 (2015).
45. Xie, L., Gong, K., Liu, Y. & Zhang, L. Strategies and Challenges of Identifying Nanoplastics in Environment by Surface-Enhanced Raman Spectroscopy. *Environ. Sci. Technol.* **57**, 25–43 (2023).
46. Boiko, D. A., Kozlov, K. S., Burykina, J. V., Ilyushenkova, V. V. & Ananikov, V. P. Fully Automated Unconstrained Analysis of High-Resolution Mass Spectrometry Data with Machine Learning. *J. Am. Chem. Soc.* **144**, 14590–14606 (2022).
47. Keiner, R. *et al.* Raman Spectroscopy—An Innovative and Versatile Tool To Follow the

Respirational Activity and Carbonate Biomineralization of Important Cave Bacteria. *Anal. Chem.* **85**, 8708–8714 (2013).
48. Reference database of Raman spectra of biological molecules - De Gelder - 2007 - Journal of Raman Spectroscopy - Wiley Online Library. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jrs.1734.
49. Bui, D. T., Tsangaratos, P., Ngo, P.-T. T., Pham, T. D. & Pham, B. T. Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. *Science of The Total Environment* **668**, 1038–1054 (2019).
50. Cai, H. *et al.* Analysis of environmental nanoplastics: Progress and challenges. *Chemical Engineering Journal* **410**, 128208 (2021).
51. Memon, Q. U. A. *et al.* Health problems from pesticide exposure and personal protective measures among women cotton workers in southern Pakistan. *Science of The Total Environment* **685**, 659–666 (2019).
52. Yu, T., Fu, Y., He, J., Zhang, J. & Xianyu, Y. Identification of Antibiotic Resistance in ESKAPE Pathogens through Plasmonic Nanosensors and Machine Learning. *ACS Nano* **17**, 4551–4563 (2023).
53. The Elements of Statistical Learning: Data Mining, Inference, and Prediction | SpringerLink. https://link.springer.com/book/10.1007/978-0-387-21606-5.
54. Smer-Barreto, V. *et al.* Discovery of senolytics using machine learning. *Nat Commun* **14**, 3445 (2023).
55. Winham, S. J. *et al.* SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics* **13**, 164 (2012).
56. Pham, D. T., Dimov, S. S. & Nguyen, C. D. Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* **219**, 103–119 (2005).
57. Gómez-Ramírez, J., Ávila-Villanueva, M. & Fernández-Blázquez, M. Á. Selecting the most important self-assessed features for predicting conversion to Mild Cognitive Impairment with Random Forest and Permutation-based methods. 785519 Preprint at https://doi.org/10.1101/785519 (2019).
58. Ciarella, S. *et al.* Finding defects in glasses through machine learning. *Nat Commun* **14**, 4229 (2023).
59. Lee, W. *et al.* Spread spectrum SERS allows label-free detection of attomolar neurotransmitters. *Nat Commun* **12**, 159 (2021).
60. Isabel Pérez-Jiménez, A., Lyu, D., Lu, Z., Liu, G. & Ren, B. Surface-enhanced Raman spectroscopy: benefits, trade-offs and future developments. *Chemical Science* **11**, 4563–4577 (2020).

61. Cai, H. *et al.* Analysis of environmental nanoplastics: Progress and challenges. *Chemical Engineering Journal* **410**, 128208 (2021).
62. Yu, E.-S. *et al.* Real-Time Underwater Nanoplastic Detection beyond the Diffusion Limit and Low Raman Scattering Cross-Section via Electro-Photonic Tweezers. *ACS Nano* (2022) doi:10.1021/acsnano.2c07933.
63. Leppänen, I. *et al.* Capturing colloidal nano- and microplastics with plant-based nanocellulose networks. *Nat Commun* **13**, 1814 (2022).
64. Azoulay, A., Garzon, P. & Eisenberg, M. J. Comparison of the Mineral Content of Tap Water and Bottled Waters. *Journal of General Internal Medicine* **16**, 168–175 (2001).
65. Gong, M., Yang, G., Zhuang, L. & Zeng, E. Y. Microbial biofilm formation and community structure on low-density polyethylene microparticles in lake water microcosms. *Environmental Pollution* **252**, 94–102 (2019).
66. Fan, G., Wasuwanich, P., Rodriguez-Otero, M. R. & Furst, A. L. Protection of Anaerobic Microbes from Processing Stressors Using Metal–Phenolic Networks. *J. Am. Chem. Soc.* **144**, 2438–2443 (2022).
67. Fukushima, K., Abbate, C., Tabuani, D., Gennari, M. & Camino, G. Biodegradation of poly(lactic acid) and its nanocomposites. *Polymer Degradation and Stability* **94**, 1646–1655 (2009).
68. Chang, L. *et al.* Nanowell-enhanced Raman spectroscopy enables the visualization and quantification of nanoplastics in the environment. *Environ. Sci.: Nano* **9**, 542–553 (2022).
69. Xu, G. *et al.* Surface-Enhanced Raman Spectroscopy Facilitates the Detection of Microplastics <1 μm in the Environment. *Environ. Sci. Technol.* **54**, 15594–15603 (2020).
70. Jeon, Y. *et al.* Detection of nanoplastics based on surface-enhanced Raman scattering with silver nanowire arrays on regenerated cellulose films. *Carbohydrate Polymers* **272**, 118470 (2021).
71. Hu, R., Zhang, K., Wang, W., Wei, L. & Lai, Y. Quantitative and sensitive analysis of polystyrene nanoplastics down to 50 nm by surface-enhanced Raman spectroscopy in water. *Journal of Hazardous Materials* **429**, 128388 (2022).
72. Vergara, J. R. & Estévez, P. A. A review of feature selection methods based on mutual information. *Neural Comput & Applic* **24**, 175–186 (2014).
73. Dharavath, R., MadhukarRao, G., Khurana, H. & Edla, D. R. t-SNE Manifold Learning Based Visualization: A Human Activity Recognition Approach. in *Advances in Data Science and Management* (eds. Borah, S., Emilia Balas, V. & Polkowski, Z.) 33–43 (Springer, Singapore, 2020). doi:10.1007/978-981-15-0978-0_3.

905 **Table of Contents**

906



907

908

909