The Planetary Science Journal, 5:280 (23pp), 2024 December https://doi.org/10.3847/PSJ/ad90b6

Autonomous Detection of Mineral Phases in a Rock Sample Using a Space-prototype LIMS Instrument and Unsupervised Machine Learning

Salome Gruchola[1], Peter Keresztes Schmidt[1], Marek Tulej[1], Andreas Riedo[1,2], Klaus Mezger[3], and Peter Wurz[1,2]
[1] Physics Institute, Space Research and Planetary Sciences, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
[2] NCCR PlanetS, University of Bern, Gesellschaftstrasse 6, 3012 Bern, Switzerland
[3] Institute for Geological Sciences, University of Bern, Baltzerstrasse 1+3, 3012 Bern, Switzerland
Received 2024 January 12; revised 2024 November 05; accepted 2024 November 08; published 2024 December 27

Abstract

In situ mineralogical and chemical analyses of rock samples using a space-prototype laser ablation ionization mass spectrometer along with unsupervised machine learning are powerful tools for the study of surface samples on planetary bodies. This potential is demonstrated through the examination of a thin section of a terrestrial rock sample in the laboratory. Autonomous isolation of mineral phases within the acquired mass spectrometric data is achieved with two dimensionality reduction techniques: uniform manifold approximation and projection (UMAP) and density-preserving variation of UMAP (densMAP), and the density-based clustering algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Both densMAP and UMAP yield comparable outcomes, successfully isolating the major mineral phases fluorapatite, calcite, and forsterite in the studied rock sample. Notably, densMAP reveals additional insights into the composition of the sample through outlier detection, uncovering signals from the trace minerals pyrite, rutile, baddeleyite, and uranothorianite. Through a grid search, the stability of the methods over a broad model parameter space is confirmed, revealing a correlation between the level of data preprocessing and the resulting clustering quality. Consequently, these methods represent effective strategies for data reduction, highlighting their potential application on board spacecraft to obtain direct and quantitative information on the chemical composition and mineralogy of planetary surfaces and to optimize mission returns through the unsupervised selection of valuable data.

Unified Astronomy Thesaurus concepts: Time-of-flight mass spectrometry (2222); Laser ablation (2245); Dimensionality reduction (1943)

1. Introduction

In the context of space exploration, the efficient and effective management of data acquisition and data transmission is a fundamental necessity. As scientific missions continue to expand our understanding of the solar system and beyond, the importance of unsupervised operation and sophisticated data reduction techniques is growing. Optimal utilization of limited onboard resources enhances the scientific return of space missions. Both automation as well as data reduction are feasible with machine learning (ML), and while ML has become a state-of-the-art data analysis technique on Earth, space-based ML applications are still rare. Space missions operate in highly challenging environments under severe computational constraints. Radiation-hard onboard hardware is expensive and therefore often very limited in its capacity, complicating the use of ML in space. Additionally, the high cost of failure of space missions puts further constraints on the reliability of onboard software, which should under no circumstances pose a threat to the spacecraft's core operations, station keeping, and maintenance (V. Kothari et al. 2020).

Nonetheless, interest in the field of in situ ML applications is rapidly growing because ML offers the autonomy for spacecraft that is highly sought after. Any task that can be autonomously performed without interaction from Earth saves time and increases the chance of valuable data collection, as most spacecraft have limited lifetimes due to component degradation, consumable depletion, or budgetary constraints (A. McGovern & K. L. Wagstaff 2011). Mission return can further be improved with ML through autonomous categorization of high- and low-quality data on board the spacecraft prior to data transmission. This helps mitigate downlink limitations due to factors like spacecraft distance from Earth, antenna size, transmitter power, selected frequency band, data priority, and mission duration. Certain ML techniques, including k-means and support vector machines, were deemed resilient enough to withstand the Earth's orbit environment without the need for rad-hard memory (K. L. Wagstaff & B. Bornstein 2009; F. Gieseke et al. 2012). Further development of ML techniques less sensitive to bit flips will ease the application of ML in space.

To date, ML in space research has primarily been applied to postanalyses of space-acquired data, and far less during space operations. In space, ML is currently more prominently used for spacecraft operation than for onboard scientific data analysis. Examples for in situ applications include image classification on Mars rovers, navigation, landing, and autonomous target selection (R. Castano et al. 2006; M. Bajracharya et al. 2008; A. E. Johnson et al. 2015; R. Francis et al. 2017; B. Gaudet et al. 2020; N. Abcouwer et al. 2021). Further applications lie predominantly in the field of Earth-orbiting satellites (R. Castano et al. 2005), or robotics, where processes are automatized with artificial intelligence (R. Doyle et al. 2021). In particular, CubeSats and microsatellites provide an optimal test bed for ML applications in space due to their relatively low cost of failure compared to larger satellites (G. Mateo-Garcia et al. 2021; D. A. Zeleke & H. D. Kim 2023). To date, two ESA missions (OPS-SAT and ϕ-Sat-1) have been

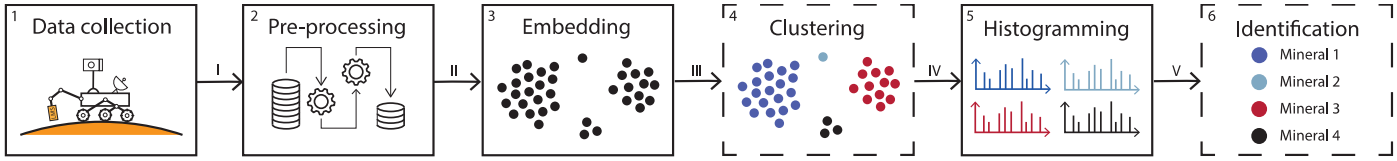

Figure 1. Pipeline for onboard ML analysis. (1) Data are collected on board a spacecraft. (2) Subsequently, the data are preprocessed on board the spacecraft. (3) Dimensionality reduction techniques can be used to embed the collected data into a low-dimensional space. (4) Clustering of the embedded data groups them according to their similarity in chemical composition. (5) Spectra of each cluster can be histogrammed to obtain a representative spectrum with greater dynamic range. (6) Analysis of the representative spectra allows for mineral identification and trace element detection.

launched with the aim of performing onboard ML analyses (G. Giuffrida et al. 2022; G. Labreche et al. 2022).

Supervised and reinforcement learning techniques dominate current ML applications in space. However, as the acquired data often lack labels, and manual label assignment is resource intensive, the demand for unsupervised techniques is growing (E. Kalinicheva et al. 2020; M. Shirobokov et al. 2021). In unsupervised ML, only the data themselves are used as input to the model, without any corresponding labels. The goal is to uncover patterns present within the data by grouping similar samples (clustering), finding the distribution of data within the input space (density estimation), or projecting high-dimensional data into a lower dimensional space (dimensionality reduction) (M. C. Bishop 2006). Unsupervised ML has been tested successfully on resource-limited hardware (V. Růžička et al. 2022). Various onboard unsupervised ML applications have been proposed, including an approach based on k-means and t-SNE for the autonomous detection of satellite fault diagnosis (S. K. Ibrahim et al. 2020). Further applications include dimensionality reduction techniques for health monitoring of satellites (T. Yairi et al. 2017) or applications during swarm missions (J. C. Davis & H. J. Pernicka 2020). In general, a significant rise of ML applications on spacecraft in the future is apparent (S. Kumar & R. Tomar 2019; A. Russo & G. Lax 2022).

This study presents an example of how unsupervised ML techniques could be used on board future space missions to enhance mission return. The goal is to filter data on board the spacecraft, selecting and transmitting only relevant data while discarding uninteresting or low-quality information.

2. Approach

A possible outline of the described procedure for data reduction is shown in Figure 1. This procedure will be presented in greater detail in subsequent sections of the paper.

The procedure commences with data collection on a rover, lander, or other spacecraft equipped with, e.g., a mass spectrometric instrument (Figure 1, panel (1)). To mimic the process as closely as possible, data in this study were collected from a geological thin section using a space-prototype laser ablation ionization mass spectrometer (LIMS). Subsequent data preprocessing reduces data complexity and noise through computationally inexpensive techniques like averaging, low-pass filtering, and downsampling or mass integration (Figure 1, panel (2)). Next, a dimensionality reduction technique based on unsupervised ML is employed to find a low-dimensional embedding of the data aimed at the autonomous isolation of mineral phases, grouping similar samples in close proximity (Figure 1, panel (3)). This embedding, represented as a low-dimensional matrix (embedding dimensions $\times$ number of samples), might be transmitted to Earth for an initial assessment of data diversity. From the embedding, feedback on parameter selection as well as chemical composition can be retrieved, in addition to the identification of possible outliers. On Earth, clustering of the embedding can be carried out with fewer computational constraints, allowing for a more thorough analysis of clustering stability (Figure 1, panel (4)). The number of retrieved clusters should ideally reflect the number of sampled minerals with a distinct chemical composition. The obtained labels can then be sent back to the spacecraft to choose representative spectra for each cluster or histogram data from the clusters, which reflect the chemical composition of similar spectra (Figure 1, panel (5)). Alternatively, clustering may be performed directly on the spacecraft. Transferring the representative spectra to Earth enables the identification of various mineral phases present within the sampled material (Figure 1, panel (6)). This then allows the minerals of interest to be selected, and the data of interest can be downloaded faster as the uninteresting data can be discarded beforehand.

In this paper, two dimensionality reduction techniques, uniform manifold approximation and projection (UMAP) and a density-preserving variation of UMAP (densMAP), were applied to the collected data (L. McInnes et al. 2018; A. Narayan et al. 2021). A comparative analysis explores the respective advantages and limitations of the two methods and discusses their relevance to space research. Subsequently, the clustering algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was employed to identify and label distinct clusters within the embedding.

While UMAP has previously been applied to LIMS data (R. R. Lukmanov et al. 2021, 2022), the application of densMAP as a density-preserving dimensionality reduction technique is novel. In this contribution, both techniques are compared, along with an evaluation of different levels of data preprocessing. Furthermore, a grid search was performed across a broad parameter space to assess clustering stability. Special emphasis is put on the impact of data preprocessing and parameter selection on the analysis outcome, as it is necessary to anticipate situations where limited computational resources prevent iterative parameter tuning. Determining suitable levels of preprocessing and model parameter selection can be challenging, given the absence of straightforward criteria to evaluate the validity of unsupervised ML models. To validate the clustering results, a comparative analysis was performed with information from both optical microscopy and element maps of the integrated data to establish the ground truth of the data set. This approach enables the manual isolation of chemically distinct phases, reflecting the objective of the ML approach.

3. Data Collection

3.1. LIMS Instrument

The LIMS instrument used in this study is a miniature space-prototype. This specialized instrument was built at the

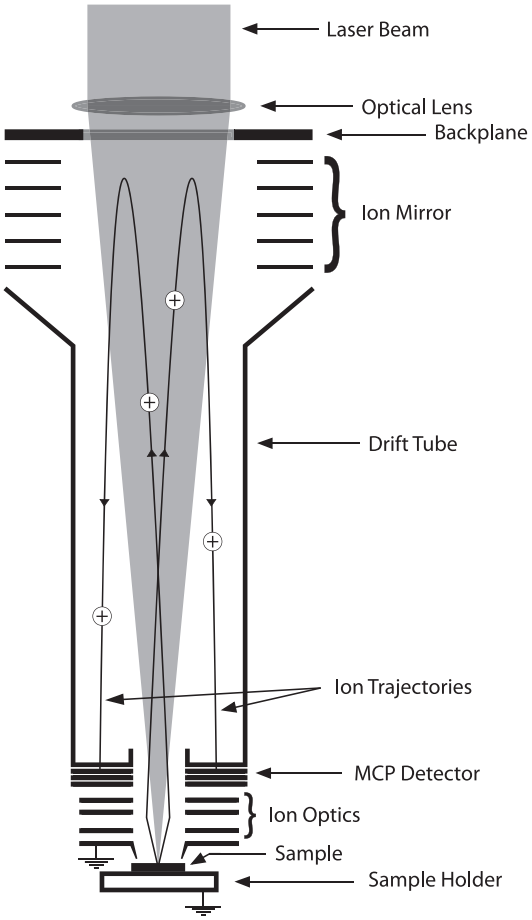

Figure 2. Schematics and principles of the operation of the LIMS instrument.

University of Bern, with its initial design aimed at in situ investigations of chemical and isotopic surface compositions of celestial bodies in our solar system. The LIMS instrument consists of a compact (160 mm × ∅ 60 mm) reflectron-type time-of-flight mass analyzer, a chirped pulse amplification (CPA)-2110 femtosecond laser system ($\lambda = 258$ nm, $f = 1$ kHz, $\tau \sim 180$ fs pulse duration; Clark-MXR, Dexter MI, USA) and a microchannel plate (MCP) detector system coupled to a U5303A high speed analog-to-digital converter (ADC) (12 bit, 3.2 GS s^{-1}; Acqiris SA, Switzerland). The operation of the LIMS instrument involves the generation of positively charged ions through laser ablation, followed by their guidance through the mass analyzer via ion optical elements and their detection with the MCP detector system. The analytical capabilities of the LIMS instrument extend to both chemical imaging of surfaces and three-dimensional depth profiling. The instrument reaches sensitivities for the abundance of chemical elements down to the lower parts per million level (weight fractions), while offering spatial resolutions at the micrometer scale laterally and submicrometer scale vertically. The schematics and principles of operation are illustrated in Figure 2. Further details of the instrument can be found in previous publications (U. Rohner et al. 2003; V. A. Riedo et al. 2013; V. Grimaudo et al. 2015, 2017; A. Neubeck et al. 2016; M. Tulej et al. 2021). To enhance the ion yield and reduce the risk of isobaric interferences of polyatomic species with element species, the laser system of the LIMS instrument was operated in double pulse (DP) mode (M. Tulej et al. 2018; A. Riedo et al. 2021).

3.2. Materials

Data were collected from a thin section of an ultramafic phoscorite rock from the Phalaborwa Carbonatite Complex, located in the Limpopo Province, South Africa (S. C. Eriksson 1984; L. M. Heaman 2009; S. Decrée et al. 2020).

Phoscorite is a rare plutonic ultramafic rock predominantly associated with carbonatite both spatially and temporally, forming a multiphase phoscorite-carbonatite series. The essential rock-forming minerals of phoscorite are typically defined as olivine (commonly serpentinized), apatite, and magnetite (H. D. Russell et al. 1954; S. C. Eriksson 1984; N. I. Krasnova et al. 2004; L. Milani et al. 2017). Olivine is a solid-solution series with the end-members forsterite and fayalite, defined by the stochiometric formula of $(Mg, Fe)_2SiO_4$. The olivine in the phoscorite is typically forsteritic (Fo_{79}–Fo_{91}) and contains sub–parts per million concentrations of rare-earth elements (REEs). Only a few minor elements exceed tens of parts per million concentrations, with Mn being the most abundant among them (L. Milani et al. 2017). Apatite ($Ca_5(PO_4)_3(F, Cl, OH)$) is the dominant constituent of the rock sample. Apatite in carbonatites is enriched in REEs, with the degree of enrichment being contingent on the formation conditions.

