JAAS

PAPER

View Article Online
View Journal | View Issue

Cite this: *J. Anal. At. Spectrom.*, 2025, **40**, 3473

Exploring high-dimensional LA-ICP-TOFMS data with uniform manifold approximation and projection (UMAP)

Katharina Kronenberg, *[a] Hennes Rave, [b] Nassim Ghaffari-Tabrizi-Wizsy, [c] Danae Nyckees, [d] Matthias Elinkmann, [a] Dalial Freitak, [d] Lars Linsen, [b] Raquel Gonzalez de Vega [e] and David Clases [a]

Spectral imaging generates information-rich datasets comprising a large map of pixels that each contain a comprehensive spectrum. A specific form of mass spectral imaging is laser ablation-inductively coupled plasma-time-of-flight mass spectrometry (LA-ICP-TOFMS). This technique enables elemental imaging of almost the entire periodic table. The large number of isotopes per pixel leads to high-dimensional data posing major challenges for visualisation, pattern recognition and interpretation. To decrease this complexity, dimensionality reduction techniques, such as uniform manifold approximation and projection (UMAP), provide powerful tools to transform high-dimensional datasets into low-dimensional representations aiming to preserve data point relationships and visualise spectral similarities. This study provides a detailed introduction to UMAP for analysing LA-ICP-TOFMS data. By transforming high-dimensional MS imaging data into two-dimensional spaces, UMAP facilitates automated visualisation to identify spectral clusters. UMAP's utility to reveal spectrally distinct regions and tissue heterogeneity is demonstrated for a chicken embryo and a honeybee specimen. For detailed cluster analysis, a hierarchical strategy is introduced involving iterative UMAP applications, first to the global dataset, and then to resulting clusters. This approach helps uncover subtle chemical patterns hidden in the initial global UMAP application. Furthermore, the influence of the most relevant UMAP hyper-parameters is discussed, providing guidance for selecting critical parameters for further datasets. Overall, this study introduces UMAP as an exploratory and versatile tool for targeted and non-targeted analysis of complex LA-ICP-TOFMS data. Its integration into imaging workflows supports spectral clustering, image segmentation, hypothesis generation, and rapid analysis of large and high-dimensional spectral data from biological and environmental specimens.

Received 31st May 2025
Accepted 17th September 2025

DOI: 10.1039/d5ja00215j

rsc.li/jaas

[a]*NanoMicroLab, Institute of Chemistry, University of Graz, Universitätsplatz 1, 8010 Graz, Austria. E-mail: katharina.kronenberg@uni-graz.at*
[b]*Institute of Informatics, University of Münster, Einsteinstraße 62, 48149 Münster, Germany*
[c]*Otto Loewi Research Center, Division of Immunology, Medical University of Graz, Neue Stiftingtalstraße 6, Graz, 8010, Austria*
[d]*Institute for Biology, University of Graz, Universitätsplatz 2, Graz, 8010, Austria*
[e]*TESLA-Analytical Chemistry, Institute of Chemistry, University of Graz, Universitätsplatz 1, 8010 Graz, Austria*

Katharina Kronenberg

Katharina Kronenberg is an analytical chemist and postdoctoral researcher in the group of Prof. David Clases at the University of Graz (Austria). She began her academic career studying Chemistry at the University of Münster (Germany), where she developed a keen interest in Analytical Chemistry. Her initial research focused on laser ablation-inductively coupled plasma-mass spectrometry (LA-ICP-MS) bioimaging. After earning a Master's degree in 2020, she joined Prof. Uwe Karst's research group at the University of Münster as a PhD student. Her approach combined molecular and elemental analysis techniques and provided comprehensive insights into cancer tissue. In these studies, she integrated diverse imaging techniques, including LA-ICP-MS, micro X-ray fluorescence, matrix-assisted laser desorption ionisation (MALDI)-MS, and infrared microspectroscopy. In 2024, she completed her PhD and joined the University of Graz, where she expands her expertise in bioimaging by employing time-of-flight-based ICP-MS and single particle ICP-MS. Additionally, her research emphasizes collaborative and interdisciplinary studies at the intersection of life science, computer science, and analytical chemistry to address current biological questions.

1 Introduction

Elemental mass spectrometry imaging (MSI) techniques, such as laser ablation-inductively coupled plasma-mass spectrometry (LA-ICP-MS), generate pixel-based datasets of spatially resolved elemental signals across complex biological or geological structures.[1,2] A variety of mass analysers are available for ICP-MS. In the past, quadrupole-based instruments have dominated LA-ICP-MS.[1] Here, the number of recorded isotopes is inherently constrained to a handful of mass-to-charge ratios (m/z). This limitation arises from the quadrupole's sequential m/z scanning and the need to analyse the transient signal of a short aerosol pulse with sufficient dwell time for each m/z.[3,4] To date, one of the most relevant innovations is time-of-flight (TOF) instrumentation, which has substantially advanced the field of elemental imaging. Unlike quadrupole analysers, TOFMS allows for the rapid and simultaneous acquisition of a full mass spectrum per pixel.[5,6] In combination with LA and rapid aerosol introduction systems, LA-ICP-TOFMS enables fast pixel acquisition with rates of up to 1 kHz,[7] while detecting virtually every element at adjustable spatial resolution down to 1 μm.[8] Each pixel in an LA-ICP-TOFMS image contains a mass spectrum with an intensity value for each m/z. Typically, LA-ICP-TOFMS data cover a mass range from ^{7}Li to ^{238}U resulting in mass spectra with intensities for over 200 m/z. This technique generates highly complex datasets in short time.

Traditionally, LA-ICP-MS data has mostly been examined manually. This involves a systematic but manual navigation through images of individual m/z to identify elemental or spatial patterns.[9] However, with the advancement of ICP-TOFMS, it becomes increasingly impractical and less efficient as dataset complexity and information density grow. When employing the full potential of a TOF-based instrument, the available mass range easily exceeds 200 m/z and an exhaustive manual analysis is no longer feasible in practice. This is especially the case in non-targeted analyses, where no *a priori* knowledge or expectation exists and increasing dataset complexity greatly amplifies the risk of overlooking relevant spatial and spectral features. In particular, it is challenging to segment spatial regions with distinct multi-elemental signatures or to visualise and understand overarching patterns in the data. A more effective strategy to identify spatial and elemental patterns should consider several m/z jointly.

To solve this challenge, the structure of the data needs to be considered to derive better strategies to handle and visualise complex data. In LA-ICP-TOFMS, the intensity of each recorded m/z represents a distinct dimension. In an exemplary case of 200 recorded m/z, each pixel contains an intensity value for all 200 m/z and is positioned accordingly in a 200-dimensional scatterplot. This represents the elemental image in a high-dimensional space. Recently, powerful dimensionality reduction algorithms have been introduced that are ideally suited for such datasets. A relevant selection is presented in the following. A glossary, containing definitions of key terms used in dimensionality reduction, can be found at the end of this article.

1.1 Dimensionality reduction

To handle complex data with high dimensionality, several processing approaches have been investigated for visualisation and clustering of spectral imaging data. These include for example k-means clustering,[10,11] principal component analysis (PCA),[12,13] t-distributed stochastic neighbour embedding (t-SNE)[12,14] or uniform manifold approximation and projection (UMAP).[10,15,16] While the k-means algorithm is exclusively concerned with the process of clustering the original data in the high-dimensional space, PCA, t-SNE and UMAP are classified as dimensionality reduction techniques. A dataset comprising numerous dimensions is difficult to visualise directly, as it is challenging to display more than a few dimensions simultaneously. Dimensionality reduction techniques such as PCA, t-SNE and UMAP aim to visualise high-dimensional data in a low-dimensional (2D or 3D) space. In t-SNE and UMAP, this low-dimensional space is also referred to as 'embedding'. The following section will explain PCA, t-SNE and UMAP in more detail, highlighting the similarities and differences between these dimensionality reduction techniques.

