JAAS

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *J. Anal. At. Spectrom.*, 2025, **40**, 2471

Self-organizing maps for the detection and classification of natural nanoparticles, nanoparticle systems and engineered nanoparticles characterized using single particle ICP-time-of-flight-MS

C. W. Cuss, *[a] M. F. Benedetti, [b] Carla Costamanga, [c] Lucas Mesnard[d] and M. Tharaud [b]

The development of single-particle inductively coupled plasma time-of-flight mass spectrometry (spICP-ToF-MS) heralds a breakthrough in our ability to measure the multi-elemental composition of natural nanoparticles and colloids (NPs), and to characterize the dynamics, responses, and impacts of systems of natural NPs (NNPs). However, further developments and associated comparisons across studies and research groups are hindered by the lack of a consistent, reliable and comparable approach for detecting and differentiating NPs and NNPs. Self-organizing maps (SOM, aka Kohonen networks) are single-layer artificial neural networks that are widely used for pattern recognition and classification in the natural sciences and beyond. The SOM is a nonparametric statistical method which adapts to data structures and is robust to noise, outliers, and sparse data, making it especially suitable for peak detection and particle classification using raw spICP-ToF-MS time-series. This article provides a brief review of SOM and their outputs before demonstrating their ability to detect particles in spICP-ToF-MS time-series, and to characterize and compare NNPs. Additional considerations and research directions for the application of SOM to spICP-ToF-MS and particle data are then discussed. The raw data and algorithms used in this study are provided in the SI to facilitate the testing of SOM across research groups, and for comparing their performance with other methods.

Received 2nd May 2025
Accepted 17th July 2025

DOI: 10.1039/d5ja00179j

rsc.li/jaas

1. Introduction

Natural nanoparticles and colloids (NNPs) play key roles in the speciation, transport and bioaccessibility of trace elements (TEs), nutrients and contaminants in surface waters, soil solutions and groundwaters.[1–3] Historically, technological limitations constrained the study of NNPs to the bulk phase and extrapolating the properties of particle populations with similar characteristics into the highly diverse and dynamic natural environment.[4–6] These strategies were exceptionally valuable for determining relationships between the properties and functioning of various bulk phase compositions and classes of NNPs;[7,8] however, the particle-scale composition, dynamics, and functionality of systems of diverse NNPs remain largely unexplored.

While constraints on measuring the properties of individual nanoparticles (NPs) were overcome more than two decades ago by advances such as atomic force microscopy (AFM) and scanning/transmission electron microscopies with energy-dispersive X-ray spectroscopy (S/TEM-EDXS), these methods are not suitable for characterizing the properties of the hundreds-of-thousands to hundreds-of-millions of NNPs typically found in each liter of natural waters. Similarly, bulk-phase measurements cannot account for particle-scale diversity since the combined properties of contributing particles can sum to the same value under a functionally infinite number of permutations of various particle populations with various properties that are present in various proportions. Furthermore, interactions between NNPs may alter functionalities at the scales of particles and populations without altering the measured bulk-scale properties.

Since early 2010, advances in single-particle (sp) inductively coupled mass spectrometry (ICP-MS) have facilitated rapid measurement of one or two elements in NNPs in a range of matrices, including natural waters.[9–12] More recently, specialized time-of-flight (TOF) mass analyzers with on-board data processing have been used to measure the complete mass

[a]*Laboratory for Environmental and Analytical Nanogeochemistry, Memorial University of Newfoundland (Grenfell Campus), Canada. E-mail: ccuss@mun.ca*
[b]*Université Paris Cité – Institut de Physique du globe de Paris, CNRS, F75005 Paris, France*
[c]*Instituto de Química Física de Materiales, Ambiente y Energía (INQUIMAE), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*
[d]*Université PSL (Paris Sciences & Lettres), France*

spectrum every 25 μs, allowing the simultaneous quantification of nearly all elements in each individual nanoparticle in natural systems.[13–18] This breakthrough marks a revolution in the capacity to measure the elemental composition of NNPs and systems of NNPs for relating to their functions in the environment and elsewhere, contributing to entire new research areas such as environmental nanobiogeochemistry.[19,20]

A solution containing nanoparticles/colloids in suspension is continuously introduced to the plasma in spICP-TOFMS. The time-resolved recording of this introduction provides an element-specific time series (counts *vs.* analysis time) in which the elements present in nanoparticles produce a relatively fast and elevated transient signal (*i.e.* a peak). It is generally accepted that 200–1100 μs are required for the cloud of ions arising from a single particle to pass the detector.[9,21,22] For TOFMS, three 25-μs acquisitions of the entire mass spectrum are typically averaged to improve the signal-to-noise ratio, such that 3–15 averaged measurements (*i.e.* data points) are obtained for each particle.[13,17] In some cases, these acquisition periods are combined into datapoints spanning time periods up to 3 ms.[23] Samples are typically diluted to reduce the background concentration and minimize the probability of two particles passing into the plasma at the same time such that roughly a couple thousand particles are measured within a 120-s analysis. Hence, ~2% of the data is associated with particles, while the rest is associated with the background signal (max. 15 measurements per particle × 0.000075 s per measurement × 2000 particles = 2.25 s of measurement time). Multiplying this large number of data points by the number of elements measured, it is readily apparent that data processing challenges becomes a primary concern. Two such challenges are: how to distinguish particle events/from the background, and how to identify and compare different populations of NPs within and between samples?

Since the number of ions of an element produced by a particle is proportional to the amount of the element present therein, an inability to distinguish the particle signal from the background also imposes a lower limit on the mass, and consequently the size of a mono-elemental particle that can be measured, and on the measurable mass of an element within a multi-elemental particle. This critical challenge is typically addressed by setting a threshold based on one or more assumed or measured distributions.[9,15,18,21,22,24–27] While this efficient approach provides a reliable confidence measure, reducing false positive detections based on a threshold necessarily increases false negatives and thus excludes a disproportionately large number of smaller particles. For example, applying a threshold with 98.7% confidence (α = 1.3%) based on a compound Poisson distribution produced a false negative rate of 32% (*i.e.* 32% of known particles were excluded).[23] Considerable information is thus lost when thresholding is used to identify particles, which may heavily bias results and impede the detection of differences between systems of particles since particles containing the lowest masses which may have differing compositions from larger particles are preferentially excluded.

The identification and comparison of particle populations within and between samples are vital in environmental nanogeochemistry, as they are required to connect particle populations to their ecosystem functioning, and to measure system-level changes.[19] Typically, particle populations are grouped and compared using (semi-)supervised approaches such as classification algorithms[18,24,28–30] and clustering algorithms.[16,30,31] Classification requires training algorithms to classify particles, which is not applicable for untargeted analysis in unknown NNP systems (NNPs) with diverse particle populations. Clustering requires subjective input, and the clustering of particles into groups is in part a function of the degree of difference between particles and their number within a given dataset. Clustering also becomes less effective as dimensionality (*i.e.* the number of elements measured) increases, and many algorithms perform poorly for sparse data (*i.e.* when there are many concentrations of elements that are zero), for noisy data and outliers (*i.e.* when measurements are heavily impacted by random variation and when a limited number of particles with unusual compositions are measured), and for skewed and/or multimodal distributions (*i.e.* when the contribution of one or more element is not distributed symmetrically as a function of the number of particles and/or has more than one maximum).[32–37] While newer clustering approaches such as the Gaussian mixture model, DBSCAN and hierarchical agglomerative clustering have advantages for complex data, they also suffer from shortcomings associated with particle mass and composition data due to non-Gaussian distributions, variable cluster densities, and high-dimensional, noisy data with outliers (Khan *et al.*, 2014; Wani, 2024).[38,39]

Kohonen networks are unsupervised single-layer artificial neural networks with excellent visualization capabilities which are broadly used for pattern recognition and classification, more commonly known as self-organizing maps (SOM).[40] The SOM algorithm performs an unsupervised topology-preserving projection using adaptive cluster centers, which can be formally delineated by meta-clustering on the converged map. It can thus be considered a prototype-based clustering approach similar to fuzzy clustering and mixture models;[35] however, the SOM is more suitable for complex data due to its ability to adapt cluster centers to data properties.

The SOM algorithm is nonparametric, robust to noise and outliers, and generally superior to clustering methods for sparse matrices and high-dimensional data with unusual distributions, including nonlinear and compositional data.[34,41–45] The SOM has been widely used in signal- and image-processing applications, including: feature identification and classification for content-based image retrieval from the internet;[46] identifying patterns in satellite imagery;[47] matrix effect correction in EDXRF analysis;[48] preprocessing and analyzing electrocardiogram signals;[49] recognizing and classifying bedforms on Earth and Mars,[50] and; reconstructing signals with sparse events in brain–machine interfaces.[51] Applications of SOM to environmental studies are similarly numerous, including: assessing multi-elemental emissions from industry;[52] assessing spatial and temporal patterns of pollutants in environmental compartments;[53] classification of river water quality using large

environmental data;[54] determining relationships between the molecular mass and optical properties of organic matter over time and across species;[55] identifying ocean current patterns and their connection to weather forcing;[42] resolving fluorophores in excitation–emission matrices and measuring variation in the corresponding composition of organic matter,[45,56,57] and; several geochemical applications.[43,44,58,59]

Following a brief introduction to Kohonen networks and their features, this study provides a proof-of-concept implementation of SOM for analyzing data generated by the analysis of NNPs using spICP-ToF-MS. The use of SOM for detecting and resolving NNP peaks from the background in spICP-ToF-MS time-series is first demonstrated. The SOM algorithm is then used to distinguish between NNPs extracted from different soils, and to detect ENPs in these NNPs. The benefits, shortcomings and considerations for future application of SOM to characterize NNPs using data generated by spICP-ToF-MS are then discussed.

2. Outputs and visualization of self-organizing maps

To facilitate the use of SOM and their comparison with other methods for detecting and comparing NPs and NNPs, the algorithms, commands, and raw data used in this study are provided in the SI. A comprehensive explanation of SOM and their optimization is extraneous to demonstrating their effectiveness, but useful for researchers wishing to test or implement the SOM; hence, these are provided in Section S1. The following section describes the major features of SOM and defines associated terms.

Excellent visualization of relationships between samples, best-matching units (BMUs), clusters, and the corresponding distributions of variables on the trained map is a major advantage of SOM. The following visualizations are especially relevant for spICP-TOFMS data (Fig. 1).

2.1 Best-matching unit (BMU)

The hexagon on the trained SOM for which the underlying vector is most similar to a given input vector is the BMU for the input vector, where similarity is typically expressed as Euclidean distance. This can be considered a primary and unsupervised clustering since several input vectors may have the same BMU.

