Electronic Supporting Information

Exploratory analysis of hyperspectral FTIR data obtained from environmental microplastics samples

Lukas Wander,*[a b] Alvise Vianello,[c] Jes Vollertsen,[c] Frank Westad, [d] Ulrike Braun[a] and Andrea Paul [a]

[a.] *Bundesanstalt für Materialforschung und -prüfung, Richard-Willstätter-Straße 11, 12489 Berlin, Germany*
[b.] *Humboldt-Universität zu Berlin, Department of Chemistry, Brook-Taylor-Str. 2, 12489 Berlin, Germany*
[c.] *Aalborg University, Department of Civil Engineering, Thomas Manns Vej 23, 9220 Aalborg, Denmark*
[d.] *Department of Engineering Cybernetics, Norwegian University of Science and Technology, Norway*

Collection and preparation of the sample (Dat1)

The sample Dat1 is a soil sample collected in the vicinity of a recreational boatyard (small harbour) located in Aalborg, Denmark. The sample was collected from an area of 2.25 m^2 (1.5 x 1.5 m), collecting five soil sub-samples (one sample at each corner of this square area and one at the centre of the area) which were then mixed together and stored in a glass jar. The soil was transferred in a 2 litres beaker and pre-oxidised using a 10 % hydrogen peroxide solution (H_2O_2), then the liquid was evaporated, and the sample was dried in an oven at 55 ˚C for seven days. A sub-sample of 50 g of the dry weight was submitted to air-assisted density separation using zinc chloride solution (density 1.9 g cm^{-3}). Briefly, the sample was poured into a glass separation funnel (1 L capacity) and the funnel inlet was connected to a dry compressed air supply. The air was introduced into the funnel opening the stopcock and the sample was aerated for thirty minutes to ensure a proper mixing of the soil matrix, also helping to detach the particles from the environmental matrix. After the mixing step, the sample was let to settle overnight, and the bottom part was discarded. The top part including the floating particles was then filtered through a pre-muffled steel mesh filter (diameter 47 mm; mesh size 10 µm) and flushed with ultrapure water using a glass filtration unit connected to a vacuum pump. The material collected on the filter was removed by flushing it into a beaker (1 L) with 200 mL of ultrapure water. The sample was then submitted to a catalysed oxidation using hydrogen peroxide (H_2O_2) and Iron Sulphate ($FeSO_4$) (Fenton Reaction). After 24 h the sample was filtered onto a 10 µm steel filter (the same used for the previous steps) and submitted to a second flotation with $ZnCl_2$ solution in a small separation funnel (100 mL capacity) following the same procedure described above. After this step, the sample was split in two using a 500 µm sieve. The fraction smaller than 500 µm was filtered again and the material was flushed into a 150 mL beaker using 50 % v.v. Ethanol. The liquid containing the sample was then transferred in small aliquots into a 10 mL glass headspace vial and evaporated using a gentle flow of nitrogen, until the whole sample was transferred and dried in the vial. A known amount of 50 % v.v. Ethanol (5 mL) was then added into the vial. A sub-sample (200 µL) was deposited onto a Zinc Selenide window (13 mm diameter x 2 mm thickness) using a compression cell (Pike Technologies, Fitchburg, WI, USA) and let it dry overnight prior to submit it to FPA-Imaging-µFTIR analysis.

The analysis was carried out using FPA-Imaging-µFTIR (Cary 620-670 FT-IR microscope, Agilent Technologies, Santa Clara, CA, USA). A background scan was collected before each sample scan on a clean window using 120 co-added scans in the spectral range of 3750 – 850 cm^{-1} at 8 cm^{-1} resolution. Subsequently, the entire area of the sample's window was scanned using 30 co-added scans applying the same settings as for the background scan. A 15x Cassegrain objective was used, resulting in a pixel size of 5.5 µm on the 128x128 Mercury Cadmium Telluride (MCT) FPA detector. Afterwards, we analysed the acquired infrared map with an in-house built software called siMPle[1, 2], an updated software derived from MPhunter, previously used by Liu et al.[1], Simon et al.[3] and Vianello et al.[4]. The software allows for the analysis of each pixel constituting the map, comparing the spectra obtained from the sample's scan to a library of reference spectra applying Pearson correlation. The library was composed of 427 spectra divided in 75 groups, including polymers and natural materials. The software assigns the material with the highest correlation score to each pixel and "builds" the particle based on the scores attributed to adjacent pixels.

Analysis of Dat1

Removal of the fluctuating CO_2 band and noisy spectra

Both FTIR datasets were acquired under ambient conditions with a fluctuating CO_2 concentration over the measurement period of several hours. This adds variance to the data not related to the target of the analysis: finding and identifying particles. As proposed by Primpke *et al*.[5] and shown in Fig. S1 a), in an initial pre-processing step, the resulting CO_2 band was replaced by a straight line between 2420 cm^{-1} and 2200 cm^{-1}. For an improved estimation of the local baseline, 15 neighbouring points were used in this approximation. Using the CO_2 corrected raw data, inFig. S1 b) spectra with high noise level were removed from Dat1. A 2nd order polynomial was used to subtract the baseline between 2673 cm^{-1} and 2441 cm^{-1} prior to calculating the standard deviation in this range. Spectra exceeding a threshold of 0.005 were removed.

Results for a random spectrum taken from Dat1 are shown in Fig. S1. Whereas in a) the CO_2 absorption is the most significant feature in the original spectrum, this is not the case after removing the band. The noise level determined in b) for this spectrum is below the threshold.