PCA transforms a high-dimensional dataset with potentially correlated variables into a set of new uncorrelated variables, known as principal components. Principle components are linear combinations of the initial variables.[17] By identifying the directions (principal axes) along which the variance of the data is maximised, PCA projects the original data onto a lower-dimensional space with new orthogonal axes while preserving as much of the original variability as possible. As such, PCA is a linear and deterministic dimensionality reduction technique.[18]

t-SNE, on the other hand, reduces dimensionality in a non-linear way. The main objective of t-SNE is to preserve the local structure of high-dimensional data during the transformation into a low-dimensional space. In this context, 'preserving local structure' means that t-SNE aims to ensure that data points close to each other in the original high-dimensional space remain close in the low-dimensional embedding, thereby maintaining immediate neighbourhood relationships. To achieve this, t-SNE first calculates how similar each data point is to every other point in the high-dimensional space, expressing these distances between data points as probabilities that reflect similarities.[19] It then attempts to create a similar probability-based representation in the low-dimensional space, where the arrangement of points reflects the same local relationships. The algorithm minimises the difference (Kullback–Leibler divergence) between these two sets of probabilities using an optimisation method known as gradient descent. This process helps t-SNE to maintain the relative distances between nearby points while embedding the data into a low-dimensional space that is easier to interpret visually. It is important to note that the axes generated in the low-dimensional space are not interpretable. They represent abstract dimensions that have been optimised for visualisation, rather than the original data features. Additionally, it is also important to acknowledge that a stochastic algorithm is employed. Consequently, multiple executions may yield a different low-dimensional embedding,

unless the random seed is fixed within the algorithm. *t*-SNE is particularly effective for uncovering patterns in large and complex datasets, allowing it to capture non-linear structures and clusters that linear techniques like PCA cannot. By focusing on preserving the immediate neighbourhood of each point, rather than the global arrangement of all points, *t*-SNE provides an effective way to explore the underlying structure of high-dimensional data. The underlying mathematical details are discussed by van der Maaten and Hinton.[19]

In 2018, UMAP was introduced as a new non-linear dimensionality reduction algorithm by McInnes *et al.* and has been gaining increasing attention over the past years.[20] UMAP and *t*-SNE are very similar in their functionality. Both methods are stochastic and reduce high-dimensional data by embedding it into a low-dimensional, visually interpretable space, while aiming to preserve the structure of the original data. However, UMAP is frequently regarded as a more effective method due to several advantages.[21–23] Literature demonstrating and explaining UMAP compared to *t*-SNE can be found elsewhere.[20,21,24] The UMAP algorithm is based on a distinct theoretical foundation with several improvements over, *e.g.*, *t*-SNE. First UMAP provides better scalability and computational speed especially for datasets with a large number of data points (*i.e.*, the number of pixels in MSI) and dimensions (*i.e.*, the number of *m*/*z* in MSI). Second, UMAP offers improved preservation of the global structure of the original high-dimensional data while also maintaining the local structure. Thus, the algorithm performs better in balancing between local and global structure.[24] As a consequence, the low-dimensional embedding generally visualises a large amount of meaningful information hidden in the original data. However, it is worth to emphasise that, like *t*-SNE, UMAP inherently distorts the high-dimensional structure of original data during the transformation into the low-dimensional space. Consequently, the axes or distances between clusters in lower dimensions are not quantitatively interpretable, particularly in comparison to linear techniques such as PCA. An in-depth introduction to the UMAP algorithm, explaining its operational mechanisms, application methodologies, and interpretation, is provided in the recent publication by Healy and McInnes, recommended for those starting to use UMAP.[25] Those seeking to develop an intuitive understanding of UMAP are also directed to the UMAP reference documentation,[26] Understanding UMAP by Google,[24] or view the instructional videos "UMAP explained"[27] as short introduction or "UMAP: Main Ideas"[28] and "UMAP: Mathematical Details"[29] by StatQuest. Readers interested in the mathematical foundation of UMAP are referred to the original article by McInnes *et al.*[20] The following sections will only address the most essential aspects of UMAP necessary for applying it to LA-ICP-MS data. With regard to terminology, note that the structural shape of high-dimensional data is also referred to as topology.

1.2 UMAP algorithm

The UMAP algorithm is designed to preserve the relevant topological structure of the original data in the high-dimensional space. This is achieved by dividing the algorithm into two different tasks. During the first stage, the data points (pixels) in the high-dimensional space are connected in a manner that accurately represents the topological structure of the original data. This topology is represented by a graph consisting of nodes (data points) and edges (connections between points), used in UMAP to model the relationships between data points in the high-dimensional space. In the second stage, the objective is to identify a low-dimensional embedding that optimally reflects this topological structure.[25] The outcome of the UMAP algorithm strongly depends on the user-defined input parameters (or hyper-parameters). The two most important parameters are *n_neighbours* and *min_dist*.

1.3 UMAP hyper-parameters

n_neighbours plays a pivotal role in the first algorithm stage and controls the number of nearest neighbours that UMAP considers for each point when it constructs the topological structure of the data in high-dimensional space. This parameter is intended to achieve a balance between the preservation of local structure and that of global structure in the data. In this regard, smaller values (*e.g.*, 5–15) prioritise local structure and are well-suited for fine-grained groupings. Larger values (*e.g.*, 50–200) have been shown to preserve more of the global structure, such as larger-scale groupings. *n_neighbours* must be chosen with care: if too small, UMAP will capture too much noise or small variations; if too large, it will over-smooth the data, missing local patterns.[24,25] Depending on whether local patterns or global data structure is desired for data visualisation, it is recommended to test different values for this parameter. Furthermore, the parameter must be empirically re-evaluated for each new dataset. The reason is that its determination is dependent not solely on the number of points or the number of dimensions, but inherently on the distribution of the points, which can vary greatly between datasets of equivalent size and dimensionality. Here, the high-speed processing capabilities of UMAP for large datasets are beneficial, as new embeddings can be computed sufficiently fast. The computational speed can be increased further by using a so-called landmark-based UMAP approach, where only a random subset of data points (*i.e.*, the landmarks) are fed into the UMAP algorithm, while the remaining data points are added to the landmark embedding by interpolation resulting in a final embedding comprising all data points.

The parameter *min_dist* influences the second algorithm stage of optimising the low-dimensional embedding and describes the minimum distance between embedded points. It controls how densely points within clusters are arranged in the low-dimensional embedding, thereby shaping the density and visual appearance of the final output embedding. The value for *min_dist* ranges from 0 to 1. Low values lead to tightly packed clusters probably providing a more accurate representation of the topological structure. Large values result in spreading points further apart.[25]

To date, the utilisation of UMAP has been extensively implemented across numerous scientific domains, including single cell biology,[21] transcriptomics[30] and population

genetics.[31] In the context of elemental mass spectrometry, the use of UMAP was described for analysing single cells labelled with metal-conjugated antibodies,[32,33] as well as for molecular MSI.[16,34] This study, however, demonstrates the application of UMAP for dimensionality reduction and subsequent image segmentation of data obtained by label-free elemental MSI data. To illustrate the utility of UMAP, thin sections from two exemplary biological specimens, a chicken embryo and a honeybee, containing various and complex biological structures, were analysed using LA-ICP-TOFMS.

