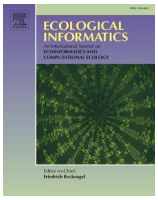
Ecological Informatics 84 (2024) 102891

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Automated identification of soil functional components based on NanoSIMS data

Yahan Hu [a,*], Johann Maximilian Zollner [b], Carmen Höschen [a], Martin Werner [b], Steffen A. Schweizer [a]

[a] TUM School of Life Sciences, Technical University of Munich, Emil-Ramann- Straße 2, 85354 Freising-Weihenstephan, Germany
[b] TUM School of Engineering and Design, Technical University of Munich, Lise-Meitner-Straße 9, 85521 Ottobrunn, Germany

ARTICLE INFO

Keywords:
NanoSIMS
Soil spatial arrangement
Pre-processing tool
Unsupervised segmentation
Organo-mineral interactions

ABSTRACT

NanoSIMS technique allows to investigate the micro-spatial organization in complex structures in multiple scientific fields such as material science, cosmochemistry, and biogeochemistry. In soil biogeochemistry applications, NanoSIMS-based approaches aim to disentangle the interactions of organic matter (OM) and mineral phases in the heterogeneous soil microstructure. Investigating the spatial arrangement of distinct organic and mineral functional components is necessary to understand how these components interact and contribute to biogeochemical processes in soil systems. Identifying soil functional components within NanoSIMS measurements necessitates advanced and efficient data processing tools capable of accessibility and automation. We have developed a pre-processing tool to streamline NanoSIMS data preparation and handling. The tool is provided as an open-source software toolbox (NanoT). In addition, a two-step unsupervised segmentation method was developed to identify soil functional components based on NanoSIMS analyses. To illustrate the segmentation method, here we describe its application to two exemplary NanoSIMS measurements. This allows to distinguish mineral- and OM-dominated regions, as well as different mineral phases. To improve the detection of iron oxides and aluminosilicates, the $^{56}Fe^{16}O^-$ channel was separately processed. The presented NanoSIMS-based processing workflow helps to disentangle functional components within a biogeochemically-diverse microstructure in soils and further warrants applications to a wide range of complex environmental samples.

1. Introduction

Nanoscale secondary ion mass spectrometry (NanoSIMS) is an imaging technique providing the spatial distribution of elements and isotopes at a submicron spatial resolution (~50–150 nm) of environmental, cosmic, and synthetic solid samples. In cosmochemistry, NanoSIMS offered information on the composition of presolar grains and lunar samples, and further on the deduction of the origin of presolar grains by spatial characterization of deuterium and oxygen isotopes (Barnes et al., 2013; Hoppe et al., 2013; McKeegan et al., 2006). NanoSIMS in material science was used to measure the composition of microstructural features in alloys, semiconductor devices or other samples (Li et al., 2020; Pedrazzini et al., 2018). NanoSIMS and isotopic techniques enable the visualization of environmental microorganisms involved in biological processes such as nitrogen fixation or the microbial transformation of specific organic compounds in the soil, providing insights into nutrient exchange within the plant root-microbe-soil interaction. (Brunet et al., 2022; Nuñez et al., 2018; Pett-Ridge and Weber, 2022). In soil science, NanoSIMS is applied to resolve the spatial heterogeneity of soil microstructures and investigate organic matter dynamics, mineral composition and their interactions by analyzing the spatial arrangement of diverse functional components at a high resolution (Mueller et al., 2023). These functional components include plant-derived and microbial-derived organic matter (OM) and mineral particles like silicates, oxides, carbonates, sulfides, and other compositions that interact, providing an intricate soil structure with diverse biogeochemical composition at the microscales (Kleber et al., 2021; Solomon et al., 2012; Wan et al., 2007). The allocation of different functional soil components, i.e., clay minerals, pedogenic oxides, or OM, locally provides biogeochemical interfaces that influence the function of soils in the ecosystem (Heckman et al., 2018; Rasmussen et al., 2018). Using secondary ion signatures and stable isotope labels, such as ^{13}C, ^{15}N, ^{29}Si, or ^{57}Fe, several groups have been able to identify the spatial arrangements of soil OM and minerals and to better understand biogeochemical cycles

* Corresponding author.
E-mail address: yahan.hu@tum.de (Y. Hu).

https://doi.org/10.1016/j.ecoinf.2024.102891
Received 17 May 2024; Received in revised form 4 October 2024; Accepted 10 November 2024
Available online 12 November 2024

at the microscale (Amelung et al., 2023; Hao et al., 2020; Hoppe et al., 2013; Remusat et al., 2012). Analyzing soil structures at the microscale has been requested by several authors to obtain further information on the heterogeneous composition of morphological, mechanical, and biological properties in order to improve our understanding of soil structure changes, microbial activities and other soil processes (Baveye et al., 2018, 2019; Pot et al., 2022). In addition to studying soil microstructures by NanoSIMS, further techniques with lower resolution such as computed tomography (Houston et al., 2017; Portell et al., 2018) and energy-dispersive X-ray spectroscopy (Allegretta et al., 2022; Hapca et al., 2015), or at finer resolution such as scanning transmission X-ray microscopy (Lehmann et al., 2008) and electron energy loss spectroscopy (Possinger et al., 2020), complement our understanding of biogeochemical processes in heterogeneous soil structures.

With increasing applications and the generation of larger datasets by NanoSIMS, challenges of data processing and data-driven identification of functional components became more critical. Various spatial analysis software tools were specially designed for NanoSIMS data processing, such as the open-source and multiplatform software tools OpenMIMS (Gormanns et al., 2012) on Fiji ImageJ, Look@NanoSIMS (Polerecky et al., 2012) on the MATLAB platform, as well as L'Image (L. R. Nittler, Carnegie Institution of Washington) and WinImage (CAMECA), which are both proprietary software. Most of the existing software tools lack further automation due to the manual operation. With the increasing amount of NanoSIMS data and image-based interpretations, fulfilling the 'FAIR' principles, the findability, accessibility, interoperability, and reusability of the data, becomes more crucial (Wilkinson et al., 2016).

The application of data-based image processing methods can help to distinguish the functional components from the microstructure images based on image datasets under different research aims, for instance, threshold models, supervised pixel classification and unsupervised segmentation methods. Threshold methods have been applied for the analysis of soil structure (Baveye et al., 2010; Hapca et al., 2013; Houston et al., 2013). For NanoSIMS measurements, an improved local Otsu's thresholding method identified soil organic particles on $^{12}C^-$, $^{13}C^-$ and $^{56}Fe^{16}O^-$ ion channels (Hao et al., 2020). Another threshold method, called Canny edge detection, generates a binarization mask and was described as a morphological segmentation (Renslow et al., 2016). Since soil components differ in multiple elemental compositions, threshold methods can be limited in identifying more than two components in soil microstructures.

To disentangle the microstructure of soil, an increasing number of machine-learning-based approaches provide novel opportunities. These approaches encompass both supervised and unsupervised learning, with supervised classification being predominantly applied in several studies. Building on this trend, Steffens et al. (2017) adapted methods from remote sensing to investigate functional components in soil, applying a pixel classification method based on the spatial-spectral endmember extraction and the spectral angle mapper algorithm. However, for further applications using spectral endmember extraction, the amount of different functional components and the required amount of reference materials can be quite large and difficult to define due to the spatially heterogeneous transformation. Other approaches of pixel classification based on a machine-learning algorithm used a segmentation toolkit called ilastik (Berg et al., 2019). This provides the possibility to segment background, mineral- and OM-dominated regions which was used to quantify microspatial patterns related to OM dynamics over time or quantify the co-location of OM with pedogenic oxides (Inagaki et al., 2020, 2023; Schweizer et al., 2018; Wilhelm et al., 2022). This approach enabled the identification of similar functional components from the same dataset that are partially needed for training the classifiers. However, such an approach provides a limited applicability to detect similar functional components in other datasets. This impedes the possibility to process larger datasets and derive a general understanding across a diverse range of soil samples.

Automated and unsupervised learning method like K-Means clustering was applied to NanoSIMS measurements for P distribution and their co-occurring with Al and Fe (Werner et al., 2017). K-Means clustering is optimized for identifying groups with convex shapes where no interior angle exceeds 180 degree. It may produce different results when applied to data containing reflex interior angles, known as non-convex polygons. In the context of soil microstructure images, this means that some functional components may be assigned into incorrect clusters when interior angles are present (Mitra et al., 2003). Beyond NanoSIMS data processing, unsupervised segmentation methods were applied to tomography measurements of soil structure to analyze pore morphology (Chauhan et al., 2016; Malik et al., 2022). In addition, deep-learning methods have been applied in digital soil mapping to analyze plot-scale gradients of soil properties and provide further possibilities to enhance NanoSIMS data analysis (Wadoux et al., 2020).

