ARTICLE   OPEN

Check for updates

Toward nanoscale molecular mass spectrometry imaging via physically constrained machine learning on co-registered multimodal data

Nikolay Borodinov[1], Matthias Lorenz [ID][1,2], Steven T. King[1], Anton V. Ievlev[1] and Olga S. Ovchinnikova [ID][1][✉]

Mass spectrometry imaging (MSI) plays a pivotal role in investigating the chemical nature of complex systems that underly our understanding in biology and medicine. Multiple fields of life science such as proteomics, lipidomics and metabolomics benefit from the ability to simultaneously identify molecules and pinpoint their distribution across a sample. However, achieving the necessary submicron spatial resolution to distinguish chemical differences between individual cells and generating intact molecular spectra is still a challenge with any single imaging approach. Here, we developed an approach that combines two MSI techniques, matrix-assisted laser desorption/ionization (MALDI) and time-of-flight secondary ion mass spectrometry (ToF-SIMS), one with low spatial resolution but intact molecular spectra and the other with nanometer spatial resolution but fragmented molecular signatures, to predict molecular MSI spectra with submicron spatial resolution. The known relationships between the two MSI channels of information are enforced via a physically constrained machine-learning approach and directly incorporated in the data processing. We demonstrate the robustness of this method by generating intact molecular MALDI-type spectra and chemical maps at ToF-SIMS resolution when imaging mouse brain thin tissue sections. This approach can be readily adopted for other types of bioimaging where physical relationships between methods have to be considered to boost the confidence in the reconstruction product.

INTRODUCTION

Recent developments in mass spectrometry imaging (MSI) have enabled characterization of localized molecular composition in signal transduction[1], drug delivery[2], disease progression[3], and forensics[4]. In particular, MSI is an invaluable tool for pharmacological applications that pinpoints spatial distribution of biologically relevant molecules such as proteins[5], peptides[6], lipids[7], and pharmaceuticals[8,9] across the tissue. This allows to visualize metabolic processes and generate direct insights into the biological action of novel drug candidates[9]. Potential impact of the ability to generate mass spectra of the intact species will have a pronounced effect on the cellular and subcellular metabolome research[10]. This subsequently drives the need for novel analytical tools offering higher sensitivity, and detailed chemical information coupled to high-spatial resolution modes.

Much attention has been focused on both developing nanometer scale molecular MSI platforms and combining molecular mass-spectrometry-based chemical imaging with high spatial resolution techniques, such as scanning probes[11,12], optical microscopy[13,14], and electron/ion systems[15]. High-spatial resolution molecular mass spectra are then generated by using fusion-based algorithms[16–18]. However, these approaches are limited by the fundamental difference between physical mechanisms of image generation for MSI and techniques used for data upsampling. As the correlation between information channels is assumed but not constrained by any known relationships, the output of such algorithms is prone to reconstruction errors[19]. Therefore, development and implementation of combining several information channels for scientific use will require integration of physical constraints into data processing workflows.

Here, we apply this principle for the automated prediction of molecular mass spectra with sub-micrometer spatial resolution by incorporating known relations between matrix-assisted laser desorption/ionization (MALDI) and time-of-flight secondary ion mass spectrometry (ToF-SIMS) signals. Both MALDI and ToF-SIMS chemical approaches are widely used for studying biological systems; however, they produce very different information about the chemical composition of the sample due to the difference in physical nature of the imaging methods[20]. In MALDI, analyte species are mixed with a matrix compound that facilitates primary charge carrier formation upon laser irradiation. Charge transfer from the matrix to the analyte molecule enables the preservation of a singly charged intact molecular species for identification. At the same time, the use of laser ionization in standard commercially available geometries limits the spatial resolution to 5–50 μm[20,21]. In contrast, ToF-SIMS uses focused ion beams to release secondary ions from analyte species, enabling sub-micrometer spatial resolutions[22–24] but resulting in significant fragmentation of molecular compounds, which can complicate the interpretation of mass spectra[25]. The combination of both techniques provides complementary information about the sample[26] enabling predictive chemical imaging of intact molecular species with sub-micrometer spatial resolution. This approach can offer critical insights for research in Alzheimer's disease[27,28], cancer[29–31], tuberculosis[32], drug action[33,34], and mechanisms of protein posttranslational modifications[35]. Here, we assume linear relationship between peaks in ToF-SIMS and MALDI spectra and determine the coefficients of their correlation thus enforcing the functional connection between those channels expected from physics.

---

[1]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [2]University of Tennessee, Knoxville, TN, USA. [✉]email: ovchinnikovo@ornl.gov
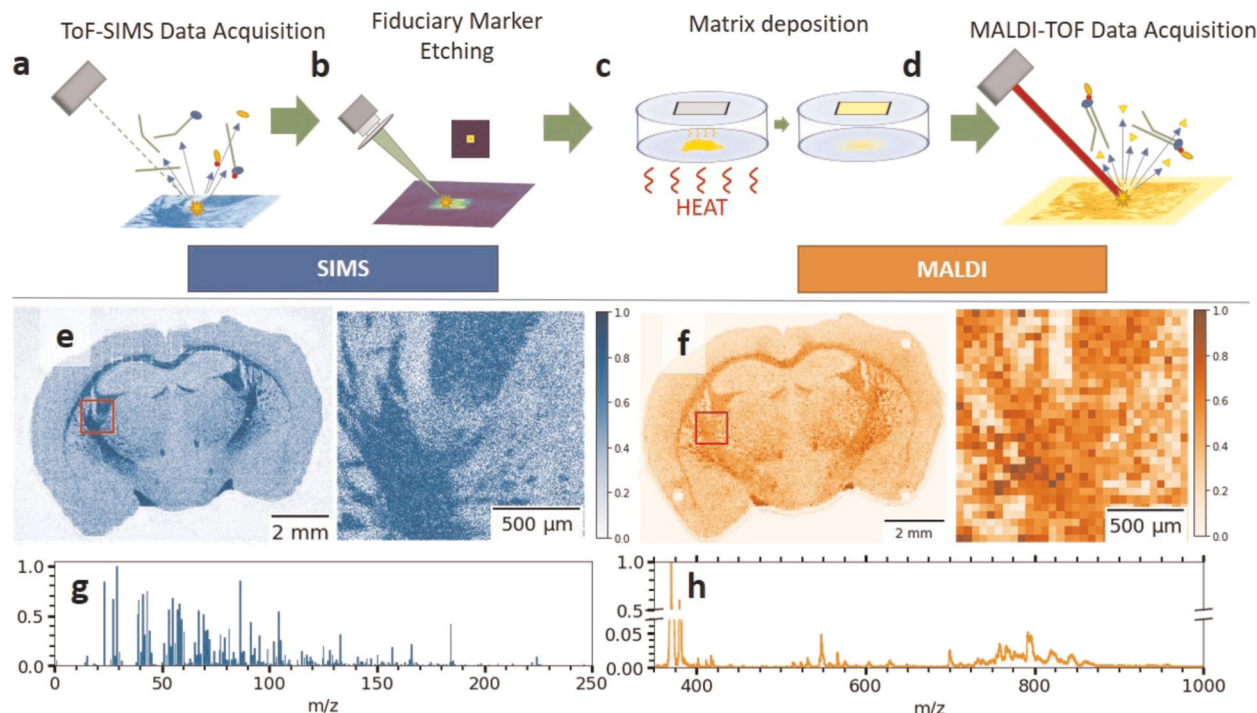
Fig. 1 The data acquisition workflow. a ToF-SIMS sample analysis is followed by the **b** sputtering of fiduciary markers and then by **c** application of the matrix and **d** subsequent MALDI imaging. The exemplary **e** ToF-SIMS and **f** MALDI ions maps for *m/z* 369.35 (MALDI) and *m/z* 95.22 (ToF-SIMS) as well as **g**, **h** corresponding point spectra are shown. A spatially aligned dataset is formed on the next step, SIMS is downsampled to MALDI resolution.