Optical images of the studied fluorapatite thin section are shown in Figure 3. Sample preparation for in situ element analyses encompassed surface polishing, fixation of the thin section with epoxy on a glass plate, and subsequent application of a thin gold layer, approximately 10 nm thick, via sputter coating (Figure 3, panels (A) and (B)). This thin conductive layer helps mitigate charge buildup and improves spectral quality (S. Gruchola et al. 2023). When held against a light source, the thin section is transparent even with the gold layer applied (Figure 3, panel (B)). The location of the studied area is marked with red dots.

Optical microscopy images of the thin section taken with a Zeiss optical microscope are shown in Figure 3, panels (C)–(K), obtained with transmitted plane-polarized light (PPL; panels (C) and (E)), reflected cross-polarized light (XPL; panels (G) and (J)), and transmitted XPL (panels (D), (F), (H), and (K)). The sampled area is marked with a red rectangle. Images shown in panels (G)–(K) were taken after LIMS measurements. Panels (J) and (K) show close-ups of two visibly distinct regions of the sampled area. The thin section was rotated under transmitted XPL light to find the optical axis of the fluorapatite, shown in Figure A1 in Appendix A.

The sampled area was further analyzed with scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDX) using a Zeiss Gemini SEM 450 instrument located at the Department for Chemistry, Biology, and Pharmacy at the University of Bern. The SEM-EDX element maps for C, P, Ca, O, Mg, and Si are shown in Figure 4, panels (A)–(F). A composite map of the six elements is shown in panel (G). Even though the sample was gold coated, the gold layer was partially removed during LIMS analysis, leading to charging effects during SEM-EDX analysis and lower spatial resolution of the image shown in Figure 4, panel (G). Nevertheless, regions of different chemical compositions are discernible. EDX spectra of selected locations together with an SEM image of the sampled area can be found in Appendix A, Figure A2.

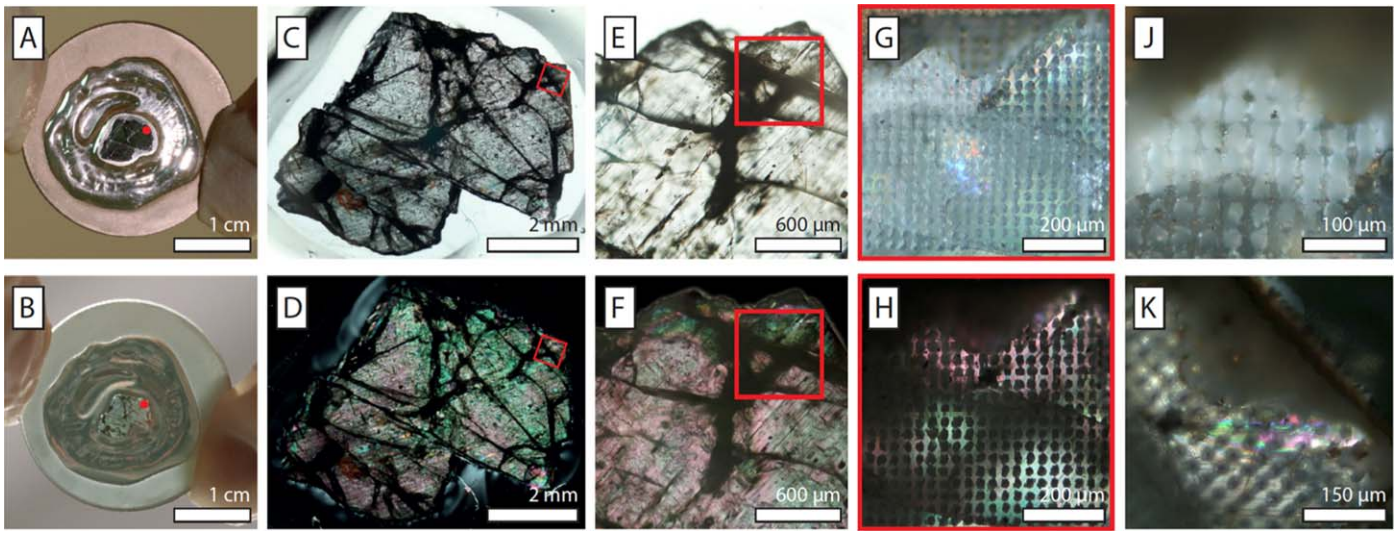

Figure 3. Optical (microscopy) images of the sampled fluorapatite thin section. (A)–(B) Sample glued on glass plate, with gold layer visible (A) and transparency of thin section highlighted (B). The red dots mark the location of the sampled area. (C)–(F) Optical microscopy images taken with transmitted plane-polarized light (PPL; C, E) and cross-polarized light (XPL; D, F). The sampled area is marked with a red rectangle. (G)–(H) Sampled area after LIMS measurements, taken with reflected PPL (G) and transmitted XPL (H). An area of 600 μm $\times$ 600 μm was sampled at 20 $\times$ 20 different locations with a pixel size of 30 μm. (J)–(K) Close-up images taken with reflected PPL (J) and transmitted XPL (K) of two visibly distinct regions of the sampled area.

3.3. Measurement Campaign

This step in the data collection and analysis procedure corresponds to Figure 1, panel (1). An area of 600 μm $\times$ 600 μm was sampled at 20 $\times$ 20 different locations with a pixel size of 30 μm. At each location, a total of 450 laser bursts were applied with a burst count of 96 single laser shots (43,200 laser shots applied per position). Data obtained in a burst were summed up during acquisition and stored in one single time-of-flight (TOF) spectrum. In accordance with the number of laser bursts applied per location, 450 TOF spectra were obtained per location.

The laser system was operated in DP mode (M. Tulej et al. 2018; A. Riedo et al. 2021), with laser pulse energies of E1 = (1.89 0.01) μJ and E2 = (2.19 $\pm$ 0.01) μJ for the first and second laser pulses applied to the sample surface, respectively. The time delay between the two pulses was set to 70 ps. The laser craters, remnants of the laser ablation process, are visible in Figure 3, panels (G)–(K). The plane of polarization of the linearly polarized laser beam was oriented at an angle of approximately 40°–50° with respect to the optical axis of the fluorapatite crystal (see Figure A1).

The sampled area comprises distinct mineral phases, i.e., solids with distinct chemical compositions and physical properties, each characterized by a distinct coloration, as evidenced in Figure 3. Notably, the uppermost region of the studied sample area, which appears blurred in the optical microscopy images (Figure 3, panels (G) and (J)), corresponds to the edge of the thin section that partially broke off during polishing. Nevertheless, it was decided to sample this region of the rock due to the occurrence of other mineral phases apart from the known primary component, fluorapatite. Based on optical microscopy analysis, the remaining part of the thin section is predominantly composed of fluorapatite. The acquisition of data from multiple mineral phases was imperative for the subsequent ML analysis.

Through the measurement of matrix-matched reference materials, the data can be quantitatively analyzed using element-wise correction with relative sensitivity coefficients (RSCs). These coefficients account for various factors like element dependence on laser wavelength, ionization potentials, and matrix effects, which affect signal intensities and cause nonstoichiometric representation of species in the raw mass spectra (M. B. Neuland et al. 2016). RSCs are comparable to calibration factors of other quantitative instrumentation, such as k-ratios for EPMA and energy-dispersive X-ray spectroscopy (EDS) analyses (R. G. Hurley & R. L. Goss 1978; X. Llovet et al. 2021) and *H*-values for PIXE analyses (J. A. Maxwell et al. 1995), which are used to correct the measured intensities to obtain quantitative results. Given the potential divergence between measured and actual compositions, selecting appropriately matched reference materials requires consideration of contextual factors, such as other mineral phases present and formation conditions. While a quantitative composition analysis represents a potential extension of the current study, it was not its primary objective, and the subsequent ML analysis can be performed using both corrected and uncorrected data. Consequently, no reference materials were subjected to measurement.

4. Data Analysis

The goal of using unsupervised ML for the analysis of an LIMS data set is to autonomously cluster the recorded spectra based on their chemical composition. While it is technically possible to use raw data for ML analysis, it is rarely advisable. Raw data often includes irrelevant features, which can hinder clustering effectiveness. Therefore, extracting, selecting, or enhancing relevant features through data preprocessing is crucial for improving clustering outcomes. Additionally, the performance of the ML model can be significantly influenced by the choice of parameters.

The following sections outline the data analysis procedure, beginning with data preprocessing (corresponding to Figure 1, panel (2)), followed by ML analysis (corresponding to Figure 1, panels (3) and (4)). Different levels of preprocessing are applied to the initial data set to assess the impact of

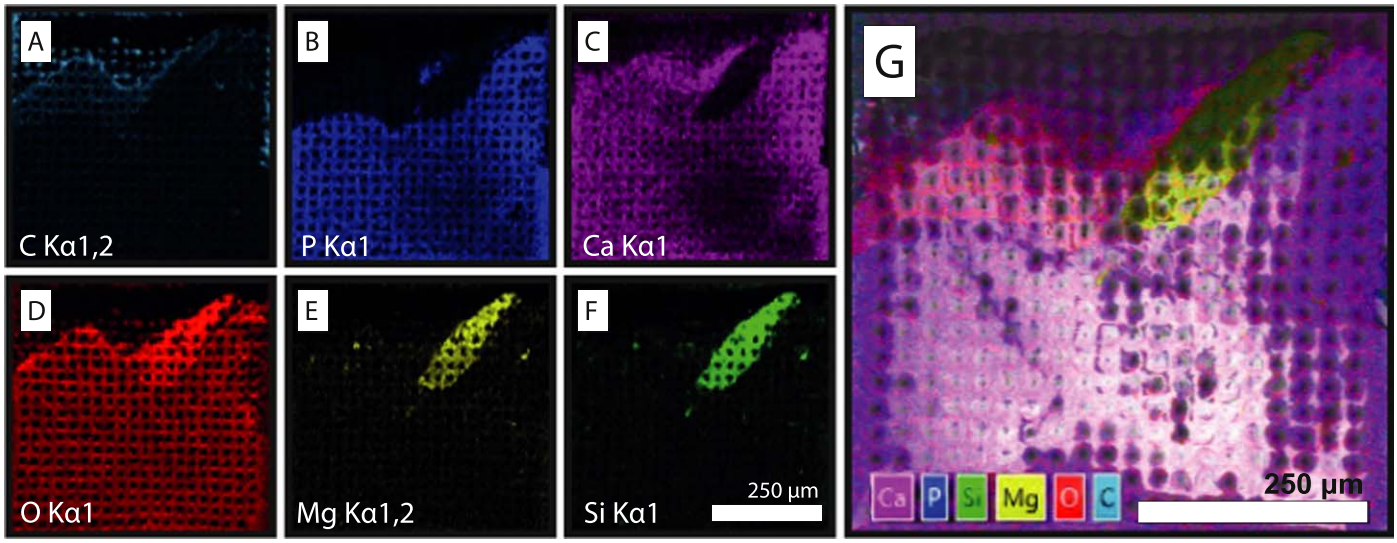

Figure 4. SEM-EDX element maps (accelerating voltage: 10 kV, probe current: 1.5 nA, magnification: 100×, backscattered electron detector). (A)–(F) Element maps of C, P, Ca, O, Mg, and Si, respectively. The corresponding emission lines are indicated in each panel. (G) Composite map of the six elements.

Table 1
Preprocessing Steps Applied to Averaged LIMS Data, Resulting in Three Differently Processed Data Sets

Processing Function	Data Set #1 (DS1)	Data Set #2 (DS2)	Data Set #3 (DS3)
Averaging	Yes	Yes	Yes
Baseline correction	No	Yes	Yes
Low-pass filtering	No	Yes	Yes
Data reduction	Downsampling	Downsampling	Mass integration
Log transformation	No	Yes	Yes
Normalization	No	Yes (1–100)	Yes (1–100)
# Features	280 (mass 5–240)	280 (mass 5–240)	240 (mass 1–240)

Note. Downsampling of DS1 and DS2 was performed with a factor of 100, meaning that every 100 consecutive features were averaged and stored as one.

preprocessing on clustering effectiveness. Clustering stability is evaluated through a grid search over a broad parameter space.

4.1. Data Preprocessing

From the raw LIMS data, three processed data sets labeled DS1, DS2, and DS3, with increasing levels of preprocessing, were generated to study the influence of preprocessing on the clustering stability. For a summary of preprocessing steps applied to DS1–DS3, please refer to Table 1. The preprocessing pipeline is visualized in Figure 5. The individual steps are explained in detail in the following, in order of application (if applied). Run times for the individual preprocessing steps on the currently used workstation (Intel Core i7-9700, 3.00 GHz base frequency, eight cores, 16 GB RAM) are summarized in Appendix C, Table A3.

4.1.1. Averaging of Data

In a first step of data preprocessing, the total number of recorded TOF spectra was reduced from 180,000 to 400 by averaging spectra at each location. Spectra recorded in bursts 1–50 from each site were discarded due to crater formation that occurs during the first few thousand laser shots, which reduces spectral quality (R. Wiesendanger et al. 2018). The remaining bursts per location, 51–450, were averaged, yielding one representative spectrum for each location. Summation of bursts increases signal-to-noise ratios (SNRs), which increases the chance of identifying species with low abundances and hence serves as feature enhancement. Furthermore, a smaller data set size speeds up subsequent preprocessing steps and the ML procedure.

4.1.2. Baseline Correction

Spectral data contain a baseline signal, which is not related to the actual sample but results from background noise or instrument drift. Baseline correction helps to subtract this unwanted background and makes it easier to detect and analyze peaks, especially those of low intensity.

4.1.3. Low-pass Filtering

SNRs can be further increased through low-pass filtering of the spectral data, where signals with a frequency higher than a selected cutoff frequency are attenuated.

4.1.4. Downsampling

The recorded raw TOF spectra consist of 64,000 initial features each (20 μs acquisition time at a sampling rate of 3.2 GS s^{-1}). For ML to work efficiently, the number of features needed to be reduced significantly, ideally without compromising relevant features. Downsampling the signal by a factor of M approximates the signal as if it were acquired at a sampling rate

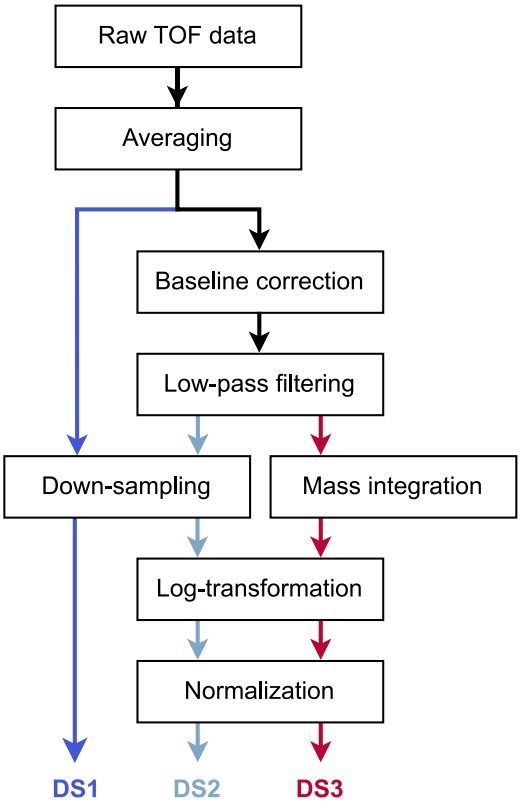

Figure 5. Pipeline of preprocessing steps applied to raw TOF data, resulting in three differently processed data sets DS1, DS2, and DS3.