Although previous methodologies have proven valuable for NanoSIMS data pre-processing and image analysis, further automation, accessibility of data processing tools, and implementation of data-driven approaches are warranted. Thus, our study aims to provide a NanoSIMS processing toolbox with automated pre-processing and data-driven segmentation to streamline the identification of functional components in soil microstructures (Fig. 1), such as OM and mineral phases. A better understanding of the soil microstructure promotes further insights of organic matter dynamics and other soil processes.

2. Pre-processing methodology

2.1. NanoSIMS measurements of soil science case study data

The NanoSIMS 50 L (CAMECA) is equipped with one positive and one negative primary ion source, namely Cs^+ and O^- primary ion beams, and up to 7 secondary ions can be measured simultaneously (Herrmann et al., 2007). By bombarding the surface of a solid sample by primary ions at an energy of 16 keV, most elements are detected as secondary ions measured by the mass spectrometer unit (Nuñez et al., 2018; Wilson, 1995).

To illustrate the steps of our proposed image processing scheme, two NanoSIMS measurements done on soil particles from Ap (0–10 cm) horizon of Podzol in Lohne were chosen as a case study, whose organic carbon was around 22.6 $mg \cdot g^{-1}$. Detailed information of the soil can be found in Urbanski et al. (2022). Prior to the NanoSIMS measurement, particles of a soil fine fraction <20 μm were suspended in deionized water, deposited on GaAs wafers, and were let to dry in a desiccator. Regions of interest for subsequent NanoSIMS measurements were identified using Scanning Electron Microscope images to locate areas on the wafer containing organic and mineral components. Two NanoSIMS measurements were conducted using the Cs^+ primary source and to measure seven ion channels, namely $^{16}O^-$, $^{12}C^{12}C^-$, $^{12}C^{14}N^-$, $^{28}Si^-$, $^{31}P^-$, $^{27}Al^{16}O^-$, and $^{56}Fe^{16}O^-$, chosen to detect the organic matter and the minerals. These two measurements utilized 256 × 256 pixels arranged in rows and columns for a field of view of 30 μm × 30 μm, with a dwell time of 1 ms per pixel and involved 30 repeated scans.

2.2. Fundamentals of the preprocessing

The raw NanoSIMS data have the shape of N × N × M × S, where N represents the number of columns and rows, M denotes the number of ion channels, and S is the number of scans for each ion channel. NanoSIMS data are essentially a two-dimensional raster, wherein each raster cell gives the abundance of ions across the analyzed area. Usually, multiple scans are conducted for individual measurements to enhance counting statistics. This results in S slices for each ion channel. The raw NanoSIMS data are typically acquired in the Indexed Mesh (IM) format. By default, IM files are accompanied by an additional information file called CHK_IM, which contains metadata.

In the following, the pre-processing procedure for NanoSIMS data is presented. The raw data processing was split into seven sequential steps:

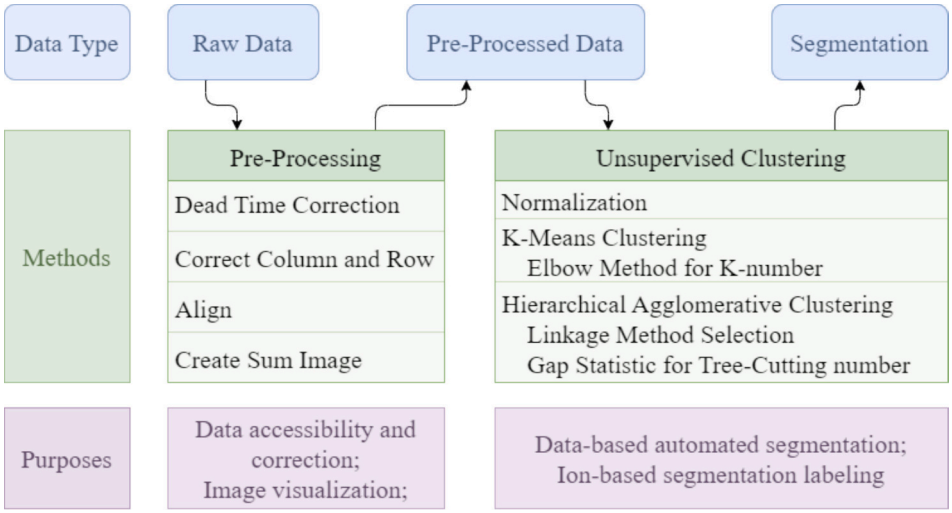


Fig. 1. NanoSIMS data processing scheme. The blue boxes represent the data types, the green blocks list the data processing methods, and the purposes of our method are in the purple blocks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

import, dead-time correction, raster correction, alignment, sum image creation, and export a simplified visualization of the pre-processing scheme is presented in Fig. 2.

In the first step of our procedure, the raw data were imported. Metadata on the measurement and various technical aspects, including the motor positions and NanoSIMS calibration, were extracted from the header of the IM files. The second step corrected the NanoSIMS measurements for the dead-time, which is the time frame after each detection of an ion signal where the system cannot record another event. Although the dead-time may vary between 20 and 84 ns (Nuñez et al., 2018), the standard practice is to use 44 ns. Nevertheless, the exact dead-time can be extracted from the header of an IM file. It is possible to account for the dead-time of each detector separately, if further accuracy is needed. The dead-time was corrected using a linear correction model:

$$C_{cor} = \frac{C_{raw}}{(1 - \tau \times C_{raw})}$$

where C_{raw} is the raw counts of secondary ions, C_{cor} is the corrected counts, and τ is the detector's dead time in ns (Nuñez et al., 2018). Third, the raster for each scan was corrected. In case the first column and row are incorrectly designated as the last column and row, they are re-assigned. Note that the column is corrected first followed by the row, and all scans must be processed in the same way. Next, the scans were aligned, which is also known as image registration, since the NanoSIMS ion beam might drift between scans, due to environmental factors such as temperature fluctuations and vibrations during the measurement. The alignment enables to increase the sharpness of imaged features when summing up scans subsequently. The algorithm implemented here by Thevenaz et al. (1998), is widely applied for example in medicine (Badimon et al., 2020) and biology (Haruwaka et al., 2019), and also used in the software tool OpenMIMS (Gormanns et al., 2012). Here, one mass with distinct boundaries across the whole field of view, such as oxygen, is used as a reference and then the resulting transformation is applied to all other ion channels and slices. Following, the slices can be accumulated by summing up the aligned scans of each ion channel. Finally, the data had the shape N × N × M and was exported in a file

