Electronic Supplementary Material (ESI) for Analytical Methods.

Supplementary material: Random Forest single channel classifier for FT-IR Microplastic hyperspectral images

In this work, a Random Decision Forest model is built for fast identification of Fourier-transform infrared spectra of the eleven most common types of microplastics in the environment. The Random Decision Forest input data is reduced to a combination of highly discriminative single wavenumbers selected using a machine learning classifier. This dimension reduction allows input from systems with individual wavenumber measurements, and decreases prediction time. The training and testing spectra are extracted from Fourier-transform infrared hyperspectral images of pure-type microplastic samples, automatizing the process with reference spectra and a fast background correction and identification algorithm. Random Decision Forest classification results are validated using procedurally generated ground truth. The classification accuracy achieved on said ground truths are not expected to carry over to environmental samples as those usually contain a broader variety of materials.

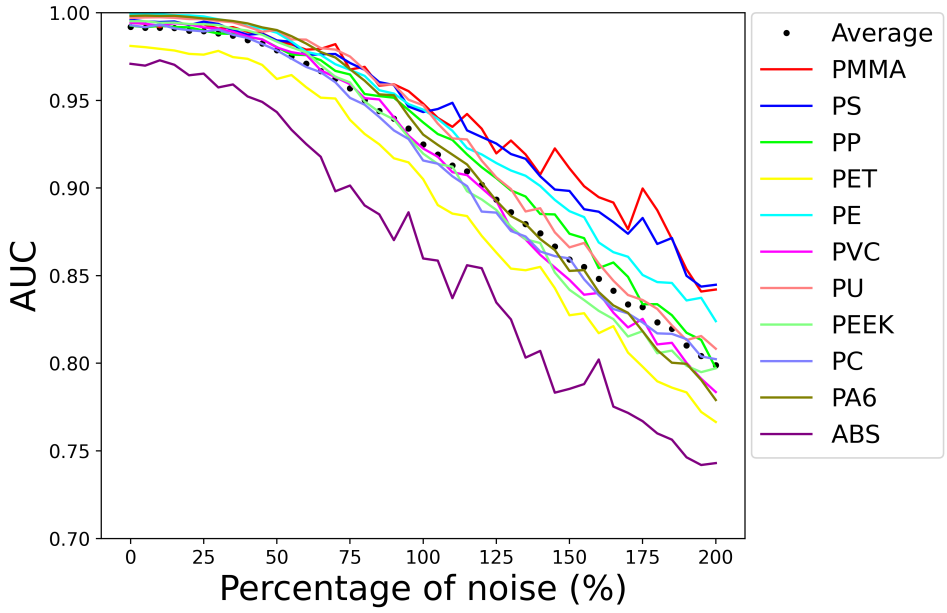
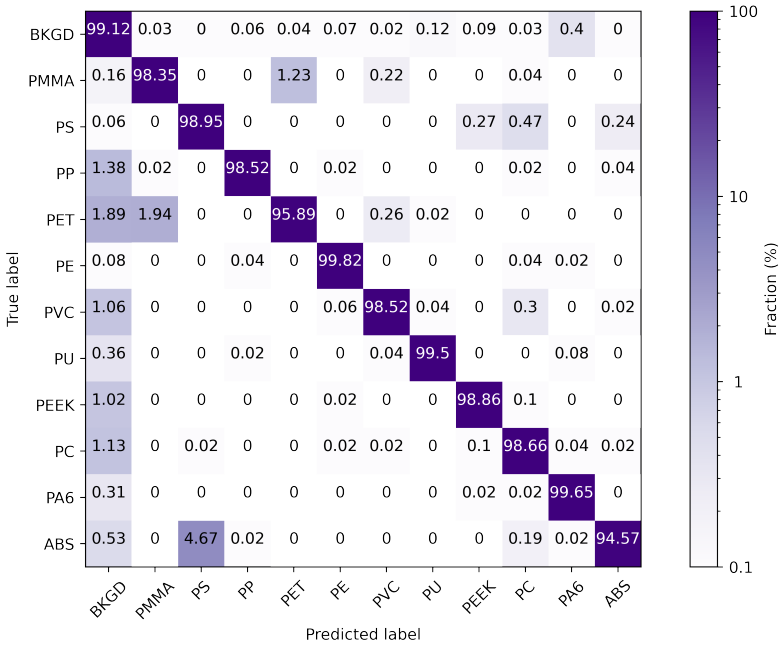
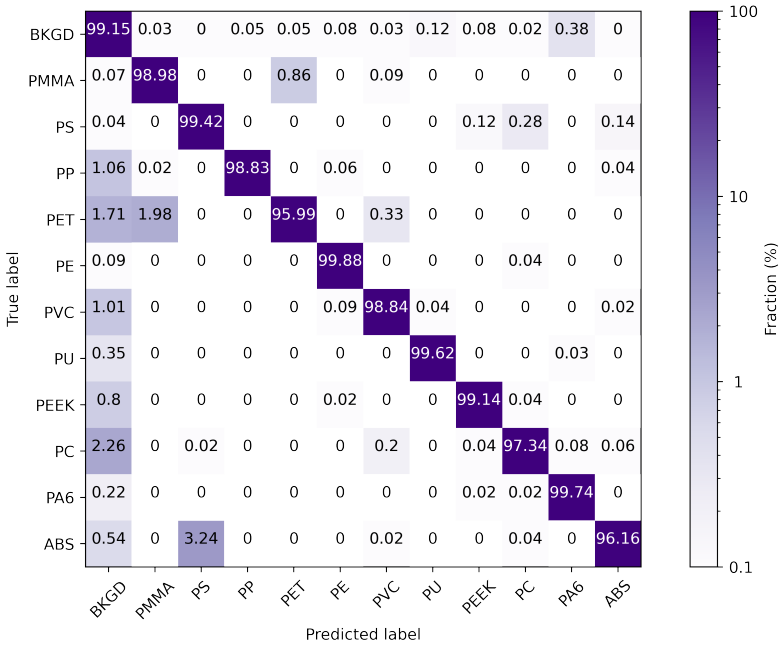