2.2 Unified distance matrix (Umatrix)

A measure of similarity between neighbouring neuron vectors on the trained map, typically expressed as Euclidean distance. This facilitates the identification of outliers on the trained map, and a visual assessment of the soundness of meta-clustering.

2.3 Meta-clusters (Mclust)

The clusters resulting from a *k*-means clustering on the trained SOM. While the SOM performs an unsupervised organization of sample vectors into clusters around their respective BMUs, the meta-clustering of the corresponding neurons requires input to determine the optimal number of clusters.

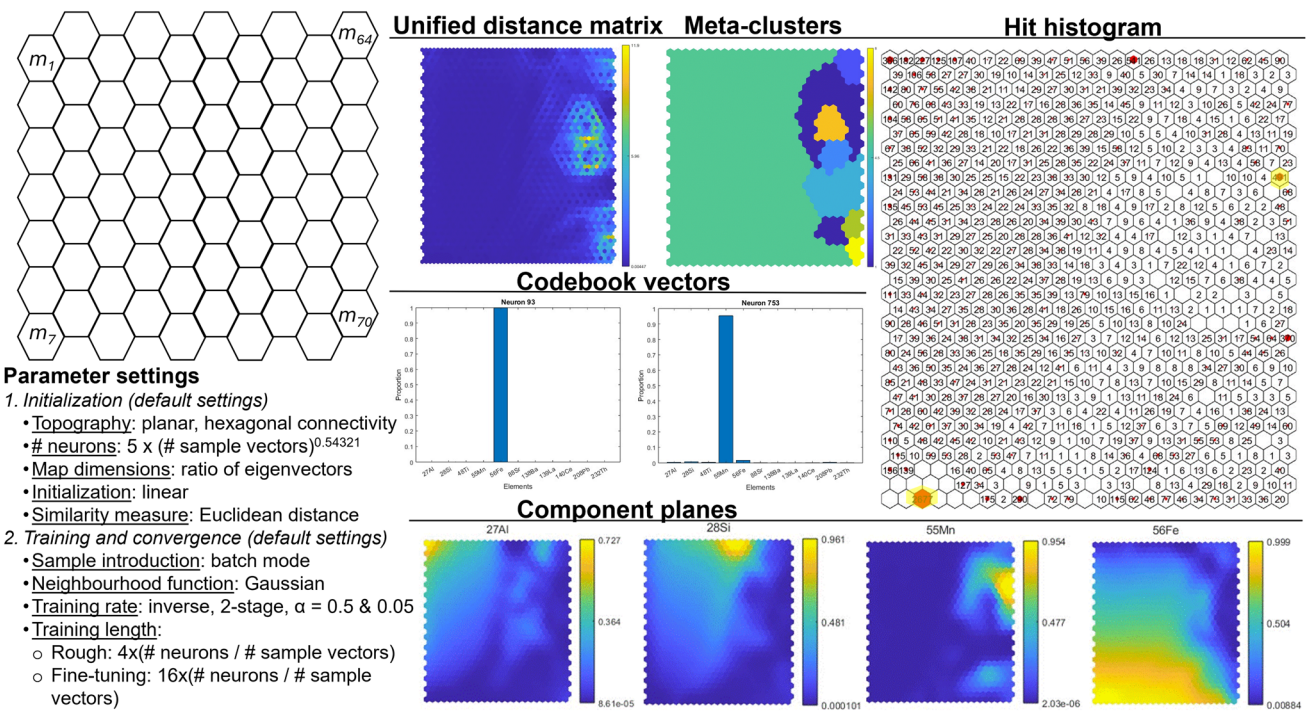

Fig. 1 Schema of self-organizing map parameter settings used in this study, numbering of neurons, and selected outputs from the trained particle differentiation comparing triplicate GS and VS extracts.

2.4 Hit histogram (HH)

Displays a number on each neuron corresponding to the number of sample vectors for which it is the BMU along with a coloured dot with size proportional to the number. Neurons that are not BMU for any sample vectors are left blank. Blank neurons occur where additional space between neighbours is needed to optimize the topology-preserving projection and thus indicate large differences between neighbours separated by blank neurons. Hence, the HH may also be useful for assessing the soundness of a given meta-clustering.

2.5 Codebook vectors (Cvects)

Display the values of each variable in a chosen neuron vector.

2.6 Component planes (Cplanes)

Display the distributions of variables in all neuron vectors across the map and thus may also be useful for assessing the soundness of a given meta-clustering.

The neurons can also be identified according to the samples for which they are BMUs by placing the sample names on the corresponding map hexagons; however, this is not useful for SOM with millions or even thousands of samples because the map becomes cluttered with overlapping text. Alternatively, this information can be printed in Matlab or output to a text file for further processing.

3. Materials and methods

3.1 Collection and preparation of samples and standards

Soil samples were collected near the cities of Gambaiseuil and Monchauvet in June of 2023, from beside the respective headwaters of the Vaucouleurs (48.885736°N, 1.624297°E) and Ponts Quentin (48.7575°N, 1.735392°E) rivers. Information about these watersheds and soils are detailed elsewhere.[60] The lithology of Gambaiseuil soils are 100% sands, whereas soils collected near the Vaucouleurs are 67% limestone/chalk parent lithology, and 33% sand.

Particles were extracted from soils in triplicate on the following day by ultrasonicating 20–25 g in 500.0 mL of ultrapure Milli-Q water (MQW; ≥18.2 MΩ cm, Millipore-Sigma, Massachusetts, USA). Extracts were then centrifuged for 5 min. At 4000 rpm in 50-mL centrifuge tubes (Fisher Scientific, Massachusetts, USA), using a 5810R centrifuge (Eppendorf, Hamburg, Germany). The corresponding Vaucouleurs (VS) and Gambaiseuil (GS) soil NNPs were isolated by filtering the supernatant through acid-rinsed 1.2 μm PTFE filters Minisart (Sartorius, Göttingen, Germany) and diluting 500-fold using MQW. The NNPs were refrigerated at 4 °C until analysis on the following day.

Tri-elemental ENPs were created using a mixture of FeCoNi and FeCoZn nanoparticles with 40 nm nominal diameter (US Research Nanomaterials, USA). The ENPs had a theoretical mole ratio of 67%, 17% and 17% for Fe, Co, and Ni/Zn, respectively. The preparation and analysis of these ENPs are detailed elsewhere.[16] Briefly, a few μL of each suspension was diluted in 50 mL of MQW and sonicated for 3 minutes. Prior to spICP-TOFMS analysis, the mixture was prepared and highly diluted to avoid NP coincidences.

3.2 spICP-TOFMS analysis and instrument settings

Samples were analyzed by scanning the mass range of 23–245 amu every 27.5 μs using a Vitesse ICP-TOFMS analyzer (Nu Instruments, North Wales, UK). Three spectra were accumulated and averaged, requiring a total of 83 μs for each measurement in single particle mode. The following isotopes were used in the present study: (1) Particle detection and comparison of NNPs: ^{27}Al, ^{28}Si, ^{55}Mn, ^{56}Fe, ^{48}Ti, ^{88}Sr, ^{138}Ba, ^{139}La, ^{140}Ce, ^{208}Pb, ^{232}Th, and; (2) Differentiation of NNPs and ENPs: ^{56}Fe, ^{59}Co, ^{60}Ni, ^{66}Zn, ^{107}Ag, ^{197}Au. A mixture of 15% H_2 and 85% He was used as the collision cell gas. Complete instrument operating conditions are described elsewhere.[16]

The time-series of all measured isotopes (*i.e.* counts of each isotope measured every 83 μs) were exported as text files directly from Nu CoDaq software that is used to operate the Vitesse. Text files were uploaded to Matlab R2024a (The MathWorks Inc., Natick, Massachusetts) for further data analysis and Figure rendering. Raw data text files for these analyses are provided in the SI. The SOM algorithm and *k*-means clustering were implemented in Matlab using the SOM Toolbox version 2.0.[61] The algorithms and commands used for data transformation, analysis and visualization are provided as SI.

3.3 Data pre-processing

Effective pattern identification using artificial neural networks requires data presented in a form emphasizing relevant differences. This includes pre-processing to eliminate noise and irrelevant information and transforming data into variables describing properties of interest. To this end, data was processed differently to identify and differentiate/group nanoparticles (Fig. 2).

After testing several combinations, time-series for each isotope were first smoothed using five passes of three-point averaging. This also duplicates the default settings of the software typically used to process time-series measured using the Vitesse (NuQuant, Nu Instruments). The background signal was corrected for drift and removed by subtracting the average counts over the entire time-series. Baseline drift was determined by averaging the first and last 50 000 points on each time-series, dividing the difference by the total number of time points, and subtracting the product of this increment and the number of the corresponding time point. Negative values were set to zero. Noteworthy, this over-corrects for the background because peaks corresponding to particles are included in averages; however, this overcorrection is negligible because dilution to avoid particle coincidence ensures the number of data points corresponding to peaks is very low relative to those corresponding to the background.

For each sample, every data point of the smoothed, background-corrected time-series was transformed into three variables differentiating particle peaks from the background: the number of elements with a signal greater than zero (NumElts), the combined counts for all isotopes (TotSig), and

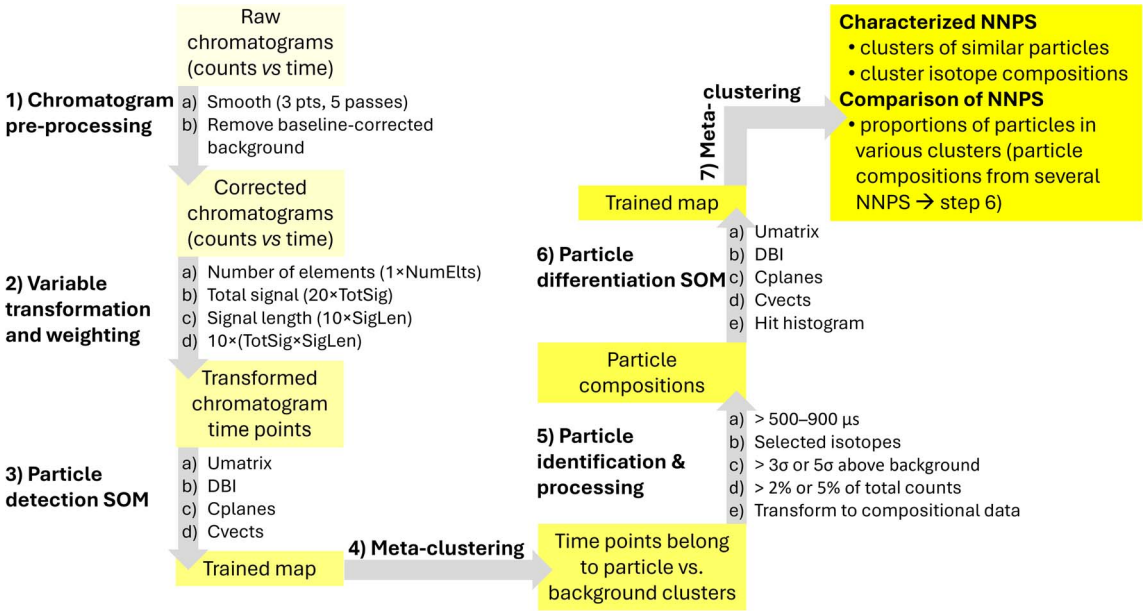

Fig. 2 Data processing and analysis flow diagram illustrating steps and processes used to detect and differentiate particles, and to distinguish between particle systems.

the number of consecutive data points with >0 counts, maximized across all elements (SigLen). The product of TotSig and SigLen was added as a fourth variable to emphasize the importance of their correlation in peak detection.