2 Experimental

2.1 Sample preparation

2.1.1 Chicken embryo assay and honeybee model. Fertilised white Lohmann chicken eggs were obtained from a local hatchery (Schropper GmbH, Gloggnitz, Austria). The outside of eggs was sterilised. Subsequently, the eggs were incubated at 37.6 °C and 40–60% humidity. At embryonic development day (EDD) 3, eggs were cracked in sterile weighting boats and covered with a square plastic lid. A heartbeat, as well as a primitive vessel system, proved the viability of the embryos. Embryos were incubated at 37.6 °C and 50–70% humidity throughout the experiments. On EDD 14, embryos were placed on ice and sacrificed by decapitation. Immediately afterwards, the torso was frozen and stored at −80 °C until cryosectioning. From a regulatory perspective, the chicken embryo assay is recognised as an *in vivo* assay that falls outside the scope of animal experimentation. In the specific experimental timeframe of embryonic development utilised in this study, regulatory frameworks in Europe (EU directive 2010/63/EU)[35] do not categorise the assay as animal experimentation.

A group of 30 worker honeybees (*Apis mellifera carnica*) were fed *ad libitum* with 50% sugar solution and were kept in an incubator at 34 °C and 70% humidity. After six days, the bees were frozen at −20 °C until cryosectioning.

2.1.2 Cryosectioning. Both the chicken embryo and honeybee were cryosectioned using a SLEE MNT cryotome (SLEE medical GmbH, Nieder-Olm, Germany). The chicken embryo torso was mounted with the back facing towards the cryostats sample holder using CryoGlue (SLEE medical GmbH, Nieder-Olm, Germany) as cryo-embedding medium. The torso was sectioned in a coronal plane with a section thickness of 40 μm and employing a chamber temperature and object temperature of −22 °C and −20 °C, respectively. The cryosections were placed on glass slides, dried, and were stored frozen −20 °C until LA-ICP-TOFMS analysis.

The honeybee was cryosectioned in a sagittal plane with a thickness of 50 μm and employing a chamber temperature and object temperature of −20 °C and −17 °C, respectively. The bee was cut in half and then the inner cavities of the bee were filled with CryoGlue (SLEE medical GmbH, Nieder-Olm, Germany). This procedure was implemented to ensure an improvement in the integrity and quality of the sections. Otherwise, the sections exhibited a high degree of brittleness and structural distortion. The resulting sections were mounted on glass slides, thoroughly dried, and stored frozen under a N_2 atmosphere until LA-ICP-TOFMS analysis was conducted.

2.2 LA-ICP-TOFMS analysis

For LA-ICP-TOFMS, an Analyte G2 excimer (193 nm) LA system (Teledyne Photon Machines, Omaha, NE, USA) equipped with an aerosol rapid introduction system (ARIS, Teledyne Photon Machines, washout time: 10 ms) was used and operated with the Chromium software (version 2.4., Teledyne Photon Machines, Omaha, NE, USA). The LA system was hyphenated to a Vitesse ICP-TOFMS instrument (Nu Instruments, Wrexham, UK), which was operated with the software Nu CoDaq (version v2.1.8728.1, Nu Instruments, Wrexham, UK). The LA-ICP-TOFMS instrument was tuned for maximum sensitivity before each measurement using the reference material NIST 612 "Trace Elements in Glass". Torch assembly, lenses, and gas flows were tuned before analysis for optimal sensitivity and single pulse response (SPR) profiles as well as a ThO/Th ratio below 2.5%. A detailed list of ICP-TOFMS tuning and method parameters is given in the SI in Tables S1 and S2. For ablating the chicken embryo sample, a spot size of 35 μm, a dosage of 4, a repetition rate of 100 Hz, a laser fluence of 0.34 J cm^{-2} and a 500 mL min^{-1} He gas flow as the sum of cup and cell gas flow were selected. In the ICP-TOFMS method, a mass range of m/z 7.0–240.0 was selected. Two mass regions (m/z 25.0–41.5 and 55.2 to 64.6) were blanked, which means that ions with a m/z within these ranges were scattered by a Bradbury–Nielsen gate to prevent high ion currents from reaching the detector. For the honeybee specimen, the following parameters were selected for the laser ablation: a spot size of 20 μm, a dosage of 4, a repetition rate of 200 Hz, a laser fluence of 1.50 J cm^{-2}, and a flow rate of 500 mL min^{-1} of He as the transport gas. In the ICP-TOFMS method, a mass range of m/z 30.0–240.0 was selected with two blanked mass regions (m/z 37.5–49.5 and 110.0 to 125.0).

2.3 Data analysis

All LA-ICP-TOFMS image and UMAP visualisations were carried out with the self-developed software Multiscale Image Analysis (**https://github.com/hennesrave/multiscale-image-analysis**) based on python and C++ using the official open-source UMAP implementation.[36] The software imports the raw data files generated by the Vitesse ICP-TOFMS instrument and subtracts the gas blank baseline for each recorded m/z. For all UMAP calculations, the selected metric was cosine similarity to compute similarity of the spectra and a fixed random seed with a value of 42 was set to force the stochastic algorithm to generate reproducible UMAP embeddings (UMAP software version 0.5.6). Further user-defined parameters include *n_neighbours*, *min_dist*, the selection of recorded m/z, the selection of an image segment and a subsampling of pixels. The parameter 'subsampling of pixels' refers to the landmark-based UMAP approach and was set to 0.5 (50%) for all UMAP analyses. The subsampling of pixels was implemented for large datasets and employs a two-step approach. First, UMAP is trained on a randomly selected subset of all pixels to learn the manifold structure. If the subset is sufficiently large, this provides

a reasonable approximation of the true structure of the entire dataset. The remaining data points can then be inserted into the existing embedding based on the locations and similarities to their respective nearest neighbours in the training subset. This approach decreases the computational cost, as the initial graph construction in the UMAP algorithm is performed on a smaller subset of the sample, while still enabling a consistent embedding for the whole dataset. The following example illustrates the computational time required for UMAP embeddings: utilising an 8 GB RAM and an Intel Core i5 CPU (4 cores, 8 threads), 200 dimensions, 500 000 pixels, *n_neighbours* = 5, and the landmark-based approach with a subsampling of pixels set to 0.1, a UMAP embedding was calculated within 3 min and 30 s. Data points in UMAP embeddings are displayed with 20% opacity to visualise density. Mass spectra can be interactively visualised within the Multiscale Image Analysis Software. In this study, all displayed mass spectra were averaged over the corresponding image segment or cluster and exported as CSV file. For figure preparation, the exported spectra were subsequently imported into OriginPro 2024 (OriginLab Corporation, Northampton, MA, USA).

3 Results and discussion

In LA-ICP-TOFMS, each ablation spot (or pixel) yields a mass spectrum enabling multi-elemental imaging. Each recorded *m*/*z* value within the mass spectrum represents a separate dimension in the MSI dataset. This creates a high-dimensional space where the number of dimensions correspond to the number of *m*/*z* recorded in the dataset. Each pixel is represented by a point in this high-dimensional space defined by the intensity values across all *m*/*z* dimensions. This study demonstrates the potential of UMAP to reduce the high-dimensional data generated by LA-ICP-TOFMS imaging to a low-dimensional embedding, while preserving both local and global data structure. To showcase the versatility of UMAP, the algorithm is applied to two exemplary biological specimens containing diverse biological structures: a cross-section of a chicken embryo and a honeybee with various internal organs. These examples illustrate how UMAP can support diverse analytical objectives in high-dimensional elemental imaging.