Fig. S1. Area Under the Curve showing the accuracy evolution depending on the percentage of noise of the test data set. The RDF model is trained with 15 channels and a subtraction descriptor at 1839 cm^{-1}. The noise is added to the Ground Truth spectra and corresponds to a normal distribution of random noise. The percentage of noise corresponds to the fraction of standard deviation of the spectra that is used as the standard deviation of the noise normal distribution, where a percentage of 0% means that no noise is added and a percentage of 200% means that the standard deviation of noise is 2 time the standard deviation of the spectra.



(a)



(b)

Fig. S2. Confusion Matrices of RDF trained with 15 channels and a subtraction descriptor at 1839 cm^{-1}. Both figures show the results have been predicted and validated using a Ground Truth with an additional percentage of noise with a standard deviation equivalent to 10% of the spectra standard deviations. The predicted Label corresponds to the RDF prediction and the True label is the Ground Truth. a) Graphical representation of the confusion matrix using a RDF prediction trained with the defined training data set; b) Graphical representation of the confusion matrix using a RDF prediction trained with the defined training data set combined with this same data set with a noise of 10% as augmented data.

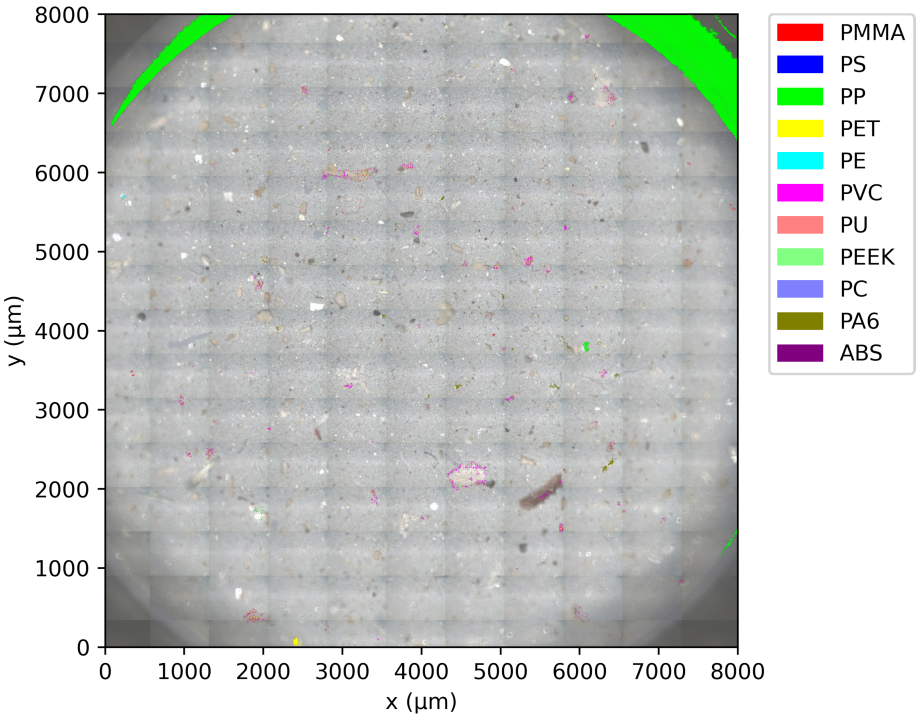


Fig. S3. RDF classification over a river sediments sample. The background and non polymer particles have no coloration showing the visible image of the sample. The RDF model has been trained with 15 channels and a subtraction descriptor at 1839 cm^{-1}. A manual comparison of the spectra of small particles with reference spectra of the corresponding material shows a good accordance. This representation shows proper fitting for small particles but the presence of particles not included in the training data set implies misclassification of those particles focused on the edges of the particles (where spectra have less intensity). The edges of the anodisc, mostly in the bottom and top left of the image, are out of focus, explaining why PP anodisc plastic ring is not identified in the bottom edges.