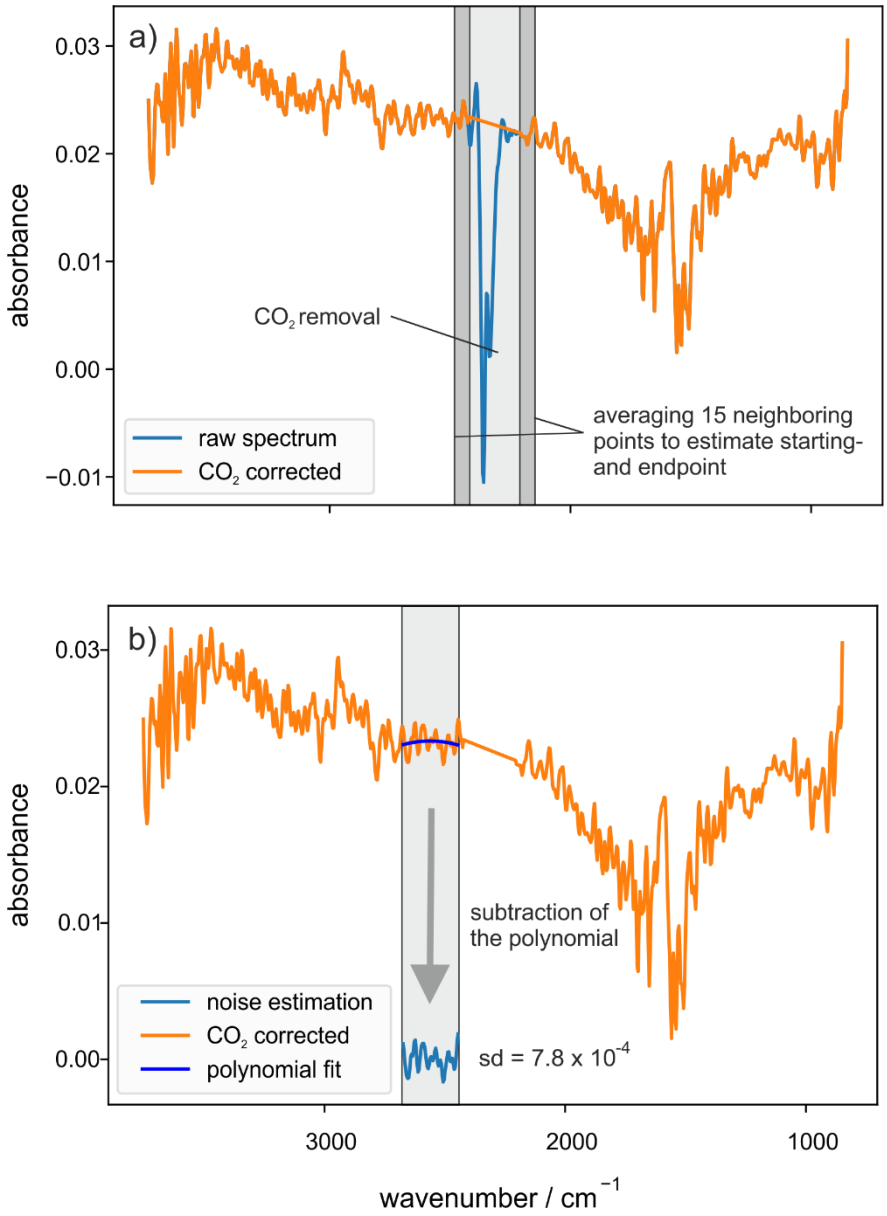

Fig. S1 a) Removal of fluctuating CO_2 band prior to analysis by introducing a straight line between 2430 cm^{-1} and 2200 cm^{-1}. b) Estimation of the noise level by calculating the standard deviation of baseline corrected data between 2673 cm^{-1} and 2441 cm^{-1}.

iPCA with the entire dataset

Spectra with a significant contribution to the variance of the data (particles) were separated from spectra carrying little information (substrate, blank) using iPCA.
Iinitially iPCA (n_components = 20, batch_size = 500). was performed on all the spectra of Dat1 using Savitzky-Golay first derivative spectra smoothed over three datapoints with a 2nd order polynomial. Noisy spectra with high absorbance levels are the main contributors to the global variance of the dataset and thereby dominate the iPCA results. The scores plot in Fig. S2 a) show no formation of groups, the loadings Fig. S2 b) possess no spectral features and the cumulative variance over the first 20 principal components rises very slowly and linearly, reaching only about 10 % (Fig. S2 c)).

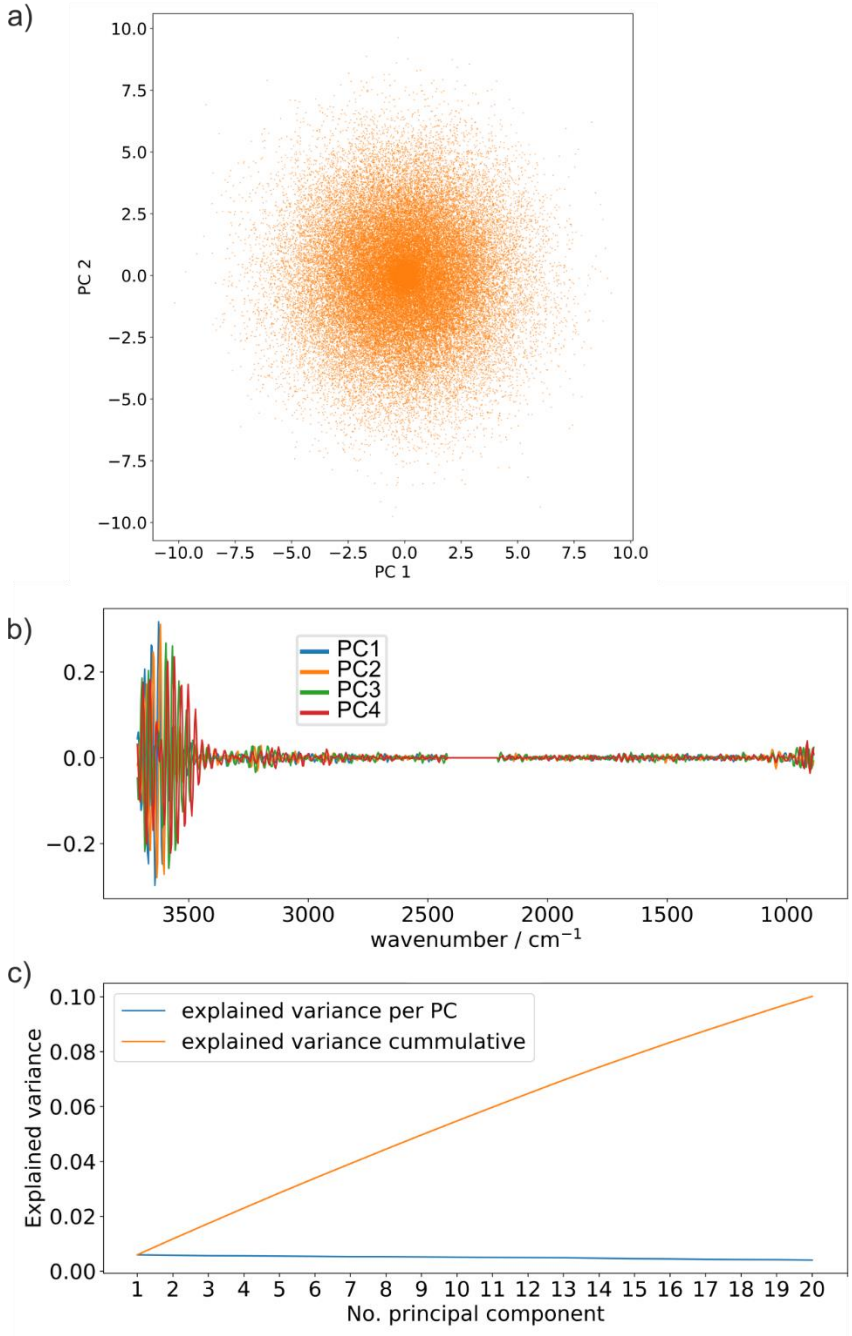

Fig. S2 Dat1: iPCA with the entire datasets. a) scores plot of PC 1 versus PC 2. shows no structure in the data. b) loadings of the first four principal components show no spectral features. c) the first 20 principal components cover merely 10 % of the variance.