M times lower. Here, $M = 100$ was chosen, meaning that every 100 consecutive features were averaged and stored as one. This approach has the advantage that it is fast and does not require prior knowledge of the TOF spectra. However, a disadvantage arises due to the quadratic relationship between the time and mass domains, resulting in a higher density of features in the lower mass range than in the higher mass range. This imbalance gives greater weight to the lower mass range during subsequent clustering analysis.

Features up to 4.2 μs (m/z of 5) were discarded due to significant electronic noise present at the beginning of each TOF spectrum. Features up to a mass-to-charge ratio (m/z) of 240 were selected, as no mass peaks were detected or are of interest at higher m/z values, which resulted in a total of 280 features per data sample.

4.1.5. Mass Integration

Unlike downsampling, mass integration reduces the number of features in the mass domain. This method has the advantage of assigning equal weight to each mass value. Although mass integration can be performed directly in the time domain, it is still necessary to determine the mass calibration constants, which requires mass calibration of at least one TOF spectrum. These calibration constants need to be recalculated if the analysis is conducted under different instrument settings.

Mass peak integration was performed according to S. Meyer et al. (2017) using integer mass integration. Each integer mass was integrated using Simpson's rule for numerical integration. Since features up to an m/z of 240 were selected, each mass-integrated data sample consisted of 240 features.

4.1.6. Log Transformation

While the major components of the studied minerals dominate the mass spectra, minor peaks can also contain valuable information. To ensure that these less abundant species contribute more meaningfully to the subsequent ML analysis, a base-10 log transformation was applied. This transformation increases the relative weight of the minor peaks, balancing their influence with the more dominant spectral components.

4.1.7. Normalization

The data samples were normalized to a range of 1–100, ensuring uniformity across the data set. This normalization facilitates easier comparison between data samples and enhances compatibility with data from other data sets when needed.

4.2. Dimensionality Reduction and Clustering

Two dimensionality reduction algorithms, UMAP and densMAP, were applied to the preprocessed data. Both algorithms aim to find a low-dimensional embedding while preserving local and global data structure. Similar samples are grouped together based on a selected metric. In this study, the cosine metric was selected, as it was found to generally provide better results than the default Euclidean metric and has been previously applied to cluster LIMS data (R. Lukmanov et al. 2022). Samples correspond to spectra in the data set with peak intensities as features. A sample with N features (e.g., $N = 240$ in DS3) can be thought of as a vector in N-dimensional vector space. Mass spectra obtained from the same mineral exhibit a similar peak pattern and hence similar features. In contrast, spectra obtained from distinct minerals with different chemical compositions display different peak patterns with less similarity. UMAP and densMAP therefore group the samples into clusters with similar chemical compositions.

DensMAP uses a modified version of the UMAP cost function to incorporate density information. The "local radius," i.e., the density around each point in the high-dimensional space, is added as an additional parameter to find a density-preserving low-dimensional embedding of the data. Unlike UMAP and the widely used t-SNE method, where the cluster size is dominantly influenced by the number of data points in the cluster, densMAP defines cluster size based on the variance of the samples within a cluster (A. Narayan et al. 2021). Dense regions of the original data remain dense in the low-dimensional densMAP embedding, and sparse regions remain sparse. While UMAP focuses on preserving structure, densMAP builds on UMAP and preserves both structure and density, making it useful when the relative density of the data is important for interpretation. The added complexity for preserving density makes densMAP computationally slower than UMAP.

After dimensionality reduction with UMAP or densMAP, the density-based clustering algorithm HDBSCAN was used to isolate clusters within the embedding. HDBSCAN can identify clusters of different shapes and sizes and is particularly useful for data sets with irregular cluster structures and variable cluster densities (Campello et al. 2013; L. McInnes et al. 2017). The same label is assigned to all spectra within a cluster. Additionally, HDBSCAN assigns a probability measure to the samples, reflecting their degree of membership within their

assigned cluster, normalized to the range from zero to one. Outliers that were not assigned to a cluster receive the label −1 and a probability of zero (L. McInnes et al. 2017). The probability measure is strongly influenced by cluster density.

Both dimensionality reduction and clustering are influenced by parameter selection. For dimensionality reduction, the most impactful parameters are the metric, n_components (the number of dimensions in the low-dimensional embedding), min_dist (the minimum distance allowed between points in the low-dimensional embedding), and n_neighbors (the size of the local neighborhood used to compute the data structure). These parameters are summarized in Table A1 in Appendix A, along with their default values and valid parameter ranges. Similarly, for HDBSCAN, the parameters found to be most influential are min_samples (the minimum number of points required to form a core cluster point. A point is considered a core point if it has at least min_samples other points within its reach) and min_cluster_size (the minimum number of points required to form a cluster).

UMAP and HDBSCAN were employed in previous LIMS studies on a 1.88 Ga old metasediment from the Gunflint formation (R. R. Lukmanov et al. 2021, 2022), where distinct clusters representing the host material (chert), organic material from embedded microfossils, and other minor minerals were retrieved. Notably, the emphasis in these earlier investigations was on achieving optimal separation between chemically distinct groups, which involved extensive efforts in parameter selection for the ML analysis, as well as rigorous data cleaning and preprocessing prior to applying the ML procedure.

4.3. Histogramming and Mineral Identification

To identify the mineral phases represented by the retrieved cluster, a representative spectrum may be selected for comparison with spectra from known minerals. Alternatively, summed histograms from the spectra within a cluster may be generated for comparison (corresponding to Figure 1, panel (5)). Summed histograms offer the advantage that they summarize information from several spectra, improving SNRs. However, clusters may comprise spectra that were assigned with a low probability, as they may contain signals from different phases, particularly in the case of spectra recorded at the boundary between two phases. Consequently, summed histograms may provide a less distinct representation of the actual cluster composition, which can challenge the identification.

Once the representative spectra have been selected, they can be used to identify the corresponding mineral phase by analyzing the element composition (corresponding to Figure 1, panel (6)). For an analysis performed in space, the representative spectra (either summed or randomly selected spectra from clusters) can be transmitted, which then provide an overview over the distinct chemistry that is present in situ, while corresponding to a significant reduction in data volume needed for obtaining such an overview.

5. Clustering Stability

The clustering stability is influenced by both data preprocessing and parameter selection for the ML models. Suboptimally selected parameters can yield a poor clustering, even for well-processed data. The same applies to poorly processed data with optimally chosen parameters. As parameter selection is a crucial step in unsupervised ML, there are different approaches to validate the clustering quality, which differ if labels are available (with ground truth) or not (without ground truth) (A. Baraldi et al. 2005; J. J. Ha et al. 2011).

For the given analysis, labels were not initially given, but as the clustering was supposed to separate the data based on the chemical composition, analyzing said composition allowed for the definition of ground truth labels. When ground truth is available, extrinsic methods can be used to assess clustering quality, which assign a score, $Q(C, Cg)$, to a clustering, C, given the ground truth, Cg (J. J. Ha et al. 2011).

Using this score, the influence of model parameter selection and data preprocessing on the cluster association was assessed, aiming to identify stable approaches that are applicable in environments where iterative parameter adjustments for improving model outcomes are limited or impossible, e.g., on board a spacecraft.

In the following, the determination of the ground truth labels as well as the definition of the scoring function are presented.

5.1. Ground Truth

To evaluate the autonomously obtained clustering outcome, clusters were manually established as a reference, the so-called ground truth of the data set. In this context, the objective is to categorize the data set through clusters that reflect distinct chemical compositions.

Manual cluster assignment was performed using the chemical composition obtained from the optical microscopy images (Figure 3, panels (G)–(K)), the SEM-EDX element maps (Figure 4), and LIMS element maps (Figure 6, panels (A)–(E)). The microscopy images show the presence of four major distinct regions that also show up in the element maps of both EDX and LIMS. Based on this information, the chemically distinct regions were determined, and the derived ground truth is summarized in Figure 6, panel (F). Manual label assignment was only possible due to the rather small size of the data set. Minor mineral phases can also be identified from the element maps, as exemplified in panel (B) of Figure 6 for Zr, where enrichment is evident approximately in the middle of the sampled area. However, such minor phases were not individually marked as distinct phases in Figure 6, panel (F), as the major phases are prominently evident in the concerned regions. The unsupervised ML technique is expected to identify the same chemical composition patterns.

5.2. Scoring Function

To assess the quality of a clustering, a scoring function was defined as follows:

$$Q = \frac{N_C + N_O}{N}. \quad (1)$$

Here, N is the total number of samples (the size of the data set), N_C is the number of correctly labeled samples, and N_O is the number of samples labeled as outliers. To determine N_C, the clustering is compared to the ground truth. The function is defined in a way to ensure that samples labeled as outliers would not contribute to the count of wrong labels, unless two or more phases were primarily labeled as outliers, hindering phase separation. This approach avoids penalizing the model when a small number of outliers is found, as they may entail information about small mineral inclusions. Such minerals

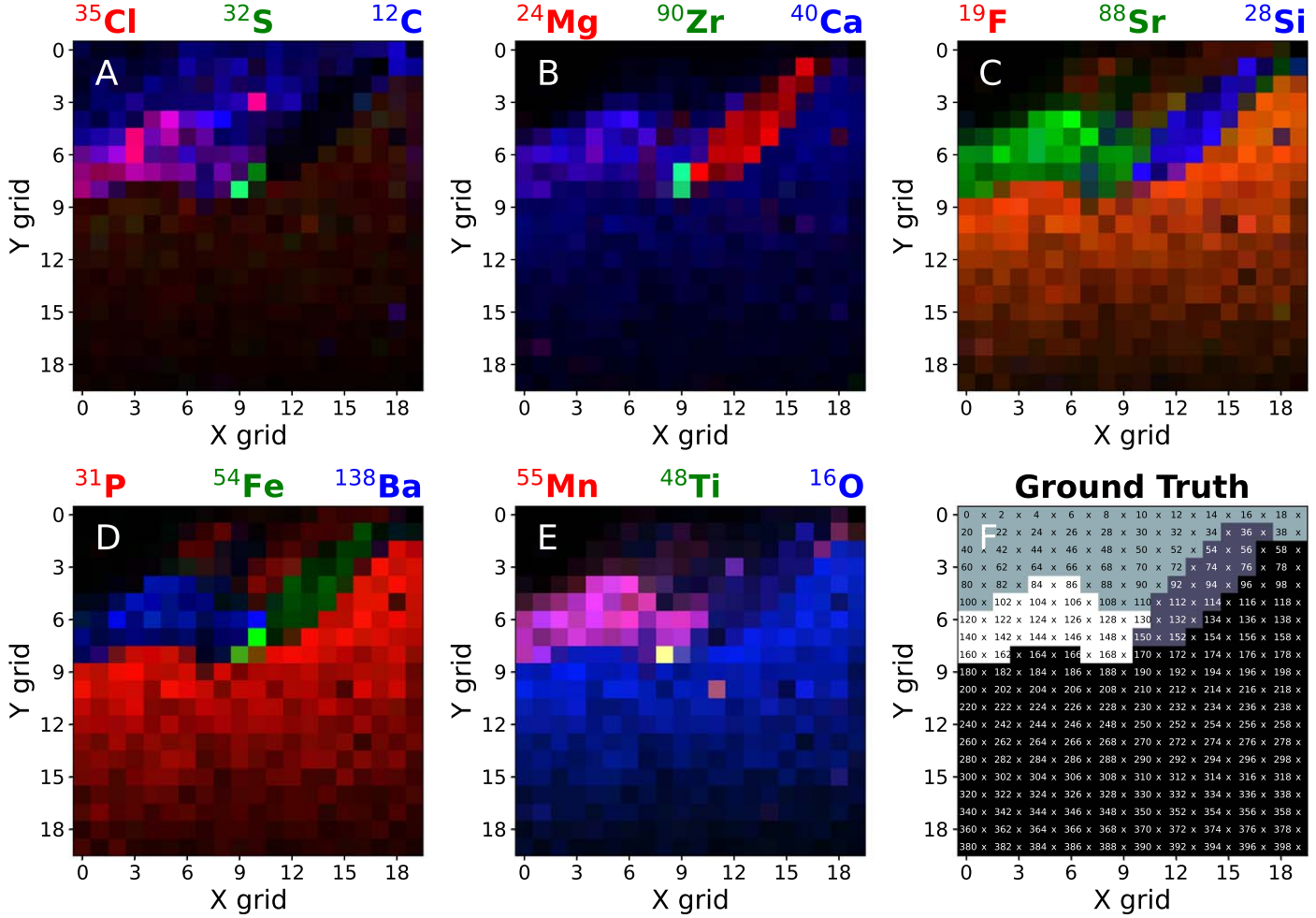

Figure 6. Element maps (panels (A)–(E)) and sampled location numbers colored according to chemically distinct regions (panel (F)). Peak intensities of each element were normalized across the sampled area. *x*-axes and *y*-axes tick marks help guide the eye to find the sampled location in the grid of 20 × 20 locations.

might be present in single locations only; however, one single sample cannot form a cluster and will therefore not be assigned an individual label, but rather be labeled as an outlier.

For spectra acquired from regions where two or more mineral phases are in contact, labels from all neighboring phases are accepted. Spectral signatures from multiple phases may be present in a border sample, and a clear assignment to a single phase is not feasible. If two or more clusters that should remain separate according to the ground truth get merged, all labels from the corresponding clusters, including any outlier labels, will be considered wrongly assigned. When phases are isolated but split into subclusters, only the biggest subcluster will be considered correctly assigned. In the special case where all samples were to be labeled as outliers, which could occur due to highly unfavorable parameter choices, the score would be assigned a value of zero. None of the tested parameter sets resulted in this particular case.

Minor differences between autonomously and manually assigned labels are to be expected, as neither clustering can be achieved with absolute certainty. Apart from spectra obtained from border regions of mineral phases, discrepancies can typically be attributed to suboptimal model parameter selections. This can cause underclustering, i.e., the merging of chemically distinct phases, or overclustering, i.e., the separation of phases that demonstrate a high degree of chemical homogeneity. Examples for poor cluster assignments as well as cluster assignments containing outliers are given in Figure A3 in Appendix A.

The possibility that the ground truth may not be entirely accurate needs to be considered, as this would impact the effectiveness of comparing the autonomously obtained clustering to the manual one using the scoring function defined in Equation (1). However, considering that the chemical composition is depicted by the element maps, which reflect the data used as input for the clustering algorithm, clustering outcomes that reveal patterns significantly distinct from those observed in the element maps are not anticipated.

5.3. Grid Search

After having defined the ground truth and a scoring function, the influence of parameter selection on the quality of clustering was assessed. Parameter selection is a crucial step for unsupervised ML methods, and the stability of a clustering with respect to parameter selection can be controlled through grid searches across the parameter space. Furthermore, this approach helps to identify parameters that exert a substantial impact on the clustering and distinguish them from parameters that have little to insignificant influence on the results.