3.1 UMAP reveals elemental and spatial patterns in LA-ICP-TOFMS data

The first exemplary dataset was obtained from analysing a thin section of a chicken embryo torso, as depicted in Fig. 1(A1). The embryo torso was sectioned in a coronal plane and was chosen as an ideal dataset as it provides great heterogeneity of various biological tissues and inner organs, including the heart, liver, gall bladder, gizzard, and bones. Following LA-ICP-TOFMS analysis (Fig. 1(A2)), raw data was analysed with UMAP resulting in a low-dimensional 2D embedding (Fig. 1(B1)). In this embedding, pixels (represented as points) with similar mass spectral profile are positioned closer together, allowing for comprehensive visualisation of elemental similarities and clustering of tissue regions. After interactively colouring all point clusters in the UMAP embedding *via* a lasso selection tool (Fig. 1(B1)), corresponding pixels in the LA-ICP-TOFMS image appear in the same colour (Fig. 1(B2)). A comparison of this clustered image with the optical tissue structure (Fig. 1(A1 and A2)) reveals that UMAP effectively distinguished biologically relevant regions based on their spectral similarities with each tissue type forming a distinct cluster in the 2D projection: the liver is represented by dark red, the gallbladder by yellow, the heart tissue by orange, the heart ventricle by light red, parts of the gizzard by light green, and the bones by dark green and pink. The remaining tissue is depicted in blue tones.

To obtain a more detailed understanding of the UMAP results, a closer investigation was conducted on the underlying mean mass spectra of three clusters labelled C1 (light blue, tissue), C2 (dark green, bone), and C3 (light red, heart ventricle) in Fig. 1(B1). A figure demonstrating the mean mass spectra of

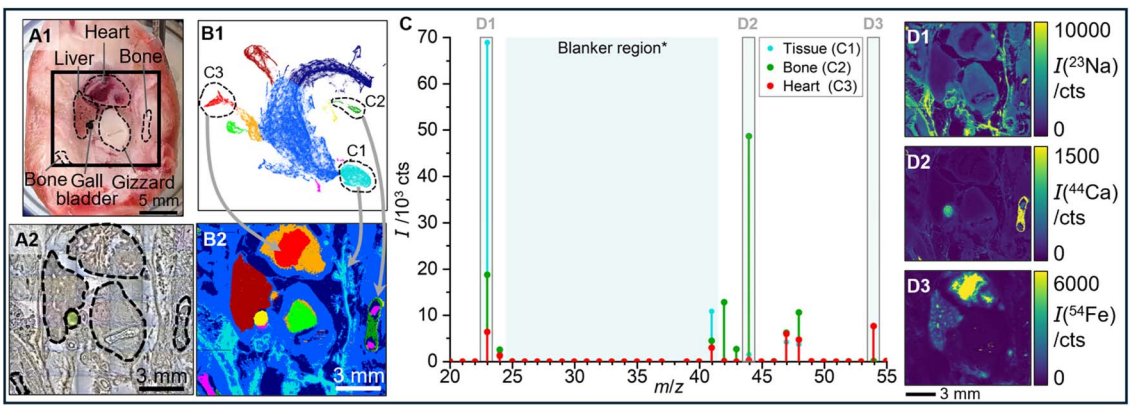

Fig. 1 LA-ICP-TOFMS data of a chicken embryo torso analysed by UMAP. A cross-section of the embryo torso (A1) and a microscopic image of the ablated tissue thin section (A2) are shown. The UMAP embedding of the LA-ICP-TOFMS data (B1) was generated using *n_neighbours* = 20, *min_dist* = 0.1, all pixels and all recorded *m*/*z*. Clusters were selected with an interactive lasso tool and coloured, resulting in the corresponding segmented LA-ICP-TOFMS image (B2). Mean mass spectra (C) are presented for three selected clusters: tissue (light blue, C1) bone (dark green, C2) and heart ventricle (light red, C3). *During TOFMS data acquisition, ions within an indicated mass range were scattered to prevent detector overload. Intense *m*/*z* signals in the mass spectra of C1–C3 are shown as images: ^{23}Na (D1), ^{44}Ca (D2), ^{54}Fe (D3).

all clusters is presented in the SI, Fig. S1. Examining mass spectra provided insight into interesting trends or differences regarding the elemental composition underlying each cluster. The mean mass spectra of clusters C1–C3 are displayed in Fig. 1(C), with each cluster exhibiting a unique elemental profile. Note that two blanker regions (m/z 25.0–41.5 and 55.2 to 64.6) were included in the recorded mass range in the ICP-TOFMS method. Ions with m/z within these ranges were scattered to prevent high ion currents from reaching the detector. This is reflected in low signal intensities for all m/z in blanker regions. To understand the spectral differences between the clusters C1, C2 and C3, the corresponding mean mass spectra were examined. For each cluster, the dominant m/z values were selected and displayed as images in Fig. 1(D1–D3). This reveals, for example, that in the dark green cluster C2, corresponding to the bone, the ^{44}Ca signal is more intense compared to the heart ventricle or the remaining tissue. This is confirmed when selecting ^{44}Ca for image display (Fig. 1(D2)). In contrast, the heart ventricle shows the highest ^{54}Fe signal compared to cluster C1 and C2 in the mass spectra. ^{54}Fe was monitored instead of the main isotope ^{56}Fe, as the latter was intentionally blanked due to exceeding the acceptable intensity levels of the detector. When selecting ^{54}Fe for image display (Fig. 1(D3)), the spatial distribution demonstrates high Fe accumulation in the heart ventricle relative to the bone and the remaining tissue regions. In addition, the liver also exhibits elevated Fe levels. However, in the UMAP embedding, the liver is separated from the heart ventricle based on differences in multiple other elemental signals. This highlights the advantage of UMAP's ability to incorporate the full spectral information from each pixel, rather than just interpreting single isotopic images. A further drawback of inspecting single isotopic images by eye is that the recognition of clusters strongly depends on the chosen colour scale and its contrast. For instance, the ^{44}Ca image is dominated by the intense signal of the bone, and usually imaging software by default adjust its colour scale to that region. However, lowering the maximum value of the scale reveals that Ca is distributed across the entire image (Fig. S2 in the SI). UMAP, in contrast, considers all intensity values simultaneously and therefore avoids overlooking such distributions that may remain hidden to the human eye due to colour scale settings. The advantage of UMAP is that a preselection or inspection of the dataset is not required. Instead, all m/z values can be included directly, and the results can be evaluated afterwards. As such, UMAP is a valuable tool for exploratory analysis, enabling rapid assessment of spectral similarities and differences within the dataset. It is particularly useful for non-targeted investigations in which tissue sections are analysed without *a priori* knowledge or expectation. However, it can also support hypothesis-driven research by facilitating comparisons between known spatial structures and data-driven clustering results as demonstrated in Fig. 2. Here, regions of interest (ROIs) in the image were defined based on known biological structures within the tissue (Fig. 2(A1 and A2)). These ROI segments were then transferred onto the UMAP embedding with corresponding colours (Fig. 2(A3)), enabling interactive exploration of whether anatomically defined regions were also clustered based on their spectral profiles. Conversely, segmentation can also be performed directly on the UMAP embedding, with the resulting clusters mapped back onto the spatial image to assess their anatomical relevance (Fig. 2(B1–B3)). This bidirectional approach facilitates the validation of expected spatial

