

simsMVA: A tool for multivariate analysis of ToF-SIMS datasets

Running title: simsMVA: A tool for multivariate analysis of ToF-SIMS datasets

Gustavo F. Trindade^{a,b*}, Marie-Laure Abel^a, John F. Watts^a

The Surface Analysis Laboratory, Department of Mechanical Engineering Sciences, University of Surrey, Guildford, Surrey, GU2 7XH, UK.

University of Nottingham, Nottingham, NG7 2RD, UK

* Electronic mail: Gustavo.FerrazTrindade@nottingham.ac.uk

Abstract: Imaging mass spectrometry datasets are every year larger and more complex, with unsupervised multivariate analysis (MVA) becoming a routine procedure for most researchers. Moreover, the increasing interdisciplinarity of the field demands the development of software for rapid and accessible MVA for researchers or various backgrounds. This paper presents a MATLAB-based software for performing principal component analysis (PCA), non-negative matrix factorisation (NMF) and k -means clustering of large analytical chemistry datasets with a particular focus on of time-of-flight secondary ions mass spectrometry (ToF-SIMS). All five modes of operation (spectra, profiles, images, 3D and multi) are described with a few examples of typical applications at The Surface Analysis Laboratory of the University of Surrey: point spectra analysis of wood growth regions, depth profiling of a metallic multi-layered sample, imaging of an organic coating on a metal substrate and 3D characterisation of an automotive grade polypropylene.

Contents

1	Introduction	2
2	Algorithms and typical running times	3
3	Main features of the GUI	8
3.1	Spectra mode	10
3.2	Profiles mode	13
3.3	Images mode	15
3.4	3D mode	19

3.5	Multi mode.....	21
4	System requirements and availability	24
5	Declaration of independent implementation	24
6	Exporting data from SurfaceLab 6	25
	Acknowledgements	28
	References	28

Keywords: SIMS; ToF-SIMS; mass spectrometry; MATLAB; GUI; multivariate analysis; principal components analysis; *k*-means clustering; non-negative matrix factorisation; software; toolbox.

1 Introduction

Secondary ion mass spectrometry (SIMS) is based on the detection of ionised atoms, molecules, or molecular fragments generated as a consequence of the bombardment of primary ions onto the surface of the sample under analysis. SIMS has its roots in the characterisation of materials in the semiconductors industry and evolved to be one of the most powerful techniques for the analysis of organic and inorganic materials [1,2]. Modern instruments will contain primary ion probes capable of rastering the samples surfaces and time-of-flight (ToF) detection systems with high speed electronics, which enable parallel detection of a large range of masses with very high sensitivity and specificity [2]. Most surface analysis laboratories have facilities capable of running ToF-SIMS in dual beam depth profiling mode, which will typically generate hyperspectral datasets distributed throughout a 3D cuboid containing more than 256 x 256 x 500 voxels with each voxel containing from 20,000 to 2,000,000 spectral channels.

Over the last twenty years, the use of multivariate analysis (MVA) methods has increased significantly within the SIMS community enabling the processing of large amounts of complex data in a reasonable amount of time and at the same time extract the maximum chemical information from the data. Such a spread of MVA has demanded standardization of the methodology and appropriate software, with a number of reviews and tutorials that set the recommended way of dealing with multivariate data within the SIMS community [3–10]. In terms of software, the most popular spectrometer manufacturers (IONTOF, Ionoptika, Physical Electronics) do not provide a complete set of MVA tools in their analysis software, which requires that researchers go for independently developed alternatives. The three most

widely employed software for MVA within the SIMS community are the PLS.MIA toolbox by Eigenvector research [11], the NBtoolbox, developed by Graham [12] and the MCR-ALS toolbox developed by Jaumot, Gargallo, de Juan and Tauler [13]. However, none of these solutions covers all cases and many research groups adopt data analysis routines developed in house. This was in fact the case of researchers at The Surface Analysis Laboratory of The University of Surrey, where MATLAB routines were developed to solve specific problems. With time, these routines became useful for more group members and were ultimately assembled in a software package called simsMVA. During development, the feedback from colleague researchers dealing with various different samples helped simsMVA to become a versatile tool that is capable of processing all kinds of data typically generated by ToF-SIMS spectrometers, with various data visualisation tools and the ability to handle large and sparse datasets.