This approach can be extended to incorporate other channels of information and/or reconstruction constraints and serves as a glimpse of powerful capabilities offered by extracting additional information from cross correlating and combinatorial processing of the captured signals for biomedical research. Incorporating of a physical model into the algorithm gives a much clearer outlook on the validity of the workflow output and interpretability of the reconstruction process.

RESULTS

Workflow overview and data collection

Overall, the workflow for processing of the coregistered MALDI and ToF-SIMS signals includes the following steps: (1) co-registration of spectral inputs; (2) dimensionality reduction using non-negative matrix factorization (NMF); (3) identification of the physical relations between channels, using canonical correlation analysis (CCA); and (4) reconstruction of molecular mass spectra at high spatial resolution based on a prediction using the identified physical relationship between MALDI and ToF-SIMS channels. Once data pre-processing is done, the rest of the workflow does not require significant computational capabilities and can be performed on a desktop computer.

MALDI and ToF-SIMS MSI datasets were collected for 10-μm-thick fresh-frozen mouse brain tissue sections that were desic-cated prior to the imaging experiments. ToF-SIMS imaging was performed using a Bi_3^+ ion beam with 2-μm pixel size (Fig. 1a). The Bi_3^+ ion beam was then used to sputter square fiducial markers to simplify co-registration with subsequent MALDI imaging (Fig. 1b). Following ToF-SIMS analysis and fiducial marker imprinting (Fig. 1c), α-cyano-4-hydroxycinnamic acid (CHCA) matrix was applied by sublimation and MALDI imaging performed with 50-μm pixel size (Fig. 1d). Fiducial markers can be clearly seen in the MALDI imaging data (Fig. 1f), which ensured that the

datasets could be correctly co-registered. Details of the spatial co-registration process are discussed in "Methods" section and shown in Supplementary Fig. 1. The resulting data represent two 3-dimensional datasets with matching spatial and spectral (*m/z* binned mass spectra) coordinates but different spatial resolutions (Fig. 1e, f display exemplary ion maps at *m/z* 95.22 and 369.35 and g and h show the point spectra).

Data processing

While the relationship between ToF-SIMS and MALDI point spectra is not explicitly known, it is possible to outline some physical constraints and correlations, which can be used in the development of the reconstruction workflow. Specifically, we assume the following: (1) each compound in the sample has a specific localization independent on imaging technique, (2) those compounds have characteristic non-negative ToF-SIMS and MALDI spectra for the chosen SIMS/MALDI/sample preparation conditions, (3) the mass spectrum in each point represents a linear combination of mass spectra of all compounds presented in this point. While these simplifications are applicable in most cases, the real measurement may be influenced by the matrix effect in ToF-SIMS[36]. These cases have to be addressed separately, however, if the assumptions are satisfied, we can enforce linearity on the data transformation relating the information gathered from ToF-SIMS and MALDI.

The assumptions outlined above allowed us to design a workflow by combining several mathematical transformations of the ToF-SIMS and MALDI data to establish a predictive link between them. First, we reduced data dimensionality using NMF. In addition, ToF-SIMS data are downsampled to MALDI MSI spatial resolution. NMF assumes that each mass spectrum in each point can be represented as a linear combination of a small number of archetypical non-negative spectra referred to as endmembers,
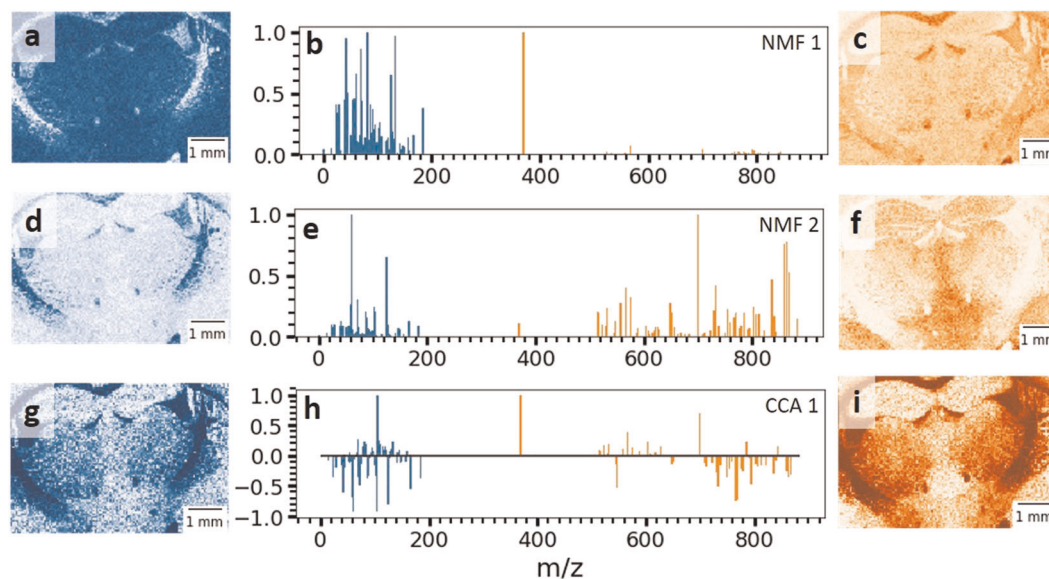
Fig. 2 Multicomponent analysis of the ToF-SIMS and MALDI data. The NMF outputs highlight the abundance maps (**a**, **d** for ToF-SIMS, **c**, **f**, for MALDI) and the endmembers **b** and **e** which are subsequently processed though CCA to establish the relationship between two methods. The loading map for **g** linear combinations of ToF-SIMS and **i** MALDI are very similar suggesting good applicability of CCA for this case. As a result, **h** the pairing between spectral endmembers is also established.