Substantial testing revealed that weighting variables improved peak identification by stretching their range to improve differentiation and increasing the magnitude of variation. This contrasts with the typical process of normalizing the variation of all variables so that none exerts greater influence. Variables were weighted by multiplying them by a constant, as: 20 × SigLen, 1 × NumElts, 10 × TotSig, and 10 × (SigLen × TotSig). These variables were inputs for the particle detection SOM, trained using the parameter settings in Fig. 1. By way of example, the raw data and corresponding transformed variables input to the particle detection SOM for the tri-elemental ENPs are provided in the SI. A visualization and comprehensive discussion of the process and interpretation of outputs is also provided in Section S2.

3.4 Data post-processing

Outputs of the particle detection SOM were post-processed as follows to describe each sample in terms of particle compositions. Then, the particle composition data was input to the particle differentiation SOM.

A *k*-means meta-clustering was performed on the self-organized particle detection SOM to distinguish data points corresponding to particles and background. This meta-clustering introduces a semi-supervised component to the process; however, once it is fixed it should be applied to all samples that will be compared to avoid potential bias. Solutions with 2–10 meta-clusters were assessed by comparing the Cplanes, Umatrix, Davies-Bouldin index (DBI),[62] changes in cluster distribution, and visual inspection of obviously false positives and false negatives on time-series. To minimize the risk of selecting a local maximum, 500 clusterings were performed from random starting conditions for each of the 2–10 cluster solutions. While particle signals may be as brief as 200 μs, the corresponding lower limit of ≥ three datapoints (*ca.* 249 μs) with signal >0 has a significant probability of random occurrence due to noise and indeed resulted in numerous false positives. Hence, the identification of particles was also constrained to require a time point sequence of >500–900 μs (*i.e.* >6–11 datapoints) with signal >0 for any one of the isotopes ^{27}Al, ^{28}Si, ^{48}Ti, ^{56}Fe, ^{139}La, or ^{140}Ce, which belonged to clusters representing obvious particles during visual inspection (Fig. 3).

To prevent noise and instrument fluctuations from being included as minor elements in particles, all data points less than three (for NNPs comparison) or five (for NNPs and ENP differentiation) standard deviations above the corrected background were removed. The corrected background and standard deviations were calculated after removing time points identified as particles. Isotopes contributing less than two (for NNPs comparison) or five (for NNPs and ENP comparison) percent of the total counts for a particle were also removed to account for masses with backgrounds that were frequently zero, such that the counts from random events/noise at even a single time point could be more than five standard deviations above the mean. Although these constitute cutoffs, they are applied only after particle detection has occurred and thus do not produce the typically high rate of false negatives in particle detection that is noted above. The remaining particle data was transformed so that the counts of each isotope were represented as a proportion of the total particle counts (*i.e.* compositional data), and introduced to the particle differentiation SOM.

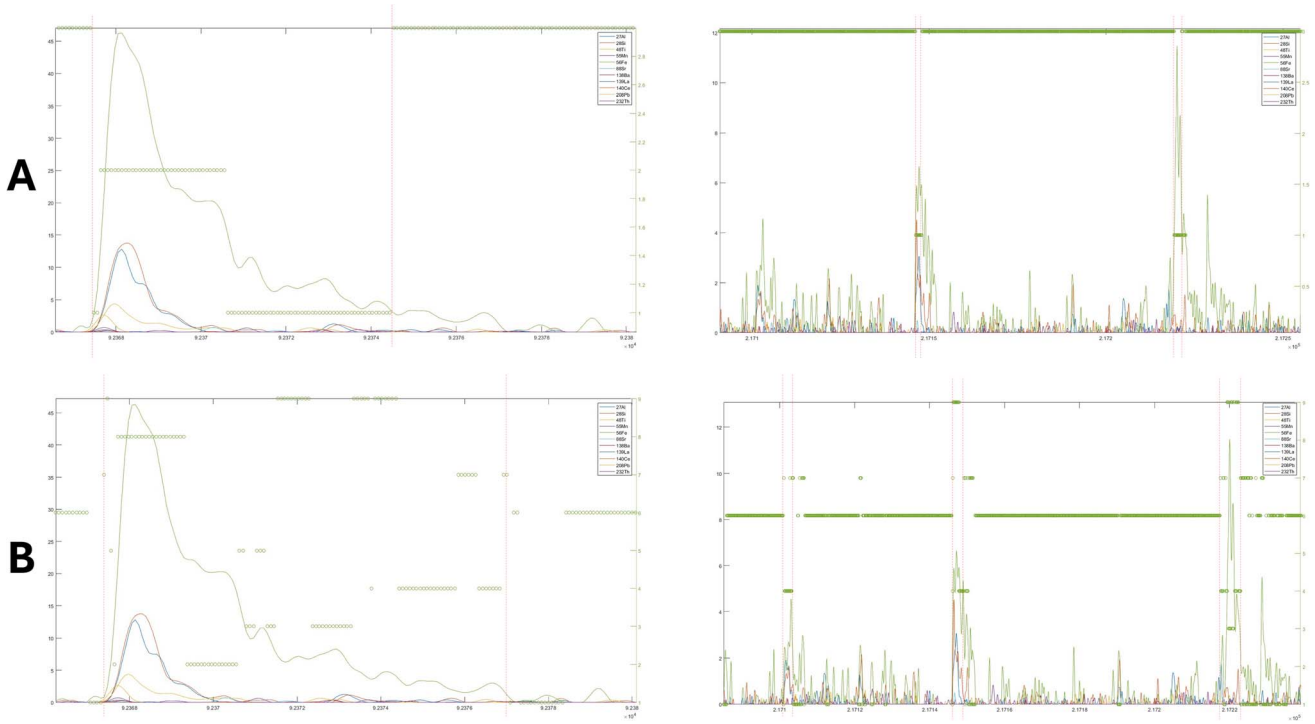

Fig. 3 Example of particle identification for various peak types for sample GS-1, using: (A) three (meta-clusters 1 and 2 denote time points corresponding to peaks), and; (B) nine meta-clusters (all meta-clusters but one and six denote time points corresponding to peaks). Dotted red dotted lines show where peak identification begins and ends; allocation of a single time point to a non-peak cluster signals the end of the peak. Peak identification with three meta-clusters was constrained to require at least seven consecutive time points (~560 μs) to identify a peak, whereas this parameter was set to 10 time points (~800 μs) for the nine meta-cluster solution to minimize the probability of false positives (*x*-axis: counts; *y*-axis: time in seconds). Corresponding SOM outputs are shown in Fig. 4.

A *k*-means meta-clustering was performed on the trained particle differentiation map to group particles according to similarities/differences in the proportions of various elements. Solutions with 2–15 meta-clusters were assessed by comparing the Cplanes, Umatrix, hit-histogram, and DBI. To minimize the risk of selecting a local maximum, 1000 clusterings were performed from random starting conditions for each of the 2–15 cluster solutions. Particle systems were compared by applying the SOM and *k*-means meta-clustering to a combined *in silico* mixture of all particles from the associated samples and comparing the proportions of particles within each sample that belonged to the resulting meta-clusters.

4. Results

4.1 Particle detection

The trained SOM readily differentiated large peaks from the background when the three meta-cluster solution (3 MCS) was applied; however, a higher number of clusters was required to detect smaller particles (Fig. 3). The 9 MCS identified 3341 particles compared to 1190 for the 3 MCS, and minimized early termination of peak detection as tails approached the background. On the other hand, the 9 MCS included several questionable time points with counts close to the background as corresponding to peaks. Peak identification with the 9 MCS was thus constrained to require at least 10 time points (~830 μs) compared to seven consecutive time points (~580 μs) for the 3 MCS, which eliminated apparent false positives.

The SOM outputs from particle detection indicated that increasing the number of meta-clusters from 2–9 MCS iteratively moved the threshold of peak detection to lower values of TotSig, SigLen, NumElts and TotSig × SigLen (Fig. 4). The highest values of all variables were in the bottom right-hand corner of the trained map, clearly corresponding with very different signal behaviour according to the Umatrix. A second maximum was evident in the bottom left-hand corner for SigLen and NumElts, with clear gradients towards the lowest values for all variables at the top of the map. The 10 MCS deviated from this behaviour by introducing a new cluster that extended close to the upper left-hand corner of the map and adding another gradient in the lower left-hand corner, greatly reducing the map area that had been associated with the background (Fig. S9). Visual inspection of time point membership in meta-clusters of the 10 MCS also indicated numerous false positive particle detections. Interestingly, both variable distributions and this multivariate thresholding effect were virtually identical for all six soil samples.

The DBI analysis was not conclusive, indicating that the optimal solution included 3–10 clusters on multiple iterations (Fig. S10). After combining these assessments with visual inspection, the 9 MCS was selected to compare NNPs from GS and VS using the particle differentiation SOM.