2 Algorithms and typical running times

The most complex dataset one can generate with ToF-SIMS is a 3D dataset. Using a dual beam configuration, it is possible to have lateral and depth information arranged in a *3D hyperspectral cuboid* [1]. In order to perform matrix factorisation methods such as principal components analysis (PCA) and non-negative matrix factorisation (NMF) on those datasets, the data must first be unfolded in a manner that enables the final results to be folded back without loss of spatial information. Fig. 1 shows a flowchart describing how the unfolding process is done for a 3D dataset containing n_z levels with maps sized n_x by n_y pixels with m spectral channels (or variables) per voxel. In case of imaging datasets there are no further levels and only level 1 is unfolded and processed whereas for depth profiling data, the lateral information is collapsed into one data point per level. For spectra analysis, every measurement is regarded as a sample and technical repeats are acquired often to increase statistics.

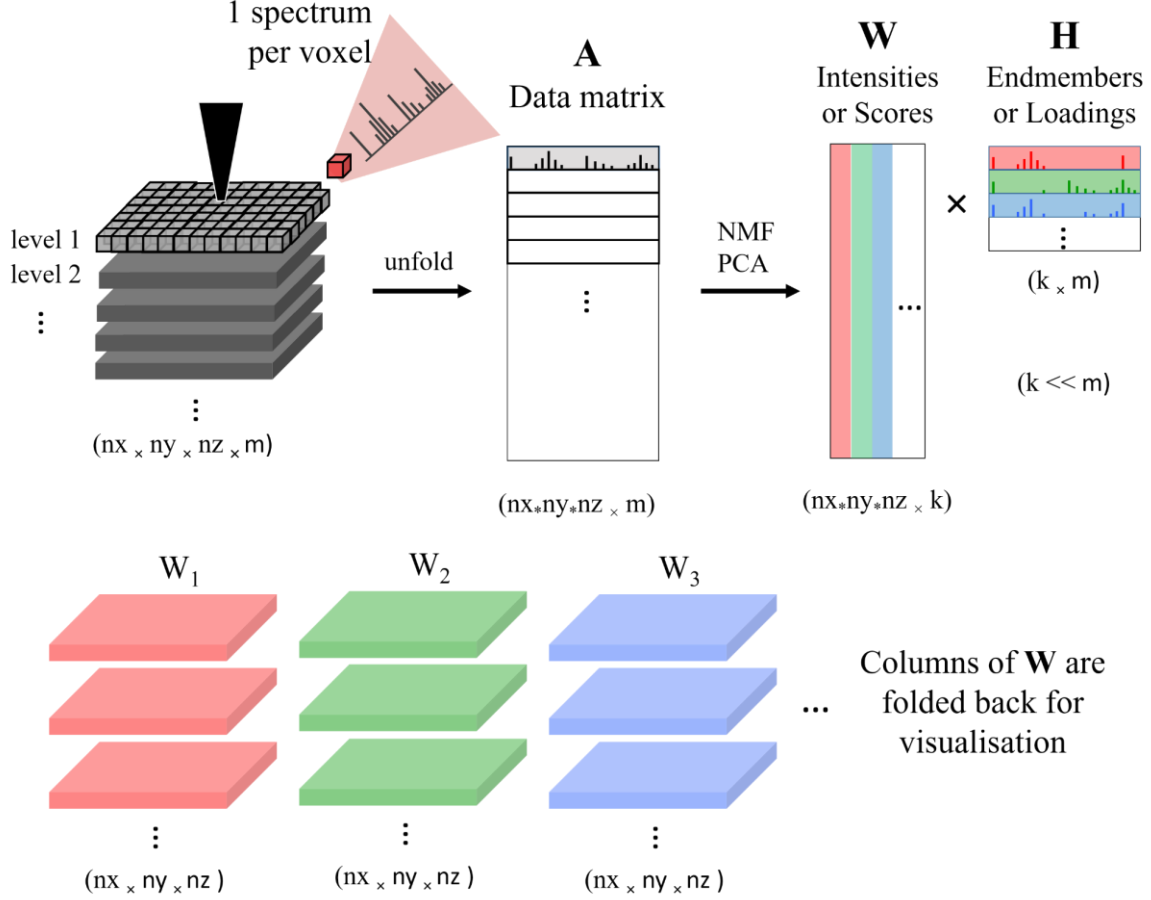


FIG. 1. The unfolding/folding process of a 3D dataset.