Grid searches can be computationally expensive, as the same computations are rerun many times with slightly adapted parameters. Especially in resource-limited environments, grid

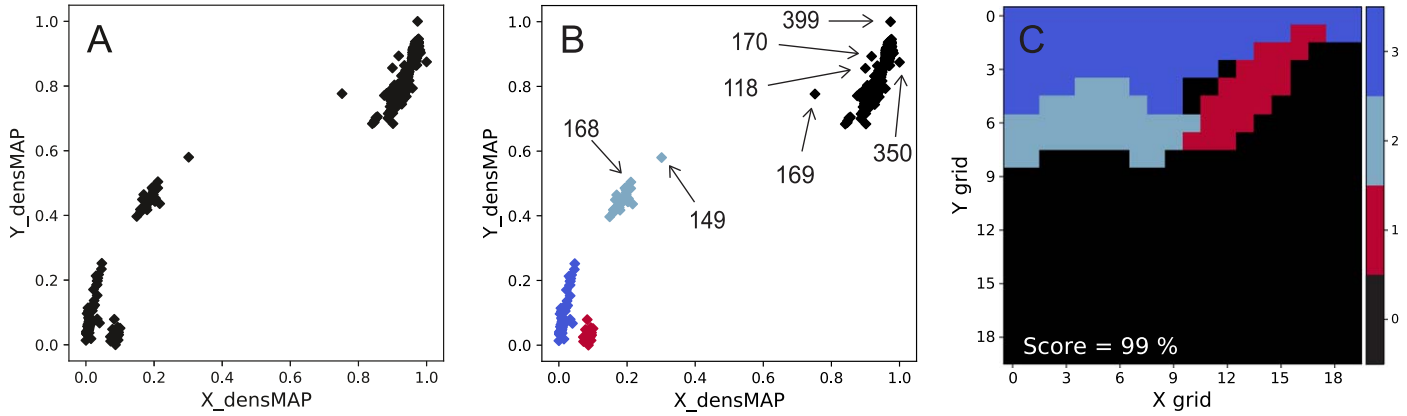

Figure 7. Data embedding using densMAP (A). The first two densMAP dimensions are shown. Subsequently, the embedding was clustered using HDBSCAN (B). Outliers are marked with the respective sampled location. Panel (C) shows the sampled area colored according to the retrieved HDBSCAN labels.

searches might be impossible or limited; hence, suitable parameter estimation for new, unseen data becomes crucial. Alternatively, preparatory steps such as parameter testing on, e.g., synthetic data in environments with fewer resource constraints can be taken.

Table A1 in Appendix A presents a summary of the UMAP and densMAP parameters min_dist, n_neighbors, and n_epochs, and the HDBSCAN parameters min_samples and min_cluster_size that were varied during the conducted grid search. A total of 7596 parameter sets were tested. The explored parameter space is summarized in Table A2 in Appendix A. These parameters play a crucial role in shaping the data embedding and subsequent clustering. They are interdependent, with embedding quality influencing cluster isolation. All parameter sets were tested on the three data sets DS1–DS3, offering further insights into the influence of preprocessing on clustering stability. The grid search was conducted both with UMAP and densMAP. Due to computational constraints, not all parameter values were varied in relation to one another; instead, each target parameter was tested across a consistent set of combinations for the other parameters to observe specific trends efficiently.

6. Results and Discussion

6.1. Data Embedding and Clustering

The retrieved two-dimensional densMAP embedding for data set DS3 is shown in Figure 7, panel (A). When clustering this embedding using HDBSCAN, the samples are assigned a label, depicted with different cluster colors in Figure 7, panel (B). As each data point in the embedding corresponds to a sampled location, the sampled area can be colored according to the retrieved labels, as shown in Figure 7, panel (C). For the given data set, four clusters were retrieved, pointing toward the presence of four distinct mineral phases.

In the low-dimensional embedding (Figure 7, panel (B)), outlier samples are marked. These samples are only loosely connected with their associated cluster. While denser clusters might initially seem preferable, detection of outliers is of great interest, as they provide additional constraint into the mineralogy of a rock sample. Their loose association with specific clusters suggests the presence of spectral signatures similar to those of other samples in the cluster. However, they must in addition contain spectral signatures that set them apart. Consequently, these outliers commonly indicate the presence of minor mineral phases in the sampled material. Locations can be retrieved from Figure 6, panel (F).

6.2. Phase Identification

The next step in the analysis sequence concerns the identification of the mineral phases of each of the retrieved clusters. For this, a representative mass spectrum was selected from each cluster and is displayed in Figure 8, panels (A)–(D). The corresponding cluster areas are indicated in small inserts in the upper-right corners. By analyzing the element species present in the mass spectra combined with educated guesses about which minerals are likely to co-occur, the clusters were identified as fluorapatite (A), forsterite (B), and calcite (C). The last spectrum (D) shows contributions from both calcite and fluorapatite, but as previously discussed, this region corresponds to the edge of the thin section analyzed under a nonoptimal laser focus, reflected by the low spectrum intensity. No further detailed analysis will be conducted for this region.

Fluorapatite is the only F-bearing rock-forming mineral containing O, P, and Ca without significant amounts of other elements. Minor traces of Cl were also found below the percent level, supporting the identification of fluorapatite with minor contributions of chlorapatite. Apatite is the most ubiquitous phosphate mineral found in sedimentary, igneous, and metamorphic rocks (P. M. Ihlen et al. 2014; M. Ouabid et al. 2021). Although apatite is usually just an accessory mineral, it can be a major rock-forming mineral in certain rocks, including phoscorites and carbonatitic rocks. In phoscorites, inclusions of carbonatite are very common, as shown in Figure 8, panel (C). The carbonatite phase is calcite. Phoscorites may further contain olivine and magnetite, and the composition of panel (B) (major Mg, Si, and O, with some Ca and minor Na, and K) strongly suggests the presence of olivine. The absence of Al rules out phlogopite, and the low Fe content is indicative of the dominance of the forsterite component in the olivine. Olivine contains minor amounts of Ca, which is common in olivine crystallized in a carbonatite melt (G. Libourel 1999). Minor chondrodite might be present because of the F-rich melt. The detected forsterite crystal has the approximate dimensions of 400 mμm $\times$ 120 μm. No magnetite has been detected.

Closer analysis of the outlier spectra found in the densMAP embedding (see Figure 7, panel (B)) corresponding to locations 118, 149, 168, 169, 170, 350, and 399 reveal the presence of minor mineral phases. These include pyrite (FeS_2, locations 169 and 170), rutile (TiO_2, locations 168 and 169), baddeleyite

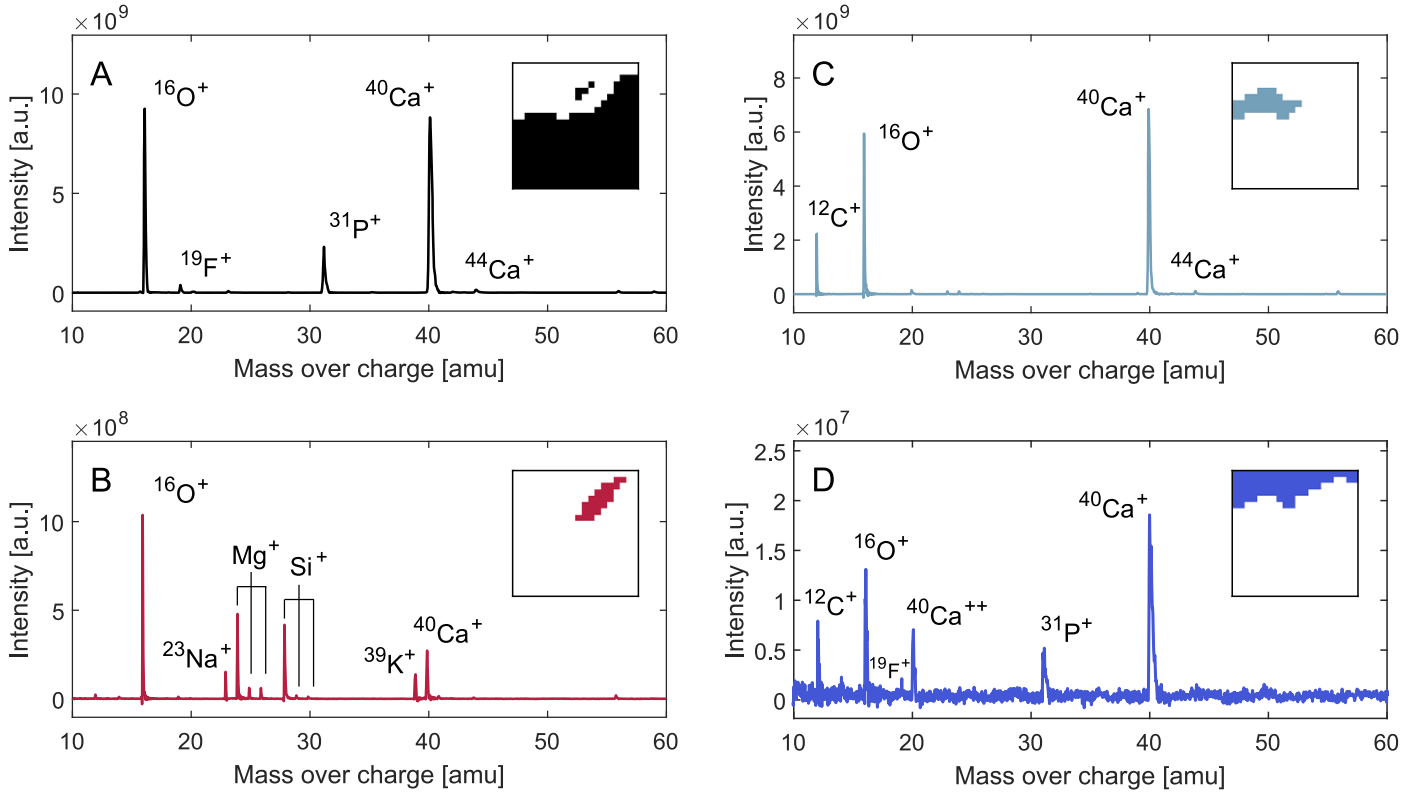

Figure 8. Identification of major mineral phases and minor inclusions. Selected mass spectra for the four major clusters. (A) Location 203. (B) Location 151. (C) Location 126. (D) Location 41. Major peaks are labeled. Note the different *y*-axis ranges.

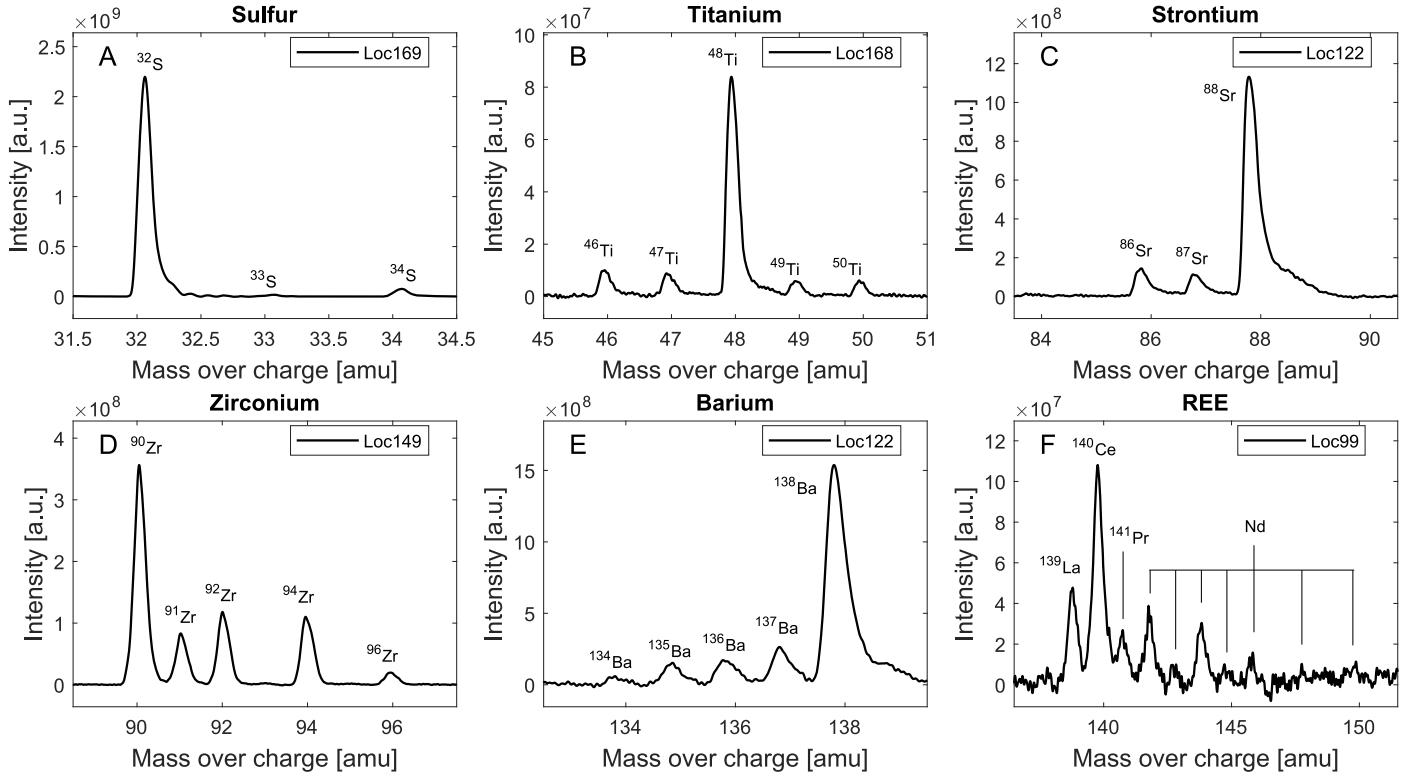

Figure 9. Isotope patterns found in mass spectra labeled as noise spectra, as well as Sr, Ba, and REEs.

(ZrO_2, locations 149 and 170), and uranothorianite ((Th, U)O_2, locations 350 and 399, minor traces in 118 and 170). Mineral phases detected from the corresponding mass spectra are shown in Figure 9, panels (A), (B), and (D). Additionally, mass peaks of REEs (Ce, La, Pr, Nd) detected primarily in the fluorapatite region are displayed (Figure 9, panel (F)), in addition to Sr and

Ba primarily found in the carbonatitic phase (Figure 9, panels (C) and (E), respectively). For spectrum numbers, see Figure 6, panel (F). These spectra contain spectral signatures of trace mineral phases, which provide geological and geochemical information on the formation processes of the host rock and preserve crucial information on its geological history (S. Ferrero & R. J. Angel 2018; W. M. White 2020). If such an analysis were to be conducted on board a spacecraft, transmitting spectral data that includes both major and trace element information would be immensely informative for the remote exploration of other planetary bodies.

If the data presented in this study were collected on a spacecraft, four spectra would have to be sent back for the four major phases, and three to four spectra for the interesting locations that were found containing spectral signatures of minor mineral phases. These seven to eight spectra give a complete overview of the encountered mineralogical complexity of the rock sample, corresponding to a significant reduction in data volume. If technically feasible, the subsequent transmission of raw data associated with clusters of interest could substantially enhance the collection of valuable data.