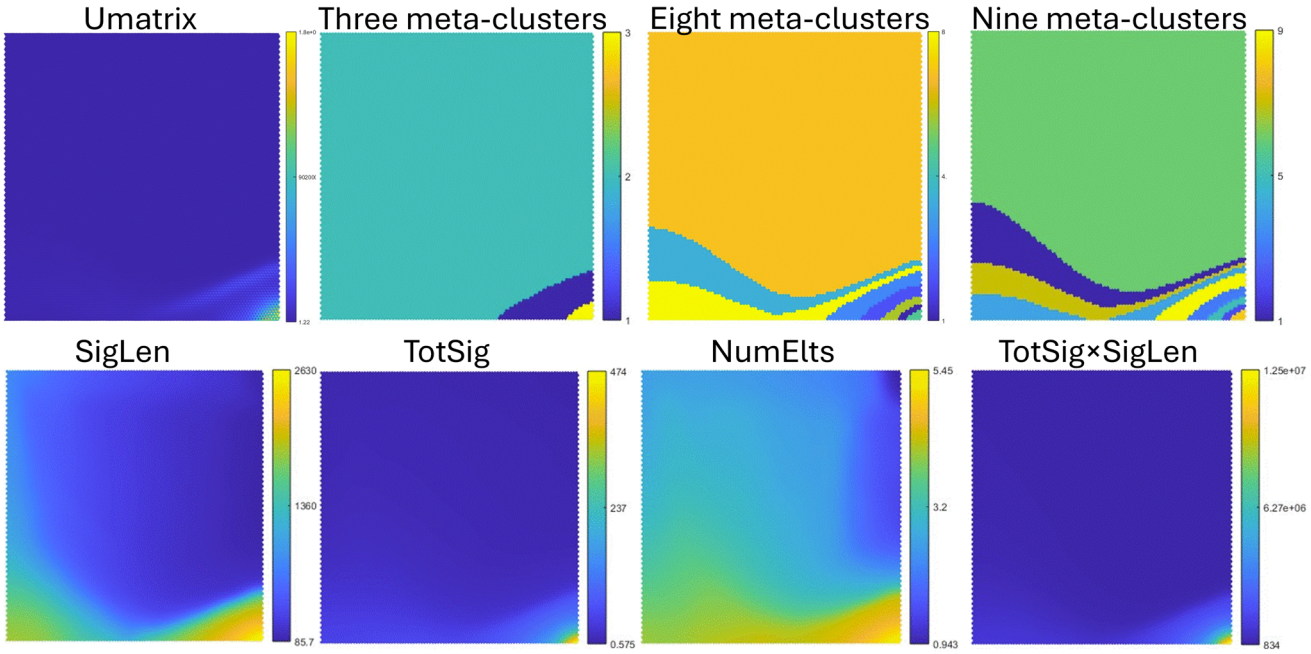

Fig. 4 Outputs from the particle detection SOM of sample GS-1. Meta-clusters one and three corresponded to particles in the 3 MCS, and all but meta-clusters one and six corresponded to particles in the 9 MCS.

The behaviour of meta-clusters and associated SOM outputs were similar for the soil samples and the FeCoNi/Zn ENPs, with the exception that the highest values and associated patterns were focused in the bottom left-hand corner of the map, and high values of NumElts and SigLen were more disperse (Fig. S5–S7). Hence, the 9 MCS was also used to identify particles in the FeCoNi/Zn ENP time-series.

4.2 Particle system characterization and differentiation

4.2.1 Tri-metallic ENP system. The composition of the FeCoNi/Zn ENP system was characterized prior to combining with the GS and VS samples, to assess the effectiveness of subsequent ENP detection and characterization within the mixture. The SOM outputs from the characterization of these particles with known composition exemplify the value of their corresponding visualization capabilities in detecting and distinguishing particles (Fig. 5). The HH contained numerous empty neurons, clearly dividing the trained map into five groups. When 1000 starting conditions were used to determine the optimal clustering for each of 2–15 meta-clusters, the outcome fluctuated and suggested anywhere between two and ten clusters as optimal. More extensive exploration of the solution space to find the global optima for 2–10 meta-clusters revealed that the DBI decreased monotonically with an increasing number of clusters with a sharp decrease between 2 and 3 clusters, suggesting that the 3 MCS was optimal (red line on DBI chart). The 3 MCS and 4 MCS had the same DBI, but the 3 MCS and 4 MCS also included "empty" clusters such that meta-cluster groupings and elemental compositions were identical for the 2 MCS, 3 MCS and 4 MCS. The 5 MCS included meta-clusters that did not align with the HH or Umatrix, despite being suggested as a viable solution. Despite apparently aligning with the distribution of particles on the HH, the 5 MCS also produced meta-clusters two and four with nearly the same composition of Fe/Co NPs, and meta-cluster five which was also primarily Fe/Co NPs with minor contributions from both Ni and Zn. The 2 MCS was therefore chosen as the optimal representation of the ENP system.

Noteworthy, some saturated peaks were identified in the ENP mixture (*e.g.* Fe in Fig. S8). Saturation biases the composition of particles, such that the particle shown in Fig. S8 was 42% Fe + 45% Co + 13% Ni in terms of counts but should be 67% Fe + 17% Co + 17% Ni or Zn according to the manufacturer. Interestingly, the BMU for this particle was neuron 260 in the bottom right corner of the HH (Fig. 5, highlighted with a green circle). This corresponded with the region of the map which had the greatest proportion of Co according to the Cplanes, and was in meta-cluster 1, which had a mean elemental composition of 73% Fe + 24% Co + 3% Ni. While the data are expressed in counts and therefore the clusters are not expected to represent elemental composition, this outcome does highlight the ability of SOM to correctly classify particles despite substantial "noise" and associated deviations from flawless particle measurements.

4.2.2 Differentiating NNPs. One of the 2720 identified particles in the ENP system included 10% Au. This suggests the inclusion of noise since Au was not present in the mixture; however, the signal was associated with a peak constituted by nine consecutive time-series points (747 μs) with >0 counts and a maximum of 25 counts, suggesting that it could also be a rare particle coincidence or case of two particles that were sorbed together, arising due to carryover from the size/mass calibration. Despite being clearly different from the other particles according to the HH and Umatrix, this particle was consistently

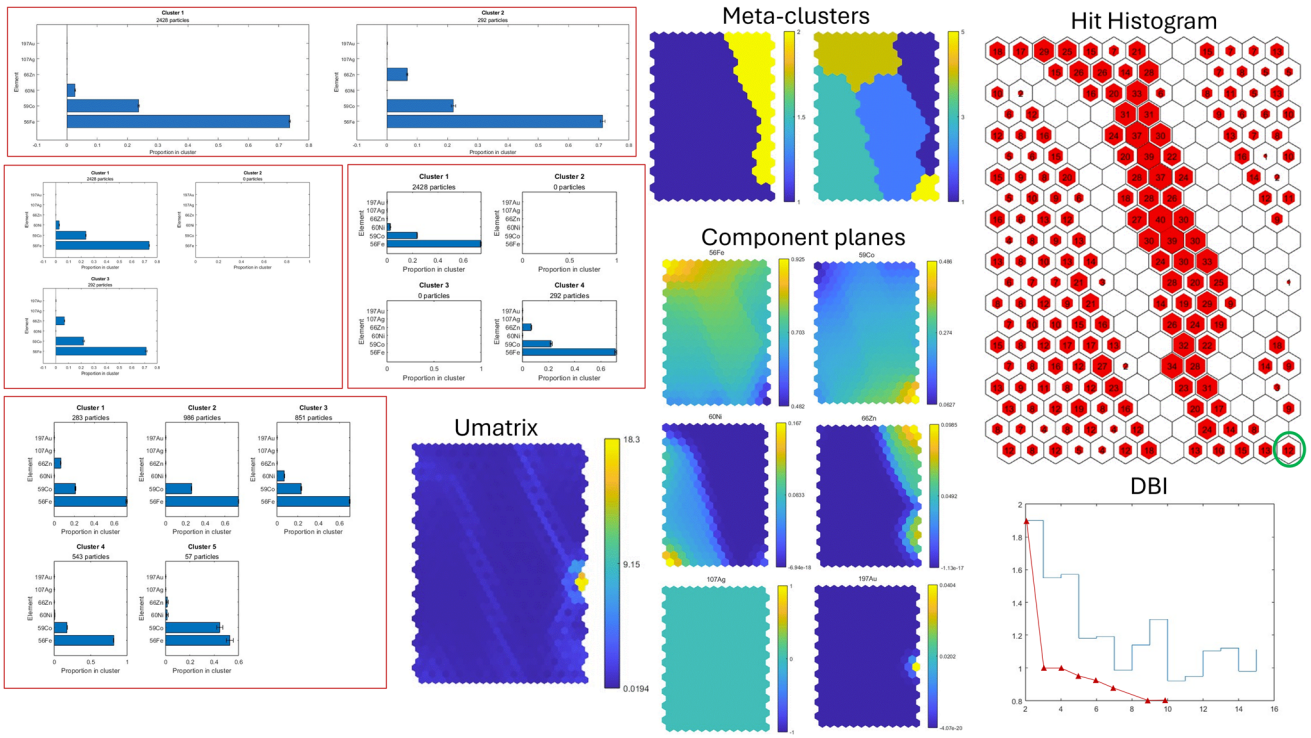

Fig. 5 Particle differentiation SOM outputs, DBI analysis results, and elemental composition of FeCoNi/Zn ENP meta-clusters based on 9 MCS particle detection som.

grouped with the meta-cluster that identified FeCoZn ENP. However, its negligible contribution to the 292 particles in that meta-cluster led the Au proportion to average out to near zero, illustrating the ability of SOM to both identify outliers and "smooth out" noise.

The DBI and SOM outputs for the comparison of triplicate GS and VS extracts suggested that the 8 MCS was optimal for grouping particles according to their elemental compositions (Fig. 1 and S11). The proportion of particles belonging to various clusters in the triplicates of GS and VS each varied substantially, so triplicates were combined to describe each NNPs as a single sample. Eighty-seven and 83% of particles respectively belonged to meta-cluster 5 (total of 20 124 particles) for GS and VS, with 6 and 8% (1714 particles) respectively belonging to meta-cluster 1, and 4 and 5% (1051 particles) respectively belonging to meta-cluster 4 (Fig. 6). While differences in the allocation and hence compositions of particles from GS and VS were limited to 1–4% for these three clusters, the difference of 4% in meta-cluster 5 alone represents more than 800 particles, illustrating the ability of SOM to distinguish NNPs from different sources.

Most particles were associated with meta-cluster 5, with an elemental composition dominated by Fe (~60%) and equal parts Al and Si comprising the remainder. On the other hand, Mn made up most of meta-cluster 1 (~65%), with minor contributions from Fe (10%) and several other elements. Meta-cluster 4 was composed of 80% Ti and 10% Fe, while the remaining three meta-clusters were comprised of several elements in varying proportions (total of 883 or 3.7% of particles).