For all kinds of datasets, simsMVA offers a series of data pre-processing steps that may be necessary for achieving effective MVA. These are: *scaling*, *binning*, *smoothing*, *normalisation* and *mean-centring*. The major reason for ToF-SIMS data *scaling* is the fact that variables will have different amounts of noise and error (*heteroscedasticity*). The statistics of secondary ion detection in ToF-SIMS can be described, at each spectral channel, by the Poisson probability distribution, thus, considering a dataset arranged in a matrix **M** (as in Fig. 1), a straightforward way of “normalising” the error across all variables is by dividing each value of data matrix **M** by its square root:

$$\mathbf{M_s} = \mathbf{M} / \mathbf{M}^{\frac{1}{2}}$$

Where “./” denotes MATLAB element-wise division. However, typical ToF-SIMS imaging and 3D datasets will have very low counting rates that reflect in large relative uncertainties and make the square root estimation inaccurate. For this reason, Keenan and Kotula [10] proposed a scaling approach that takes into account the relationship among all variables by using the row-wise and column-wise means of the data matrix. A scaled dataset is then written as:

$$\mathbf{Ms} = (\mathbf{G}^{-1/2})\mathbf{M}(\mathbf{H}^{-1/2})$$

Where \mathbf{G} is a diagonal matrix with the unfolded mean image (row-wise mean) along its diagonal and \mathbf{H} is a diagonal matrix with the mean spectrum along its diagonal (column-wise mean). This method of scaling is nowadays widely used in the SIMS community and is known as “Poisson scaling”.

Normalisation consists of dividing all elements of each row of \mathbf{M} by a common value. In other words, it is a correction that is applied to the individual spectra of the dataset. The values used for normalisation can be a specific variable (intensity of a specific peak of the spectrum) or a combination of a number of variables, such as the total ion intensity of a spectrum or the total intensity of all variables present in a dataset (after variables selection). Typically, the main reason for normalisation is to account for differences in acquisition conditions of individual spectra. These conditions can be, for example, different acquisition times, different primary ion beam current or secondary ion yield hindered by charging and topography effects. *Mean-centring* consists of subtracting each column of \mathbf{M} by their average so that the values are distributed around zero. Another scaling method that is more commonly used in other fields (but sometimes applied to ToF-SIMS data) is known as “auto-scaling” and consists of mean-centring the data and dividing each element by its column-wise standard deviation. In other words, auto-scaling converts each value to its distance to the mean in units of standard deviation. *Binning* is only applicable for imaging and 3D datasets and consists of a simple image compression for each ion map (at each level for 3D data). simsMVA uses the *imresize* function of MATLAB’s *statistics and machine learning toolbox*. *Smoothing* can be applied to profiles, images or 3D data. For each data structure, simsMVA uses a specific smoothing routine: a moving-average function for profiles, a Gaussian-kernel filter for imaging and a combination of both for 3D.