Identification of noisy spectra

A 2nd order polynomial was used to subtract the baseline offset between 2673 cm^{-1} and 2441 cm^{-1} prior to calculating the standard deviation in this range as shown in Fig. S1. Spectra exceeding a threshold of 0.005 were identified as noisy spectra.

In Dat1 151268 noisy spectra were found and are highlighted in Fig. S3. Noisy spectra are located mainly in the corners of the imaged area. Radiation is almost completely blocked in this region yielding very high absorbance values >> 1 and therefore also very noisy spectra. In addition, with this approach some spectra originating from particles were removed. They are mainly located in areas of particle agglomerations such as the centre of the image. Spectra were measured in transmission which causes thick particles or agglomerates of particles to absorb almost all the light. Reflection and scattering[6, 7] may in addition contribute to higher absorbance and lower signal to noise ratio.

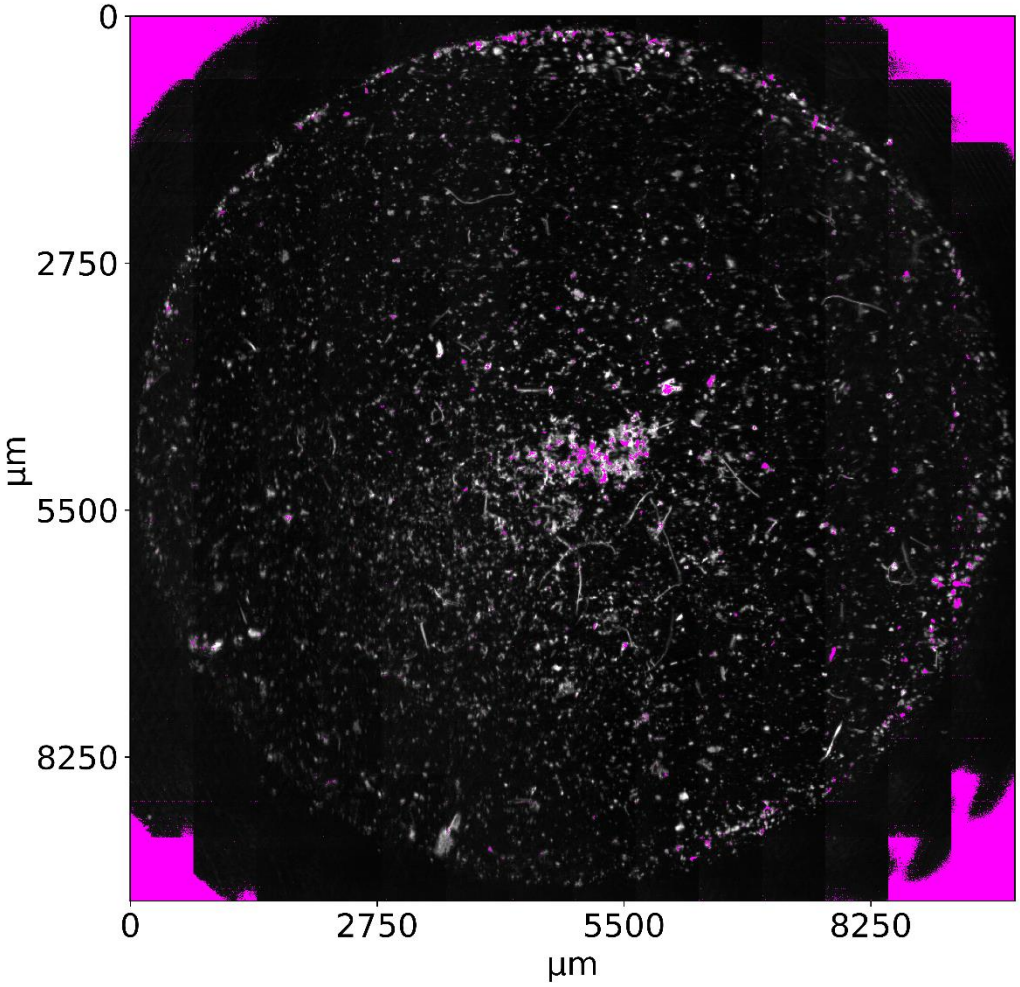

Fig. S3 Dat1: spectra with a high noise level (pink) are found on some particles and in the corners of the image.

iPCA selection of particle spectra after elimination of noise

Dimensionality reduction with iPCA (n_components = 20, batch_size = 500). was performed on the spectra of Dat1 remaining after removing the noisy data. Savitzky-Golay first derivative spectra smoothed over three datapoints with a 2nd order polynomial were used. The scores plot in Fig. S4 a) show some structure in the data compared to Fig. S2 and the loadings Fig. S4 b) possess spectral features in the ranges expected for microplastics and natural organic residue (C-H stretching between ca. 2800 cm^{-1} and 3000 cm^{-1}, some functional groups such carbonyls or amides between ca. 1800 cm^{-1} and ca. 1500 cm^{-1} and the fingerprint region at wavenumbers smaller than ca. 1500 cm^{-1}). The cumulative variance over the first 20 principal components accounts for 49 % of the total variance in the data, but for the selection of particle spectra only 12 PCs explaining 43 % of the variance were used.