Modern spacecraft provide a service called "selective downlink" of data that would serve this purpose. However, to apply this approach on board a spacecraft to respective data, modifications would need to be undertaken to adapt the method to perform well on resource-constrained hardware. If onboard storage will suffice to keep the raw data alongside the processed data until after clustering has been performed, the clustering can be used to identify regions of interest on the studied sample. Subsequently, more data could be collected from these interesting regions and transmitted as raw spectra for more detailed analysis.

6.3. Comparison of UMAP with DensMAP

Clustering results obtained both with UMAP and densMAP for data set DS3 are presented in Figure 10 in panels (U1) and (D1), respectively. The retrieved embeddings are color coded by HDBSCAN labels. Figure 10, panels (U2) and (D2), show the sampled area colored according to the corresponding HDBSCAN probabilities for UMAP and densMAP, respectively. The sampled area, colored according to HDBSCAN labels, is identical for both methods and agrees with Figure 7, panel (C). The same model parameters were used to obtain the embeddings (n_components = 3, metric = cosine, min_dist = 0.1, n_neighbors = 50, n_epochs = 2500), as well as the HDBSCAN clustering results (min_samples = 5, min_cluster_size = 20). A detailed description of these parameters is given in Table A1 in Appendix A.

The comparison of the embeddings obtained with UMAP and densMAP reveals a high degree of agreement, with a total of four clusters retrieved and identical cluster labels assigned. A more prominent distinction emerged in the sizes of the clusters. Specifically, the UMAP-derived embedding (Figure 10, panel (U1)) yielded clusters with higher densities compared to the corresponding embedding obtained with densMAP (Figure 10, panel (D1)). The correlation between cluster size and number of associated samples is much more apparent in the UMAP-generated embedding than in the densMAP-generated one. This has a direct impact on the associated HDBSCAN probabilities, as samples in denser clusters tend to be associated with a higher probability, with the corresponding clusters compared to samples in more dispersed clusters. This can be directly observed in Figure 10, panels (U2) and (D2), respectively. Hence, the higher cluster densities in the embedding obtained with UMAP lead to higher overall clustering probabilities compared to densMAP. Higher probabilities do not necessarily imply a more accurate embedding, as they are influenced by cluster density. As only densMAP preserves density, the probability parameter of the density-based clustering algorithm HDBSCAN seems to have less significance in combination with UMAP compared to densMAP.

Another notable difference between the UMAP and densMAP embeddings is the presence of outliers in the latter, which consequently appears noisier than the UMAP embedding, where clusters are densely packed. While denser clusters might initially seem preferable, detection of outliers is of great interest, as they provide additional insights into the mineralogy of a rock sample, as previously discussed. As outliers scatter far from their associated cluster, the corresponding HDBSCAN probabilities are low as well (see Figure 10, panel (D2)). In the corresponding panel for the UMAP embedding (Figure 10, panel (U2)), certain locations with lower probabilities coincide with those detected as outliers with densMAP. Finding spectra with spectral signatures of different mineral phases using UMAP can therefore not be completely ruled out. However, in this particular case, certain interesting locations found with densMAP, such as 350 and 399, could not be retrieved with UMAP.

Spectra acquired from regions situated between two different minerals may encompass signatures from both phases. These spectra, along with those with minimal or no signal intensity, may be less reliably embedded and only loosely attached to a cluster, resulting in low HDBSCAN probabilities. This is particularly evident in most spectra of the dark-blue cluster in the uppermost region of the sampled area (Figure 10, panels (U2) and (D2)). During sample preparation, this edge region of the rock partially broke off while polishing the surface, leaving it uneven. The thin section was positioned such that the main, flat area was in optimal laser focus, while the uneven edge region was not. Since the laser focus was not adjusted throughout the measurement campaign, the edge region was sampled outside the optimal laser focus. This resulted in predominantly low-intensity spectra and consequently lower HDBSCAN probabilities.

6.4. Validation of the Clustering Stability

To validate that the obtained clustering results are not specific to the selected parameters but can be reproduced over a broad parameter space, the clustering stability was assessed through a grid search. The score function defined in Equation (1) was used to quantify the clustering quality. Simultaneously, the influence of preprocessing on clustering quality was studied.

The model parameters selected for the grid search include the UMAP/densMAP parameters min_dist, n_neighbors, and n_epochs, and the HDBSCAN parameters min_samples and min_cluster_size. The parameters are summarized and described in Table A1 in Appendix A. The explored parameter space is summarized in Table A2 in Appendix A.

Grid search results are summarized in Figure 11. A total of 7596 parameter sets were evaluated. Each panel in Figure 11 displays mean scores achieved for a selected parameter for combinations with the other four parameters, both for UMAP

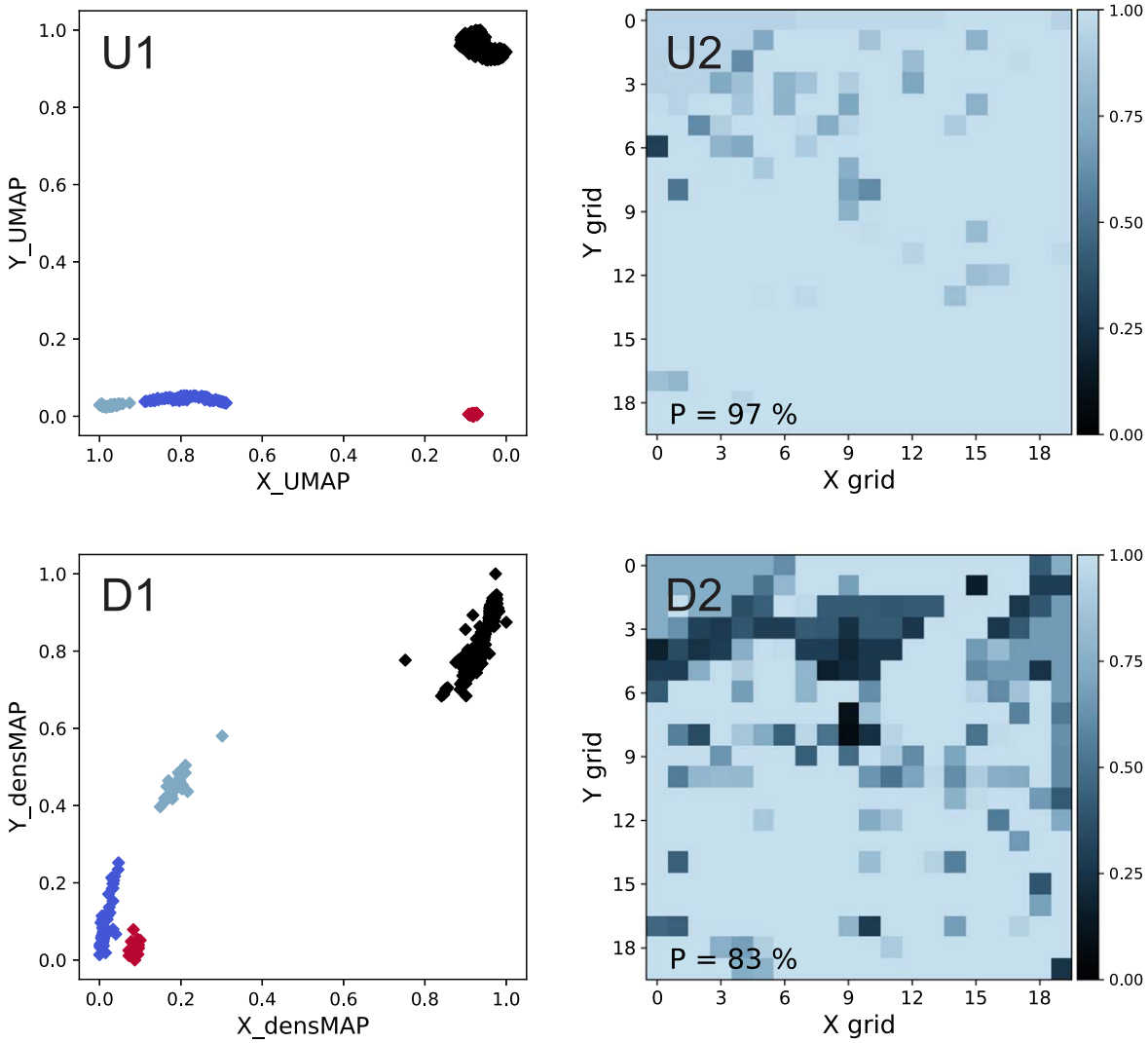

Figure 10. Clustering results of the dimensionality reduction analyses using UMAP (U1) and densMAP (D1) for data set DS3. The corresponding UMAP and densMAP embeddings are shown using the first two UMAP and densMAP dimensions, respectively (i.e., X_UMAP refers to the first UMAP dimension). Panels (U2) and (D2) show the sampled area colored according to the assigned HDBSCAN probabilities, with mean probabilities annotated in the lower-left panel corner.

(upper panels) and densMAP (lower panels). A corresponding figure displaying the HDBSCAN probabilities instead of mean scores is presented in Appendix A, Figure A4.

6.4.1. Stability Over Parameter Space

The clustering score is stable over a large parameter space, with drop-offs occurring toward the higher end of the searched space for the parameters min_dist and min_cluster_size (Figure 11, panels (A1) and (A2), and (E1) and (E2), toward the lower end of the searched space for n_neighbors (Figure 11, panels (B1) and (B2)), and on both ends of the searched space for min_samples (Figure 11, panels (D1) and (D2)). The score shows no trend with the parameter n_epochs (Figure 11, panels (C1) and (C2)). The score drop-offs occur due to under-clustering (the loss of clusters through merging) or over-clustering (more clusters identified than what might be considered meaningful or reflective of the underlying structure). Examples for under- and overclustering are provided in Figure A3 in Appendix A, panels (D1) and (D2), and (C1) and (C2), respectively.

When min_dist is set too high, clusters in the low-dimensional embedding become more dispersed, leading to increased separation between points and a loss of local structure. This results in reduced cluster compactness, resulting in merged or overlapping clusters in the embedding, and hence leads to underclustering (Figure 11, panels (A1) and (A2), and Figure A3, panels (D1) and (D2)). If n_neighbors is set to a too small value, the embedding may focus excessively on local structures, leading to increased sensitivity to noise and fragmentation of clusters. This can result in the identification of many small, disconnected groups while overlooking larger, more meaningful patterns in the data (Figure 11, panels (B1) and (B2), and Figure A3, panels (C1) and (C2)). If min_samples is set too low, the algorithm may create overly dense, disconnected clusters leading to overclustering (Figure 11, panels (D1) and (D2)). Conversely, if set too high, more points may be classified as noise, resulting in larger clusters and potentially missing smaller significant groups (underclustering). Underclustering also occurs for too large values of min_cluster_size, as the algorithm may overlook smaller, meaningful clusters, resulting in fewer, larger clusters (Figure 11, panels (E1) and (E2)).

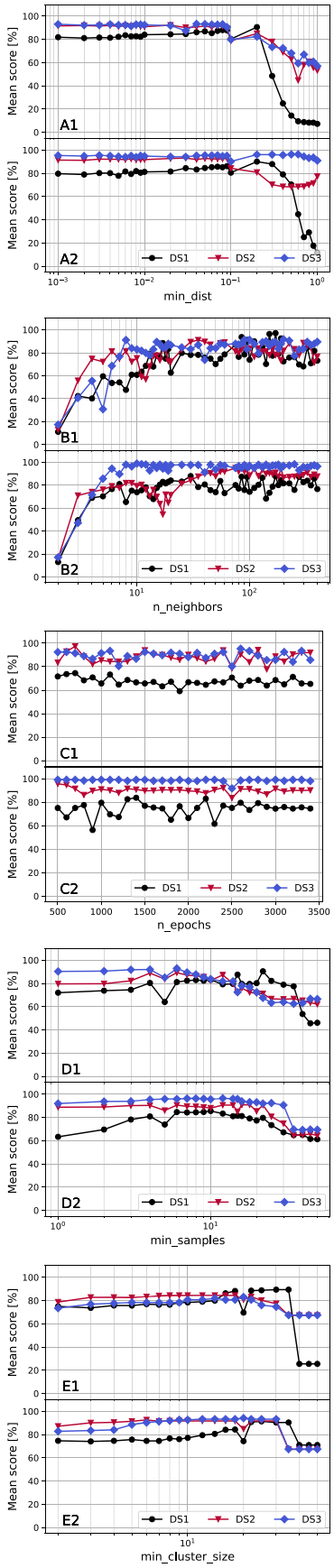

Figure 11. Summary of the grid search. Mean scores are shown for all three data sets DS1–DS3, with UMAP (upper panels) and densMAP (lower panels). The *x*-axis labels indicate the varied parameter (A: min_dist, B: n_neighbors, C: n_epochs, D: min_samples, and E: min_cluster_size).

Optimally, for an analysis conducted on board a spacecraft, the obtained embedding could be transferred to Earth, where various parameter values for min_samples and min_cluster_size could be tested to find an optimal clustering. Alternatively, the data embedding could be rerun with different parameters on board the spacecraft. Overall, low scores were predominantly encountered toward extreme parameter choices, e.g., min_dist > 0.1, n_neighbors < 10, min_samples > 20, and min_cluster_size > 30. For moderate parameter choices, the overall clustering exhibited a high level of stability.

6.4.2. Stability with Respect to Preprocessing Levels

Mean scores demonstrate a correlation with data-preprocessing levels, reflecting the impact of the noise level on the clustering. Higher preprocessing levels (DS1 $\rightarrow$ DS2 $\rightarrow$ DS3) reduce noise and homogenize data more efficiently, yielding higher scores. While mean scores obtained with DS2 and DS3 show similar parameter dependencies (e.g., Figure 11, panel (C1)), reflecting their similar preprocessing levels, including noise reduction through low-pass filtering, DS1 received overall lower scores (e.g., Figure 11, panel (A1)). The lower mean scores and lower score stabilities for DS1 suggest model sensitivity to noise independent of parameter selection. Figure 11, panel (C1), shows that the score does not depend on the selection of n_epochs but has a strong dependence on preprocessing levels as values obtained for DS1 are overall lower and less stable than for the other two data sets. Therefore, poor data preprocessing can only be remedied to a limited extent with optimal parameter selection, while well-processed data shows more resilience toward suboptimal parameter selection.

Table 2 summarizes the number of parameter sets yielding scores $\geqslant$95% with both UMAP and densMAP for the three data sets DS1, DS2, and DS3. The total scores depict the mean scores obtained for a data set, independent of the dimensionality reduction technique. Higher preprocessing levels generally yield scores $\geqslant$95% for more parameter sets. DS1, with the least amount of preprocessing, yielded overall very low scores, particularly when analyzed with densMAP. No low-pass filtering or other noise reduction was applied to DS1; therefore, the data remained relatively noisy compared to DS2 and DS3. As densMAP takes the variance of the samples into account when constructing the embedding, a higher dependence on noise for densMAP than UMAP is expected. For data with low SNRs, UMAP is therefore expected to yield more stable results. The decision whether to apply UMAP or densMAP to a data set is therefore case dependent. Conversely, DS3, with the highest preprocessing level, achieved the highest overall scores with densMAP. This highlights the stability of the method when coupled with effective data set preprocessing. DS2, despite being downsampled like DS1 and not mass integrated like DS3, achieved more similar scores to DS3. This is noteworthy, as log transformation, applied to DS2 and DS3 and aimed at giving low-abundant species more weight in the clustering process, simultaneously gives noise values more weight, which could potentially have a negative impact on the results. Although not encountered in this study, this needs to be kept in mind, especially when working with low SNR data.