with weights forming abundance maps of the spatial distribution:

$$V = W \times H + R, \tag{1}$$

where V is the original data with the shape $m \times n$ (m is the number of samples or spatial points, n is the number of data points in the mass spectra); H is a matrix of archetypical spectra with size of $p \times n$, (p is number of components); W is matrix of weights with size of $m \times p$; and R is a residual matrix.

The maps of corresponding weights W are further referred to as abundance maps. Here, the algorithm seeks to decrease the reconstruction error R given the amount to components and simultaneously enforces its non-negativity. The latter lies in perfect agreement with the nature of mass spectrometry data. Under this model, the random shots cannot be fit by NMF decomposition and are cast out into the residual matrix R. This allows to preserve the spatial details while improving signal-to-noise ratio. It is important to highlight that one NMF component does not necessarily contain one compound but rather a mixture of compounds that spatially coexists and cannot be separated by linear unmixing. For example, if compounds A and B always appear together in the tissue, their spectra will appear in the same component as well. See Supplementary Figs 1 and 2 for all 20 components for each mode.

This way both ToF-SIMS and MALDI datasets are simplified to a limited number of endmembers and corresponding abundance maps. Results of NMF with 20 components are presented in Fig. 2 and in Supplementary Figs. 1 and 2. Here each endmember (such as Fig. 2b, e, endmember spectra for both modes are displayed on the same axis) corresponds to a certain compound or a group of chemical compounds that spatially appear together (the maps are shown in Fig. 2a, d for ToF-SIMS and c, f for MALDI) and hence cannot be automatically unmixed by NMF.

A single MALDI ion may correspond to multiple ToF-SIMS ions due to molecular fragmentation in ToF-SIMS, similarly a single ToF-SIMS fragment can correspond to several MALDI peaks. Relations between ToF-SIMS and MALDI data are expected to be linear under one of our assumptions. To identify those relations, we utilized CCA. This algorithm looks to maximize the correlation between two column vectors X and Y by identifying the vectors a and b in such a way that $a^T \times X$ and $b^T \times Y$ are correlated (see Supplementary Fig. 3 for all ten CCA components). Here, we use

abundance maps generated by NMF to find the correlation. In particular, using of NMF components instead of the full data allows to reduce the amount of noise in the data, in addition, the peaks with similar abundance are combined into the same NMF component which helps to increase the influence of the peaks with unique spatial distribution. Each pixel in NMF-transformed datasets is characterized by a 20-long vector showing the intensities of the abundance map in that point. ToF-SIMS NMF loading are used as X and MALDI as Y.

This decomposition yields a series of loading map pairs, first of which is displayed in Fig. 2—a ToF-SIMS map (Fig. 2g) is very close to the MALDI map (Fig. 2i). The weights of this decomposition form CCA components and show which linear mixture of ToF-SIMS NMF components would correspond to which linear mixture of MALDI NMF components (displayed in Fig. 2h). Further details on the CCA can be found elsewhere[37]. Overall, the spatial correlation between MALDI and ToF-SIMS NMF abundance maps is used to identify the relationship between corresponding endmembers and, ultimately, the sample mass spectra in both modes. At the same time, for each component (consisting of one or more analyte compounds) present in the sample the localization of abundance maps has to be correlated between the two methods. We use CCA with ten components (Supplementary Information Fig. 3) to link co-registered ToF-SIMS and MALDI MSI datasets.

Prediction of high-spatial resolution molecular spectra

The ultimate goal of our workflow is to predict high-spatial resolution molecular spectra. To achieve this, we need to establish the coefficients of all linear unmixing and linear combination operations, which are done during the training step of the algorithm. This step is performed on the co-registered ToF-SIMS and MALDI MSI datasets brought to the same spatial resolution. These two datasets separately undergo NMF to extract abundance maps, which are then paired through CCA establishing the cross-correlation matrix G. This matrix can be further used to reconstruct either of the dataset. Effectively, the latter allows to infer MALDI-type molecular mass spectra by ToF-SIMS dataset at original high spatial resolution. This dataset is transformed by the NMF using previously calculated endmembers, then a trained CCA transformation infers the MALDI abundance maps via multiplication by
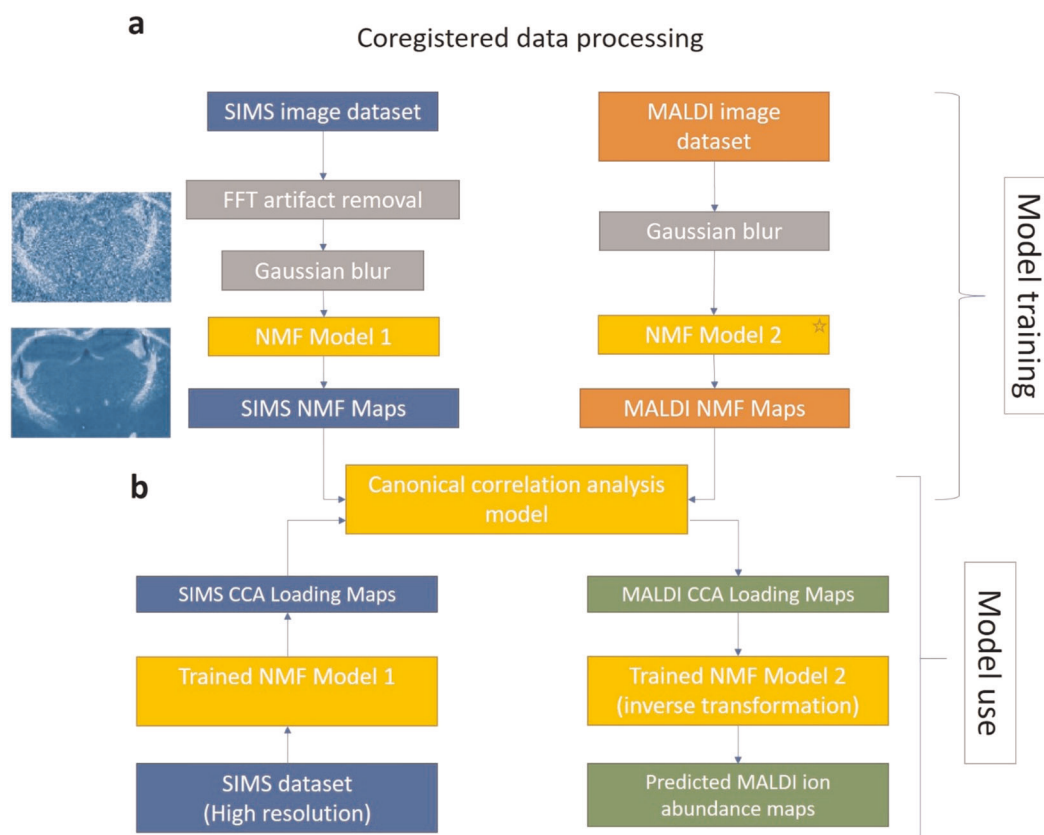
a Coregistered data processing