Interestingly, meta-cluster 5 took up most of the space on the trained map and included the maximum proportions of major colloidal carriers Al, Si and Fe. The remaining seven meta-clusters were associated with the maximum proportions of other elements such as Mn, Ce, La, Ti, Pb and Th (Fig. 1 and 6). These latter seven clusters may contain NPs that are associated with one or many minerals or may be considered as "outlier clusters" when their contribution to the overall NNPs is very low. Indeed, the Umatrix for the *in silico* mixture of particles extracted from GS and VS also suggests that clusters 3 and 6–8 have highly different compositions respectively marked by large proportions of La, Pb, Ce and Th. Their inclusion and identification as highly different clusters may thus impair the effectiveness of the *k*-means algorithm in detecting finer differences within meta-cluster 5. This is particularly striking for the 2677 particles with very high proportions of Fe that are clearly isolated in the lower left corner of the HH, and thus are quite different from neighbouring BMUs. This is likely caused by a misalignment between the primary variables and corresponding axes created when these 11 elements were projected into two dimensions, and the fact that the elements with high proportions in these four clusters were largely absent from the other clusters. This suggests potential limits on the ability of the combined SOM and meta-clustering to deal effectively with sparse data.

4.2.3 Detecting ENPs in NNPs. Based on the DBI, HH, Umatrix, Cplanes and associated meta-clustering, the 6 MCS was determined as the optimal meta-clustering on the trained particle differentiation SOM for the *in silico* mixture of 2720

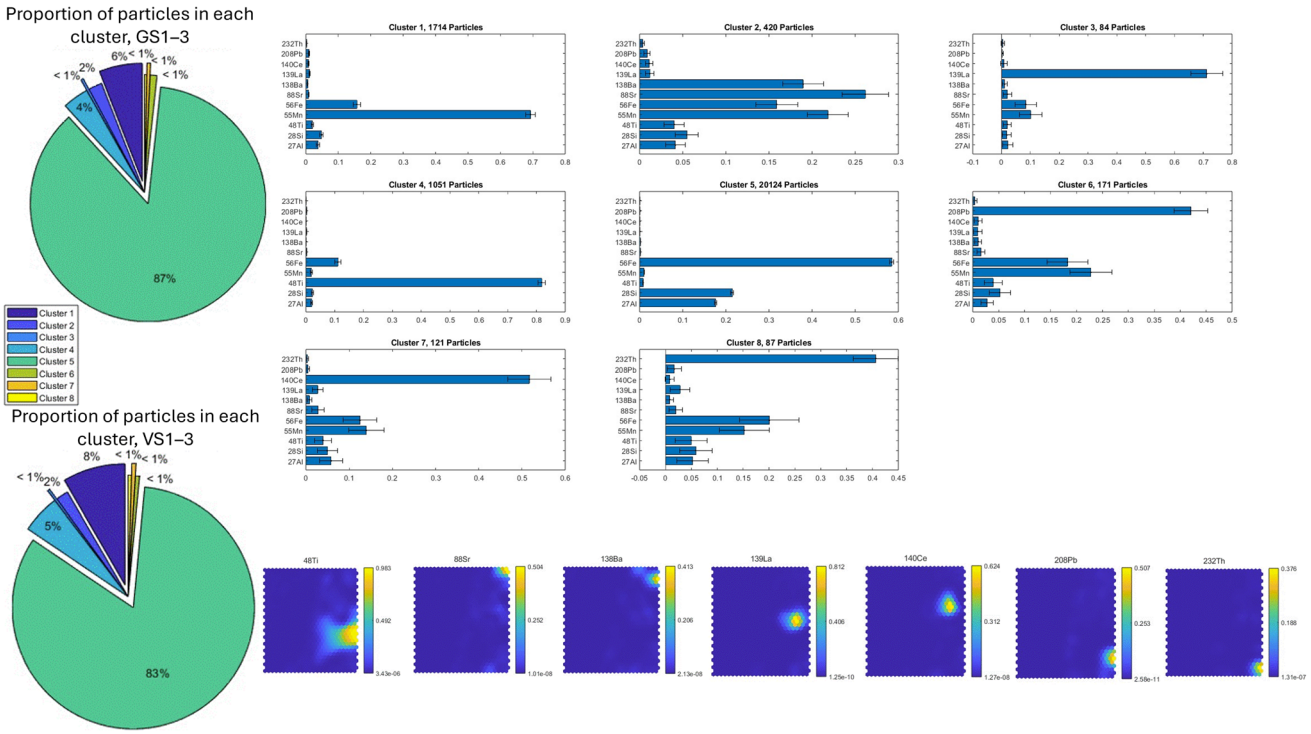

Fig. 6 NNPs differentiation for extracts of soils collected near the gambaiseuil and vaucoulers rivers using six meta-clusters. The corresponding meta-cluster distributions, hh, umatrix, and remaining cplanes (^{27}Al, ^{28}Si, ^{55}Mn and ^{56}Fe) are shown in Fig. 1.

FeCoNi/Zn ENP and 23 772 NNP from GS and VS (Fig. 7 and S12). Using this solution, 96% of FeCoNi/Zn NPs (2611 particles) were classified as belonging to meta-cluster 1 with a composition resembling the combination of meta-clusters 1 and 2 in the independent characterization of these ENP (Fig. 5). Four percent of the ENP were also assigned to meta-cluster 4

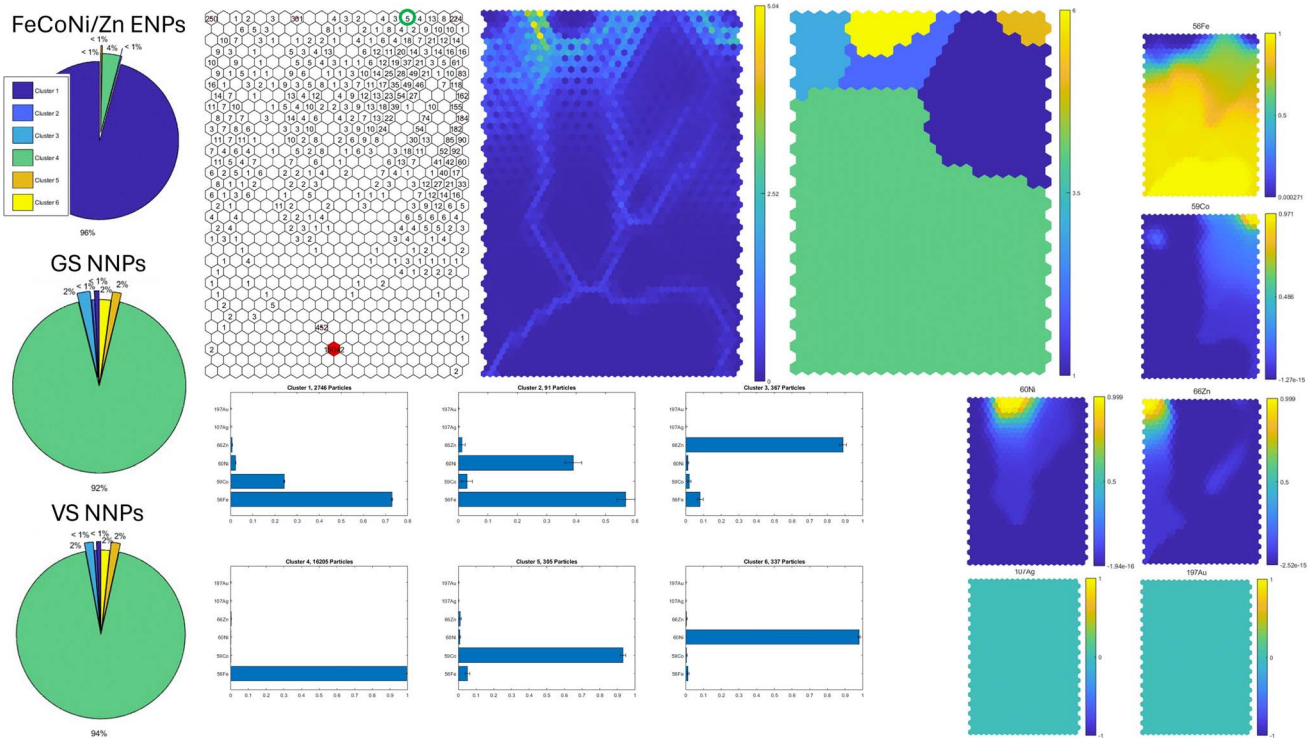

Fig. 7 Example of distinguishing multi-elemental ENPs from soil NNPs using six meta-clusters with corresponding Umatrix, Cplanes, HH, and cluster compositions.

6 MCS

		True classification	
		ENP	NNP
Predicted classification	ENP	2611 (96%)	135 (0.6%)
	NNP	109 (4%)	23,637 (99.4%)

10 MCS

		True classification	
		ENP	NNP
Predicted classification	ENP	2666 (98%)	510 (2.1%)
	NNP	54 (2%)	23,262 (97.9%)

Fig. 8 Confusion matrix for the classification of tri-metallic Fe–Co–Ni/Zn ENPs in NNPs extracted from soils using six and ten meta-clusters.

with negligible Co and Ni relative to Fe. While these 4% are incorrect classifications, it is expected that the mass of Co and Ni in FeCoNi/Zn ENP will decrease to levels below the detection limit as particle size decreases, in which case these multi-elemental particles will appear to contain only Fe. On the other hand, <1% of NNP (135 particles) were also associated with meta-cluster one. Overall, this represents 4% false negatives and 0.6% false positives in the unsupervised classification of ENPs within an NNPs containing approximately ten times more particles (Fig. 8).

Comparing only the Umatrix and HH, clear differences are evident between groups of particles that were assigned to meta-cluster 4 which could facilitate breaking this into a further 5 or more meta-clusters. The corresponding 10 MCS classified 95% of the ENP as belonging to the same meta-cluster as for the 6 MCS, 3% as belonging to a new meta-cluster with greater proportions of Fe and Zn but less Co and negligible Ni, and < 1% belonging to each of the remaining eight clusters (Fig. S13). Similarly, 2% of GS and VS were again classified as belonging to this second cluster, with <1% of each belonging to the ENP-dominated meta-cluster. Designating both of these clusters as ENP classifiers, the false negative rate for ENP detection is reduced to 2%, while the false positive rate increases to 2.1%. Hence, the 6 MCS is clearly superior to the 10 MCS as indicated by the DBI.

Interestingly, neuron 529 was the BMU for the ENP particle event with the saturated Fe peak, nearest to the edge of meta-cluster 1 in the 6 MCS (highlighted with a green circle in Fig. 7). This location correctly classifies the particle as an ENP, albeit in the region of meta-cluster 1 with the lowest ratio of Fe/Co, again highlighting the ability of the SOM to account for substantial noise.