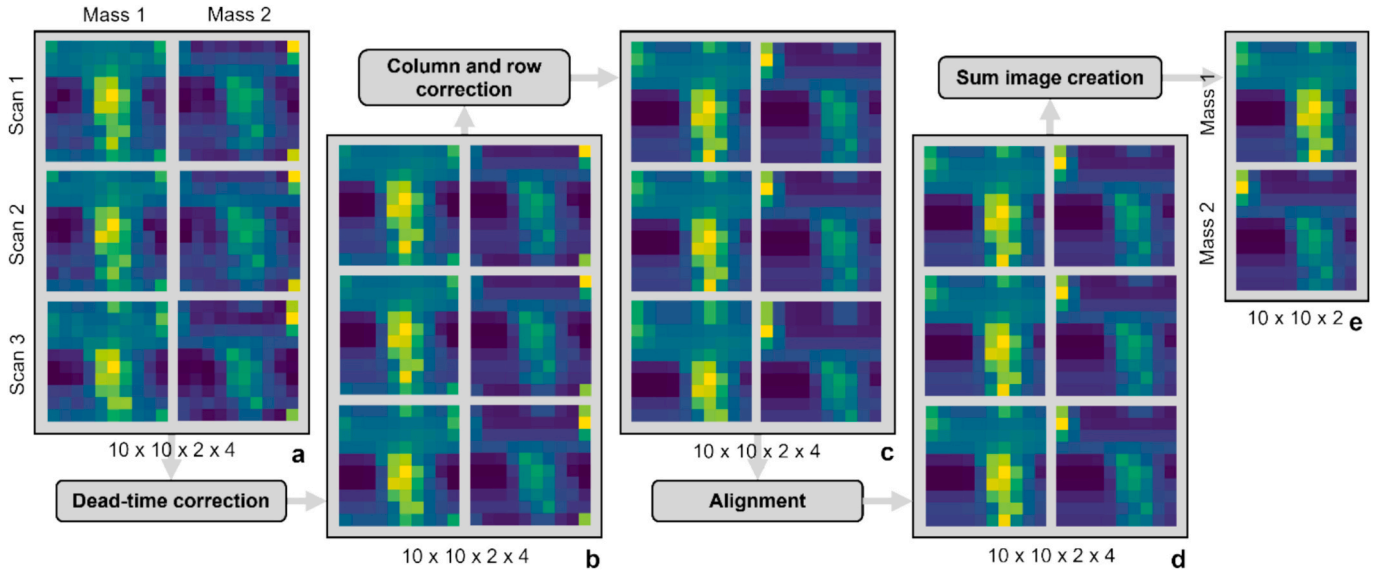


Fig. 2. A simplified visualization of the pre-processing of raw NanoSIMS data with N = 10 rows and columns, M = 2 measured masses, and S = 3 scans. Imported data has the shape of N × N × M × S, for which the scans are summed up to provide a two-dimensional map for each mass as N × N × M. The last row and column are corrected by shifting from panel b to c, and the drift of different scans are aligned in the last scan from c to d.

format of choice.

2.3. Discussion of the pre-processing

The labor and time requirements of the necessary pre-processing of NanoSIMS data result in a fragmentation, that is datasets isolated to individual projects containing a limited number of measurements. Although, numerous studies have applied NanoSIMS, the fragmentation, diverse pre-processing methods, and diverse file formats limit the analysis across larger datasets from different studies. Regarding data formats, NanoSIMS data is often saved as Tagged Image File (TIF) and Nearly Raw Raster Data (NRRD) in addition to the raw data in IM format. Hence, integrating large datasets into automated workflows can result in redundancy and compatibility challenges. However, the analysis of larger datasets across various studies holds the potential to obtain a more comprehensive understanding of how spatial patterns of similar structures evolve in different ecosystems. The dynamic evolution of NanoSIMS applications in different fields increasingly requires a standardized pre-processing scheme and a data format characterized by both robust computational performance and enduring usability.

To streamline the pre-processing procedure and simplify batch processing of NanoSIMS data, we introduce a novel open-source software tool: the TUM NanoSIMS Toolbox (NanoT, https://go.tum.de/122733). NanoT extends existing software tools as presented in (Nuñez et al., 2018) by enabling automatic processing to efficiently handle large NanoSIMS datasets. The toolbox enables the processing scheme presented in Section 2.2 and the generic file format, the Hierarchical Data Format (HDF). The HDF data format can store extensive N-dimensional data and associated metadata within a singular file accessible by various platforms (Folk et al., 2011). These features might be advantageous for batch-processing of large NanoSIMS datasets. The toolbox aims to promote the findability, accessibility, interoperability, and reusability by providing open-source and user friendly automatable processing algorithms for NanoSIMS data in accordance with the FAIR protocol (Wilkinson et al., 2016). Future work regarding NanoT is the adaption of methodologies from remote sensing, such as feature engineering, denoising, segmentation, and classification (Zhang and Zhang, 2022), as well as novel image registration methods from medical imaging for non-rigid structures (Haskins et al., 2020) for NanoSIMS image processing.

3. Unsupervised segmentation

To identify functional components constituting soil microstructure, here we present an automatic segmentation on the pre-processed NanoSIMS data using a two-step unsupervised clustering method. When measurements of reference materials for the training of classifiers are unavailable, data-driven methods such as the presented one provide an automated segmentation through clustering.

3.1. Data normalization

Due to the diverse range of ion counts across the seven measured masses, a normalization is needed to facilitate the comparison of individual ion channels. To avoid changing the shape of the original distribution (Cabello-Solorzano et al., 2023), a Min-Max normalization is applied to scale the signal from 0 to 1 in individual channels, calculated by:

$$x_{norm} = (x - x_{min})/(x_{max} - x_{min})$$

Where x is the initial pixel count from one ion channel, while x_{min} and x_{max} are the minimum and maximum signal counts of the ion channel over the whole dataset (cell 10 and 11 in the R code of supplementary material).

3.2. First step of the unsupervised segmentation: K-means clustering

To identify functional components with similar ion compositions, we conducted a two-step unsupervised segmentation method first starting with K-Means clustering. Since the NanoSIMS measurements for this method explanation are on GaAs wafers, we use the ion channels: $^{12}C^{12}C^-$, $^{12}C^{14}N^-$, and $^{16}O^-$ to distinguish mineral-dominated and OM-dominated regions. Pixels with high $^{16}O^-$ counts were defined as mineral-dominated regions, while OM-dominated regions were composed of pixels with high $^{12}C^{12}C^-$ and low $^{12}C^{14}N^-$ and $^{16}O^-$. Pores and background were indicated by pixels with low $^{12}C^{12}C^-$, $^{12}C^{14}N^-$ and $^{16}O^-$.

The pre-defined number of clusters for K-Means clustering (K-number) determines the performance of the segmentation result and was estimated by the first selection for this unsupervised segmentation method. This is because that K-Means clustering assigns pixels to the corresponding K-number of centroids, which is the center point of a cluster, and the movement of centroids terminates at the closest Euclidean distances between the pixels and the centroid (Ikotun et al., 2023; Macqueen, 1967; Nainggolan et al., 2019). The K-number provides the opportunity to obtain a number of data-driven clusters that can help to capture the multiple elemental compositions of organic and mineral components in the soil microstructures. This leads to the estimation of K-number by data-driven tools such as the elbow method. The elbow method calculates the total Within Sum of Square for an increasing number of clusters and provides the most favorable K-number at the elbow point, where the significant decrease in values of two clusters occurs (Demidenko, 2018; Guyeux et al., 2019; Nainggolan et al., 2019). However, the elbow point is hard to recognize from the elbow plot (in Fig. 3 (b)) and probably does not contribute to any explainable segments. Meanwhile, as in NanoSIMS measurements, the shapes of functional components are always irregular polygons with at least one reflex interior angle, which is called non-convex polygons. K-Means clustering is not the most supportive method for dealing with non-convex polygons. Therefore, the segmentation strategy changes for NanoSIMS data by increasing the K-number for overfitting results and sending the overfitted clusters to the second step for regrouping.

To optimize the segmentation, the K-number is first defined by the elbow method and set for the overfitting clusters as 70 (the K-number range in the elbow plot from 1 to 70; Cell15 in R code). Then, K-Means clustering, conducted by the “Lloyd” algorithm (Lloyd, 1982) with the 70 centroids, presents the preliminary clusters with highly similar ion counts (Cell 18 in R code).

3.3. Second step of the unsupervised segmentation: re-grouping of clusters

To solve the overfitting preliminary clusters and merge the clusters with similar mass signatures, re-grouping serves as the second step of this unsupervised segmentation using Hierarchical Agglomerative Clustering (HAC). The concept of HAC starts from each cluster centroid and then iteratively merges the closest clusters until one cluster contains all clusters (Davidson and Ravi, 2005; Kaufman and Rousseeuw, 1990). Thus, two selections for HAC are placed, linkage method selection for the similarity of clusters of the optimal re-grouping determination and gap statistic for optimal number for the final segmentation.