Fig. 3 Co-registered data processing. a During the training step, ToF-SIMS and MALDI data at the same resolution undergo NMF and filtering and are linked through CCA. The trained NMF and CCA models are later used **b** to process high spatial resolution ToF-SIMS and generate desired reconstructed dataset of molecular intact species.

the matrix G, which is followed by an inverse NMF which generates a MALDI-type mass spectral array at ToF-SIMS imaging resolution (Fig. 3).

DISCUSSION

To estimate the accuracy of this molecular mass spectra reconstruction process we utilized a spectral angle mapper (SAM) to compare actual MALDI spectra in each pixel and the output of the workflow—reconstructed MALDI-type data. SAM values range from 0 to 1, where 0 corresponds to a perfect reconstruction, to 1 corresponding to no correlation between. For our test dataset the SAM values averaged for all pixels across the image was found to be 0.268, which indicates sufficient quality of the spectral processing. To further evaluate reconstruction quality, we compared a MALDI ion map (m/z 369.35) (Fig. 4a), with a reconstructed molecular high spatial resolution map generated by the workflow (Fig. 4b) and a ToF-SIMS ion map of cholesterol (Fig. 4c) and its fragment (m/z 95.07)[38] (Fig. 4d). One can see a close resemblance of the ion maps acquired by both modes and the workflow product. While it is possible to detect some molecular ions in ToF-SIMS mode, as exemplified here, these maps have overall low ion count and some fine details cannot be resolved (e.g., missing features from Fig. 4d). Using the entire ToF-SIMS spectra rather than a singular peak allows to efficiently reconstruct the distribution of the intact molecular species that was almost absent in the original SIMS spectra, which highlights the strength of our algorithm. In addition, phospholipid MALDI maps clearly correlate with phosphocholine SIMS ion maps and appear in the same CCA component with the same sign (negative) (Supplementary Fig. 4 for details). This fact indicates that the workflow puts the signal generated by the same compound in the

same CCA components thus recognizing its cooccurrence. It is important to notice that there are some small differences between the central region of Fig. 4b, d. This particular region of interest is very small and features apparently distinctive spectrum. It is experiencing reconstruction errors outpacing the rest of the image. We believe this could be caused by the fact that when calculating the scores for NMF and CCA the contribution of this region to the overall reconstruction error is too small, so it is not being considered on par with larger regions featuring their unique spectra. We believe that this could be addressed by performing training of NMF models with respect to this apparent class imbalance. The simplest method to tackle this would be to select a number of regions of interest, which would have different areas with distinctively different spectra and then use this dataset for endmember extraction. Then, trained models could be deployed for the full image. This issue is going to be addressed in future publications.

The approach developed here can be used to reconstruct spatial maps of any molecular species distribution with sub-micrometer spatial resolution under the conditions that they are visible in MALDI spectra and a reliable correlation of those peaks with signals in ToF-SIMS spectra can be found. Figure 5a, b shows the MALDI MSI data for m/z 513.76 and its reconstructed high spatial resolution image, respectively. The reconstructed image clearly resolves fine features, which are barely visible in the original data. Similarly, comparison of the original and reconstructed images for m/z 676.45 (Fig. 5c, d, respectively) highlights fine features of the tissue, which are smaller than pixel size in the MALDI dataset. The reconstructed point spectra (point spectrum for MALDI, averaged over 25×25 pixel area for reconstructed) (Fig. 5f) clearly show molecular signals, which are consistent with the original MALDI data (Fig. 5e). This showcases that the
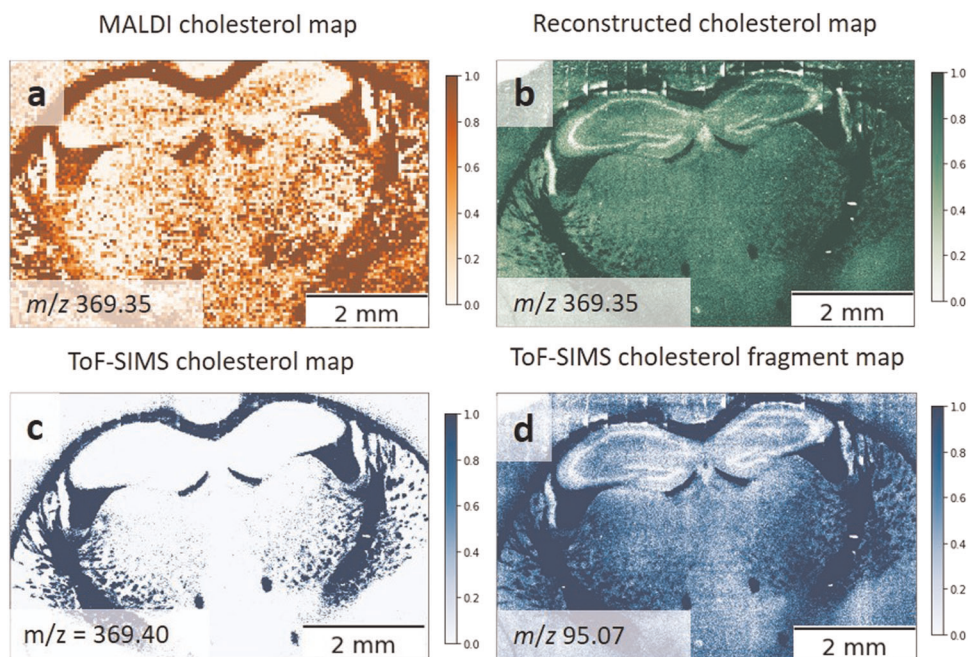
MALDI cholesterol map   Reconstructed cholesterol map
a *m/z* 369.35   2 mm   **b** *m/z* 369.35   2 mm