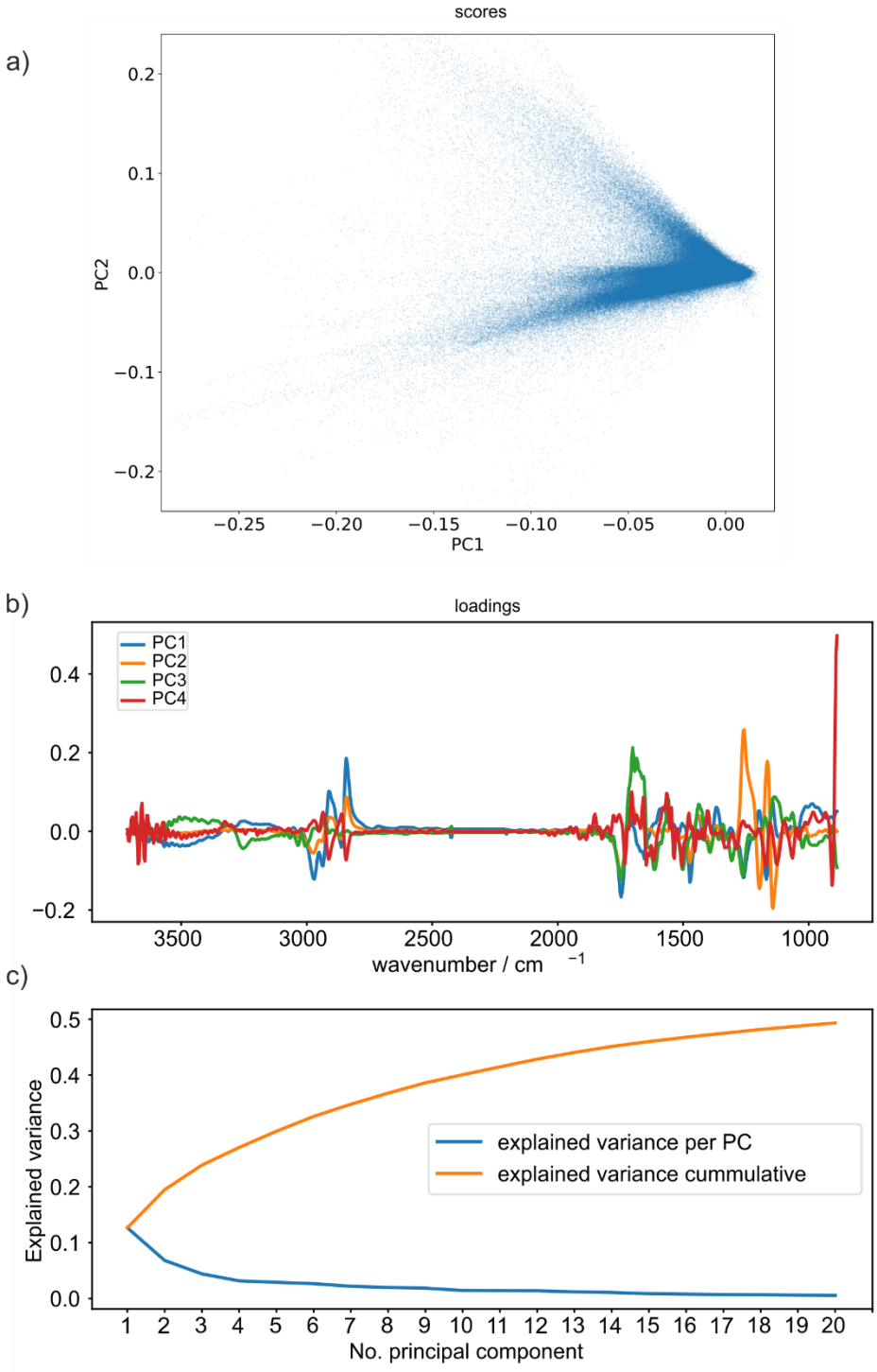

Fig. S4 Dat1: iPCA after the removal of noisy spectra. a) scores plot of PC 1 versus PC 2 shows structure in the data. b) loadings of the first four principal components show spectral features. c) the first 20 principal components cover 49 % of the variance.

iPCA of selected spectra

Spectra of selected particles were subject to iPCA (n_components = 20, batch_size = 500). Based on the CO_2 corrected raw data, Savitzky-Golay first derivative and smoothing over 15 datapoints using a 2nd order polynomial was used. Subsequently, normalization in the range from 3714 cm^{-1} to 883 cm^{-1} to unit length was performed. The scores values of the first 20 principal components were used in k-means (n_clusters = 8) clustering (Fig. S6). The first 20 principal components explain 85 % of the variance of the data, much more than before particle selection (Fig. S4 b)). The corresponding loadings of the first four PCs are shown in Fig. S5 a)

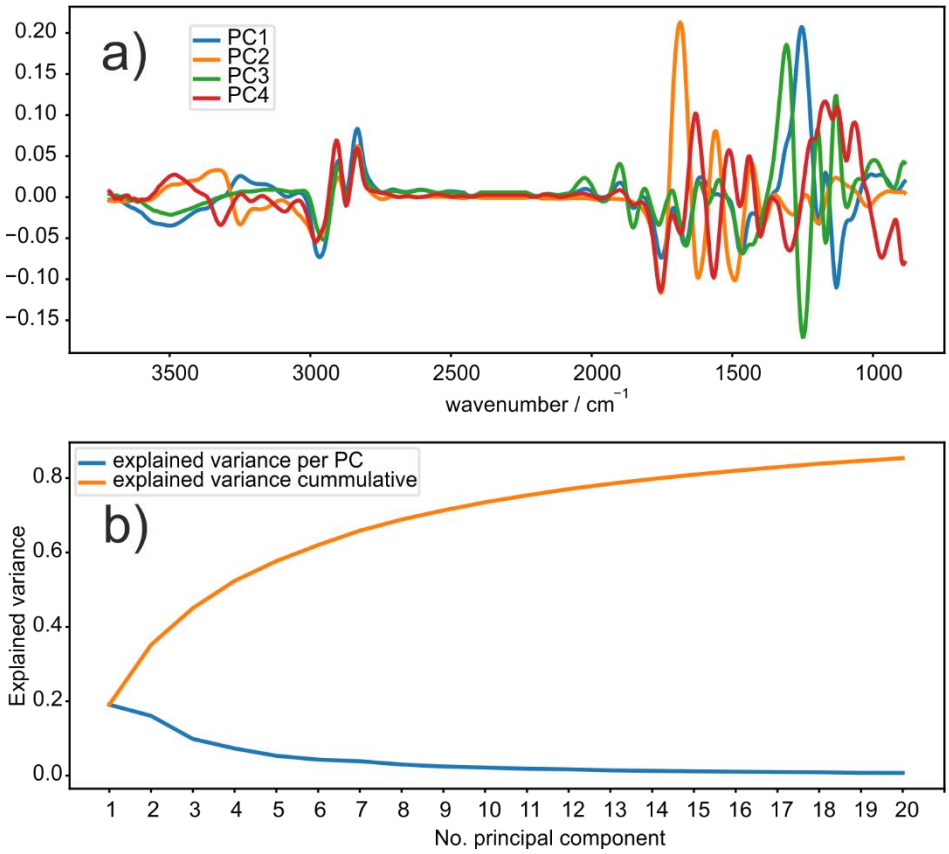

Fig. S5 Dat1: iPCA results of the selected particle spectra. a) The loadings of the first four principal components show spectral features. b) Explained variance of the dataset depending on the number of principal components considered.