Four different linkage methods for HAC determine which clusters can be merged by computing their Euclidean distances. “Single” linkage is the smallest dissimilarity and “complete” linkage is the largest dissimilarity between two clusters. “Average” linkage calculates the average of the dissimilarities, while “ward’s” linkage aims to minimize the sum of squares of distances within each cluster (Table S1). To determine the optimal linkage method, the agglomerative coefficient is introduced to evaluate the strength of the clustering structure (Kaufman and Rousseeuw, 1990; Pandove et al., 2019; cell 20 in R code). A high agglomerative coefficient indicates the most robust hierarchical tree structure (Fig. 3 (h)). We used the centroid values from previous K-

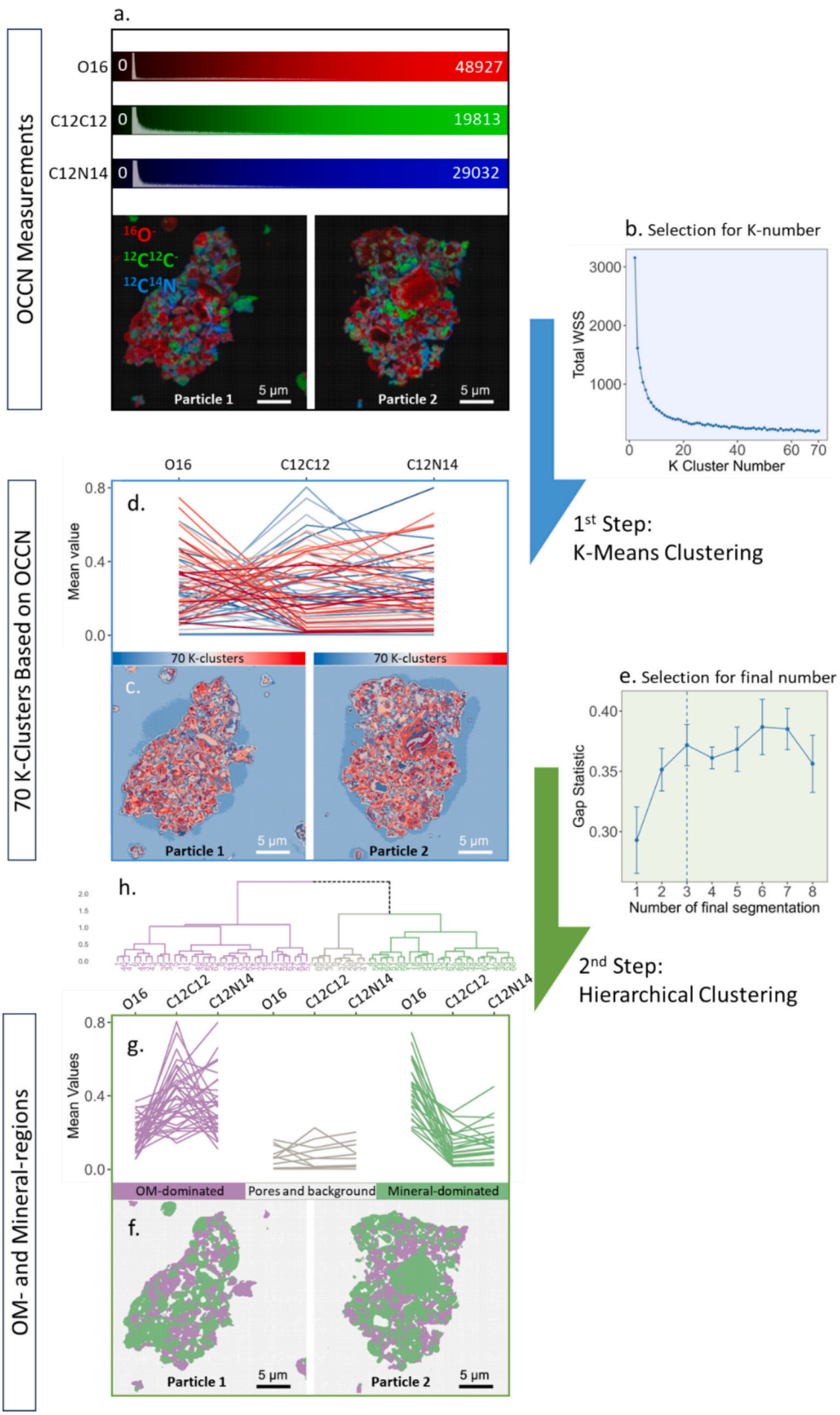


Fig. 3. Unsupervised segmentation working scheme on $^{16}O^-$, $^{12}C^{12}C^-$ and $^{12}C^{14}N^-$ channels in NanoSIMS measurements. (a) the pre-processed image data with $^{16}O^-$ (red), $^{12}C^{12}C^-$ (green) and $^{12}C^{14}N^-$ (blue). (b) the elbow plot exhibits the total Within Sum of Square (WSS) with K-numbers ranging from 1 to 70. (c) preliminary unsupervised segmentation results with the K-number as 70. (d) Mean values of preliminary 70 K-clusters. (e) The gap statistic plot exhibits the optimal final clustering number selection from 1 to 12. (f) final unsupervised segmentation results with 3 clusters, labels of these 3 clusters derived from (g). (g) Mean values of 3 final clusters, each curve representing the K-clusters. (h) the hierarchical tree being cut into 3 final clusters according to (e). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clusters as the base tree branches and the tree was formed based on the "ward's" linkage method (Cell 20 in R code). To cut the hierarchical tree into the optimal final clusters, the determination optimal number is determined using the Gap Statistic (Tibshirani et al., 2001). The Gap Statistic provides a null reference distribution and identifies as the smallest clustering number with the largest gap between its null reference which is similar to but advanced compared with the elbow method (Yang et al., 2020; Fig. 3 (b); cell 21 and 22 in R code). Based on this clustering number for re-grouping, the hierarchical tree is segmented into respective sections (Fig. 3 (h)), and the resulting clusters are subsequently assigned to their corresponding pixels (Fig. 3 (f)).

After the selection and re-grouping of K-Means clusters, the final segmentation shows three different segments: Cluster 1 with high $^{12}C^{12}C^-$ and $^{12}C^{14}N^-$ counts resembling OM-dominated regions, Cluster 2 with low ion counts in $^{12}C^{12}C^-$, $^{12}C^{14}N^-$ and $^{16}O^-$ channels resembling the background, and Cluster 3 with high $^{16}O^-$ counts resembling mineral-dominated regions.

3.4. *Unsupervised segmentation for highly skewed ion distribution*

Here a second exemplary application of an unsupervised segmentation is provided to identify different soil mineral phases and gain further information on how these are related with OM. To distinguish mineral phases, an altered two-step segmentation of the regions left after the exclusion of the regions identified as pore and background improved. Then the $^{28}Si^-$, $^{27}Al^{16}O^-$, and $^{56}Fe^{16}O^-$ distributions were used to distinguish the major mineral phases. The distribution of $^{56}Fe^{16}O^-$ counts was relatively more skewed than other ion channels. This was evaluated by the skewness, which represents the difference between the medium and mean counts of pixels within channels, and the z-score, as related to the skewness standardized to the mean skewness of all channels (Fig. 4a and Fig. S1, Joanes and Gill, 1998; Milligan and Cooper, 1988). This led to an unreliable segmentation result (Fig. 4c) which cannot reflect the $^{56}Fe^{16}O^-$ channel in its corresponding position (Fig. 4a). Together with the mean value curves in Fig. 4d, the $^{56}Fe^{16}O^-$ channel was not the primary factor for this segmentation. To improve the detection of Fe-rich phases, a separate second step HAC on the $^{56}Fe^{16}O^-$ channel was conducted, and the second step HAC of $^{28}Si^-$, $^{27}Al^{16}O^-$ channels were conducted together. This separated HAC points out with high and low $^{56}Fe^{16}O^-$, and $^{28}Si^-$ and $^{27}Al^{16}O^-$ segments (Fig. S2). By combining the two separated segments, a clear difference from Fig. 4c to 4e emerges where a part of Al-Si-Fe-dominated regions were assigned to high-count aluminosilicates. According to the mean value of ion counts curves (Fig. 4f), the high $^{56}Fe^{16}O^-$ points out the