The comparison with Figure A4 in Appendix A portraying the same grid search as in Figure 11, but evaluated using the HDBSCAN probability measure, reinforces the claim made previously, that the probability parameter of HDBSCAN

Table 2
Number of Parameter Sets Yielding Scores ⩾95% in Conducted Grid Search

Data Set	UMAP	densMAP	Total
DS1	24%	8%	16%
DS2	47%	49%	48%
DS3	55%	83%	69%

Note. Total scores correspond to mean scores obtained for a data set, independent of UMAP or densMAP. Scores are rounded to integer precision.

appears to have less significance when combined with UMAP rather than with densMAP. Overall, the probabilities obtained with UMAP are higher than with densMAP; however, this does not necessarily imply that UMAP embeddings are more accurate, as revealed by a comparison with the obtained scores. The probability measure displays less variability across the different parameters, creating a misleading sense of stability, as low scores were still obtained even when mean probabilities were high.

To validate the approach, a cross-check with another data set was performed and can be found in Appendix B. A run-time analysis was conducted for the various preprocessing and ML steps and can be found in Appendix C, with results presented in Table C1. This analysis showed that being economical by limiting preprocessing does not pay off as mean scores obtained for DS1 were rather low, while total processing times per spectrum are only slightly faster compared to DS3.

7. Conclusion

The acquisition of data from space is often limited by downlink rates, and transmitting data that is of low quality or not of interest reduces mission return. Prefiltering of data on board spacecraft optimizes the use of limited resources such as onboard storage and time. This study presents a possible analysis approach for mass spectrometric data collected on board spacecraft from geological samples using unsupervised ML for autonomous mineral phase separation with the objective to improve mission return.

Data were acquired from a thin section of a terrestrial rock containing different mineral phases dominated by fluorapatite using a space-prototype LIMS instrument. The collected data were preprocessed and subsequently dimensionality reduced and clustered. This approach autonomously isolated four major clusters, which were then identified as the mineral phases fluorapatite, calcite, and forsterite-rich olivine, as well as a region corresponding to the edge of the thin section. In addition, outliers corresponding to minor mineral phases, present in single locations, were found and identified as pyrite, rutile, baddeleyite, and uranothorianite.

The separation of the data into distinct groups and their subsequent identification through the selection of representative mass spectra provides a preview of the diverse mineralogical compositions and mineral distributions of the studied thin section. Allowing for the detection, localization, and identification of major and minor phases through the analysis of only a handful of mass spectra corresponds to a significant reduction in data volume. For the current study, only four spectra are required for the identification of the major mineral phases. The application of the given or a similar analysis technique would therefore be useful on a spacecraft, particularly when looking for low-abundant features. If a distinction between spectra of high and low interest could be achieved already on the spacecraft, downlink capacity limitations would become less problematic and onboard storage could be optimized.

The clustering stability was assessed through a grid search over a broad parameter space, and the influence of data preprocessing, the dimensionality reduction technique, and parameter selection was analyzed. Through a parallel manual chemical composition analysis, the ground truth was determined to score the clustering. Testing the approach on three data sets with different levels of preprocessing demonstrated that higher levels of preprocessing stabilize the clustering results over a larger parameter space. A significant difference in clustering stability was observed between data that underwent noise reduction and homogenization steps, such as low-pass filtering and baseline correction, and data that did not, with the latter yielding less stable clustering results. As these data-preprocessing steps are computationally much less expensive than dimensionality reduction and clustering, saving a relatively small amount of time through very minimal preprocessing is not advised.

The best results were achieved with a mass-integrated data set (DS3), for which preprocessing took only slightly longer than for the corresponding downsampled data set (DS2). However, for a corresponding analysis conducted on board a spacecraft, communication with the spacecraft would be required to mass calibrate data, potentially slowing down the analysis significantly. In such a case, downsampling of the data might be preferred, at the cost of slightly lower clustering stability. For analyses conducted on Earth, however, mass integration of the data is the preferred method. The grid search showed that, for mass-integrated data, the clustering results are more stable over a broad range of parameters compared to downsampled data. The stability of the method was further tested on an additional data set, where phase isolation was successfully achieved using the same preprocessing approach as for DS3. For much larger data sets, the adaptation of certain parameters is anticipated.

This study demonstrated that using the clustering algorithm HDBSCAN on data reduced with UMAP or densMAP has effectively isolated the major mineral phases. Minor mineral phases were only found when using densMAP. As densMAP is a density-preserving algorithm that considers the variance of data samples within a cluster when constructing the embedding, corresponding cluster densities were found to be lower compared to UMAP. This is an advantage for outlier detection, but a disadvantage for data exhibiting low SNRs, as densMAP is more sensitive to noise. As log transformation of the data increases the weight of noisy values on the clustering outcome, this preprocessing step can become problematic when coupled with densMAP for data exhibiting small SNRs. Depending on the data set at hand, one of the other methods may be preferred. It also needs to be noted that densMAP, being an extension of UMAP, is computationally more resource intensive.

For the future, the development of a computationally less expensive clustering approach applicable on resource-constrained hardware should be pursued. Furthermore, studies with supervised models might be considered as well. For instance, a model could be trained on Earth on a subset of data collected on a spacecraft and later used on board to

directly classify the remaining data. No onboard training would be required.

Acknowledgments

This research was funded by the Swiss National Science Foundation (SNSF) grant No. 200020_207409. Special thanks go to Beatrice Frey from the Department of Chemistry, Biochemistry, and Pharmaceutical Sciences at the University of Bern, Switzerland, for operating the sputter coater and the SEM-EDS instrument.

Appendix A

Figure A1 shows optical microscopy images of the fluorapatite thin section taken with transmitted cross-polarised light (XPL) at different angles of orientation. Figure A2 shows an SEM image of the sampled area and the different chemical compositions of four different regions marked on the sample can be deduced from the EDX spectra shown. Table A1 summarizes the main UMAP, densMAP and HDBSCAN parameters that were varied in this work, giving a description of the parameters together with their default and possible values. Table A2 summarizes the parameter space explored in the grid search. Values are given for the varied parameters min_dist, n_epochs, n_neighbors, min_samples, min_cluster_size, and densmap. Figure A3 shows a selection of different embeddings and cluster assignments of the sampled region for different ML parameter sets. The choice of parameters strongly influences the clustering results, with suboptimal parameter choices leading to under- and overclustering. Parameter selection also affects the detection of outliers. Figure A4 shows the result of the grid search for the HDBSCAN probability measure, analogous to the scores shown in Figure 11.

Table A1
Description of Major UMAP, densMAP, and HDBSCAN Parameters

Parameter	Description	Possible Values	Default Value
densMAP	Dimensionality reduction technique (UMAP (false) or densMAP (true))	True or false (Boolean)	False
n_components	Number of dimensions in the reduced space	$1 < n < N$ with $n \in \mathbb{N}$ and $N = \#$samples (integer)	2
metric	Metric used to compute distances in high-dimensional space	Euclidean, cosine, Manhattan, Jaccard, … (string)	Euclidean
n_neighbors	Number of neighbors used for local data structure	$1 < n < N$ with $n \in \mathbb{N}$ and $N = \#$samples (integer)	15
n_epochs	Number of optimization iterations during projection	$n \in \mathbb{N}_0$ (integer)	500 (<10,000 samples) 200 (>= 10,000 samples)
min_dist	Minimum distance between points in low-dimensional space	$r \in \mathbb{R}_0$ (float)	0.1
min_samples	Minimum number of samples required in a neighborhood to consider it dense	$n \in \mathbb{N}$ (integer)	"None" (same value as min_cluster_size)
min_cluster_size	Minimum number of points required to form a cluster	$n \in \mathbb{N} > 1$ (integer)	5

Table A2
Explored Parameters in Grid Search

min_dist	n_epochs	n_neighbors	min_samples	min_cluster_size	densMAP
1, 0.9, 0.8, …, 0.1, 0.09, 0.08, …, 0.01, 0.009, 0.008, …, 0.001	500, 600, 700, …, 3400	2, 3, 4, …, 12, 14, 16, …, 22, 25, 30, 35, …, 60, 70, 80, 90, 100, 120, 140, …, 200, 225, 250, 300, 350, 399	1, 2, 3, …, 10, 12, 14, 15, 16, 18, …, 22, 25, 30, 35, …, 50	2, 3, 4…, 10, 12, 14, …, 22, 25, 30, 35, …, 50	True, false

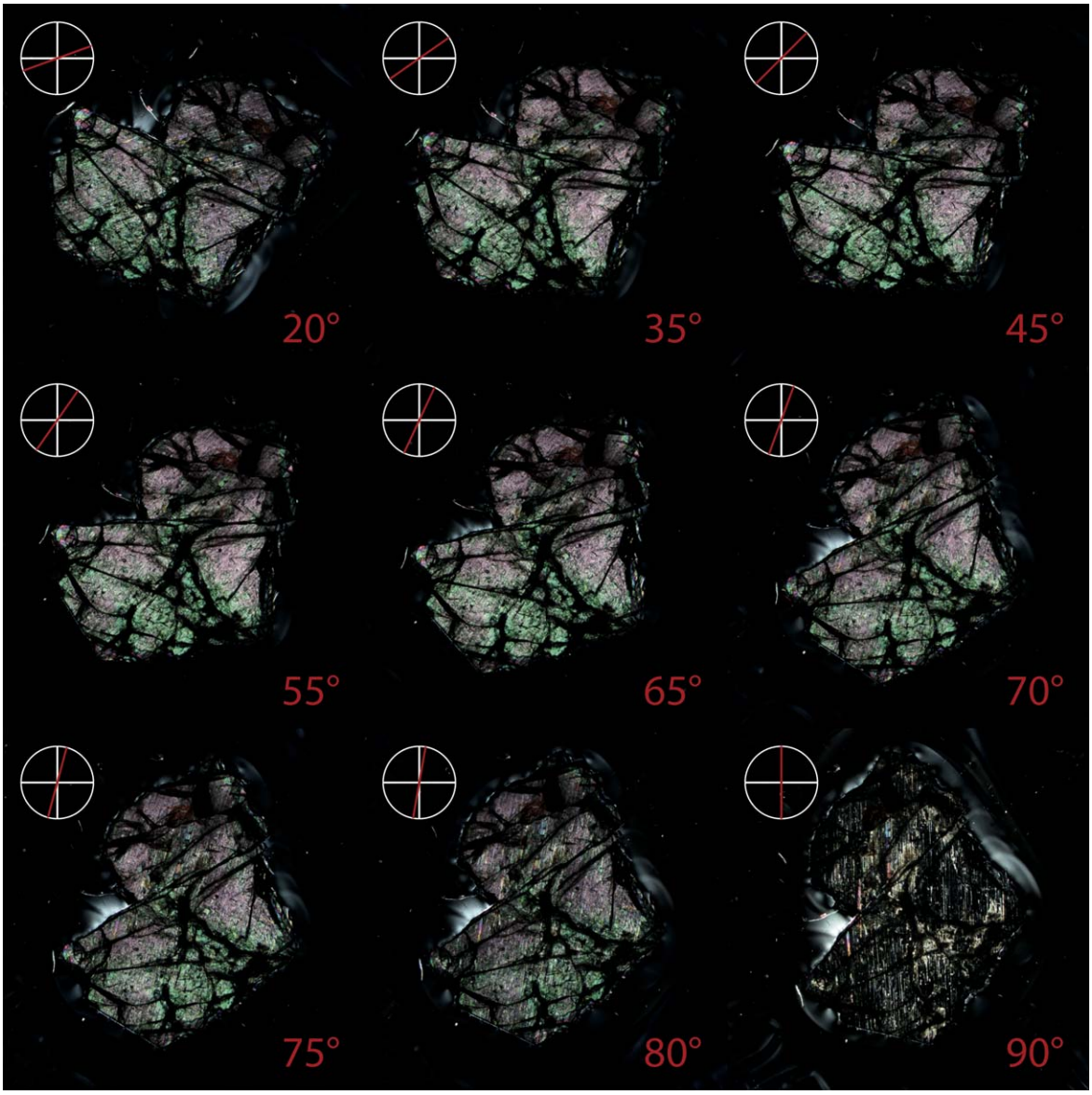

Figure A1. Optical microscopy images of the apatite thin section taken with transmitted XPL. The angle of orientation was varied between images to find the optical axis of the crystal. The brightest transmission is achieved around 45°, while at 90°, transmission is almost completely hindered. The angle between the laser polarization plane of the subsequent LIMS measurements and the optical axis was between 40° and 50°. The optical microscopy images in this figure were taken prior to LIMS analysis.

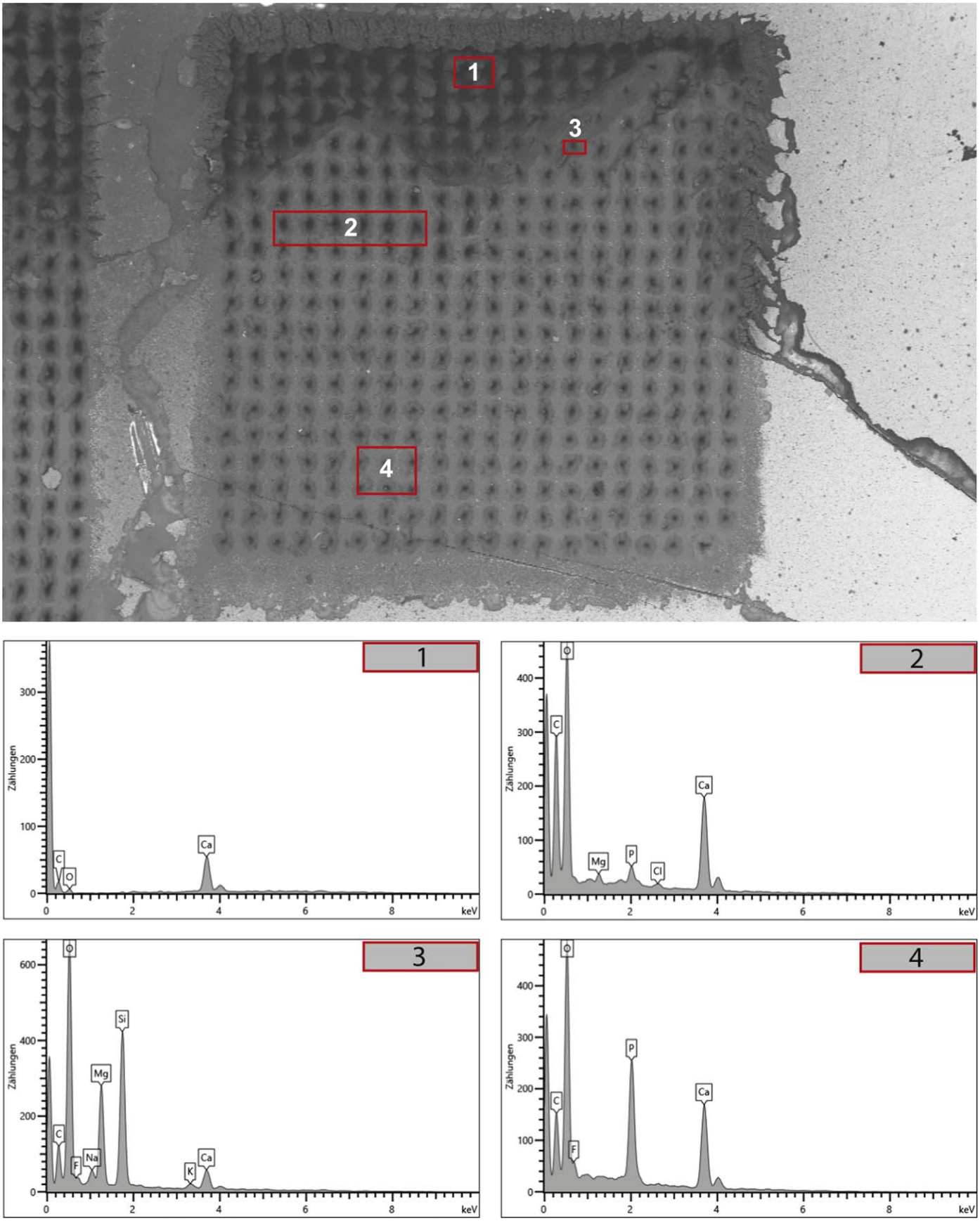

Figure A2. SEM-EDX spectra of the sampled area (accelerating voltage: 5 kV, magnification: 100x, working distance: 8.5 mm, probe current: 200 pA, backscattered electron detector). Four regions are marked on the SEM image and the corresponding EDX spectra are shown below the SEM image.