ToF-SIMS cholesterol map   ToF-SIMS cholesterol fragment map
c m/z = 369.40   2 mm   **d** *m/z* 95.07   2 mm

Fig. 4 The output of the reported workflow. a MALDI *m/z* 369.35 ion map and **b** reconstructed map, **c** ToF-SIMS map attributed to cholesterol and **d** cholesterol fragment (*m/z* 95.07) are well matched by the output of the workflow.

developed approach enables accurate spectral prediction based on MALDI spectra with ToF-SIMS spatial resolution.

The developed machine learning-based approach for correlative chemical imaging allows reconstruction of spectral data with improved spatial resolution based on coregistered multimodal imaging. This approach can shine the light on the fine details of molecular distributions within complex systems and can be used for subcellular imaging of biological systems and incorporate other channels such as Raman and FTIR imaging as long as these systems also follow known mixing rules.

METHODS

Sample preparation
Frozen coronal mouse brain sections mounted to ITO-coated glass microscope slides were purchased from Zyagen (San Diego, CA) and stored at −80 °C until analysis. Just prior to analysis, slides were retrieved from cold storage and placed into a hand-pumped vacuum desiccator alongside silica gel desiccant for transport between buildings and to allow the samples to warm to room temperature in a dry environment. Samples were not washed prior to analysis.

SIMS imaging
The ITO-coated glass slide was mounted on top of a top mount sample holder for the ToF-SIMS instrument using electrically conductive carbon tape. A conductive pathway between the stage and the ITO surface of the slide was confirmed by using a digital multimeter to gauge the resistance between the two surfaces. A TOF.SIMS 5 secondary ion mass spectrometer (IONTOF GmbH, Münster, Germany) was used with a Bi$_3^+$ primary ion beam (31 nA DC current). The instrument was operated in spectrometry mode, which provides a mass resolution of $m/\Delta m = 3000$–10,000. Charge compensation was enabled. Samples were imaged in stage scan experiments, acquiring five shots per pixel with a pixel step size of 2 μm. The patched image of 10.9 mm × 8.4 mm dimensions was recorded in about 5 h.

Fiduciary marker etching
Three 250 μm × 250 μm patches in the tissue were milled using a SIMS Bi$_3^+$ primary ion beam with 10% duty cycle, scanning 512 × 512 pixels in a sawtooth pattern. These patches were located at the vertices of an

imagined right triangle spanning the area of the tissue section. The locations of the etched markers were obtained from the stage coordinates in the ToF-SIMS dataset with 10 nm precision; however, the MALDI imaging would allow the imaging of the markers with ca. 10 μm resolution.

Matrix application and MALDI imaging
The ITO-coated slide with two tissue samples on it was scored and split in half using a handheld diamond-tipped glass cutter so that each half now bore one mouse brain section. The samples were then gently cleaned with low-velocity dry air to remove glass particulate debris. The sublimation apparatus (Chemglass Life Sciences, Vineland, NJ) was primed with 300 mg dry α-cyano-4-hydroxycinnamic acid (purchased from Sigma Aldrich, St. Louis, MO) powder and coupled to a rough vacuum pump (Pfeiffer). The mouse brain section was mounted to the underside of the coolant reservoir of the apparatus with conductive copper tape. The apparatus was then heated in a sand bath to 155 °C, as measured by a digital thermometer probe placed into the sand bath directly beneath the apparatus, then maintained at a temperature between 150 and 160 °C for 20 min.

MALDI MSI data were acquired using a Bruker Autoflex Speed (Bruker Daltonics, Bremen, Germany) instrument equipped with a Smartbeam-II laser using FlexImaging (version 4.1) and FlexControl (version 3.4). The tissue section was sampled with a laser in a rectangular grid covering the entire tissue surface area with a pixel diameter of 50 μm. Each collected spectrum was the sum of 500 laser shots randomly distributed within the 25-μm radius. Spectra were collected on the mass range *m/z* (0–7300) using a digitizer frequency of 5 GHz, resulting in a spectral resolution of 246,726 data points per spectrum. The resulting spectra were individually normalized by total ion count (TIC) to account for pixel-to-pixel signal variation caused by the uneven distribution of CHCA matrix on the sample surface.