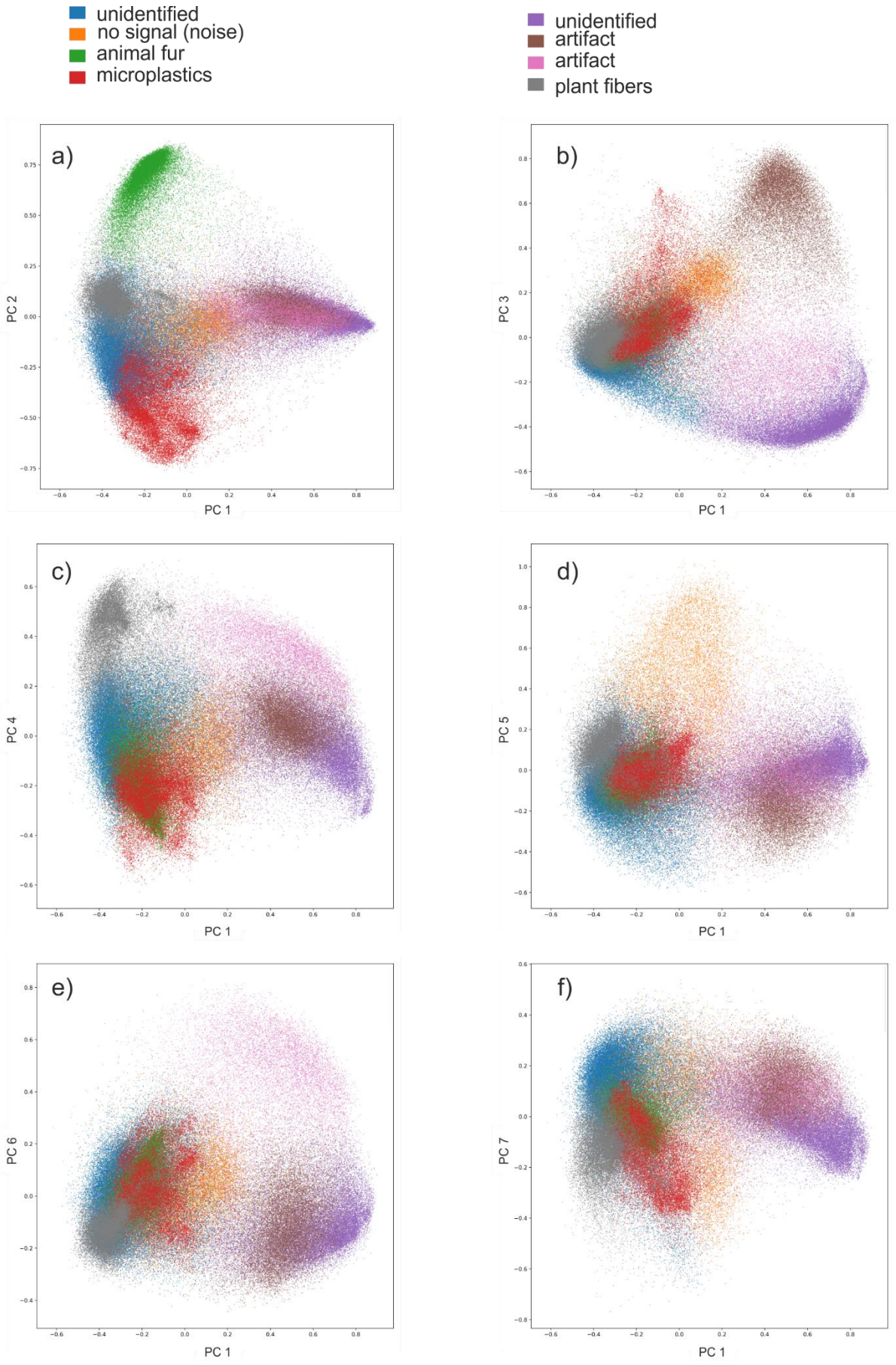

Fig. S6 Dat1: iPCA scores of PC1 versus PC2—PC7 of selected particle spectra clustered with k-means and labeled according to results obtained by comparing the mean spectrum of each cluster to a spectral library.

Evaluation of Dat1 with the siMPle software

Results obtained with the presented exploratory approach were qualitatively compared to a state-of-the-art automated library search with the siMPle software[1, 2] in Fig. S7. To give a quantitative comparison between the results is difficult because numerous parameters may be tuned by the user. The results are strongly affected by the spectral library used in the analysis and the thresholds which determine whether a match is obtained.

- The library search shows a smaller number of particles. Particles found with the exploratory approach but not by library search belong mainly to the four groups: PFTE-like (red) and alkyd (light green) unidentified (green) or spectra with artifact (light blue).
- Plant (cellulose) and protein-based material show good agreement
- Polyethylene, Polypropylene good agreement