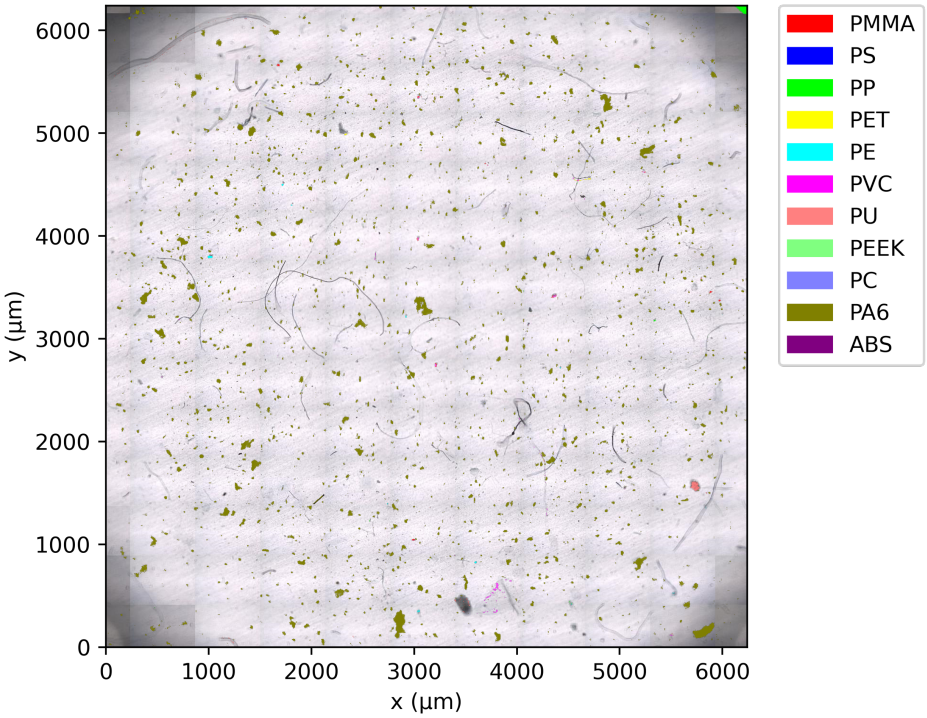


Fig. S4. RDF classification over a drinking water sample. The background and non polymer particles have no coloration showing the visible image of the sample. The RDF model has been trained with 15 channels and a subtraction descriptor at 1839 cm^{-1}. This representation showcases an abundance of PA6 particles. This behaviour is expected since organic particles may contain the Amides group in their spectra, meaning that organic materials and PA6 have similar spectra shape in the range between 1300 - 1900 cm^{-1}. This issue can be prevented by including non polymer particles (e.g. human skin particles) in the training data set, but at the same time this is a complex task which is under development.

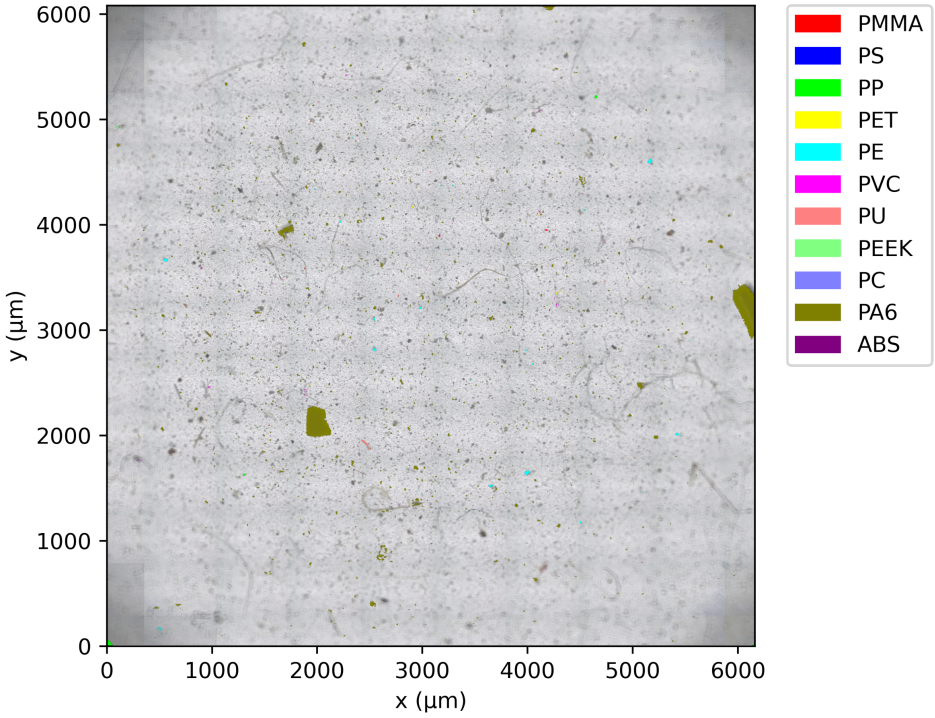


Fig. S5. RDF classification over a sea salt sample. The background and non polymer particles have no coloration showing the visible image of the sample. The RDF model has been trained with 15 channels and a subtraction descriptor at 1839 cm^{-1}. This image has a minor misclassificated particles mostly related to organic particles present as PA6.