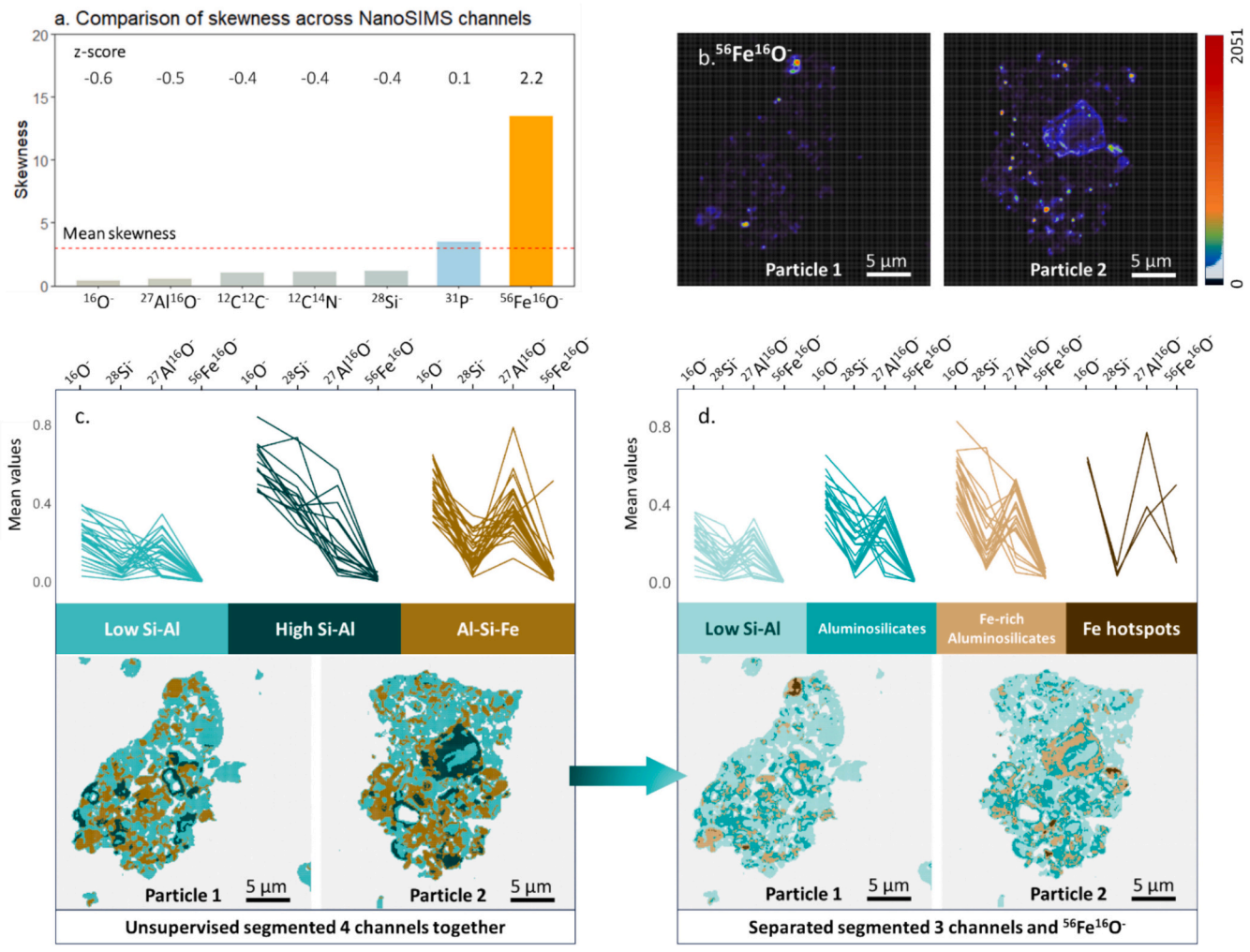


Fig. 4. Unsupervised segmentation for highly skewed $^{56}Fe^{16}O^-$. (a) Comparison of the skewness across seven ion channels (histograms shown in Fig. S1), which indicates that the $^{56}Fe^{16}O^-$ channel has a much higher skewness than other channels. The z-score in the top row indicates the difference to the mean skewness divided by the standard deviation. (b) visualization on $^{56}Fe^{16}O^-$ channel after contrast enhancement. (c) unsupervised segmentation results of $^{16}O^-$, $^{28}Si^-$, $^{27}Al^{16}O^-$, and $^{56}Fe^{16}O^-$ channels and the mean value of these 3 regrouped segments. (d) the joint result of unsupervised segmentation results of $^{56}Fe^{16}O^-$ channel (Fig. S2(c)) and of $^{16}O^-$, $^{28}Si^-$, $^{27}Al^{16}O^-$ channels (Fig. S1(e)) representing the relative occurrence of $^{56}Fe^{16}O^-$ with $^{28}Si^-$ and $^{27}Al^{16}O^-$ channels, and the mean value of 4 joint regrouped segments.

distribution of iron hotspots surround by clay minerals containing low $^{56}Fe^{16}O^-$. The integrated result of the OM- and mineral-dominated regions with specific mineral phases in Fig. 5 indicates the potential co-occurrence of Fe-rich aluminosilicates and aluminosilicates with OM-dominated regions.

3.5. Discussion on the unsupervised segmentation

Our automated unsupervised segmentation method aimed to distinguish mineral and organic components adapted to the characteristics of its NanoSIMS measurements, which are relatively large pixel data for multiple images and complex soil elements distributions. The composition of a soil sample is related to a number of factors, including parent material and the pedogenic transformation of mineral and organic components providing an intricate soil microstructure for which resolving the individual functional components can help to advance our understanding of biogeochemical cycles. To resolve the complex soil structure, here we took advantage of an unsupervised learning approach to differentiate OM- and mineral-dominated regions using a data-driven approach without the need for pre-defined components and references. Since our method is automated data-driven, it does not require experience on the visual identification of soil component structures to train the classifier. This avoids the multi-platform operation and the time to train a well-performed classifier compared with the supervised classification in Schweizer et al. (2018). Moreover, compared with only K-Means clustering, our method considered the influence of non-convex polygons on segmentation results and decreased this influence by regrouping the overfitted clusters (Mitra et al., 2003).

Due to the weakness of K-Means clustering, "regroup" is introduced to form this two-step unsupervised clustering method as the key concept. By comparing the 1st overfitting clusters with each other and grouping closely related clusters, one 2nd cluster would be regrouped. The combination of partition and hierarchical clustering techniques with the "regroup" connection could excel in performance and yield favorable outcomes when non-convex polygons occur. The order of applying K-Means clustering and HAC is irreversible; if K-Means clustering precedes, it eliminates the influence of non-convex polygons by the overfitting K-clusters, whereas if HAC is implemented first, it struggles with processing a substantial number of pixels. Consequently, the first K-Means clustering is conducted with a self-defined clustering number based on the operating central processing unit, ranging from 50 to 100 according to the elbow method (Sapegin et al., 2015). This two-step unsupervised clustering method is capable of identifying functional soil components according to its elemental signatures as illustrated using two soil NanoSIMS measurement.

To further improve our segmentation method, the ion ratios could be considered as supplementary parameters in HAC. The C/N ratio ($^{12}C^-/^{12}C^{14}N$) is widely used in NanoSIMS data analysis, to characterize the composition of organic matter (Hatton et al., 2012). Later, the normalized CN/C ratio ($^{12}C^{14}N^-/(^{12}C^{14}N^- + ^{12}C^-)$) was used to identify more N-rich regions of OM patches (Schweizer et al., 2018). Here we implemented the $^{12}C^{14}N^-/(^{12}C^{14}N^- + ^{12}C^-)$ ratio of OM-dominated regions in Fig. S2(a). Simultaneously, isotope ratios could also be included in the supplementary parameters if useful to differentiate functional components, e.g. the $^{15}N/^{14}N$ ratio (Boiteau et al., 2020; Herrmann et al., 2007) and $^{13}C/^{12}C$ ratio (Boiteau et al., 2020). In addition, apart from C/O, N/O and C/N ratios, Si/Al, Si/O, Fe/Si, Fe/Al ratios could be used used to distinguish Al-rich from Si-rich phyllosilicates phases and pedogenic metal oxides as applied previously (Kölbl et al., 2017; Li et al., 2016; Steffens et al., 2017). A normalized meaningful ratio calculated in 'channel A / (channel A + channel B)' might lead to a more precise and promising extension of the presented segmentation approach. Additionally, when the soil samples are embedded in resin, it could be used to segment the resin-filled pores in order to analyze exclusively the embedded soil structures.

The presented two-step unsupervised segmentation method provides an applicable approach when the measurements of reference materials are unavailable. When the measurements of reference materials are accessible, other machine learning approaches, such as supervised classification and deep learning (O'Shea and Nash, 2015; Ronneberger et al., 2015; Winterfeldt and Edwards, 1986), may provide an improved segmentation. This would then allow identify specific mineral phases and OM compounds, to further improve our understanding of the local role of functional soil components at the micro-scale.