Peaks of interest were manually selected from the MALDI average spectrum. This procedure was practically limited to the mass range from *m/z* (500–950) due to the overwhelming isobaric noise from matrix clusters below *m/z* 450 and contamination by optimal cutting temperature (OCT) compound, which, in addition to being a powerful ion suppressant, generated a wide band of polymeric peaks from *m/z* 1000 to *m/z* 2000, precluding detection of isobaric endogenous species. Due in part to the ion suppressing effect of the OCT contamination layer and also in part to the intentional omission of a sample washing step prior to MALDI analysis, we additionally observed no signals associated with a *m/z* greater than *m/z* 2000. In practice, these factors limited our analysis to a band of 60
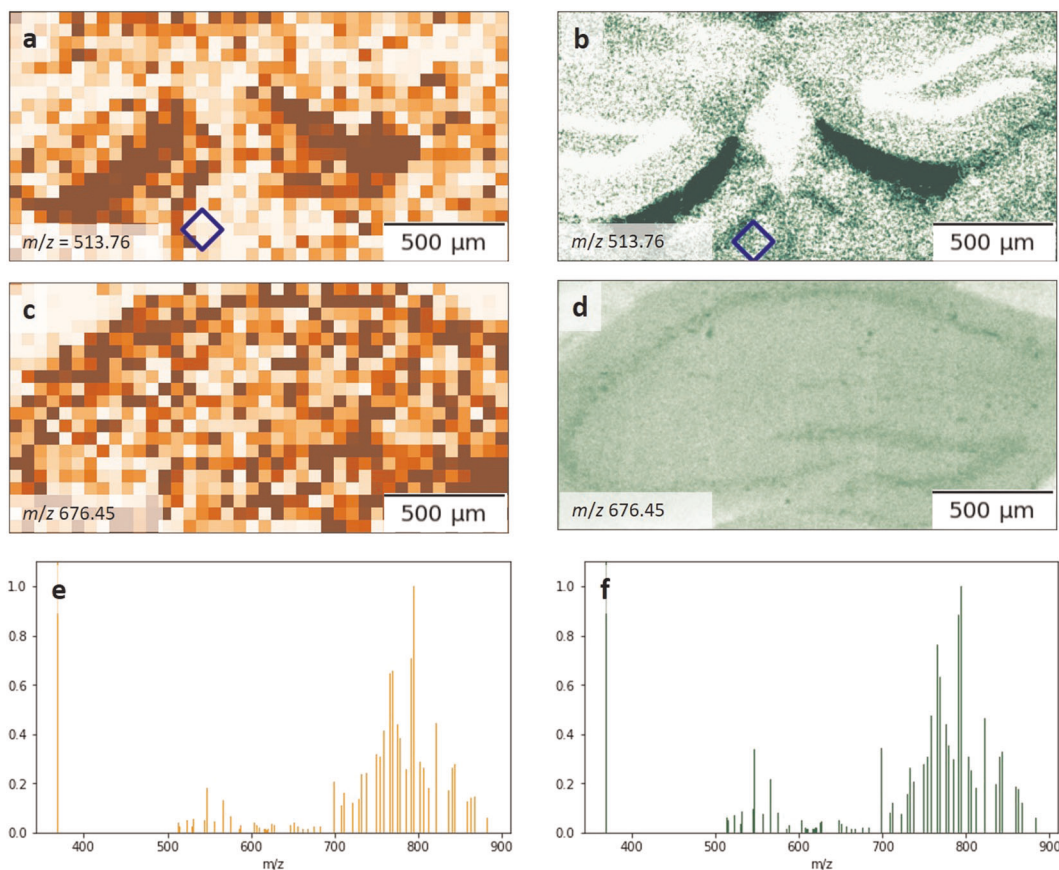
Fig. 5 The zoom-in of ion maps and exemplary point spectra. The examples of molecular spectra inferred from MALDI data are shown for peaks *m/z* 513.76 and *m/z* 676.45. It is evident that the workflow allows for drastic improvement of the spatial resolution of the MALDI data. **a**, **c** The comparison between the original data and **b**, **d** the workflow output shows that the CCA-based pipeline reveals fine details of the sample. Comparison of MALDI and reconstructed point spectra (**e** and **f**, respectively, taken from the regions of the datasets marked with blue square) demonstrate the fidelity of the reconstruction across the entire mass range.

manually identified peaks typically associated with low-mass lipids between *m/z* 500 and *m/z* 950.

MALDI and SIMS co-registration

Data processing was done using Python 3.6. Both the SIMS and MALDI datasets were converted from different proprietary file formats into a common data structure. We developed and applied Python codecs for converting and merging these dissimilar data formats into the Universal Spectroscopic Imaging Data (USID) format, a file structure based on the Hierarchical Data File format. Both instrument vendor software packages contain closed-source tools to carry out data export into an open-source format. SIMS data were exported from the IONTOF SurfaceLab 7 software as binary GRD files roughly constituting a chronologically ordered list of detected ions, their times-of-flight, and the pixel coordinates from which they originated. We reformatted these data into spectral data of the format intensity-vs.-*m/z* with fixed *m/z* bin size, then saved them as coordinate-linked lists of spectra. Due to the volume of the data contained in these files, we elected to perform spectral (64×) binning on the SIMS dataset to reduce the on-disk size of the output file. MALDI data were exported from FlexImaging in the imzML format, consisting of .ibd binary data file and .imzML plaintext metadata header file. We used pyimzML[39] to parse the imzML file pair to extract spectra, then reformatted the dataset into USID format[40].

USID-formatted SIMS and MALDI data files were coregistered by manually identifying ion maps within the MALDI dataset in which the etched fiduciary markers were visible as the dominant features of the map (Supplementary Fig. 5). Ninety-nine such maps were identified and passed to a two-part automated feature finding algorithm. A sliding window 2D correlation of the MALDI single ion images with the fiducial marker has been done to find the coarse position of the squares. Since the size of the etched marker is precisely known, the pixel dimension of the sliding window can be calculated with high accuracy. The size of the sliding window step can be relatively high

to speed up the feature finding as size matching makes it already very robust. A second stage in the co-registration algorithm is fine adjustment of the scale, physical location, and rotation of the image. Through an iterative procedure maximizing the 2D correlation between the chosen marker shape and the local image, these parameters are identified. The marker coordinates in the MALDI dataset, as well as their counterpart coordinates within the SIMS dataset (as recorded during marker etching), were saved as metadata attributes alongside their respective datasets within the merged USID data file. The linked pairs of coordinates were therefore quickly accessible for use in bidirectional co-registration of individual ion maps.

Then, SIMS is downsampled to MALDI resolution via linear interpolation to establish one-to-one pairing of the spectra. Finally, the ionization for ToF-SIMS leads to "salt-and-pepper" type of noise as a certain number of secondary ions is needed in order to get a representative spectrum. See Supplementary Fig. 6 for the effect of NMF denoising.