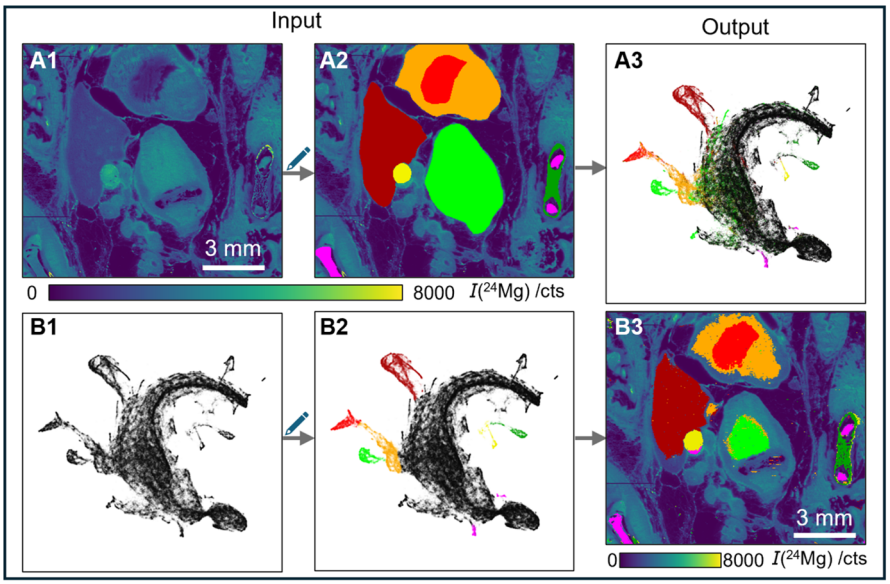

Fig. 2 Interactive LA-ICP-TOFMS image segmentation using UMAP embeddings. Segmentation performed directly on the image (A1 and A2) results in coloured pixels that correspond to equally coloured points in the UMAP embedding (A3). Alternatively, segmentation can be performed on the UMAP embedding (B1 and B2), with the resulting coloured clusters reflected in the image (B3). The black data points in UMAP embeddings A3, B1 and B2 represent data points that were not selected for segmentation and consequently do not appear as black coloured pixels in the image.

patterns and supports the identification of unexpected spectral heterogeneity. In short, UMAP provides a straightforward overview to visualise spatial structures based on their unique elemental composition obtained by LA-ICP-TOFMS.

While this example demonstrates the utility of UMAP in uncovering meaningful spectral patterns and their spatial organisation, the quality and interpretability of the resulting UMAP embeddings are highly dependent on the choice of UMAP parameters. Therefore, consideration must be given to selecting appropriate values particularly for key parameters such as *n_neighbours* and *min_dist* to ensure that both local and global data structure is adequately preserved. Here, a notable benefit is the employed landmark-based UMAP approach with its high computational speed. It was employed to facilitate rapid testing of various hyper-parameter settings to empirically identify a UMAP embedding that aligns with the desired visualisation of the data.

Fig. 3 illustrates the impact of the UMAP hyper-parameter selection on the resulting embeddings of the chicken embryo dataset. The parameter defining the number of nearest neighbours (*n_neighbours*) was systematically varied from 5 to 100, while *min_dist* was held constant at 0.1 (Fig. 3(A1–A6)). As outlined in the introduction, lower *n_neighbours* values emphasise the preservation of local data structure, which may lead to an overrepresentation of minor variations. In this dataset, increasing *n_neighbours* from 5 to 20 improved the definition and separation of clusters (Fig. 3(A1–A4)). However, further increases to 50 and 100 (Fig. 3(A5 and A6)) resulted in slightly reduced cluster separation, likely due to the connection of too many neighbouring data points. Since well-separated clusters are essential for the intended visualisation, *n_neighbours* = 20 was selected as the optimal setting for this dataset.

In addition to *n_neighbours*, the parameter, which defines the minimum distance between points in the low-dimensional embedding (*min_dist*), was varied between 0 and 1 to assess its influence on the UMAP embedding (Fig. 3(B1–B6)). Lower *min_dist* values result in more tightly packed points, enhancing the visual separation of clusters. Conversely, higher values produce more dispersed embeddings, which can improve global structure representation at the expense of local detail. While compact grouping can aid in cluster identification, excessively small *min_dist* values may lead to overplotting, obscuring point density and reducing the interpretability of dense regions. Based on these observations, the final parameters for the chicken embryo dataset were selected to be *n_neighbours* = 20 and *min_dist* = 0.1 (Fig. 3(A4)). As an example that linear dimensionality techniques such as PCA fail to represent the complexity of the data structure, a comparison of UMAP with optimised hyper-parameters and PCA is presented in Fig. S3 in the SI.

3.2 Hierarchical UMAP analysis uncovers substructures in LA-ICP-TOFMS imaging data

While the initial UMAP embedding provides a global overview of major clusters in the LA-ICP-TOFMS dataset, finer spatial or compositional differences within these clusters may remain unresolved. To explore such hidden substructures, a hierarchical UMAP approach was employed, in which individual clusters identified in the initial embedding ($UMAP^1$ level) were isolated and re-analysed using UMAP resulting in a new embedding ($UMAP^2$ level). This iterative strategy can support the identification of subtle spectral variations within specific tissue regions and enable more detailed spatial interpretation of complex biological samples. The following section demonstrates this approach using a honeybee dataset, revealing subclusters and spatial features not distinguishable in the initial UMAP embedding.

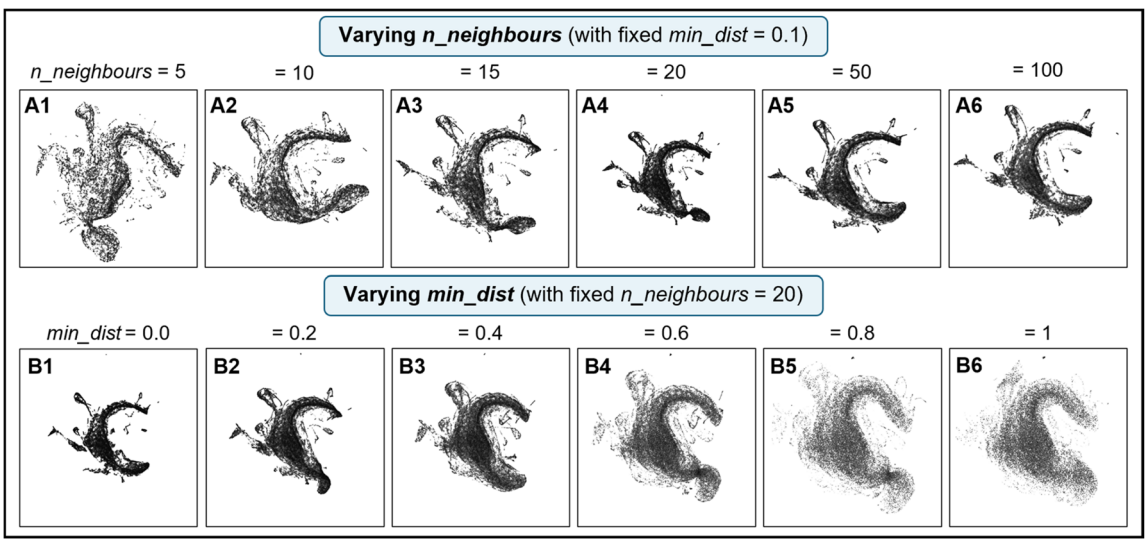

Fig. 3 Effect of varying UMAP hyper-parameters *n_neighbours* and *min_dist* on the embedding (related to Fig. 1). *n_neighbours* is varied from 5 to 100, with *min_dist* fixed at 0.1 (A1–A6). *min_dist* is varied from 0 to 1, with *n_neighbours* fixed at 20 (B1–B6). Additional input parameters included all recorded *m/z* and all pixels of the ablated area shown in Fig. 1(A2).