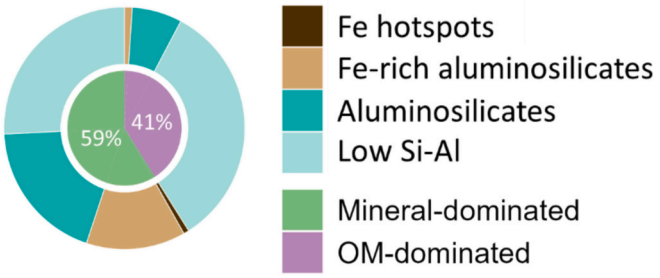


Fig. 5. Integrated result of Fig. 3 and Fig. 4. The inner pie chart indicates the proportion of soil OM- and mineral-dominated regions in these two soil particles. The out donut indicates the proportion of iron hotspots and aluminosilication in OM- and mineral-dominated regions.

4. Conclusion

In this paper, fundamental methods for the automated processing of NanoSIMS measurements were presented. We present the open-source software NanoT, an automated tool to streamline data pre-processing and improve data accessibility. By implementing pre-processing features ranging from dead-time correction to alignment, large amounts of NanoSIMS data can be rapidly analyzed and evaluated. The tool is openly accessible, enabling a growing number of NanoSIMS applications to process their data at no cost, with the potential of future extensions, such as different segmentation algorithms. In addition, we provide a data-based segmentation pipeline to identify functional components of soil that are challenging to pre-define beforehand. The unsupervised segmentation approach offers automated segmentation results for investigations based on different ion mass signatures, enabling analyses of the co-localization of individual ions with other channels. For this, we provide a way to incorporate highly skewed data distributions and successfully identify $^{56}Fe^{16}O^-$ and aluminosilicates hotspots. Altogether, the presented NanoSIMS processing scheme provides the opportunity to be transferred to other environmental matrics from cosmochemistry, material science, and biology applications. Apart from NanoSIMS measurements, our unsupervised approach could also be applied as a tool towards better understanding spatial patterns and the local interactions of functional components based on data from other techniques.

CRediT authorship contribution statement

Yahan Hu: Writing – original draft, Visualization, Software, Methodology. **Johann Maximilian Zollner:** Writing – original draft, Visualization, Software, Methodology. **Carmen Höschen:** Writing – review & editing, Data curation. **Martin Werner:** Writing – review & editing, Supervision. **Steffen A. Schweizer:** Writing – review & editing, Supervision, Conceptualization.

Acknowledgment

We thank Gertraud Harrington and Johann Lugmeier from the Technical University of Munich for their technical support in providing the NanoSIMS measurements. We acknowledge support from the Federal Ministry of Education and Research (BMBF) for the project

SoilCarbonHack [grant number: 16DKWN134].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102891.

Data availability

The R script of the unsupervised segmentation algorithm is provided in the supplementary information. The data files of the two exemplary NanoSIMS measurements are available in the figshare repository: https://doi.org/10.6084/m9.figshare.26049415 (preliminary link during peer review: https://figshare.com/s/c6a80d7af5d75a2e58f4, DOI link will be activated upon acceptance of the manuscript).

References

Allegretta, I., Legrand, S., Alfeld, M., Gattullo, C.E., Porfido, C., Spagnuolo, M., Janssens, K., Terzano, R., 2022. SEM-EDX hyperspectral data analysis for the study of soil aggregates. Geoderma 406, 115540. https://doi.org/10.1016/j.geoderma.2021.115540.

Amelung, W., Meyer, N., Rodionov, A., Knief, C., Aehnelt, M., Bauke, S.L., Biesgen, D., Dultz, S., Guggenberger, G., Jaber, M., Klumpp, E., Kögel-Knabner, I., Nischwitz, V., Schweizer, S.A., Wu, B., Totsche, K.U., Lehndorff, E., 2023. Process sequence of soil aggregate formation disentangled through multi-isotope labelling. Geoderma 429. https://doi.org/10.1016/j.geoderma.2022.116226 undefined-undefined.

Badimon, A., Strasburger, H.J., Ayata, P., Chen, X., Nair, A., Ikegami, A., Hwang, P., Chan, A.T., Graves, S.M., Uweru, J.O., Ledderose, C., Kutlu, M.G., Wheeler, M.A., Kahan, A., Ishikawa, M., Wang, Y.-C., Loh, Y.-H.E., Jiang, J.X., Surmeier, D.J., Schaefer, A., 2020. Negative feedback control of neuronal activity by microglia. Nature 586 (7829), 417–423. https://doi.org/10.1038/s41586-020-2777-8.

Barnes, J.J., Franchi, I.A., Anand, M., Tartèse, R., Starkey, N.A., Koike, M., Sano, Y., Russell, S.S., 2013. Accurate and precise measurements of the D/H ratio and hydroxyl content in lunar apatites using NanoSIMS. Chem. Geol. 337–338, 48–55. https://doi.org/10.1016/j.chemgeo.2012.11.015.

Baveye, P.C., Laba, M., Otten, W., Bouckaert, L., Dello Sterpaio, P., Goswami, R.R., Grinev, D., Houston, A., Hu, Y., Liu, J., Mooney, S., Pajor, R., Sleutel, S., Tarquis, A., Wang, W., Wei, Q., Sezgin, M., 2010. Observer-dependent variability of the thresholding step in the quantitative analysis of soil images and X-ray microtomography data. Geoderma 157 (1), 51–63. https://doi.org/10.1016/j.geoderma.2010.03.015.

Baveye, P.C., Otten, W., Kravchenko, A., Balseiro-Romero, M., Beckers, É., Chalhoub, M., Darnault, C., Eickhorst, T., Garnier, P., Hapca, S., Kiranyaz, S., Monga, O., Mueller, C.W., Nunan, N., Pot, V., Schlüter, S., Schmidt, H., Vogel, H.-J., 2018. Emergent properties of microbial activity in heterogeneous soil microenvironments: different research approaches are slowly converging, yet major challenges remain. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.01929.

Baveye, P.C., Otten, W., Kravchenko, A., 2019. Editorial: elucidating microbial processes in soils and sediments: microscale measurements and modeling. Front. Environ. Sci. 7. https://doi.org/10.3389/fenvs.2019.00078.

Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J.I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F.A., Kreshuk, A., 2019. ilastik: interactive machine learning for (bio)image analysis. Nat. Methods 16 (12). https://doi.org/10.1038/s41592-019-0582-9. Article 12.

Boiteau, R.M., Kukkadapu, R., Cliff, J.B., Smallwood, C.R., Kovarik, L., Wirth, M.G., Engelhard, M.H., Varga, T., Dohnalkova, A., Perea, D.E., Wietsma, T., Moran, J.J., Hofmockel, K.S., 2020. Calcareous organic matter coatings sequester siderophores in alkaline soils. Sci. Total Environ. 724, 138250. https://doi.org/10.1016/j.scitotenv.2020.138250.

Brunet, M.A., Gorman, B.L., Kraft, M.L., 2022. Depth correction of 3D NanoSIMS images shows intracellular lipid and cholesterol distributions while capturing the effects of differential sputter rate. ACS Nano 16 (10), 16221–16233. https://doi.org/10.1021/acsnano.2c05148.

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L.J., Tallón-Ballesteros, A., 2023. The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In: Bringas, P. García, García, H. Pérez, de Pisón, F.J. Martínez, Álvarez, F. Martínez, Lora, A. Troncoso, Herrero, Á., Rolle, J. L. Calvo, Quintián, H., Corchado, E. (Eds.), 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023). Springer Nature Switzerland, pp. 344–353. https://doi.org/10.1007/978-3-031-42536-3_33.

Chauhan, S., Rühaak, W., Khan, F., Enzmann, F., Mielke, P., Kersten, M., Sass, I., 2016. Processing of rock core microtomography images: using seven different machine learning algorithms. Comput. Geosci. 86, 120–128. https://doi.org/10.1016/j.cageo.2015.10.013.

Davidson, I., Ravi, S.S., 2005. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (Eds.), Knowledge Discovery in Databases: PKDD 2005. Springer, pp. 59–70. https://doi.org/10.1007/11564126_11.

Demidenko, E., 2018. The next-generation K-means algorithm. Stat. Anal. Data Min. 11 (4), 153–166. https://doi.org/10.1002/sam.11379.

Folk, M., Heber, G., Koziol, Q., Pourmal, E., Robinson, D.. An overview of the HDF5 technology suite and its applications. https://doi.org/10.1145/1966895.1966900.

Gormanns, P., Reckow, S., Poczatek, J.C., Turck, C.W., Lechene, C., 2012. Segmentation of multi-isotope imaging mass spectrometry data for semi-automatic detection of regions of interest. PLoS One 7 (2), e30576. https://doi.org/10.1371/journal.pone.0030576.