To finally coregister image A to image B, the stored pairs of marker coordinates are read from the merged data file and used to calculate a 2 × 3 transformation matrix. This matrix specifies scaling, rotation, translation, and shearing parameters necessary to spatially transform the supplied coordinate pairs of image A into their corresponding coordinate pairs within image B. Once spatially aligned, image A is passed through a linear interpolation filter to either upsample or downsample the image as necessary to match the resolution of image B. This results in a 1:1 translation of image A into the coordinate system of image B such that each pixel in the merged dataset now represents spatially-linked pairs of MALDI and SIMS spectra. This transformation may be performed bidirectionally, either coregistering image A to image B, or coregistering image B to image A. The current implementation of this procedure required over 40 GB of RAM to process the presented datasets in multiprocessing regime and requires Linux for the multiprocessing to be used. As a result, the processing of raw MALDI and ToF-SIMS data is recommended to be performed on a high-performance computing cluster

or cloud-based virtual machine. Single-core processing is also possible through the same package, but is not recommended due to the very long time needed to process the data.

Processing of co-registered SIMS and MALDI datasets

The ToF-SIMS instrument allows stitching of several patches of maximum field-of-view allowing to cover the entire millimeter-sized sample. Within these discrete fields of view, individual pixels are resolved by deflecting the ion beam to the analysis site using conventional ion optics. Near the edges of each field-of-view, the primary ion beam is further deflected just prior to impacting the sample by the accumulated charge of the surface. This results in the appearance of a grid-like artifact on the single ion maps, which is not related to the distribution of the analytes in the sample. To reduce this distortion, we have applied a 2D Fourier filter to decrease the intensity of the measurement artifacts (see Supplementary Fig. 7 for the comparison). A 2D Gaussian smoothing is also performed to eliminate the noise.

In the case of "salt-and-pepper" type of ionization, the mass spectrometry signal in a point is not exactly proportional to the concentration of the compounds in the sample being distorted by random shots. However, these random events are not representative of the system. To extract valuable knowledge out of the dataset, we have used NMF.

Quality of the workflow product: metrics and sanity checks

Several metrics are used to characterize the output of the spectral datasets: root mean square error (RMSE), SAM, cross correlation (CC), and peak signal-to-noise ratio (PSNR). RMSE is an overall measure of the similarity between two datasets, SAM is sensitive to the spectral distortions and CC characterizes the geometrical distortion introduced by the workflow. It is important to notice, however, that in the case where a random noise is present, the values of the metrics (especially SAM and CC) can be low, while the quality of the data processing is satisfactory. Thus, it is important to compare these metrics within the same dataset looking for the best values of workflow parameters. The ideal value of SAM and RMSE is 0 and the ideal value of CC is 1. PSNR is widely used to characterize the image compression quality and is calculated as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.

Dataset size considerations

We specifically focus in this work on the novelty of the data processing aspect of our approach and would like to point out a few decisions that enabled us to limit the amount of data to handle for the model system described here. First, ToF-SIMS allows for submicron resolution (down to 50–100 nm)[41] which would provide additional spatial fidelity, but also boost the dataset size. Same considerations led us to set an absolute threshold for the peak intensity for ToF-SIMS at 1% of maximum TIC intensity and exclude the m/z range above 250 Da. While we collected ions up to 2000 Da with ToF-SIMS (which would include some intact molecular ions), they were not used in a reconstruction process to highlight the applicability of the algorithm for cases where the molecules of interest are too heavy to appear in the secondary ion mode. Second, we decided to focus on lipids in MALDI imaging to highlight our workflow, and sublimation without recrystallization was performed during the matrix application. MALDI allows for the imaging of the species much heavier than 2000 Da, but in order to improve the ion extraction a different matrix application approach would be required. However, in our experimental design the overlap between ToF-SIMS and MALDI spectral ranges allows for an immediate assessment of the algorithm output (such as comparing the cholesterol map obtained my MALDI and cholesterol fragment map obtained by SIMS).

DATA AVAILABILITY

Both raw and processed data are available from the authors upon request.

CODE AVAILABILITY

The Python scripts are available as a supplementary material—a Jupyter Notebook at https://github.com/nickborodinov/mass_spec_paper.