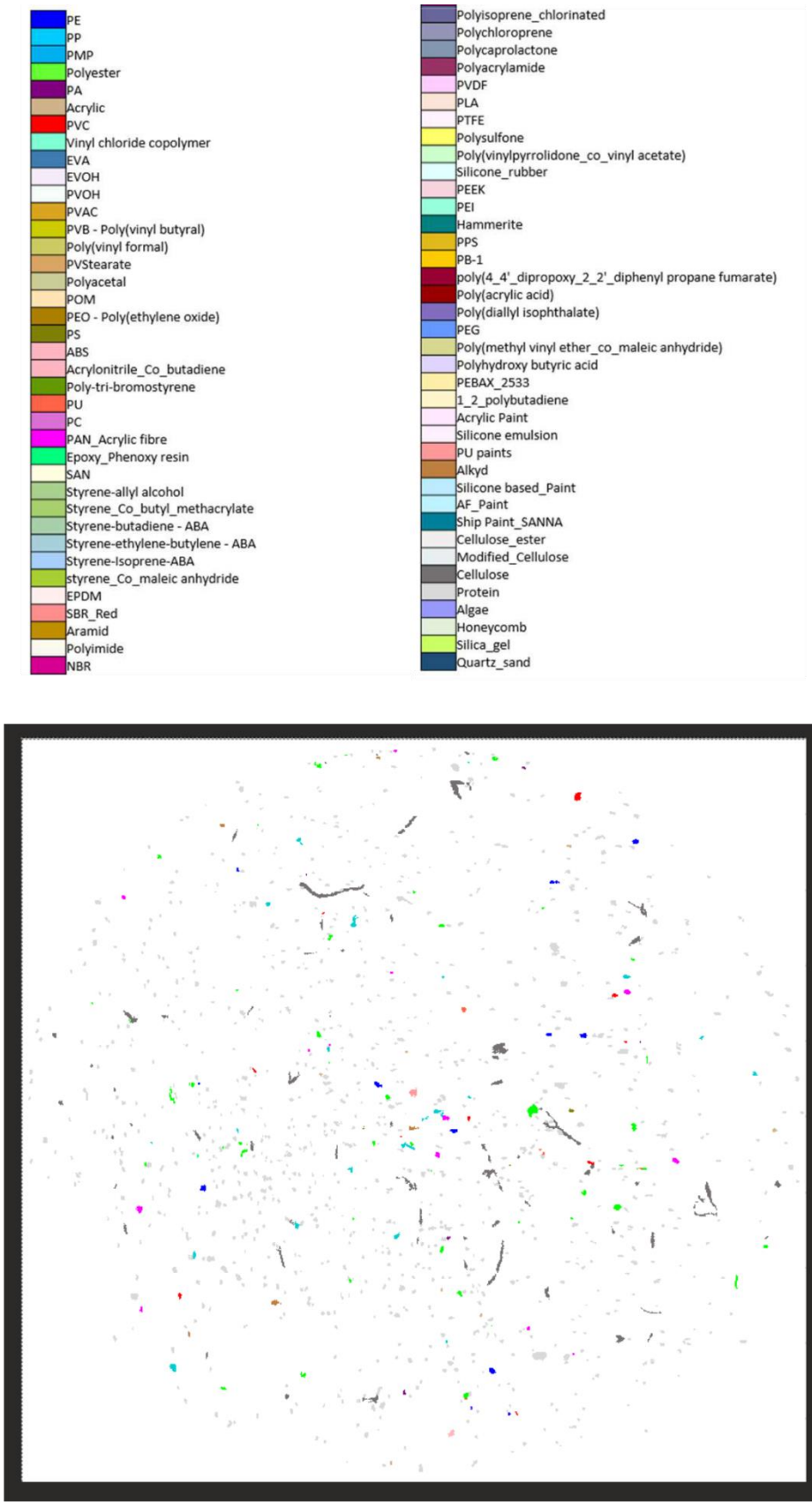

Fig. S7 Dat1: Particles identified by an automated library search with the siMPle software[1, 2].

Analysis of RevEnv2

Dataset RevEnv2 is less affected by very high absorbance and noisy spectra which dominated the initial iPCA in Dat1 shown in Fig. S2.

Selection of particle spectra was performed with the CO_2 corrected raw data. Baseline correction was achieved with Savitzky-Golay. first derivative spectra and smoothing with a 2nd order polynomial over 15 datapoints. Spectra belonging to particles were selected from iPCA (n_components = 20, batch_size = 500) scores. The first 12 PCs were considered and spectra closer to the origin than 0.06 were removed.

The first 12 principal components together explain 93 % of the variance found in the data. This allows to reduce the dimensionality of the dataset significantly without a mayor loss of information.
Selecting only the spectra significantly contributing to the variance in the data, the number of spectra in RevEnv2 was reduced by 89 %, from 1763584 to 191195.

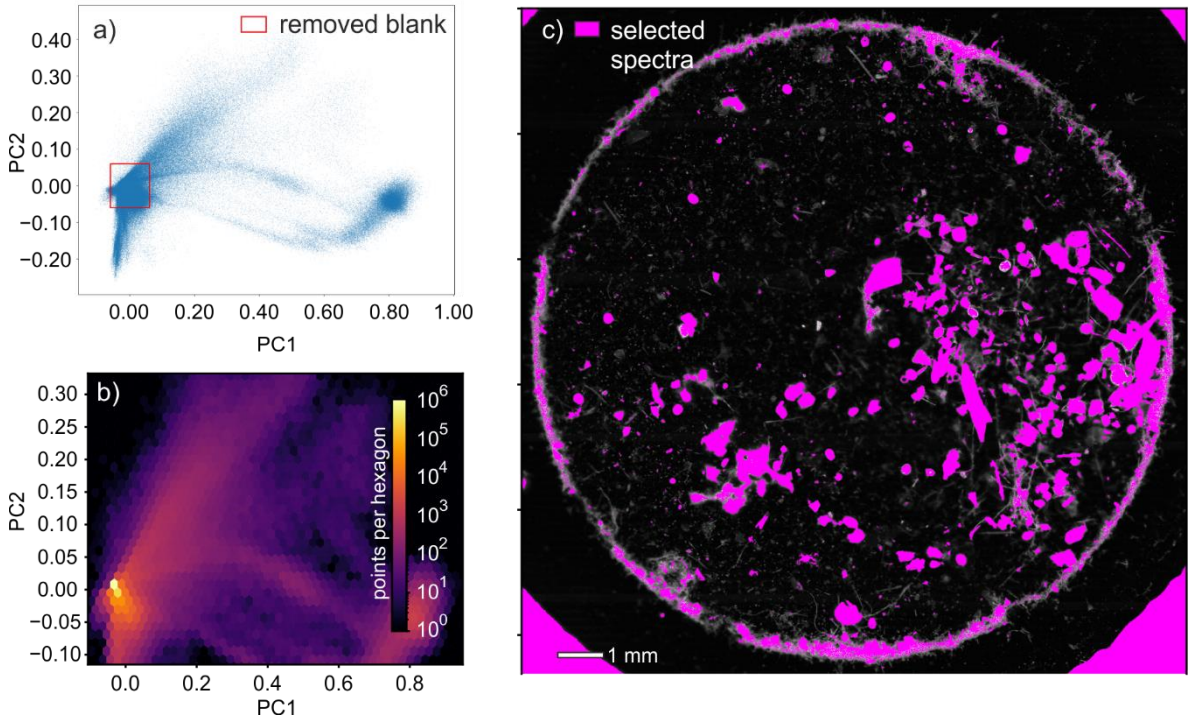

Fig. S8 RevEnv2: a) A threshold of 0.06 is used to remove blank spectra which carry little variance and are found close to the origin in the PCA scores plot. b) hexagonal binning of the scores in the PC 1 and PC 2 plane shows that the density of points is exceptionally high close to the origin. c) Spectra remaining after thresholding can be assigned to particles and the mounting of the substrate (corners).

iPCA of selected particle spectra

Selected spectra of particles were subject to iPCA. The scores plot in Fig. S9 show that several distinct groups are present in the data. k-means clustering with 8 clusters was used to systematically group the spectra. Chemical species were assigned. Due to the non-spherical shape of clusters 1 (unidentified), 6 (polyethylene / polyethylene oxidized) and 7 (polypropylene), k-means is not a suitable algorithm to separate those groups well. HDBSCAN (not shown) performs better on those clusters of different shape.