Guyeux, C., Chrétien, S., Bou Tayeh, G., Demerjian, J., Bahi, J., 2019. Introducing and comparing recent clustering methods for massive data management in the internet of things. J. Sens. Actuator Netw. 8, 56. https://doi.org/10.3390/jsan8040056.

Hao, J., Yang, W., Huang, W., Xu, Y., Lin, Y., Changela, H., 2020. NanoSIMS measurements of sub-micrometer particles using the local thresholding technique. Surf. Interface Anal. 52 (5), 234–239. https://doi.org/10.1002/sia.6711.

Hapca, S.M., Houston, A.N., Otten, W., Baveye, P.C., 2013. New local thresholding method for soil images by minimizing grayscale intra-class variance. Vadose Zone J. 12 (3), vzj2012.0172. https://doi.org/10.2136/vzj2012.0172.

Hapca, S., Baveye, P.C., Wilson, C., Lark, R.M., Otten, W., 2015. Three-dimensional mapping of soil chemical characteristics at micrometric scale by combining 2D SEM-EDX data and 3D X-ray CT images. PLoS One 10 (9), e0137205. https://doi.org/10.1371/journal.pone.0137205.

Haruwaka, K., Ikegami, A., Tachibana, Y., Ohno, N., Konishi, H., Hashimoto, A., Matsumoto, M., Kato, D., Ono, R., Kiyama, H., Moorhouse, A.J., Nabekura, J., Wake, H., 2019. Dual microglia effects on blood brain barrier permeability induced by systemic inflammation. Nat. Commun. 10 (1), 5816. https://doi.org/10.1038/s41467-019-13812-z.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31 (1–2), 8. https://doi.org/10.1007/s00138-020-01060-x.

Hatton, P.-J., Remusat, L., Zeller, B., Derrien, D., 2012. A multi-scale approach to determine accurate elemental and isotopic ratios by nano-scale secondary ion mass spectrometry imaging: accurate elemental and isotopic ratios by NanoSIMS imaging. Rapid Commun. Mass Spectrom. 26 (11), 1363–1371. https://doi.org/10.1002/rcm.6228.

Heckman, K., Lawrence, C.R., Harden, J.W., 2018. A sequential selective dissolution method to quantify storage and stability of organic carbon associated with Al and Fe hydroxide phases. Geoderma 312, 24–35. https://doi.org/10.1016/j.geoderma.2017.09.043.

Herrmann, A.M., Clode, P.L., Fletcher, I.R., Nunan, N., Stockdale, E.A., O'Donnell, A.G., Murphy, D.V., 2007. A novel method for the study of the biophysical interface in soils using nano-scale secondary ion mass spectrometry. Rapid Commun. Mass Spectrom. 21 (1), 29–34. https://doi.org/10.1002/rcm.2811.

Hoppe, P., Cohen, S., Meibom, A., 2013. NanoSIMS: technical aspects and applications in cosmochemistry and biological geochemistry. Geostand. Geoanal. Res. 37 (2), 111–154. https://doi.org/10.1111/j.1751-908X.2013.00239.x.

Houston, A.N., Otten, W., Baveye, P.C., Hapca, S., 2013. Adaptive-window indicator kriging: a thresholding method for computed tomography images of porous media. Comput. Geosci. 54, 239–248. https://doi.org/10.1016/j.cageo.2012.11.016.

Houston, A.N., Otten, W., Falconer, R., Monga, O., Baveye, P.C., Hapca, S.M., 2017. Quantification of the pore size distribution of soils: assessment of existing software using tomographic and synthetic 3D images. Geoderma 299, 73–82. https://doi.org/10.1016/j.geoderma.2017.03.025.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

Inagaki, T.M., Possinger, A.R., Grant, K.E., Schweizer, S.A., Mueller, C.W., Derry, L.A., Lehmann, J., Kögel-Knabner, I., 2020. Subsoil organo-mineral associations under contrasting climate conditions. Geochim. Cosmochim. Acta 270, 244–263. https://doi.org/10.1016/j.gca.2019.11.030.

Inagaki, T.M., Possinger, A.R., Schweizer, S.A., Mueller, C.W., Hoeschen, C., Zachman, M.J., Kourkoutis, L.F., Kögel-Knabner, I., Lehmann, J., 2023. Microscale spatial distribution and soil organic matter persistence in top and subsoil. Soil Biol. Biochem. 178, 108921. https://doi.org/10.1016/j.soilbio.2022.108921.

Joanes, D.N., Gill, C.A., 1998. Comparing measures of sample skewness and kurtosis. J. Royal Stat. Soc. Ser. D (The Statistician) 47 (1), 183–189. https://doi.org/10.1111/1467-9884.00122.

Kaufman, L., Rousseeuw, P.J., 1990. Agglomerative Nesting (Program AGNES). In: Finding Groups in Data. John Wiley & Sons, Ltd., pp. 199–252. https://doi.org/10.1002/9780470316801.ch5

Kleber, M., Bourg, I.C., Coward, E.K., Hansel, C.M., Myneni, S.C.B., Nunan, N., 2021. Dynamic interactions at the mineral–organic matter interface. Nat. Rev. Earth Environ. 2 (6). https://doi.org/10.1038/s43017-021-00162-y. Article 6.

Kölbl, A., Schweizer, S., Mueller, C., Höschen, C., Said-Pullicino, D., Romani, M., Lugmeier, J., Schlüter, S., Kögel-Knabner, I., 2017. Legacy of rice roots as encoded in distinctive microsites of oxides, silicates, and organic matter. Soils 1 (1). https://doi.org/10.3390/soils1010002. Article 1.

Lehmann, J., Solomon, D., Kinyangi, J., Dathe, L., Wirick, S., Jacobsen, C., 2008. Spatial complexity of soil organic matter forms at nanometre scales. Nat. Geosci. 1 (4), 238–242. https://doi.org/10.1038/ngeo155.

Li, K., Sinha, B., Hoppe, P., 2016. Speciation of nitrogen-bearing species using negative and positive secondary ion spectra with nano secondary ion mass spectrometry. Anal. Chem. 88 (6), 3281–3288. https://doi.org/10.1021/acs.analchem.5b04740.

Li, K., Liu, J., Grovenor, C.R.M., Moore, K.L., 2020. NanoSIMS imaging and analysis in materials science. Annu Rev Anal Chem (Palo Alto, Calif) 13 (1), 273–292. https://doi.org/10.1146/annurev-anchem-092019-032524.

Lloyd, S., 1982. Least squares quantization in PCM. In: *IEEE Transactions on Information Theory*, 28. IEEE Transactions on Information Theory, pp. 129–137. https://doi.org/10.1109/TIT.1982.1056489.

Macqueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, pp. 281–297.

Malik, J., Kiranyaz, S., Al-Raoush, R.I., Monga, O., Garnier, P., Foufou, S., Bouras, A., Iosifidis, A., Gabbouj, M., Baveye, P.C., 2022. 3D quantum cuts for automatic segmentation of porous media in tomography images. Comput. Geosci. 159, 105017. https://doi.org/10.1016/j.cageo.2021.105017.

McKeegan, K.D., Aléon, J., Bradley, J., Brownlee, D., Busemann, H., Butterworth, A., Chaussidon, M., Fallon, S., Floss, C., Gilmour, J., Gounelle, M., Graham, G., Guan, Y., Heck, P.R., Hoppe, P., Hutcheon, I.D., Huth, J., Ishii, H., Ito, M., Zinner, E., 2006. Isotopic compositions of cometary matter returned by stardust. Science 314 (5806), 1724–1728. https://doi.org/10.1126/science.1135992.

Milligan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. J. Classif. 5 (2), 181–204. https://doi.org/10.1007/BF01897163.

Mitra, P., Pal, S.K., Siddiqi, M.A., 2003. Non-convex clustering using expectation maximization algorithm with rough set initialization. Pattern Recogn. Lett. 24 (6), 863–873. https://doi.org/10.1016/S0167-8655(02)00198-8.

Mueller, C.W., Hoeschen, C., Koegel-Knabner, I., 2023. Understanding of soil processes at the microscale—Use of NanoSIMS in soil science. In: Goss, M.J., Oliver, M. (Eds.), Encyclopedia of Soils in the Environment, Second edition. Academic Press, pp. 670–680. https://doi.org/10.1016/B978-0-12-822974-3.00045-8.