5 Discussion

5.1 Benefits and limitations of SOM use with spICP-TOFMS

Results from this proof-of-concept study illustrate several benefits of SOM for detecting and differentiating particles measured using spICP-TOFMS, including.

5.1.1 Pattern recognition/grouping. The iterative SOM algorithm facilitates topology-preserving pattern recognition and classification which self-organize/adapt to the properties of the data. This adaptability is well suited to the variable data generated by spICP-TOFMS which may differ depending on the instrument and settings used, the size, number, and isotopic compositions of the particles measured, the type of aquatic system and its associated background, and both the number and identity of the masses which are used as input data. While the range of clustering algorithms may each perform well on data with particular properties, none share the adaptive capabilities of the SOM.

5.1.2 Visualization. Juxtaposing the Umatrix, Cplanes and HH on a two-dimensional map allows simultaneous visualization variable distributions, regions with high/low numbers of samples, and the degree of difference between map regions. This provides information about relationships between these variables and the overall organization of the data, including the nature and number of potential outliers. The SOM projection onto two dimensions has been likened to "pressing a flower", and as such may not reveal some higher order relationships; however, this can be beneficial for isolating and grouping according to major multivariate similarities/differences.

5.1.3 Cluster assessment. As evidenced herein, grouping particles into clusters may have complex optimization surfaces riddled with local maxima making indicators such as the DBI difficult to interpret (Fig. 5, S11 and S12). Visualizing "pressed" regions of similarity/difference and high/low sample density alongside proposed meta-clusters and variable distributions using the SOM greatly improved the evaluation of clustering solutions, limiting subjective assessment and preventing reliance upon ineffective indicators.

5.1.4 Versatility. As a combined projection and classification approach, SOM were herein applied to detect particles in time-series, to group/classify particles according to their similarities, and to distinguish between particle systems based on differences in the distribution of particles within various groups. Additional measurands may also be added to the dataset to explore relationships between a wide array of properties, cause–effect relationships, and environmental functioning.

5.1.5 Tunable parameters. Although default settings were used in this proof-of-concept study, there are a wide array of

parameters that can be tuned to optimize the performance of SOM to address different questions for various types of data and systems of interactions/relationships (Fig. 1).

5.1.6 Outlier detection and noise removal/smoothing. The SOM algorithm facilitates effective outlier identification and assessment (*e.g.* the Au-containing particle in Fig. 7). It is robust to outliers and noise due to first order smoothing associated with the use of BMUs instead of exact values to determine the association of particles with clusters (*i.e.* neurons), and the second order smoothing associated with meta-clustering.

5.1.7 Comparisons. Applying and reporting parameter settings for SOM facilitates both the comparison of results across studies and tests of reproducibility. This could be advanced through the creation of a database for SOM outputs/models and/or the associated raw spICP-TOFMS time-series, similar to OpenFluor for the analysis of fluorescence excitation–mission matrices of organic matter using parallel factor analysis.[63]

Several limitations of SOM also became apparent through this study, such as.

5.1.8 Computational demand. Even with the somewhat advanced computing capacity (8-core AMD Ryzen 7 7735HS microprocessor with a base clock of 3.20 GHz and boost up to 4.75 GHz, AMD Radeon 680M graphics card, and 64 GB of RAM), random initialization with sequential training was not feasible and even linear initialization with batch training of the particle detection SOM required >20 min. This approach does not facilitate exploration of the solution space to assess robustness of the overall pattern on the trained map. A reasonable search for the global optimum in meta-clustering also required several minutes. Thorough exploration of such complex optimization spaces with numerous local maxima requires iterations from multiple starting points, which would benefit from supercomputers, quantum computers, or more efficient algorithms.

5.1.9 Challenging to interpret. Although the many outputs of SOM are beneficial for visualizing complex data, substantial familiarity is required for straightforward and reliable interpretation. Several resources provide a gentle introduction to SOM;[45,61,64–66] however, deep understanding requires familiarity with underlying mathematics or substantial experience under various conditions.

5.1.10 Geochemical interpretation. The grouping of particles using SOM suffers from the same geochemical and mineralogical limitations as other clustering approaches when considering NNPs: the elemental composition of these groups is unlikely to correspond to a mineral or class of minerals since grouping arises in part from the extent of variation within a particular sample. This issue is widely encountered whenever treating large amounts of data from untargeted analyses of natural systems.[67–70] In the analysis of dissolved organic matter using high-resolution mass spectrometry, the problem of virtually infinite homologous series has been addressed through several approaches, including: classification and visual analysis based on units in homologous series (*i.e.* Kendrick mass defect spectra),[71,72] element ratios associated with various compound groups (*i.e.* Van Krevelen diagrams),[73–75] measures of diversity (*e.g.* chemodiversity),[76] and stoichiometric approaches.[77] Incorporating similar approaches to interpret and meaningfully organize particle information from spICP-TOFMS analyses may be helpful for characterizing and comparing NNPs and relating their properties to environmental functions.

The tendency of the SOM to identify data points corresponding to particles based on a multivariate threshold of transformed variables tends to manifest in part as a peak height threshold such that the tails of some peaks may be excluded, particularly when an inadequate number of meta-clusters are specified (*e.g.* compare the 3 and 9 MCS in Fig. 3 and S8, and the visualized transformed variables for the latter in Fig. S2). This excluded portion may comprise signal that extends too far into the background and so is unreliable, but it may also remove some of the dominant element in particles. Other particle-identification and modeling approaches relying upon a threshold to remove background noise encounter similar artefacts. Combining SOM for peak detection with other methods which model peak shape to optimize the extraction of information in regions close to the background may be helpful for overcoming this challenge.

5.2 Further considerations and next steps

Several aspects of the SOM can be further explored to improve their application to spICP-TOFMS data, including.

5.2.1 Data treatment, interpretation, and SOM optimization. The choice and weighting of 10 × SigLen, 20 × TotSig, 1 × NumElts and 10 × SigLen × TotSig as input variables for the particle detection SOM was based on extensive testing of intuitive peak indicators. Other variables, transformations, weightings, or combinations of variables may improve particle identification. These choices may also not be optimal for all analysis conditions, isotope subsets, or systems with different backgrounds. The effectiveness of a range of potential variables, weightings and combinations could be systematically explored for various conditions through (hyper)parameter optimization using the genetic algorithm or multi-layer artificial neural networks.[78] Once such an optimization algorithm is generated, it could be applied to automatically optimize these parameters for each new time series.

To demonstrate the effectiveness of SOM with minimal data treatment, raw counts were chosen to describe isotope contributions in this study. Transforming counts to mass involves multiplying by a different constant for each element, which will not change the grouping of particles on the trained particle differentiation SOM because each variable is normalized according to its variance; however, using element masses or mole percentages may provide more readily interpretable cluster compositions.

While the compositional representation of particles is beneficial for comparing composition, this simplification excludes information about particle mass and polydispersity. This information could be added to the compositional representation using a new variable such as total mass, or the mass of each isotope could be used as input. Furthermore, any number

of variables describing particle properties could be added to the SOM to provide a more integrated representation of the particle system. Similarly, the SOM can be applied to assess the likely nonlinear multivariate relationships between the properties and behaviour of NPs/NNPs with other variables by adding these variables as inputs (*e.g.* relationships between the fluorescence, size, source and age of DOM in [55]).

As noted above, it would be beneficial to explore the robustness of the overall pattern emerging from self-organization by randomly initializing the map and introducing sample vectors individually, rather than using linear initialization with batch introduction. Similarly, the combination of compositional data with a linear map structure necessarily groups particles according to the dominant isotope as the minimization of topographical error pushes the maximum for each variable to the edges of the map (Fig. 1, 4 and 5–7). Further exploration of other data formats and map structures may therefore lead to improved particle grouping. Testing various shapes and training rates for the neighbourhood function may also improve outcomes.

The *k*-means algorithm and DBI were chosen for meta-clustering because they are effective for pre-organized SOM neurons;[61] however, other meta-clustering algorithms may also be effective, such as minimum spanning trees,[79] cluster validity indices[80] and hierarchical agglomerative clustering (Vesanto and Alholniemi, 2000).[81] The DBI also was not a reliable measure for determining the optimum number of clusters determined by the *k*-means algorithm. While the other SOM outputs are helpful for informing this decision and other clustering algorithms may perform better, using or adding other clustering measures may also be helpful, such as the elbow method, silhouette indicator,[82] Caliński–Harabasz index[83] or jump method.[84] However, determining the optimum number of clusters under various conditions remains an open mathematical challenge with numerous options.[85–87]

5.2.2 Combinations and comparisons. Using SOM to organize particle systems according to prevalent differences may be used to assess NNPs dynamics by measuring the system before and after a sequence of treatments. Treatment variables may be used as inputs to the SOM to explore correlations with various particle compositions and properties within the NNPs, or may be overlain on the trained map to visualize treatment impacts on the measured particle properties.

Although not explored as part of the unsupervised and untargeted analysis approach demonstrated herein, SOM can also classify unknown particles and distinguish ENPs from NNPs by training it like a classical artificial neural network. Using this approach, the SOM is trained using test data and then classifies new particles by determining which BMU is most similar to the sample vector, and classifying it as belonging to the corresponding meta-cluster.

By recording and reporting the various parameter settings used to detect or differentiate particles in a study, SOM provide the requisite flexibility and reproducibility for adapting to various conditions and comparing results across studies. If it is not possible or desirable to include raw time-series data for direct comparison across studies, the values of variables in the set of neuron vectors (*i.e.* the codebook) from the trained SOM provide condensed metadata for comparisons. This is sufficient for comparing particle classification within various representations of NNPs across studies since the BMU and associated meta-cluster in a published study can be determined for any sample vector, so that the corresponding grouping of particles and particle systems may thus be compared. The reporting of the complete codebook also allows comparison with other data processing and clustering methods, since sample vectors from studies using different approaches can still be associated with BMUs and meta-clusters from published studies.

6. Conclusions

This report demonstrates the ability of SOM to detect and differentiate NPs and NNPs measured using spICP-TOFMS. The SOM offers substantial advantages compared to other methods due to adaptive cluster centers, excellent visualization, and flexibility for adapting to various backgrounds and analysis conditions. The coupling of SOM with other methods and combining particle data with other variables to improve understanding of the complex relationships between NPs, NNPs, and the variables governing their behaviour and impacts have been outlined as future research goals.