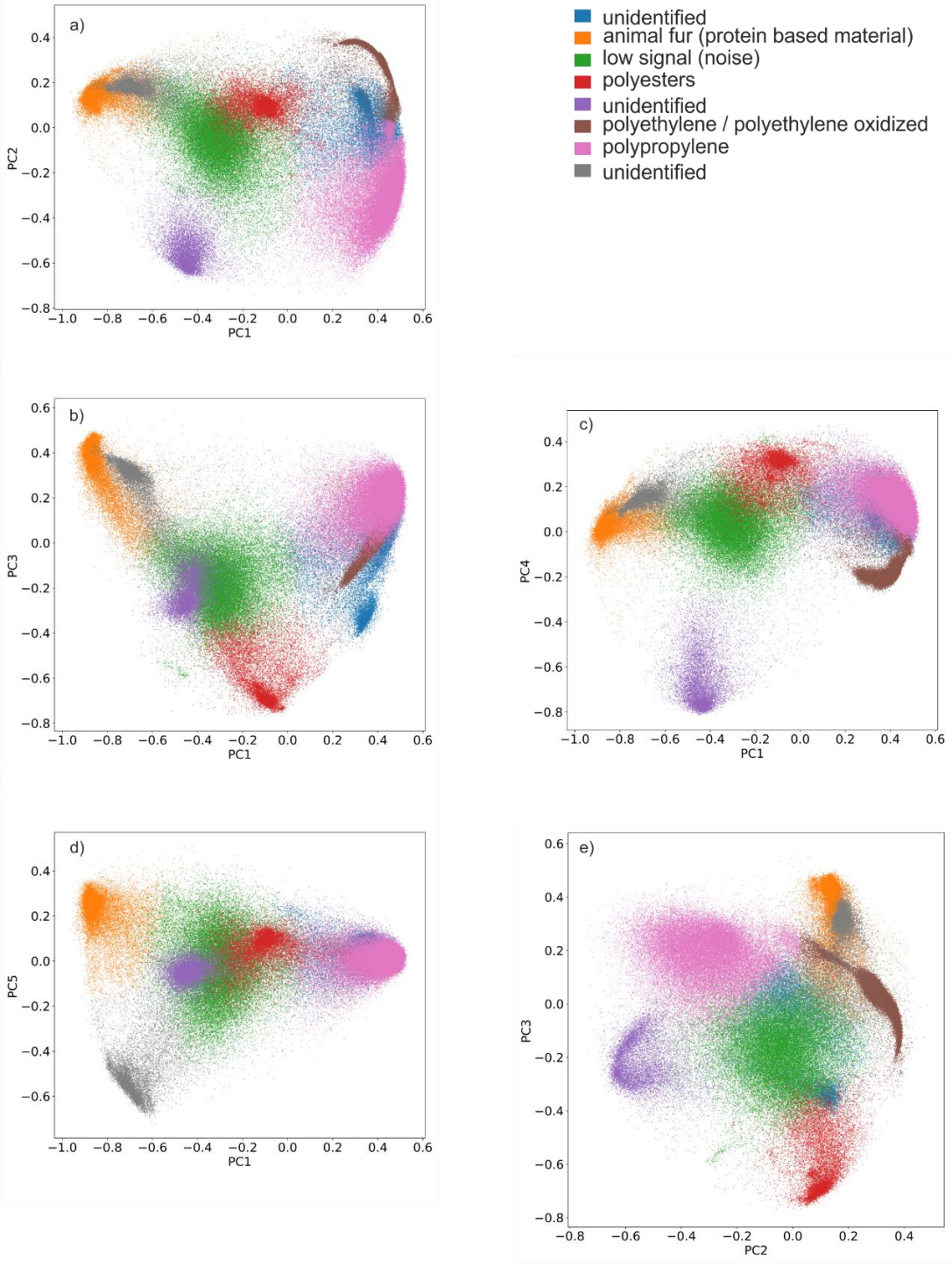

Fig. S9 RevEnv2: iPCA scores showing the clustering results with k-means for selected combinations of the principal components PC 1 to PC 5.

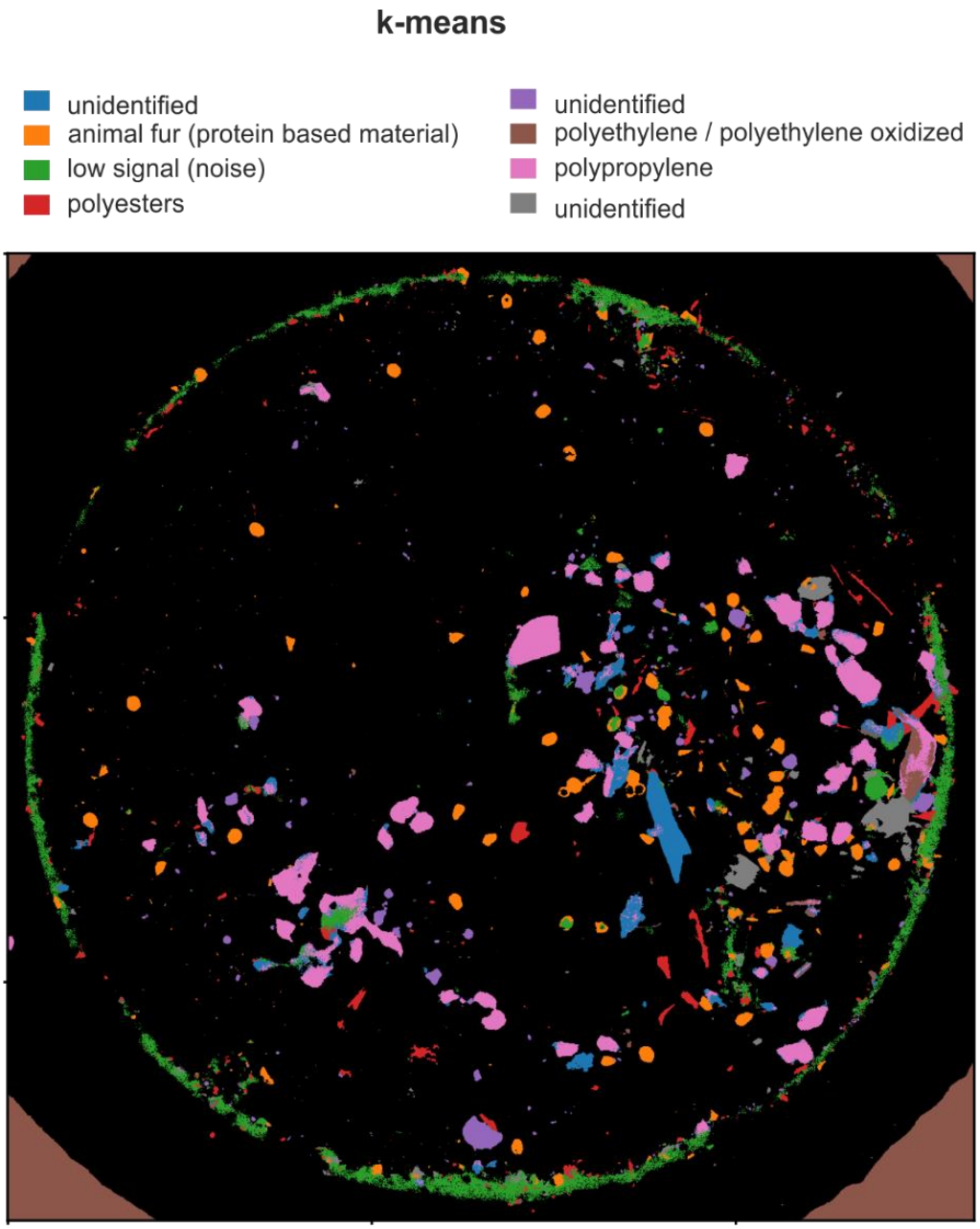

Fig. S10 RevEnv2: two-dimensional representation of the k-means clusters. Most particles are clearly assigned to one cluster. Incomplete separation between clusters unidentified (blue), polyethylene (brown) and polypropylene (pink) is also found in the two-dimensional image.