Nainggolan, R., Perangin-angin, R., Simarmata, E., Tarigan, A.F., 2019. Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the elbow method. J. Phys. Conf. Ser. 1361 (1), 012015. https://doi.org/10.1088/1742-6596/1361/1/012015.

Nuñez, J., Renslow, R., Cliff, J.B., Anderton, C.R., 2018. NanoSIMS for biological applications: current practices and analyses. Biointerphases 13 (3), 03B301. https://doi.org/10.1116/1.4993628.

O'Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks (arXiv: 1511.08458). arXiv. http://arxiv.org/abs/1511.08458.

Pandove, D., Goel, S., Rani, R., 2019. General correlation coefficient based agglomerative clustering. Clust. Comput. 22 (2), 553–583. https://doi.org/10.1007/s10586-018-2863-y.

Pedrazzini, S., Child, D.J., Aarholt, T., Ball, C., Dowd, M., Girling, A., Cockings, H., Perkins, K., Hardy, M.C., Stone, H.J., Bagot, P.A.J., 2018. On the effect of environmental exposure on dwell fatigue performance of a fine-grained nickel-based superalloy. Metall. Mater. Trans. A 49 (9), 3908–3922. https://doi.org/10.1007/s11661-018-4752-7.

Pett-Ridge, J., Weber, P.K., 2022. NanoSIP: NanoSIMS applications for microbial biology. In: Navid, A. (Ed.), Microbial Systems Biology, vol. 2349. Springer US, pp. 91–136. https://doi.org/10.1007/978-1-0716-1585-0_6.

Polerecky, L., Adam, B., Milucka, J., Musat, N., Vagner, T., Kuypers, M.M.M., 2012. Look@NanoSIMS – a tool for the analysis of nanoSIMS data in environmental microbiology. Environ. Microbiol. 14 (4), 1009–1023. https://doi.org/10.1111/j.1462-2920.2011.02681.x.

Portell, X., Pot, V., Garnier, P., Otten, W., Baveye, P.C., 2018. Microscale heterogeneity of the spatial distribution of organic matter can promote bacterial biodiversity in soils: insights from computer simulations. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.01583.

Possinger, A.R., Zachman, M.J., Enders, A., Levin, B.D.A., Muller, D.A., Kourkoutis, L.F., Lehmann, J., 2020. Organo–organic and organo–mineral interfaces in soil at the nanometer scale. Nat. Commun. 11 (1), 6103. https://doi.org/10.1038/s41467-020-19792-9.

Pot, V., Portell, X., Otten, W., Garnier, P., Monga, O., Baveye, P.C., 2022. Understanding the joint impacts of soil architecture and microbial dynamics on soil functions: insights derived from microscale models. Eur. J. Soil Sci. 73 (3), e13256. https://doi.org/10.1111/ejss.13256.

Rasmussen, C., Heckman, K., Wieder, W.R., Keiluweit, M., Lawrence, C.R., Berhe, A.A., Blankinship, J.C., Crow, S.E., Druhan, J.L., Hicks Pries, C.E., Marin-Spiotta, E., Plante, A.F., Schädel, C., Schimel, J.P., Sierra, C.A., Thompson, A., Wagai, R., 2018. Beyond clay: towards an improved set of variables for predicting soil organic matter content. Biogeochemistry 137 (3), 297–306. https://doi.org/10.1007/s10533-018-0424-3.

Remusat, L., Hatton, P.-J., Nico, P.S., Zeller, B., Kleber, M., Derrien, D., 2012. NanoSIMS study of organic matter associated with soil aggregates: advantages, limitations, and combination with STXM. Environ. Sci. Technol. 46 (7), 3943–3949. https://doi.org/10.1021/es203745k.

Renslow, R.S., Lindemann, S.R., Cole, J.K., Zhu, Z., Anderton, C.R., 2016. Quantifying element incorporation in multispecies biofilms using nanoscale secondary ion mass spectrometry image analysis. Biointerphases 11 (2), 02A322. https://doi.org/10.1116/1.4941764.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation (arXiv:1505.04597). arXiv. https://doi.org/10.48550/arXiv.1505.04597.

Sapegin, A., Gawron, M., Jaeger, D., Cheng, F., Meinel, C., 2015. High-speed security analytics powered by in-memory machine learning engine. In: 2015 14th International Symposium on Parallel and Distributed Computing, pp. 74–81. https://doi.org/10.1109/ISPDC.2015.16.

Schweizer, S.A., Hoeschen, C., Schlüter, S., Kögel-Knabner, I., Mueller, C.W., 2018. Rapid soil formation after glacial retreat shaped by spatial patterns of organic matter accrual in microaggregates. Glob. Chang. Biol. 24 (4), 1637–1650. https://doi.org/10.1111/gcb.14014.

Solomon, D., Lehmann, J., Harden, J., Wang, J., Kinyangi, J., Heymann, K., Karunakaran, C., Lu, Y., Wirick, S., Jacobsen, C., 2012. Micro- and nano-environments of carbon sequestration: multi-element STXM–NEXAFS spectromicroscopy assessment of microbial carbon and mineral associations. Chem. Geol. 329, 53–73. https://doi.org/10.1016/j.chemgeo.2012.02.002.

Steffens, M., Rogge, D.M., Mueller, C.W., Höschen, C., Lugmeier, J., Kölbl, A., Kögel-Knabner, I., 2017. Identification of distinct functional microstructural domains controlling C storage in soil. Environ. Sci. Technol. 51 (21), 12182–12189. https://doi.org/10.1021/acs.est.7b03715.

Thevenaz, P., Ruttimann, U.E., Unser, M., 1998. A pyramid approach to subpixel registration based on intensity. IEEE Trans. Image Process. 7 (1), 27–41. https://doi.org/10.1109/83.650848.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B Stat Methodol. 63 (2), 411–423. https://doi.org/10.1111/1467-9868.00293.

Urbanski, L., Schad, P., Kalbitz, K., van Mourik, J., Gehrt, E., Kögel-Knabner, I., 2022. Legacy of plaggen agriculture: high soil organic carbon stocks as result from high carbon input and volume increase. Geoderma 406, 115513. https://doi.org/10.1016/j.geoderma.2021.115513.

Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359.

Wan, J., Tyliszczak, T., Tokunaga, T.K., 2007. Organic carbon distribution, speciation, and elemental correlations within soil microaggregates: applications of STXM and NEXAFS spectroscopy. Geochim. Cosmochim. Acta 71 (22), 5439–5449. https://doi.org/10.1016/j.gca.2007.07.030.

Werner, F., Mueller, C.W., Thieme, J., Gianoncelli, A., Rivard, C., Höschen, C., Prietzel, J., 2017. Micro-scale heterogeneity of soil phosphorus depends on soil substrate and depth. Sci. Rep. 7 (1), Article 1. https://doi.org/10.1038/s41598-017-03537-8.

Wilhelm, R.C., Lynch, L., Webster, T.M., Schweizer, S., Inagaki, T.M., Tfaily, M.M., Kukkadapu, R., Hoeschen, C., Buckley, D.H., Lehmann, J., 2022. Susceptibility of new soil organic carbon to mineralization during dry-wet cycling in soils from contrasting ends of a precipitation gradient. Soil Biol. Biochem. 169, 108681. https://doi.org/10.1016/j.soilbio.2022.108681.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Mons, B., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3 (1). https://doi.org/10.1038/sdata.2016.18. Article 1.

Wilson, R.G., 1995. SIMS quantification in Si, GaAs, and diamond—an update. Int. J. Mass Spectrom. Ion Process. 143, 43–49. https://doi.org/10.1016/0168-1176(94)04136-U.

Winterfeldt, D., Edwards, W., 1986. Decision Analysis and Behavioral Research. August 29. https://www.semanticscholar.org/paper/Decision-Analysis-and-Behavioral-Research-Winterfeldt-Edwards/28b906ad525b51de39cc1d9f53580a47ee59dfed.

Yang, J., Lee, J.-Y., Choi, M., Joo, Y., 2020. A new approach to determine the optimal number of clusters based on the gap statistic. In: Boumerdassi, S., Renault, É., Mühlethaler, P. (Eds.), Machine Learning for Networking. Springer International Publishing, pp. 227–239. https://doi.org/10.1007/978-3-030-45778-5_15.

Zhang, L., Zhang, L., 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. In: IEEE Geoscience and Remote Sensing Magazine, 10(2), pp. 270–294. https://doi.org/10.1109/MGRS.2022.3145854.