The data analysis starts with an initial UMAP embedding (UMAP^1) of the full dataset, including all pixels from the ablated area and the complete set of recorded *m*/*z* values (Fig. 4(A1)). Following the previously outlined exploration strategy, the lasso selection tool was used to identify major clusters in the UMAP^1 embedding (Fig. 4(A2)), which correspond to distinct anatomical structures in the honeybee, as visualised in the clustered LA-ICP-TOFMS image (Fig. 4(A3)). These include the flight muscle (green), the crop also known as honey stomach (blue), the midgut (red), and the remaining tissue (orange). Additionally, the surrounding cryo-embedding medium (grey clusters), which covered the specimen and filled vacant spaces within, was successfully separated from biological tissue. Vertical stripe patterns, highlighted in black, are also visible in the clustered image and grouped together in the UMAP embedding. These features represent auto blank events which were triggered during data acquisition when signal intensities exceeded a certain threshold. During these events, ions in small mass ranges around the high intensity *m*/*z* were automatically scattered by the Bradbury–Nielson gate in the ICP-TOFMS instrument. After a brief interval, the blanking is deactivated, and the full mass range is once again acquired. Consequently, *m*/*z* images within the affected mass region display rows of missing data along the ablation direction. As UMAP is sensitive to variations in signal intensity across all selected *m*/*z* values, it effectively identifies and clusters these artefactual features, thereby facilitating their recognition and potential removal during data analysis. This data-dependent recognition of artefacts, such as auto blank events, provides a useful tool to mask compromised pixels for downstream analyses or quantification.

In a second step, the orange cluster from the UMAP^1 embedding was isolated and processed once again with UMAP resulting in a new UMAP embedding (UMAP^2, Fig. 4(B1)), revealing finer substructures within this region. Distinct subclusters emerge, including a yellow cluster that corresponds to the bee's brain and a brown cluster likely associated with glandular tissue (Fig. 4(B2 and B4)). These subclusters can be retrospectively visualised in the context of the initial embedding by overlaying their segmentation on the UMAP^1 embedding (Fig. 4(B3)), illustrating how these finer structures were previously unresolved.

A similar hierarchical analysis was applied to the crop region (blue cluster in UMAP^1). In this case, only the 14 most intense and artefact-free *m*/*z* signals from the crop cluster were selected as input for UMAP^2. This improved the separation of internal substructures compared to using the full spectral range (Fig. 4(C1)). The resulting subclusters reveal the crop wall and further blanking events (Fig. 4(C2 and C4)), exhibiting a strong correlation with spatial patterns observed in multiple single-element ion images (see Fig. S4 in the SI). This procedure represents a targeted application of UMAP where input parameter settings are refined until a desired spatial structure

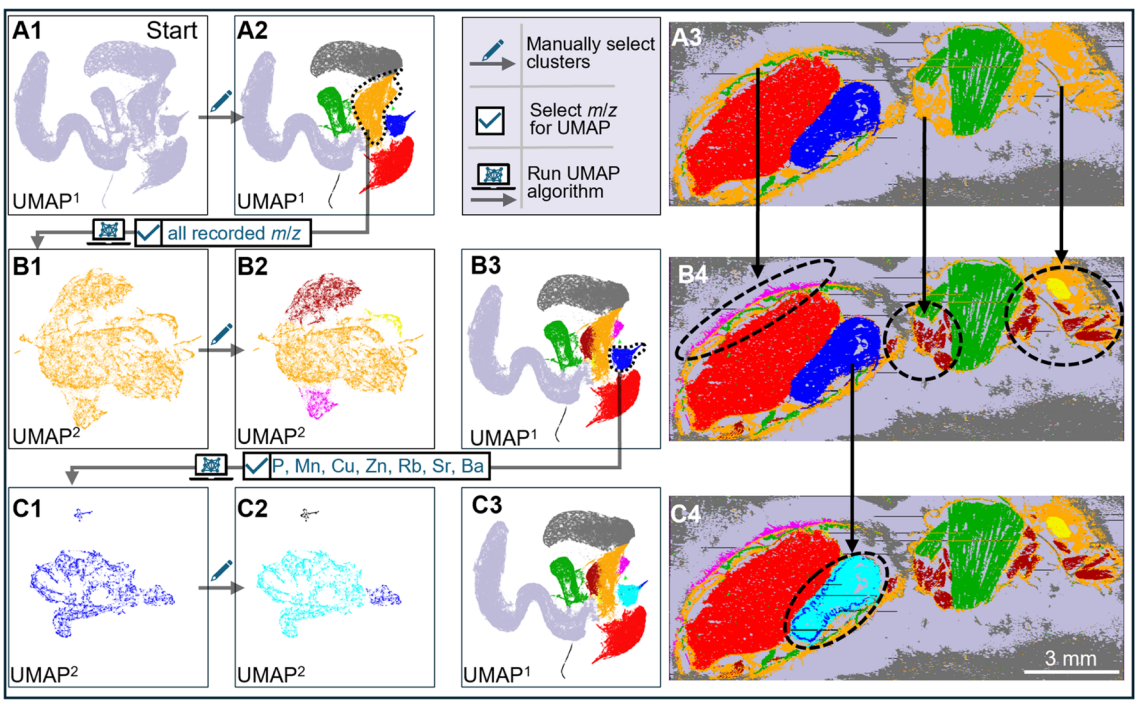

Fig. 4 Hierarchical UMAP analysis with subcluster visualisation of a honeybee thin section analysed by LA-ICP-TOFMS. An initial UMAP analysis was performed on the full dataset resulting in the initial UMAP embedding (UMAP^1, A1) to identify major clusters (A2) which correspond to distinct regions in the LA-ICP-TOFMS image (A3). The orange cluster identified in UMAP^1 was isolated for a second UMAP embedding (UMAP^2), revealing finer substructure (B1, B2 and B4) not apparent in the initial UMAP^1 embedding (B3). The same hierarchical approach was applied to the blue cluster from UMAP^1, uncovering detailed spatial organisation of the honeybee crop (C1–C4). The generation of all UMAP embeddings used *n_neighbours* = 10 and *min_dist* = 0.1. Additional parameters: UMAP^1 (A1) – all pixels of ablated area, all recorded *m*/*z*; orange cluster UMAP^2 (B1) – orange cluster pixels in UMAP^1, all recorded *m*/*z*; blue cluster UMAP^2 (C1) – blue cluster pixels in UMAP^1, specific *m*/*z* (see Fig. S4).

forms a cluster in the UMAP embedding. A targeted UMAP analysis is particularly useful in cases where substructures are visible in individual *m*/*z* images but difficult to segment by manually drawing ROIs. By restricting the input to relevant *m*/*z* values, the UMAP embedding can be refined to enhance the grouping of pixels of these specific regions. This facilitates the detection of subtle substructures compared to embeddings generated using the full *m*/*z* range. Details on the targeted subclustering of the crop and the selected *m*/*z* are provided in the SI (Fig. S4). Once more, these subclusters may be retrospectively visualised in the context of the initial embedding by superimposing their segmentation on the UMAP[1] embedding (see Fig. 4(C3)). The mean mass spectra of the final image segments are presented in Fig. S5 in the SI.

The use of a targeted UMAP analysis workflow can be particularly advantageous in scenarios where the image segmentation of known spatial ROIs is desired but manual segmentation is impractical and tedious. The fine structure in the crop of the honeybee demonstrates a visual example of this scenario. Other exemplary cases would be the segmentation of individual nanoparticles that are dispersed throughout the entire ablated area, inclusions in geological samples or biological tissues with a complex spatial structure, for example lung or cancer tissue. In such instances, the parameters as well as the spatial and spectral input for UMAP can be adjusted until desired ROIs can be readily segmented *via* the UMAP embedding rather than *via* the image directly.