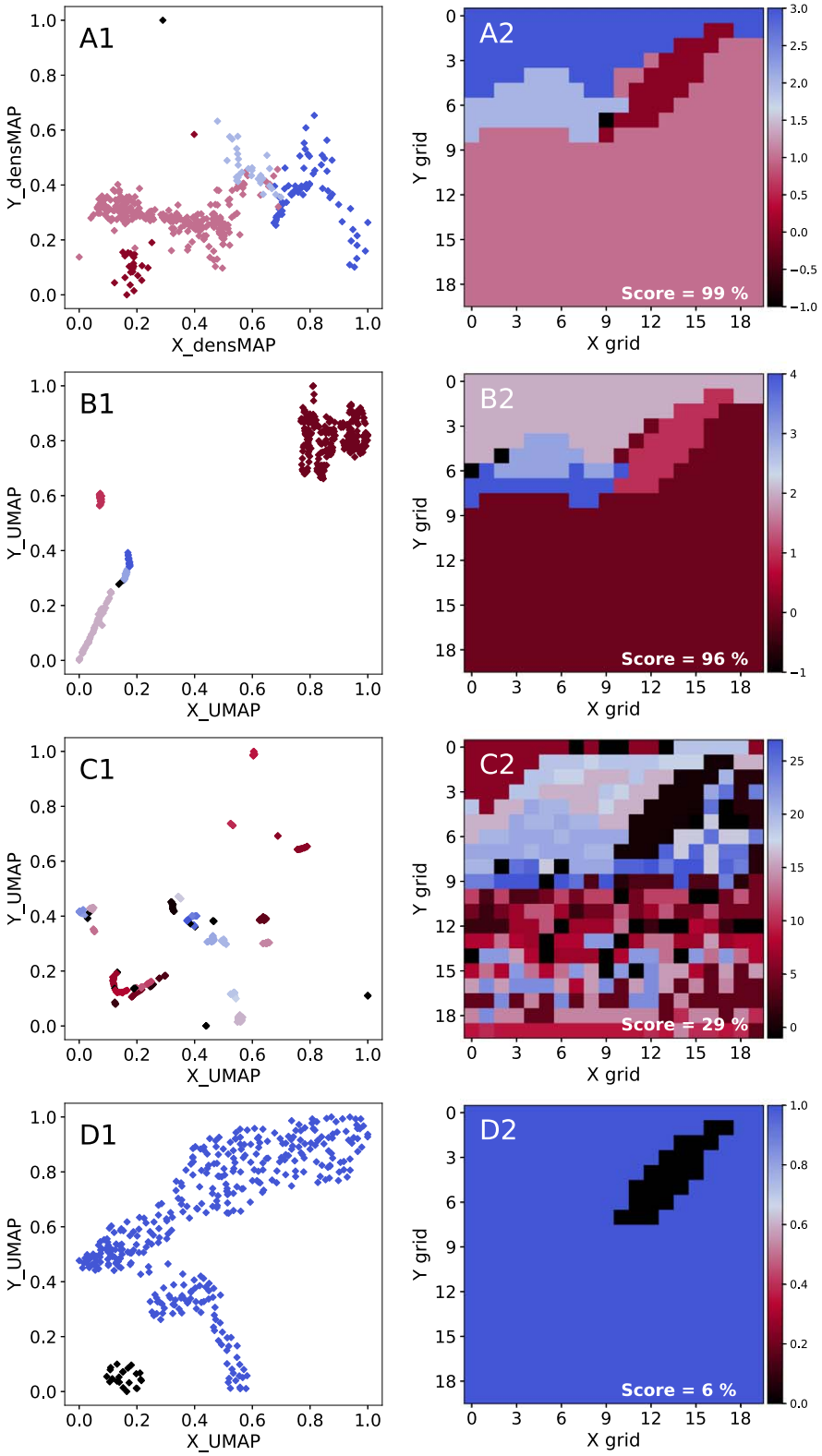

Figure A3. Selection of different embeddings (left panels) and corresponding cluster associations of the sampled areas (right panels) for various ML parameter sets. The parameters (min_dist, n_epochs, n_neighbors, min_samples, min_cluster_size) were chosen as follows. A1–A2: 0.1, 2500, 40, 15, 5; B1–B2: 0.01, 1500, 20, 5, 5; C1–C2: 0.001, 1500, 3, 5, 5; and D1–D2: 1, 1200, 50, 5, 20. Whether UMAP or densMAP were used can be derived from the corresponding axes labels. Achieved scores are provided. Panels (A1)–(A2) show the detection of a noise sample. Panels (B1)–(B2) and to an extremer extent (C1)–(C2) show the case of overclustering. Panels (D1)–(D2) provide an example for underclustering.

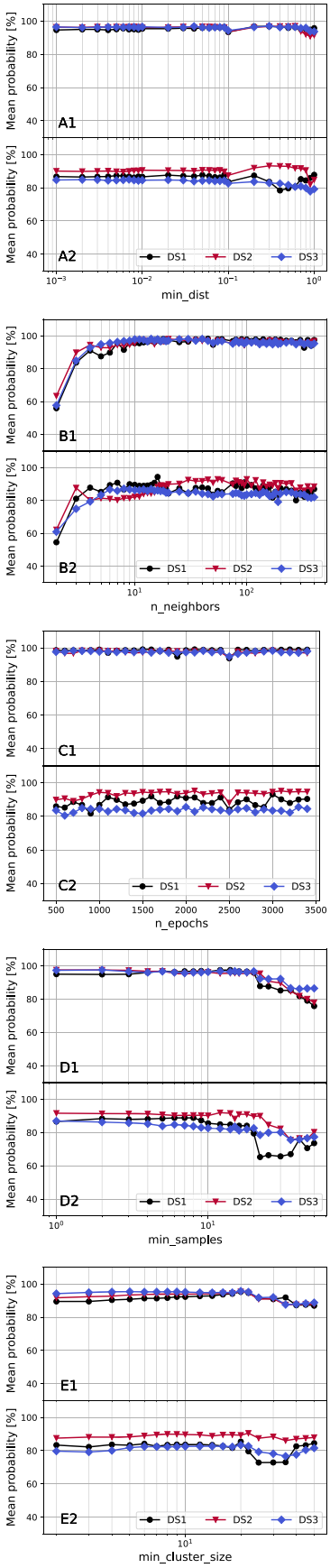

Figure A4. Summary of HDBSCAN probabilities obtained in the grid search. Probabilities are shown for all three data sets DS1–DS3, with UMAP (upper panels) and densMAP (lower panels). The *x*-axis labels indicate the varied parameter (A: min_dist; B: n_neighbors; C: n_epochs; D: min_samples; E: min_cluster_size).

Appendix B
Cross-check with Another Data Set

For the given data, the most stable approach for achieving phase separation has been realized with the combination of mass-integrated data and densMAP. To ensure that this outcome is not a result of overfitting, where the method performs well only on the specific data set at hand, the approach was applied to a previously recorded data set (M. Tulej et al. 2022). The data were obtained from the same thin section under similar measurement conditions using the same LIMS setup. An area of 30×30 locations was sampled, resulting in 900 mass spectra. The data were subjected to the same preprocessing procedure as DS3 (see Table 1) and embedded with densMAP.

Figure B1, panel (L0) displays the obtained densMAP embedding, colored according to the HDBSCAN cluster associations, whereas Figure B1, panel (P0) shows the same embedding but color coded by the HDBSCAN probability measure. Figure B1, panels (L1) and (P1) show the sampled area, colored according to the HDBSCAN labels and probabilities, respectively. The obtained clustering results are compared to both the optical microscopy image (Figure B1, panel (M0)) and selected element maps (Figure B1, panels (E0)–(E3)). These provide the required information to manually select the chemically distinct regions, as depicted in Figure B1, panel (C0). Comparing panels (C0)–(L1) and (P1), a high level of agreement is apparent. The HDBSCAN clustering found a major phase (black), a minor phase (red), and outliers (gray). Despite all outliers receiving the label -1, their local separation into two groups (upper-right and lower-left corner in Figure B1, panel (L1)) indicates that they are indeed two separate phases, which aligns with the observations in Figure B1, panel (C0). The only phase that could not be retrieved with this procedure is a semicircular phase identified in panel (C0) and shown in red. However, examination of panel (P1) reveals that the corresponding spectra received low HDBSCAN probabilities. Although better separation between phases could potentially be achieved through parameter refinement, the information conveyed in panel (C0) can be fully extracted from panels (L1) and (P1). This further highlights the stability of the method.

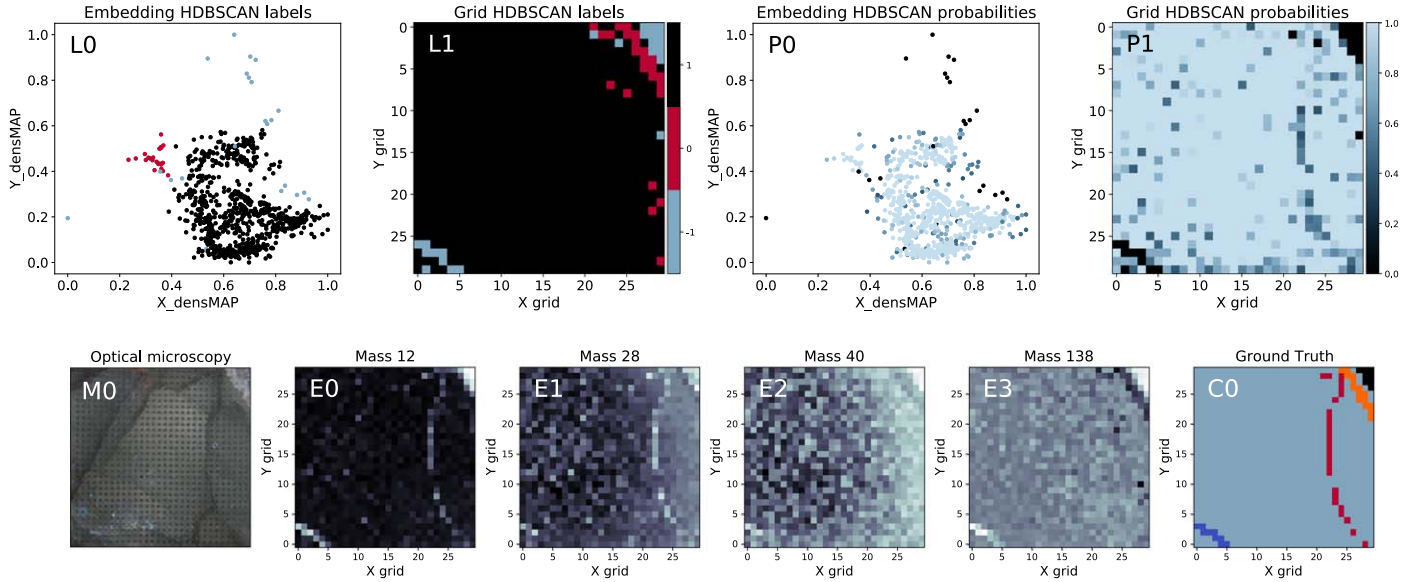

Figure B1. Results of the cluster analysis applied to a different data set. Panel (L0): densMAP embedding, colored according to the assigned HDBSCAN labels. Panel (L1): sampled area colored according to the HDBSCAN labels. Panel (P0): densMAP embedding, colored according to the assigned HDBSCAN probabilities. Panel (P1): sampled area colored according to the HDBSCAN probabilities. Panel M0: optical microscopy image under polarized reflected light. Panels (E0)–(E3): element maps of distinct element distributions. Panel (C0): chemically distinct regions inferred from panels (M0) and (E0)–(E3).

Appendix C

A substantial amount of data can be collected in a short amount of time with the current acquisition system of the LIMS instrument. However, data preprocessing and especially dimensionality reduction can take considerably longer than data acquisition. ML procedures are computationally quite expensive. Here on Earth, resources are comparably unlimited, whereas on a spacecraft, electric and computational power as well as data storage are limited. Run times on the currently used workstation (Intel Core i7-9700, 3.00 GHz base frequency, eight cores, 16 GB RAM) are summarized in Table C1. As a comparison, the NASA perseverance rover uses a RAD750 chip, a variation of a PowerPC 750 with a single core, 233 MHz base frequency, and 256 MB RAM (V. Verma et al. 2022). These times give a first impression on the order of run time and which processes are slowest. The given times are averaged values of 1000 iterations and presented as run times per sample. Preprocessing was performed with MATLAB, and the embedding and clustering (UMAP/densMAP and HBSCAN) with Python.

Preprocessing of a single spectrum took roughly 3 ms for DS1, but around 200 and 300 ms for a spectrum of DS2 and DS3, respectively (see Table C1). However, data preprocessing is negligible compared to the ML analysis, taking up more than 90% of the run time for all three data sets. Especially feature selection, downsampling, log transformation, and normalization are negligible at less than 1‰ of the total run time. HDBSCAN clustering was found to complete very quickly at approximately 1‰ of the total run time. This can be advantageous when clustering parameters need to be adapted to find optimal values that do not over- or undercluster the data. Notably, for all data sets DS1–DS3, densMAP analysis took roughly 40% longer than analysis with UMAP, e.g., 5.1 s compared to 3.6 s per DS1 sample with densMAP and UMAP, respectively. Compared to that, analysis of DS1 was about 10% faster than DS3, and approximately 3% faster than DS2.