REFERENCES

1. Rubakhin, S. S., Greenough, W. T. & Sweedler, J. V. Spatial profiling with MALDI MS: distribution of neuropeptides within single neurons. *Anal. Chem.* **75**(20), 5374–5380 (2003).
2. Muramoto, S., Gillen, G., Collett, C., Zeissler, C. J. & Garboczi, E. J. ToF-SIMS depth profiling of oral drug delivery films for 3D visualization of active pharmaceutical particles. *Surf. Interface Anal.* **52**, 76–83 (2020).
3. Kriegsmann, J., Kriegsmann, M. & Casadonte, R. MALDI TOF imaging mass spectrometry in clinical pathology: a valuable tool for cancer diagnostics (review). *Int. J. Oncol.* **46**(3), 893–906 (2015).
4. Green, F. M., Salter, T. L., Stokes, P., Gilmore, I. S. & O'Connor, G. Ambient mass spectrometry: advances and applications in forensics. *Surf. Interface Anal.* **42**(5), 347–357 (2010).
5. Spraggins, J. M. et al. Next-generation technologies for spatial proteomics: Integrating ultra-high speed MALDI-TOF and high mass resolution MALDI FTICR imaging mass spectrometry for protein analysis. *Proteomics* **16**(11–12), 1678–1689 (2016).
6. Luxembourg, S. L., Mize, T. H., McDonnell, L. A. & Heeren, R. M. High-spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface. *Anal. Chem.* **76**(18), 5339–5344 (2004).
7. Thomas, A., Charbonneau, J. L., Fournaise, E. & Chaurand, P. Sublimation of new matrix candidates for high spatial resolution imaging mass spectrometry of lipids: enhanced information in both positive and negative polarities after 1,5-diaminonapthalene deposition. *Anal. Chem.* **84**(4), 2048–2054 (2012).
8. LaBonia, G. J., Lockwood, S. Y., Heller, A. A., Spence, D. M. & Hummon, A. B. Drug penetration and metabolism in 3D cell cultures treated in a 3D printed fluidic device: assessment of irinotecan via MALDI imaging mass spectrometry. *Proteomics* **16**(11–12), 1814–1821 (2016).
9. Nilsson, A. et al. Mass spectrometry imaging in drug development. *Anal. Chem.* **87**(3), 1437–1455 (2015).
10. Petras, D., Jarmusch, A. K. & Dorrestein, P. C. From single cells to our planet-recent advances in using mass spectrometry for spatially resolved metabolomics. *Curr. Opin. Chem. Biol.* **36**, 24–31 (2017).
11. Somnath, S., Jesse, S., Van Berkel, G. J., Kalinin, S. V. & Ovchinnikova, O. S. Improved spatial resolution for spot sampling in thermal desorption atomic force microscopy—mass spectrometry via rapid heating functions. *Nanoscale* **9**(17), 5708–5717 (2017).
12. Ovchinnikova, O. S. et al. Co-registered topographical, band excitation nanomechanical, and mass spectral imaging using a combined atomic force microscopy/mass spectrometry platform. *ACS Nano* **9**(4), 4260–4269 (2015).
13. Van de Plas, R., Yang, J., Spraggins, J. & Caprioli, R. M. Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nat. Methods* **12**(4), 366–372 (2015).
14. Belianinov, A. et al. Correlated materials characterization via multimodal chemical and functional imaging. *ACS Nano* **12**(12), 11798–11818 (2018).
15. Milillo, T., Hard, R., Yatzor, B., Miller, M. E. & Gardella, J. Image fusion combining SEM and ToF-SIMS images. *Surf. Interface Anal.* **47**(3), 371–376 (2015).
16. Scupakova, K., Terzopoulos, V., Jain, S., Smeets, D. & Heeren, R. M. A. A patch-based super resolution algorithm for improving image resolution in clinical mass spectrometry. *Sci. Rep.* **9**(1), 2915 (2019).
17. Rabe, J. H. et al. Fourier transform infrared microscopy enables guidance of automated mass spectrometry imaging to predefined tissue morphologies. *Sci. Rep.* **8**(1), 313 (2018).
18. Cassese, A. et al. Spatial autocorrelation in mass spectrometry imaging. *Anal. Chem.* **88**(11), 5871–5878 (2016).
19. Vollnhals, F. et al. Correlative microscopy combining secondary ion mass spectrometry and electron microscopy: comparison of intensity-hue-saturation and laplacian pyramid methods for image fusion. *Anal. Chem.* **89**(20), 10702–10710 (2017).
20. Korte, A. R., Yandeau-Nelson, M. D., Nikolau, B. J. & Lee, Y. J. Subcellular-level resolution MALDI-MS imaging of maize leaf metabolites by MALDI-linear ion trap-Orbitrap mass spectrometer. *Anal. Bioanal. Chem.* **407**(8), 2301–2309 (2015).
21. Ly, A. et al. High-mass-resolution MALDI mass spectrometry imaging of metabolites from formalin-fixed paraffin-embedded tissue. *Nat. Protoc.* **11**(8), 1428–1443 (2016).
22. McDonnell, L. A. & Heeren, R. M. Imaging mass spectrometry. *Mass. Spectrom. Rev.* **26**(4), 606–643 (2007).
23. Ievlev, A. V. et al. Automated interpretation and extraction of topographic information from time of flight secondary ion mass spectrometry data. *Sci. Rep.* **7**(1), 17099 (2017).
24. Ievlev, A. V. et al. Nanoscale electrochemical phenomena of polarization switching in ferroelectrics. *ACS Appl. Mater. Interfaces* **10**(44), 38217–38222 (2018).

25. Bodzon-Kulakowska, A. & Suder, P. Imaging mass spectrometry: Instrumentation, applications, and combination with other visualization techniques. *Mass. Spectrom. Rev.* **35**(1), 147–169 (2016).

26. Mohammadi, A. S., Phan, N. T., Fletcher, J. S. & Ewing, A. G. Intact lipid imaging of mouse brain samples: MALDI, nanoparticle-laser desorption ionization, and 40 keV argon cluster secondary ion mass spectrometry. *Anal. Bioanal. Chem.* **408**(24), 6857–6868 (2016).

27. Nakamura, A. et al. High performance plasma amyloid-beta biomarkers for Alzheimer's disease. *Nature* **554**(7691), 249–254 (2018).

28. Cheng, K. K. et al. Curcumin-conjugated magnetic nanoparticles for detecting amyloid plaques in Alzheimer's disease mice using magnetic resonance imaging (MRI). *Biomaterials* **44**, 155–172 (2015).

29. Melo, S. A. et al. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* **523**(7559), 177–182 (2015).

30. Shroff, E. H. et al. MYC oncogene overexpression drives renal cell carcinoma in a mouse model through glutamine metabolism. *Proc. Natl Acad. Sci. USA* **112**(21), 6539–6544 (2015).

31. Balluff, B. et al. De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *J. Pathol.* **235**(1), 3–13 (2015).

32. Prideaux, B. et al. The association between sterilizing activity and drug distribution into tuberculosis lesions. *Nat. Med.* **21**(10), 1223–1227 (2015).

33. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **15**(7), 473–484 (2016).

34. Chumbley, C. W. et al. Absolute quantitative MALDI imaging mass spectrometry: a case of rifampicin in liver tissues. *Anal. Chem.* **88**(4), 2392–2398 (2016).

35. Larsen, S. C. et al. Proteome-wide analysis of arginine monomethylation reveals widespread occurrence in human cells. *Sci. Signal.* **9**(443), rs9 (2016).

36. Shard, A. G., Spencer, S. J., Smith, S. A., Havelund, R. & Gilmore, I. S. The matrix effect in organic secondary ion mass spectrometry. *Int. J. Mass Spectrom.* **377**, 599–609 (2015).

37. Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004).

38. Piehowski, P. D. et al. MS/MS methodology to improve subcellular mapping of cholesterol using TOF-SIMS. *Anal. Chem.* **80**(22), 8662–8667 (2008).

39. Bald, T. et al. pymzML–Python module for high-throughput bioinformatics on mass spectrometry data. *Bioinformatics* **28**(7), 1052–1053 (2012).

40. Somnath, S., Smith, C. R., Laanait, N., Vasudevan, R. K. & Jesse, S. USID and pycroscopy—open source frameworks for storing and analyzing imaging and spectroscopy data. *Microsc. Microanal.* **25**(S2), 220–221 (2019).

41. Kollmer, F., Paul, W., Krehl, M. & Niehuis, E. Ultra high spatial resolution SIMS with cluster ions - approaching the physical limits. *Surf. Interface Anal.* **45**(1), 312–314 (2013).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

Concept development was done by N.B., M.L., and O.S.O. Data acquisition was done by S.T.K. and M.L. Code development has been done by N.B., A.V.I., and S.T.K. All authors contributed to writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00357-9.

Correspondence and requests for materials should be addressed to O.S.O.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.