These exemplary applications of hierarchical UMAP demonstrate its potential to uncover biologically meaningful substructures within spatially complex and highly dimensional imaging datasets. By combining non-targeted and targeted UMAP analyses, both broad tissue architecture and subtle, compositionally distinct regions can be resolved with minimal manual image segmentation. UMAP is frequently used in single cell biology, for example for the analysis of high-dimensional data arising from antibody staining techniques such as flow cytometry or imaging mass cytometry.[32,33] These applications showcase the ability of UMAP to uncover cellular heterogeneity based on antibody staining. However, as demonstrated in this study, the potential of UMAP extends well beyond these established use cases. Its application to label-free LA-ICP-TOFMS imaging data rather than data obtained from metal-conjugated antibody staining demonstrates that UMAP can also effectively distinguish more subtle elemental profiles within tissues and discern larger biological morphologies.

Moreover, it is important to consider the impact of elements with multiple isotopes in UMAP analysis. A potential concern is that the inclusion of all isotopes of a multi-isotopic element could result in an overweighting of this element in the UMAP embedding relative to mono-isotopic elements. In this study, a global normalisation was applied across all *m*/*z*, thereby preserving the relative signal intensities between isotopes. An overemphasising of multi-isotopic elements would only arise if normalisation were performed individually for each *m*/*z* (or isotope), but not under global normalisation. Compared to mono-isotopic elements, multi-isotopic elements are expected to receive a lower weighting in the UMAP embedding as their abundance and overall intensity are distributed across several isotopes. Selecting only a single isotope from a multi-isotopic element further amplifies this underweighting. The degree of underweighting also depends on the chosen distance metric (*e.g.*, Euclidean distance or cosine similarity), which is a user-defined UMAP parameter. In this study, cosine similarity was selected as it results in less underemphasising of multi-isotopic elements compared to Euclidean distance. Summing the signal intensities from all isotopes of an element may appear to be a solution, but this approach introduces drawbacks in non-targeted data analysis due to isobaric and polyatomic interferences, as well as interferences from doubly charged ions. Therefore, in this study, all individual isotopes of an element were included for UMAP analysis. This strategy minimised underweighting while avoiding the need for prior knowledge of possible interferences. Furthermore, the inclusion of *m*/*z* values dominated by noise, which do not represent any relevant spatial distribution in the image, can be discussed. As these *m*/*z* values are usually low in signal intensity, they have only a minimal impact on the UMAP embedding. For the non-targeted UMAP approaches used in this study, all recorded *m*/*z* values were included, as this requires minimal prior knowledge of the dataset. However, as UMAP is based on user-defined inputs, future users may choose to include one or all of an element's recorded isotopes, to sum up multiple isotopes of elements while considering potential interferences or to filter noisy *m*/*z* before applying UMAP.

While UMAP offers clear advantages for the analysis of spectral imaging data, it is equally important to recognise the potential for misinterpretation of its results. As previously noted, the method is sensitive to hyper-parameter choices. Moreover, the axes of UMAP embeddings lack intrinsic meaning, and the distances between points do not correspond linearly to those in the original high-dimensional space due to inherent distortions introduced during dimensionality reduction. As a result, both the relative sizes of clusters and the distances between them should not be interpreted as quantitatively meaningful. In this light, UMAP should be understood primarily as an exploratory tool, effective for uncovering patterns and generating hypotheses, rather than as a definitive means of classification.

Nonetheless, UMAP remains a highly promising approach for advancing high-dimensional image analysis workflows across many diverse applications. Finally, it is worth emphasising that the workflows for applying UMAP demonstrated herein are not limited to LA-ICP-TOFMS data of biological specimens. Instead, they can be seamlessly employed in analogous ways for geological or other materials analysed with any type of high-dimensional imaging technique. Further studies could also focus on co-registering images from multimodal imaging approaches and using them together as input for the UMAP algorithm for improved image segmentation.

4 Conclusion

This study explored the utility of uniform manifold approximation and projection (UMAP) as a tool for dimensionality

reduction and for the investigative analysis in laser ablation-inductively coupled plasma-time of flight mass spectrometry (LA-ICP-TOFMS) imaging. UMAP enables the visualisation of complex high-dimensional MS imaging data in a low-dimensional embedding that facilitates the identification of pixels with spectral similarity. In this study, datasets obtained from LA-ICP-TOFMS analysis of a chicken embryo and a honeybee thin section showed that UMAP can distinguish anatomical regions and reveal meaningful substructures that are often not apparent from individual isotope images. The UMAP embedding offers a compact and information-rich representation of the data, facilitating rapid visual exploration, hypothesis generation, and downstream analysis. To support future users in achieving the desired visual appearance of the UMAP embedding with a focus on either global or local structure, a systematic evaluation of the most critical UMAP parameters *n_neighbours* and *min_dist* was exemplarily presented for the chicken embryo dataset. Here, the utilisation of a landmark-based UMAP approach was particularly beneficial for rapidly evaluating different hyper-parameter settings. Moreover, it is concluded that a hierarchical UMAP strategy serves to further enhance the segmentation of fine spatial structures and to reveal intra-tissue heterogeneity. By isolating specific clusters in an initial embedding and subjecting them to a second UMAP analysis, either with the full mass range or a selected subset of *m*/*z*, subclusters within complex tissue types such as the bee's crop and brain were resolved. Consequently, UMAP can be effectively employed in targeted analysis workflows, particularly in scenarios where the segmentation of known spatial ROIs is desired, but manual image segmentation is impractical and tedious. A visual example of this scenario was demonstrated in this study for the fine structure in the crop of the honeybee.

UMAP offers a versatile framework for both exploratory and hypothesis-driven analysis of high-dimensional LA-ICP-TOFMS imaging data. It supports the identification of spatial regions based on their distinct chemical composition, the discovery of subtle spectral variations, and the integration of complex multi-isotope information into a coherent analytical workflow.

5 Glossary

5.1 High-dimensional space

A mathematical space where each data point is described by a large number of variables (dimensions). In MSI, each *m*/*z* value represents one dimension, making each pixel a point in a high-dimensional space.

5.2 Dimensionality reduction

In a mathematical context, dimensionality reduction is the mapping of *n*-dimensional data to *m*-dimensional data with $n > m$. In UMAP, the goal commonly is to transform high-dimensional data into a lower-dimensional space while preserving structural properties.

5.3 Embedding

Technical expression used in dimensionality reduction. Denotes that the detection of structures is performed in the high-dimensional space before the structure is mapped to the low dimensional space. Used as term for the low-dimensional (2D) representation of the high-dimensional data. In MSI, UMAP embeddings visualise spectral similarities of pixels to identify chemical patterns.

5.4 Local and global structure

Local structure refers to the relationships between similar data points, while global structure refers to the overall arrangement of data points.

5.5 Topology

A branch of mathematics concerned with geometric objects and their deformation. In the context of UMAP, it refers to the study of how points are connected in a space, focusing on their relationships rather than exact positions or distances.

5.6 Graph

A mathematical data representation consisting of nodes (data points) and edges (connections between points), used in UMAP to model the topology of the data in high-dimensional space.

5.7 Random seed

UMAP is a stochastic algorithm and involves randomness. Setting a fixed random seed produces the same embedding for each execution of the algorithm on the same dataset and equal parameters.

5.8 Hyper-parameter

A configuration variable that governs the behaviour of an algorithm but is not learned from the data. In UMAP, key hyper-parameters are *n_neighbours* and *min_dist*.