Table C1
Run Times for the Three Data Sets and the Various Processing Steps

Process	Run Time DS1 (s)	Run Time DS2 (s)	Run Time DS3 (s)
Baseline correction	…	0.1460 (∼3% − 4%)	0.1460 (∼3% − 4%)
Low-pass filtering	…	0.0723 (∼1% − 2%)	0.0723 (∼1% − 2%)
Mass calibration	…	…	0.0184 (<1%)
Mass integration	…	…	0.0666 (∼1% − 2%)
Feature selection	0.0002 (<1‰)	0.0002 (<1‰)	…
Downsampling	0.0024 (<1‰)	0.0024 (<1‰)	…
Log transformation	…	0.0001 (<1‰)	0.0001 (<1‰)
Normalization	…	0.0001 (<1‰)	0.0001 (<1‰)
UMAP | densMAP	3.5755 (∼99%) | 5.1167 (∼99%)	3.6471 (∼94%) | 5.2371 (∼96%)	3.7082 (∼92%) | 5.3147 (∼95%)
HDBSCAN (U | D)	0.0056 (∼1‰) | 0.0054 (∼1‰)	0.0060 (∼1‰) | 0.0056 (∼1‰)	0.0053 (∼1‰) | 0.0053 (∼1‰)
Total (U | D)	3.5837 (100%) | 5.1247 (100%)	3.8742 (100%) | 5.4638 (100%)	4.017 (100%) | 5.6235 (100%)

Note. "(U | D)" refers to run times for UMAP and densMAP, respectively.

ORCID iDs

Salome Gruchola https://orcid.org/0000-0002-9757-1402
Peter Keresztes Schmidt https://orcid.org/0000-0002-4519-8861
Marek Tulej https://orcid.org/0000-0001-9823-6510
Andreas Riedo https://orcid.org/0000-0001-9007-5791
Peter Wurz https://orcid.org/0000-0002-2603-1169

References

Abcouwer, N., Daftry, S., Del Sesto, T., et al. 2021, Machine Learning Based Path Planning for Improved Rover Navigation, in IEEE Aerospace Conf. Proc. (Piscataway, NJ: IEEE), 1
Bajracharya, M., Maimone, M. W., & Helmick, D. 2008, Autonomy for Mars Rovers: Past, Present, and Future, Compr, 41, 44
Baraldi, A., Bruzzone, L., & Blonda, P. 2005, Quality Assessment of Classification and Cluster Maps Without Ground Truth Knowledge, ITGRS, 43, 857
Bishop, C. M. 2006, in Pattern Recognition and Machine Learning, ed. M. Jordan, J. Kleinberg, & B. Schölkop (New York: Springer)
Campello, R. J.-G. B., Moulavi, D., & Sander, J. 2013, Density-based Clustering Based on Hierarchical Density Estimates, LNCS, 7819, 160
Castano, R., Estlin, T., Gaines, D., et al. 2006, Opportunistic Rover Science: Finding and Reacting to Rocks, Clouds and Dust Devils, in IEEE Aerospace Conf. (Piscataway, NJ: IEEE), 1
Castano, R., Mazzoni, D., Tang, N., et al. 2005, Learning Classifiers for Science Event Detection in Remote Sensing Imagery in 'i-SAIRAS 2005', The 8th Int. Symp.on Artifical Intelligence, Robotics and Automation in Space, ed. B. Battrick (Noordwijk: ESA), 36
Davis, J. C., & Pernicka, H. J. 2020, Spacecraft Identification Leveraging Unsupervised Learning Techniques for Formation and Swarm Missions, in AIAA Scitech 2020 Forum (Reston, VA: AIAA), AIAA 2020–1195
Decrée, S., Cawthorn, G., Deloule, E., et al. 2020, Unravelling the Processes Controlling Apatite Formation in The Phalaborwa Complex (South Africa) Based on Combined Cathodoluminescence, LA-ICPMS and in situ O and Sr Isotope Analyses, CoMP, 175, 1
Doyle, R., Kubota, T., Picard, M., et al. 2021, Recent Research and Development Activities on space Robotics and AI, Advanced Robotics, 35, 1244
Eriksson, S. C. 1984, Age of Carbonatite and Phoscorite Magmatism of the Phalaborwa Complex (South Africa), ChGeo, 46, 291
Ferrero, S., & Angel, R. J. 2018, Micropetrology: Are Inclusions Grains of Truth?, JPet, 59, 1671
Francis, R., Estlin, T., Doran, G., et al. 2017, AEGIS Autonomous Targeting for Chemcam on Mars Science Laboratory: Deployment and Results of Initial Science Team Use, Sci Robot, 2, 4582
Gaudet, B., Linares, R., & Furfaro, R. 2020, Deep Reinforcement Learning for Six Degree-of-Freedom Planetary Landing, AdSpR, 65, 1723
Gieseke, F., Moruz, G., & Vahrenhold, J. 2012, Resilient k–d Trees: K-means in Space Revisited, Front. Comput. Sci., 6, 166
Giuffrida, G., Fanucci, L., Meoni, G., et al. 2022, The Φ-Sat-1 Mission: The First On-board Deep Neural Network Demonstrator for Satellite Earth Observation, ITGRS, 60, 3125567
Grimaudo, V., Moreno-García, P., Riedo, A., et al. 2017, Toward Three-dimensional Chemical Imaging of Ternary Cu–Sn–Pb Alloys Using Femtosecond Laser Ablation/Ionization Mass Spectrometry, AnaCh, 89, 1632
Grimaudo, V., Moreno-García, P., Riedo, A., et al. 2015, High-Resolution Chemical Depth Profiling of Solid Material Using a Miniature Laser Ablation/Ionization Mass Spectrometer, AnaCh, 87, 2037
Gruchola, S., Riedo, A., Schmidt, P. K., et al. 2023, Reduction of Surface Charging Effects in Laser Ablation Ionisation Mass Spectrometry Through Gold Coating, JAAS, 38, 1372
Han, J., Kamber, M., & Pei, J. 2011, Data Mining: Concepts and Techniques. Data Mining: Concepts and Techniques. (Burlington, MA: Morgan Kaufmann)
Heaman, L. M. 2009, The Application of U–Pb Geochronology to Mafic, Ultramafic and Alkaline Rocks: An Evaluation of Three Mineral Standards, ChGeo, 261, 43
Hurley, R. G., & Goss, R. L. 1978, Quantitative Energy-dispersive X-ray Analysis Using Relative k-ratios, XRS, 7, 70
Ibrahim, S. K., Ahmed, A., Zeidan, M. A.-E., & Ziedan, I. E. 2020, Machine Learning Techniques for Satellite Fault Diagnosis, Ain Shams Engineering Journal, 11, 45
Ihlen, P. M., Schiellerup, H., Gautneb, H., & Skår, Ø. 2014, Characterization of Apatite Resources in Norway and their REE Potential—A Review, OGRv, 58, 126
Johnson, A. E., Cheng, Y., Montgomery, J., et al. 2015, Real-time Terrain Relative Navigation Test Results from a Relevant Environment for Mars Landing, in AIAA Guidance, Navigation, and Control Conf. (Reston, VA: AIAA), 1
Kalinicheva, E., Sublime, J., & Trocan, M. 2020, Unsupervised Satellite Image Time Series Clustering Using Object-based Approaches and 3D Convolutional Autoencoder, RemS, 12, 1816
Kothari, V., Liberis, E., & Lane, N. D. 2020, The Final Frontier: Deep Learning in Space, in HotMobile '20: Proc. of the 21st Int. Workshop on Mobile Computing Systems and Applications (New York: ACM), 45
Krasnova, N. I., Petrov, T. G., Balaganskaya, E. G., et al. 2004, Introduction to Phoscorites: Occurrence, Composition, Nomenclature and Petrogenesis, in Phoscorites and Carbonatites from Mantle to Mine, ed. F. Wall & A.N. Zaitsev (Chantilly, VA: MinSoc), 45
Kumar, S., & Tomar, R. 2019, The Role of Artificial Intelligence in Space Exploration, in Proc. 2018 Int. Conf. On Communication, Computing and Internet of Things, IC3IoT 2018 (Piscataway, NJ: IEEE), 499
Labreche, G., Evans, D., Marszk, D., et al. 2022, OPS-SAT Spacecraft Autonomy with TensorFlow Lite, Unsupervised Learning, and Online Machine Learning, in IEEE Aerospace Conf. Proc. (Piscataway, NJ: IEEE), 1
Libourel, G. 1999, Systematics of Calcium Partitioning Between Olivine and Silicate Melt: Implications for Melt Structure and Calcium Content of Magmatic Olivines, CoMP, 136, 63
Llovet, X., Moy, A., Pinard, P. T., & Fournelle, J. H. 2021, Reprint of: Electron Probe Microanalysis: A Review of Recent Developments and Applications in Materials Science and Engineering, PrMS, 120, 100818
Lukmanov, R., de Koning, C., Schmidt, P. K., et al. 2022, High Mass Resolution fs-LIMS Imaging and Manifold Learning Reveal Insight Into Chemical Diversity of the 1.88 Ga Gunflint Chert, FrST, 3, 10
Lukmanov, R., Riedo, A., Wacey, D., et al. 2021, On Topological Analysis OF fs-LIMS Data. Implications for in Situ Planetary Mass Spectrometry, Front. Artif. Intell., 4, 119
Mateo-Garcia, G., Veitch-Michaelis, J., Smith, L., et al. 2021, Toward Global Flood Mapping Onboard Low Cost Satellites With Machine Learning, NatSR, 11, 1
Maxwell, J. A., Teesdale, W. J., & Campbell, J. L. 1995, The Guelph PIXE Software Package II, NIMPB, 95, 407
McGovern, A., & Wagstaff, K. L. 2011, Machine Learning In Space: Extending Our Reach, Mach Learn, 84, 335
McInnes, L., Healy, J., & Astels, S. 2017, hdbscan: Hierarchical Density Based Clustering, JOSS, 2, 205
McInnes, L., Healy, J., & Melville, J. 2018, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, JOSS, 3, 861
Meyer, S., Riedo, A., Neuland, M. B., Tulej, M., & Wurz, P. 2017, Fully Automatic and Precise Data Analysis Developed for Time-of-Flight Mass Spectrometry, JMSp, 52, 580
Milani, L., Bolhar, R., Frei, D., Harlov, D. E., & Samuel, V. O. 2017, Light Rare Earth Element Systematics as a Tool for Investigating the Petrogenesis of Phoscorite-Carbonatite Associations, as Exemplified by the PHALABORWA Complex, South Africa, MinDe, 52, 1105
Narayan, A., Berger, B., & Cho, H. 2021, Assessing Single-cell Transcriptomic Variability Through Density-preserving Data Visualization, NatBi, 39, 765
Neubeck, A., Tulej, M., Ivarsson, M., et al. 2016, Mineralogical Determination in Situ of a Highly Heterogeneous Material Using a Miniaturized Laser Ablation Mass Spectrometer with High Spatial Resolution, IJAsB, 15, 133
Neuland, M. B., Grimaudo, V., Mezger, K., et al. 2016, Quantitative Measurement of The Chemical Composition of Geological Standards With a Miniature Laser Ablation/Ionization Mass Spectrometer Designed for in situ Application in Space Research, MeScT, 27, 035904
Ouabid, M., Raji, O., Dautria, J. M., et al. 2021, Petrological and Geochemical Constraints on The Origin of Apatite Ores from Mesozoic Alkaline Intrusive Complexes, Central High-Atlas, Morocco, OGRv, 136, 104250
Riedo, A., Bieler, A., Neuland, M., Tulej, M., & Wurz, P. 2013, Performance Evaluation of a Miniature Laser Ablation Time-of-Flight Mass Spectrometer Designed for in Situ Investigations in Planetary Space Research, JMSp, 48, 1
Riedo, A., Lukmanov, R., Grimaudo, V., et al. 2021, Improved Plasma Stoichiometry Recorded by Laser Ablation Ionization Mass Spectrometry Using a Double-pulse Femtosecond Laser Ablation Ion Source, RCMS, 35, e9094

Rohner, U., Whitby, J. A., & Wurz, P. 2003, A Miniature Laser Ablation Time-of-flight Mass Spectrometer for in Situ Planetary Exploration, MeScT, 14, 2159
Russell, H. D., Hiemstra, S. A., & Groeneveld, D. 1954, The Mineralogy and Petrology of the Carbonatite at Loolekop, Eastern Transvaal, S. Afr. J. Geol., 57, 197
Russo, A., & Lax, G. 2022, Using Artificial Intelligence for Space Challenges: A Survey, ApSci, 12, 5106
Růžička, V., Vaughan, A., De Martini, D., et al. 2022, RaVÆn: Unsupervised Change Detection oF Extreme Events Using ML On-board Satellites, NatSR, 12, 1
Shirobokov, M., Trofimov, S., & Ovchinnikov, M. 2021, Survey of Machine Learning Techniques in Spacecraft Control Design, AcAau, 186, 87
Tulej, M., Ligterink, N. F.-W., de Koning, C., et al. 2021, Current Progress in Femtosecond Laser Ablation/Ionisation Time-of-Flight Mass Spectrometry, ApSci, 11, 2562
Tulej, M., Schmidt, P. K., Gruchola, S., et al. 2022, Toward in Situ Geochemical Analysis of Planetary Rocks and Soils by Laser Ablation/Ionisation Time-of-Flight Mass Spectrometry, Univ, 8, 410
Tulej, M., Wiesendanger, R., Riedo, A., Knopp, G., & Wurz, P. 2018, Mass Spectrometric Analysis oF the Mg Plasma Produced by Double-pulse Femtosecond Laser Irradiation, JAAS, 33, 1292
Verma, V., Hartman, F., Rankin, A., et al. 2022, First 210 Solar Days of Mars 2020 Perseverance Robotic Operations-Mobility, Robotic Arm, Sampling, and Helicopter, in 2022 IEEE Aerospace Conf. (AERO) (Piscataway, NJ: IEEE), 1
Wagstaff, K. L., & Bornstein, B. 2009, K-means in Space: A Radiation Sensitivity Evaluation, in Proc. of the 26th Annual International Conf. on Machine Learning, ed. L. Bottou & M. Littman (New York: ACM), 1097
White, W. M. 2020, in Geochemistry, ed. W. M. White (Hoboken, NJ: Wiley)
Wiesendanger, R., Tulej, M., Grimaudo, V., et al. 2018, A Method for Improvement of Mass Resolution and Isotope Accuracy for Laser Ablation Time-of-Flight Mass Spectrometers, J. Chemom., 33, 3081
Yairi, T., Takeishi, N., Oda, T., et al. 2017, A Data-driven Health Monitoring Method for Satellite Housekeeping Data Based on Probabilistic Clustering and Dimensionality Reduction, ITAES, 53, 1384
Zeleke, D. A., & Kim, H. D. 2023, A New Strategy of Satellite Autonomy with Machine Learning for Efficient Resource Utilization of a Standard Performance CubeSat, Aeros, 10, 78