The SOM is deterministic for a given dataset when linear initialization and the same input parameters are used, and if the SOM outputs are reported in associated publications then they can be used by other researchers for direct comparison, and to make independent assessments about the validity of the meta-clustering. If the SOM input parameter settings are also reported, then they can be also tested/compared for other data. While meta-clustering requires user input to determine the number of clusters and so is prone to the same sort of subjective assessment as other clustering methods, the combination of SOM and meta-clustering is highly flexible/adaptable for different data rather than being a straitjacket that works very well in some instances but poorly in others. At the same time, the inclusion of proper information allows replication and assessments across research groups and samples. Although the process of implementing and interpreting SOM for single-particle data is initially complicated, it quickly becomes intuitive and routine. In the case that this rare combination of flexibility, replicability and comparability overshadows the shortcoming of initially challenging interpretation, then SOM may be ideally suited for widespread application to detect and classify nanoparticles and nanoparticle systems measured using spICP-TOFMS. Further testing of the SOM is needed across research groups and datasets to assess the feasibility of adopting it as a standard approach, which the authors seek to facilitate through this work.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

All data used for this manuscript is either provided in the SI, or available upon request (the data files are too large for sharing

them all as SI, and for making available through a host website without substantial cost).

Figures/tables, as well as a tutorial, algorithms, and test data for applying SOM to data generated by spICP-TOFMS analysis. See DOI: https://doi.org/10.1039/d5ja00179j.

Acknowledgements

The authors gratefully acknowledge support from the Université Paris Cité International Relations Office, which funded this collaboration through CWC's visit to the IPGP. CWC is also grateful for ongoing Discovery grant support from the Natural Sciences and Engineering Research Council of Canada (NSERC). MFB and MT gratefully acknowledge that parts of this work were supported by IPGP multidisciplinary program PARI, and by Paris–IdF region SESAME Grant no. 12015908 and the PIREN-Seine sub project C20/1541A01. This study contributes to the IdEx Université de Paris ANR-18-IDEX-0001.

References

1 J. F. McCarthy and J. M. Zachara, Subsurface transport of contaminants, *Environ. Sci. Technol.*, 1989, **23**, 496–502.
2 M. Hassellöv and F. von der Kammer, Iron oxides as geochemical nanovectors for metal transport in soil-river systems, *Elements*, 2008, **4**, 401–406.
3 M. F. Hochella, D. W. Mogk, J. Ranville, I. C. Allen, G. W. Luther, L. C. Marr, B. P. McGrail, M. Murayama, N. P. Qafoku, K. M. Rosso, N. Sahai, P. A. Schroeder, P. Vikesland, P. Westerhoff and Y. Yang, Natural, incidental, and engineered nanomaterials and their impacts on the Earth system, *Science*, 2019, **363**, 1414.
4 J. Buffle and G. G. Leppard, Characterization of aquatic colloids and macromolecules. 1. Structure and behavior of colloidal material, *Environ. Sci. Technol.*, 1995, **29**, 2169–2175.
5 Ö. Gustafsson and P. M. Gschwend, Aquatic colloids: concepts, definitions, and current challenges, *Limnol. Oceanogr.*, 1997, **42**, 519–528.
6 M. Filella, Colloidal properties of submicron particles in natural waters, in *Environmental Colloids and Particles: Behaviour, Separation and Characterization*, ed. J. R. Lead and K. J. Wilkonson, John Wiley & Sons Ltd, Chichester, 2007, p. 687.
7 J. F. McCarthy and L. D. McKay, Colloid transport in the subsurface: past, present, and future challenges, *Vadose Zone J.*, 2004, **3**, 326–337.
8 J. R. Lead and K. J. Wilkinson, Aquatic colloids and nanoparticles: current knowledge and future trends, *Environ. Chem.*, 2006, **3**, 159–171.
9 M. D. Montaño, J. W. Olesik, A. G. Barber, K. Challis and J. F. Ranville, Single particle ICP-MS: advances toward routine analysis of nanomaterials, *Anal. Bioanal. Chem.*, 2016, **408**, 5053–5074.
10 J. Vidmar, Detection and characterization of metal-based nanoparticles in environmental, biological and food samples by single particle inductively coupled plasma mass spectrometry, *Compr. Anal. Chem.*, 2021, **93**, 345–380.
11 A. Laycock, N. J. Clark, R. Clough, R. Smith and R. D. Handy, Determination of metallic nanoparticles in biological samples by single particle ICP-MS: a systematic review from sample collection to analysis, *Environ. Sci. Nano*, 2022, **9**, 420–453.
12 S. G. Bevers, C. Smith, S. Brown, N. Malone, D. H. Fairbrother, A. J. Goodman and J. F. Ranville, Improved methodology for the analysis of polydisperse engineered and natural colloids by single particle inductively coupled plasma mass spectrometry (spICP-MS), *Environ. Sci. Nano*, 2023, **10**, 3136–3148.
13 A. Azimzada, I. Jreije, M. Hadioui, P. Shaw, J. M. Farner and K. J. Wilkinson, Quantification and characterization of Ti-, Ce-, and Ag- nanoparticles in global surface waters and precipitation, *Environ. Sci. Technol.*, 2021, **55**, 9836–9844.
14 L. N. Rand, K. Flores, N. Sharma, J. Gardea-Torresdey and P. Westerhoff, Quantifying nanoparticle associated Ti, Ce, Au and Pd occurrence in 35 U.S. surface waters, *ACS EST Water*, 2021, **1**, 2242–2250.
15 M. D. Montaño, C. W. Cuss, H. M. Holliday, M. B. Javed, W. Shotyk, K. L. Sobocinski, T. Hofmann, F. von der Kammer and J. F. Ranville, Exploring nanogeochemical environments: new insights from single particle ICP-TOFMS and AF4-ICPMS, *ACS Earth Space Chem.*, 2022, **6**, 943–952.
16 M. Tharaud, L. Schlatt, P. Shaw and M. F. Benedetti, Nanoparticle identification using single particle ICP-TOFMS acquisition coupled to cluster analysis. From engineered to natural nanoparticles, *J. Anal. At. Spectrom.*, 2022, **37**, 2042–2052.
17 F. Wang, M. Tharaud and M. F. Benedetti, Advancing surface water preparation for nanoparticle quantification and characterization using spICP-MS or spICP-TOFMS, *Microchem. J.*, 2024, **203**, 110843.
18 J. Wielinski, X. Huang and G. V. Lowry, Characterizing the stoichiometry of individual metal sulfide and phosphate colloids in soils, sediments, and industrial processes by inductively coupled plasma time-of-flight mass spectrometry, *Environ. Sci. Technol.*, 2024, **58**, 12113–12122.
19 C. W. Cuss, E. Alasonati, M. F. Benedetti, C. Churchill, S. Fernando, R. Gasco, A. Goodman, C. Moens, M. D. Montaño, V. I. Slaveykova, M. Tharaud and I. A. M. Worms, Exploring environmental nanobiogeochemistry with field-flow fractionation and ICP-MS-based tools: Background and fundamentals, *Environ. Sci. Nano*, 2025, DOI: 10.1039/D5EN00095E.
20 I. A. M. Worms, M. Tharaud, R. Gasco, M. D. Montaño, A. Goodman, V. I. Slaveykova, M. F. Benedetti, C. Churchill, S. Fernando, E. Alasonati, C. Moens and C. W. Cuss, Exploring environmental nanobiogeochemistry with field-flow fractionation and ICP-MS-based tools: Progress and frontiers, *Environ. Sci. Nano*, 2025, DOI: 10.1039/D5EN00096C.

21 A. Gundlach-Graham, Multiplexed and multi-metal single-particle characterization with ICP-TOFMS, *Compr. Anal. Chem.*, 2021, **93**, 69–101.
22 F. Laborda, A. C. Gimenez-Ingalaturre and E. Bolea, Single-particle inductively coupled plasma mass spectrometry for the analysis of inorganic engineered nanoparticles: metrological and quality issues, *Compr. Anal. Chem.*, 2021, **93**, 35–67.
23 A. Gundlach-Graham, L. Hendriks, K. Mehrabi and D. Günther, Monte Carlo simulation of low-count signals in time-of-flight mass spectrometry and its application to single-particle detection, *Anal. Chem.*, 2018, **90**, 11847–11855.
24 T. R. Holbrook, D. Gallot-Duval and T. Reemtsma, Wagner. Machine learning: our future spotlight into single-particle ICP-TOFMS analysis, *J. Anal. At. Spectrom.*, 2021, **36**, 2684–2694.
25 O. Meili-Borovinskaya, F. Meier, R. Drexel, M. Baalousha, L. Flamigni, A. Hegetschweiler and T. Kraus, Analysis of complex particle mixtures by asymmetrical flow field-flow fractionation coupled to inductively coupled plasma time-of-flight mass spectrometry, *J. Chromatogr. A*, 2021, **1641**, 461981.
26 A. Gundlach-Graham and R. Lancaster, Mass-dependent critical value expressions for particle finding in single-particle ICP-TOFMS, *Anal. Chem.*, 2023, **95**, 5618–5626.
27 R. Gonzalez de Vega, T. E. Lockwood, L. Paton, L. Schlatt and D. Clases, Non-target analysis and characterisation of nanoparticles in spirits via single particle ICP-TOF-MS, *J. Anal. At. Spectrom.*, 2023, **38**, 2656–2663.
28 A. Praetorius, A. Gundlach-Graham, E. Goldberg, W. Fabienke, J. Navratilova, A. Gondikas, R. Kaegi, D. Günther, T. Hofmann and F. von der Kammer, Single-particle multi-element fingerprinting (spMEF) using inductively-coupled plasma time-of-flight mass spectrometry (ICP-TOFMS) to identify engineered nanoparticles against the elevated natural background in soils, *Environ. Sci. Nano*, 2017, **4**, 307–314.
29 R. L. Buckman and A. Gundlach-Graham, Machine learning analysis to classify nanoparticles from noisy spICP-TOFMS data, *J. Anal. At. Spectrom.*, 2023, **38**, 1244–1252.
30 G. Wang, H. Ruser, J. Schade, J. Passig, T. Adam, G. Dollinger and R. Zimmerman, Machine learning approaches for automatic classification of single-particle mass spectrometry data, *Atmos. Meas. Tech.*, 2024, **17**, 299–313.
31 M. Baalousha, J. Wang, M. Erfani and E. Goharian, Elemental fingerprints in natural nanomaterials determined using SP-ICP-TOF-MS and clustering analysis, *Sci. Total Environ.*, 2021, **792**, 148426.
32 C. C. Aggarwal, An introduction to cluster analysis, in *Data Clustering: Algorithms and Applications*, ed. C. C. Aggarwal and C. K. Reddy, Taylor & Francis Group, Boca Raton, Florida, 2014, p. 616.
33 H. Xiong and Z. Li, Clustering validation measures, in *Data Clustering: Algorithms and Applications*, ed. C. C. Aggarwal and C. K. Reddy, Taylor & Francis Group, Boca Raton, Florida, 2014, p. 616.
34 A. C. Benabdellah, A. Benghabrit and I. Bouhaddou, A survey of clustering algorithms for an industrial context, *Procedia Comput. Sci.*, 2019, **148**, 291–302.
35 P.-N. Tan, M. Stenbach, A. Karpatne and V. Kumar, *Introduction to Data Mining (2e)*, Pearson Education Inc, New York, New York, 2019, p. 839.
36 A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke and A. A. Akinyelu, A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future researcg prospects, *Eng. Appl. Artif. Intell.*, 2022, **110**, 104743.
37 C. X. Gao, D. Dwyer, Ye Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, S. Wood and S. M. Cotton, An overview of clustering methods with guidelines for application in mental health research, *Psychiatry Res.*, 2023, **327**, 115265.
38 K. Khan, S. Ur Rehman, K. Aziz, S. Fong and S. Sarasvady, DBSCAN: Past, present and future, *Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, Bangalore, India, 2014, pp. 232–238, DOI: **10.1109/ICADIWT.2014.6814687**.
39 A. A. Wani, Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions, *PeerJ Comput. Sci.*, 2024, DOI: **10.7717/peerj-cs.2286**.
40 T. Kohonen, *Self-Organizing Maps (3e)*, Springer-Verlag, 2001, p. 501.
41 P. Mangiameli, S. K. Chen and D. West, A comparison of SOM neural network and hierarchical clustering methods, *Eur. J. Oper. Res.*, 1996, **93**, 402–417.
42 Y. Liu, R. H. Weisberg and C. N. K. Mooers, Performance evaluation of the self-organizing map for feature extraction, *J. Geophys. Res.*, 2006, **111**, DOI: **10.1029/2005JC003117**.
43 G. Žibret and R. Šajn, Hunting for Geochemical Associations of Elements: Factor Analysis and Self-Organising Maps, *Math. Geosci.*, 2010, **42**, 681–703.
44 J. A. Cortés and J. L. Palma, Geological applications of self-organizing maps to multidimensional compositional data, *Pioneer J. Adv. Appl. Math.*, 2013, **7**(2), 17–49.
45 C. W. Cuss and C. Guéguen, Analysis of dissolved organic matter fluorescence using self-organizing maps: mini-review and tutorial, *Anal. Methods*, 2016, **8**, 716–725.
46 J. Laaksonen, M. Koskela and E. Oja, PicSOM: self-organizing maps for content-based image retrieval, *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, Washington, DC, USA, 1999, pp. 2470–2473, vol. 4, DOI: **10.1109/IJCNN.1999.833459**.
47 A. J. Richardson, C. Risien and F. A. Shillington, Using self-organizing maps to identify patterns in satellite imagery, *Prog. Oceanogr.*, 2003, **59**, 223–239.
48 X. Tuo, C. Bo, M. Keliang and L. Zhe, Neural network-based matrix effect correction in EDXRF analysis, *Nucl. Sci. Tech.*, 2008, **19**, 278–281.