5.9 Landmark-based UMAP

A variant of the original UMAP algorithm that reduces computational cost by embedding a selected subset of representative data points (landmarks) and then positioning the remaining points based on their relationships to these landmarks.

5.10 Segmentation

The process of partitioning an image into non-overlapping segments of connected pixels. Typically, it is based on the similarity of the intensities stored in the pixel. In MSI, data-dependent image segmentation is based on the spectral similarity of pixels with respect to *m*/*z* and intensity. In addition, manual segmentation following visible, spatial anatomical structures is common.

Author contributions

K. K. conceptualisation, methodology, investigation, data curation, formal analysis, validation, visualisation, writing – original draft. H. R. methodology, software, data curation, formal analysis, validation, writing – review & editing. N. G. T. W. resources, methodology for chicken embryo assay, writing – review & editing. D. N. methodology for honeybee model, writing – review & editing. M. E. conceptualisation, methodology, writing – review & editing. D. F. resources, supervision of D. N., methodology for honeybee model, writing – review & editing. L. L. UMAP methodology, supervision of H. R., writing – review & editing. R. G. V. LA-ICP-TOFMS methodology, writing – review & editing. D. C. resources, conceptualisation, methodology, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

Further data is available in the supplementary information (SI) and the referenced GitHub repository. Raw and meta data is available upon request. Supplementary information: supplementary figures and tables. See DOI: **https://doi.org/10.1039/d5ja00215j**.

Acknowledgements

Parts of this study have been funded by the Deutsche Forschungsgemeinschaft (DFG) – CRC 1450 – 431460824 (to L. L.). D. C. has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation program (grant agreement No. 101165171, project acronym: NanoArchive). The authors further acknowledge financial support by the University of Graz.

Notes and references

1 P. A. Doble, R. G. De Vega, D. P. Bishop, D. J. Hare and D. Clases, *Chem. Rev.*, 2021, **121**, 11769–11822.
2 A. L. Gray, *Analyst*, 1985, **110**, 551–556.
3 J. T. Van Elteren, V. S. Šelih and M. Šala, *J. Anal. At. Spectrom.*, 2019, **34**, 1919–1931.
4 M. Tanner and D. Günther, *Anal. Chim. Acta*, 2009, **633**, 19–28.
5 D. Metarapi, A. Schweikert, A. Jerše, M. Schaier, J. T. van Elteren, G. Koellensperger, S. Theiner and M. Šala, *Anal. Chem.*, 2023, **95**, 7804–7812.
6 D. P. Myers, G. Li, P. Yang and G. M. Hieftje, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 1008–1016.
7 I. Basabe-Mendizabal, R. Maeda, S. Goderis, F. Vanhaecke and T. Van Acker, *Anal. Chem.*, 2025, **97**, 6481–6488.
8 C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther and B. Bodenmiller, *Nat. Methods*, 2014, **11**, 417–422.
9 K. Kronenberg, J. Werner, P. Bohrer, K. Steiger, R. Buchholz, M. Von Bremen-Kühne, M. Elinkmann, P. M. Paprottka, R. F. Braren, F. K. Lohöfer and U. Karst, *Metallomics*, 2023, **15**, mfad052.
10 K. Kronenberg, J. Werner, M. Seeba, H. Rave, L. Linsen, K. Steiger, A. Jeibmann, P. Bohrer, P. M. Paprottka, R. F. Braren, F. K. Lohöfer and U. Karst, *ChemRxiv*, 2023, preprint, DOI: **10.26434/chemrxiv-2023-85hbd**.
11 A. M. Oros-Peusquens, A. Matusch, J. S. Becker and N. J. Shah, *Int. J. Mass Spectrom.*, 2011, **307**, 245–252.
12 H. A. O. Wang and M. S. Krzemnicki, *J. Anal. At. Spectrom.*, 2021, **36**, 518–527.
13 H. B. Andrews, L. Hendriks, S. B. Irvine, D. R. Dunlap and B. T. Manard, *J. Anal. At. Spectrom.*, 2025, **40**, 910–920.
14 C. Schwarz, R. Buchholz, M. Jawad, V. Hoesker, C. Terwesten-Solé, U. Karst, L. Linsen, T. Vogl, V. Hoerr, M. Wildgruber and C. Faber, *ACS Infect. Dis.*, 2022, **8**, 360–372.
15 L. Wander, A. Vianello, J. Vollertsen, F. Westad, U. Braun and A. Paul, *Anal. Methods*, 2020, **12**, 781–791.
16 T. Smets, N. Verbeeck, M. Claesen, A. Asperger, G. Griffioen, T. Tousseyn, W. Waelput, E. Waelkens and B. De Moor, *Anal. Chem.*, 2019, **91**, 5706–5714.
17 H. Hotelling, *J. Educ. Psychol.*, 1933, **24**, 417–441.
18 M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos and E. Tuzhilina, *Nat. Rev. Methods Primers*, 2022, **2**, 100.
19 L. Van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
20 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018, DOI: **10.48550/arXiv.1802.03426**.
21 E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux and E. W. Newell, *Nat. Biotechnol.*, 2019, **37**, 38–44.
22 Y. Yang, H. Sun, Y. Zhang, T. Zhang, J. Gong, Y. Wei, Y.-G. Duan, M. Shu, Y. Yang, D. Wu and D. Yu, *Cell Rep.*, 2021, **36**, 109442.
23 D. Wu, J. Y. Poh Sheng, G. T. Su-EnM. Chevrier, J. L. Jie Hua, T. L. Kiat Hon, J. Chen, *bioRxiv*, 2019, preprint, 549659, DOI: **10.1101/549659**.
24 A. Coenen and A. Pearce, *Understanding UMAP*, **https://pair-code.github.io/understanding-umap/**, accessed 30 May 2025.
25 J. Healy and L. McInnes, *Nat. Rev. Methods Primers*, 2024, **4**, 82.
26 L. McInnes, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, **https://umap-learn.readthedocs.io/en/latest/**, accessed 30 May 2025.
27 L. Parcalabescu, *UMAP explained*, **https://www.youtube.com/watch?v=6BPl81wGGP8**, accessed 30 May 2025.
28 J. Starmer, *UMAP: Main Ideas*, **https://www.youtube.com/watch?v=eN0wFzBA4Sc**, accessed 30 May 2025.
29 J. Starmer, *UMAP: Mathematical Details*, **https://www.youtube.com/watch?v=jth4kEvJ3P8**, accessed 20 May 2025.

30 M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields and C. Trapnell, *Nat. Commun.*, 2020, **11**, 1537.
31 A. Diaz-Papkovich, L. Anderson-Trocmé and S. Gravel, *J. Hum. Genet.*, 2021, **66**, 85–91.
32 G. Braun, M. Schaier, P. Werner, S. Theiner, J. Zanghellini, L. Wisgrill, N. Fyhrquist and G. Koellensperger, *JACS Au*, 2024, **4**, 2197–2210.
33 L. Ferrer-Font, J. U. Mayer, S. Old, I. F. Hermans, J. Irish and K. M. Price, *Cytometry, Part A*, 2020, **97**, 824–831.
34 T. Smets, E. Waelkens and B. De Moor, *Anal. Chem.*, 2020, **92**, 5240–5248.
35 Directive 2010/63/EU of the European Parliament and of the council of 22 September 2010 on the protection of animals used for scientific purposes, **https://eur-lex.europa.eu/eli/dir/2010/63/oj/eng**, accessed 30 May 2025.
36 L. McInnes, J. Healy, N. Saul and L. Großberger, *J. Open Source Softw.*, 2018, **3**, 861.