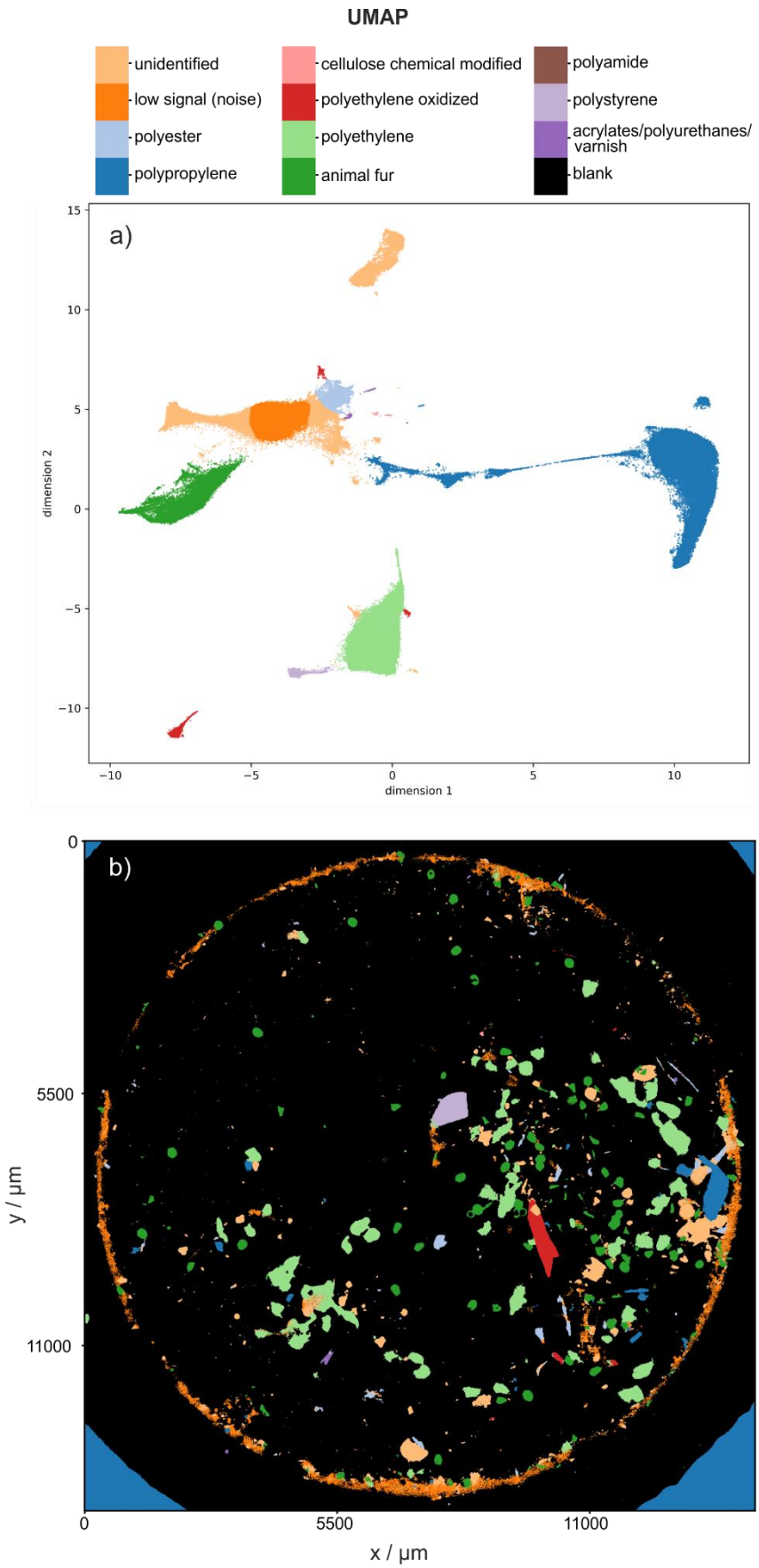

Fig. S11 RevEnv2: a) Spectra transformed into 2D space by UMAP with manual interactive clustering and substances assigned to clusters based on library search and visual examination. b) Clustering results represented as two-dimensional image

Baseline correction by an iterative polynomial fit

Python code for the custom baseline correction used for an initial visualization of the data:

```
def baseline_poly(X, order=2):
    """
    author: Frank Westad
    Baseline correction by iterative substraction of a polynomial preventing negative values
    """
    N, M = X.shape
    xw = list(range(M))    #
    Xw = np.zeros((N, M))  # temp x-values
    for i in range(N):
        Xw[i,:] = xw
    Yfit = np.zeros((N,M))

    P = np.zeros((N, order+1))
    Modpoly = X.copy()

    # iterative substraction of polynomial fit
    for k in range(50):
        for i in range(0,N):
            P[i,:] =  np.polyfit(Xw[i,:], Modpoly[i,:], order)
        for i in range(0,N):
            p = np.poly1d(P[i,:])
            Yfit[i,:] = p(Xw[i,:])
        #find values still below zero and asign them positive values
        Diff = Yfit-Modpoly
        mask = Diff < 0
        Modpoly[mask] = Yfit[mask]

    Xny = X-Yfit
    return Xny
```

Notes and references

1. F. Liu, K. B. Olesen, A. R. Borregaard and J. Vollertsen, *Science of The Total Environment*, 2019, **671**, 992-1000.
2. S. Primpke, P. A. Dias and G. Gerdts, *Analytical Methods*, 2019, **11**, 2138-2147.
3. M. Simon, A. Vianello and J. Vollertsen, *Water*, 2019, **11**.
4. A. Vianello, R. L. Jensen, L. Liu and J. Vollertsen, *Scientific Reports*, 2019, **9**, 8670.
5. S. Primpke, C. Lorenz, R. Rascher-Friesenhausen and G. Gerdts, *Analytical Methods*, 2017, **9**, 1499-1511.
6. B. Hufnagl, D. Steiner, E. Renner, M. G. J. Löder, C. Laforsch and H. Lohninger, *Analytical Methods*, 2019, **11**, 2277-2285.
7. P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas and P. Gardner, *Analyst*, 2009, **134**, 1586-1593.