49 A. Gacek, Preprocessing and analysis of ECG signals – a self-organizing maps approach, *Expert Syst. Appl.*, 2011, **38**, 9008–9013.
50 M. Foroutan and J. R. Zimbelman, Semi-automatic mapping of linear-trending bedforms using 'Self-Organizing Maps' algorithm, *Geomorphology*, 2017, **293**, 156–166.
51 J. Cho, A. R. C. Paiva, S. P. Kim, J. C. Sanchez and J. C. Príncipe, Self-organizing maps with dynamic learning for signal reconstruction, *Neural Netw.*, 2007, **20**, 274–284.
52 C. Guéguen, C. W. Cuss and S. Cho, Snowpack deposition of trace elements in the Athabasca oil sands region, Canada, *Chemosphere*, 2016, **153**, 447–454.
53 S. Licen, A. Astel and S. Tsakovski, Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review, *Sci. Total Environ.*, 2023, **878**, 163084.
54 A. Astel, S. Tsakovski, P. Barbieri and V. Simeonov, Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets, *Wat. Res.*, 2007, **41**, 4566–4578.
55 C. W. Cuss and C. Guéguen, Relationships between molecular weight and fluorescence properties for size-fractionated dissolved organic matter from fresh and aged sources, *Wat. Res.*, 2015, **68**, 487–497.
56 M. Bieroza, A. Baker and J. Bridgeman, Exploratory analysis of excitation-emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works, *J. Geophys. Res.*, 2009, **114**, DOI: **10.1029/2009JG000940**.
57 C. W. Cuss, M. W. Donner, T. Noernberg, R. Pelletier and W. Shotyk, EEM-PARAFAC-SOM for assessing variation in the quality of dissolved organic matter: simultaneous detection of differences by source and season, *Environ. Chem.*, 2019, DOI: **10.1071/EN19016**.
58 S. C. Löhr, M. Grigorescu, J. H. Hodgkinson, M. E. Cox and S. J. Fraser, Iron occurrence in soils and sediments of a coastal catchment: a multivariate approach using self organising maps, *Geoderma*, 2010, **156**, 253–256.
59 A. Bigdeli, A. Maghsoudi and R. Ghezelbash, Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran, *J. Geochem. Explor.*, 2022, **233**, 106923.
60 J. L. Wang, E. Alasonati, M. Tharaud, A. Gelabert, P. Fisicaro and M. F. Benedetti, Flow and fate of silver nanoparticles in small French catchments under different land-uses: The first one-year study, *Wat. Res.*, 2020, **176**, 115722.
61 J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, *SOM Toolbox for Matlab 5*, SOM Toolbox Team, Helsinki University of Technology, 2000, available at: **www.cis.hut.fi/projects/somtoolbox/**.
62 D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE T. Pattern Anal.*, 1979, **PAMI-1**(2), 224–227, DOI: **10.1109/TPAMI.1979.4766909**.
63 K. R. Murphy, C. A. Stedmon, P. Wenig and R. Bro, OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment, *Anal. Methods*, 2014, **6**, 658–661.
64 H. Yin, The self-organizing maps: background, theories, extensions and applications, *Stud. Comput. Intell.*, 2008, **115**, 715–762.
65 R. Brereton, Self-organizing maps for visualising and modeling, *Chem. Cent. J.*, 2021, **6**, S1.
66 M. M. Van Hulle, Self-organizing maps, in *Handbook of Natural Computing*, ed. G. Rozenberg, T. Bäck and J. N. Kok, Springer-Verlag Berlin Heidelberg, 2012, vol. 4, p. 2051.
67 N. Hertkorn, C. Ruecker, M. Meringer, R. Gugisch, M. Frommberger, E. M. Perdue, M. Witt and Ph Schmitt-Kopplin, High-precision frequency measurements: indispensable tools at the core of the molecular-level analysis of complex systems, *Anal. Bioanal. Chem.*, 2007, **389**, 1311–1327.
68 N. Hertkorn, M. Frommberger, M. Witt, B. P. Koch, Ph Schmitt-Kopplin and E. M. Perdue, Natural organic matter and the event horizon of mass spectrometry, *Anal. Chem.*, 2008, **80**, 8908–8919.
69 R. G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak and R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, *Anal. Bioanal. Chem.*, 2017, **409**, 5891–5899.
70 R. G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak and R. Tauler, Chemometrics in analytical chemistry—part II: modeling, validation, and applications, *Anal. Bioanal. Chem.*, 2018, **410**, 6691–6704.
71 E. Kendrick, A mass scale based on CH_2=14.0000 for high resolution mass spectrometry of organic compounds, *Anal. Chem.*, 1963, **35**, 2146–2154.
72 C. A. Hughey, C. L. Hendrickson, R. P. Rodgers, A. G. Marshall and K. Qian, Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra, *Anal. Chem.*, 2001, **73**, 4676–4681.
73 K. Van, Graphical-statistical method for the study of structure and reaction processes of coal, *Fuel*, 1950, **29**, 269–284.
74 S. A. Visser, Application of Van Krevelen's graphical-statistical method for the study of aquatic humic material, *Environ. Sci. Technol.*, 1983, **17**, 412–417.
75 J. R. Laszakovits and A. A. MacKay, Data-based chemical class regions for Van Krevelen diagrams, *J. Am. Soc. Mass Spectr.*, 2022, **33**, 198–202.
76 A. M. Kellerman, T. Dittmar, D. N. Kothawala and L. J. Tranvik, Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology, *Nat. Commun.*, 2014, **5**, 3804, DOI: **10.1038/ncomms4804**.
77 A. Rivas-Ubach, Y. Liu, T. S. Bianchi, N. Tolić, C. Jansson and L. Paša-Tolić, Moving beyond the van Krevelen diagram: a new stoichiometric approach for compound classification in organisms, *Anal. Chem.*, 2018, **90**, 6152–6160.

78 A. M. Vincent and P. Jidesh, An improved hyperparameter optimization framework for AutoML systems using evolutionary algorithms, *Sci. Rep.*, 2023, **13**, 4737.
79 R. Mayer and A. Rauber, Visualising Clusters in Self-Organising Maps with Minimum Spanning Trees, *ICANN'10: Proceedings of the 20th International Conference on Artificial Neural Networks: Part II*, 2010, pp. 426–431.
80 S. Wu and T. W. S. Chow, Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density, *Pattern Recogn.*, 2004, **37**, 175–188.
81 J. Vesanto and E. Alhoniemi, Clustering of the self-organizing map, *IEE T. Neural Networ.*, 2000, **11**, 586–600.
82 P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, 1987, **20**, 53–65.
83 T. Caliński and J. Harabasz, A dendrite method for cluster analysis, *Commun. Stat. A-Theor.*, 1974, **3**, 1–27.
84 C. A. Sugar and G. M. James, Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach, *J. Am. Stat. Assoc.*, 2003, **98**, 750–763.
85 T. M. Kodinariya and P. R. Makwana, Review on determining number of cluster in *k-means* clustering, *IJARCSMS.*, 2013, **1**, 90–95.
86 S. Xu, X. Qiao, L. Zhu, Y. Zhang, C. Xue and L. Li, Reviews on determining the number of clusters, *Appl. Math. Inf. Sci.*, 2016, **10**, 1493–1512.
87 A. Karanikola, C. M. Liapis and S. Kotsiantis, Investigating cluster validation metrics for optimal number of clusters determination, *Smart Innov. Syst. Tec.*, 2021, **15**, 809–824.