For very large and sparse datasets, *simsMVA* allows the user to switch the memory allocation of the data matrix to MATLAB *sparse* prior to MVA. For PCA and *k*-means, the *statistics and machine learning* toolbox functions are used (*pca* for full matrices, *svds* for sparse matrices and *kmeans* for clustering). For NMF, an in-house developed package that is able to handle sparse matrices is used. It is possible to choose among three algorithms: two based on multiplicative update rules originally proposed by Lee and Seung [14,15] and one based on alternating least squares solutions [16]. The use of sparse matrices has the advantage of saving memory but it increases the computational times for MVA. Fig. 2 shows plots of *dataset size vs. computational time* for PCA, NMF (normal multiplicative update algorithm) and *k*-means clustering (current *kmeans* implementation cannot handle sparse matrices) for the imaging ToF-SIMS dataset presented in Section 3. The size of the bubbles is proportional to the number of zero elements in the data matrix. The different dataset sizes were obtained by selecting different mass ranges (some fragments will have lower yield and therefore fewer counts per pixel) and different degrees of pixel binning. The variables in this case are areas integrated under peaks, so most of the zero elements of the original raw dataset are already excluded. When using raw datasets, depending on memory size, using sparse allocation may be the only viable option [17].

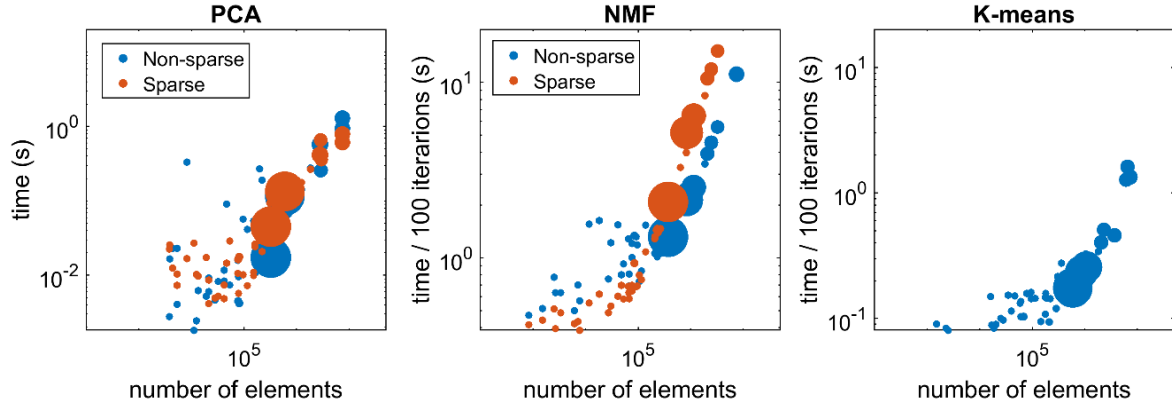


FIG.2. Computational time for PCA, NMF and *k*-means clustering for the imaging ToF-SIMS dataset presented in Section 3.

From the plots in Fig. 2, it can be observed that PCA computational time is not influenced by matrix sparsity whereas the NMF computational time is quicker for sparse matrices when the total matrix size is less than 1×10^5 , becoming slower for higher values.

For NMF, apart from computational time, an important parameter is the *percentage error*, which is defined as:

$$error = 100 \times \sqrt{\frac{\sum_j \sum_i (A_{i,j} - (WH)_{i,j})^2}{\sum_j \sum_i (A_{i,j})^2}}$$

Where \mathbf{A} is the original data matrix and \mathbf{W} and \mathbf{H} are the factorised matrices (as illustrated in Fig. 1).

As mentioned in the introduction, imaging or 3D datasets generated by modern ToF-SIMS instruments can be extremely large (millions of observations \times millions of variables), which makes it impossible to perform the required matrix manipulations and operations using conventional computers due to memory and time limitations. One approach to overcome this problem is the use of *training sets* by means of pixel/voxel subsampling [17–20]. The basic steps of the subsampling done by simsMVA are: Loadings or endmembers \mathbf{Hs} are calculated by factorising (via PCA or NMF) a reduced set of rows (\mathbf{As}) of matrix \mathbf{A} . Matrix \mathbf{W} is then calculated using the pseudo-inverse of \mathbf{Hs} and the original \mathbf{A} . Subsampling works for ToF-SIMS data because, for every impact point of primary ions, there is a fundamental volume (that usually spans more than one pixel/voxel) where all generated secondary ions will be highly correlated. However, potential problems are the discrepancy of random selection of voxels, especially on datasets containing compounds at very low concentrations. This has been addressed by Cumpson et. al [20] with the use of Sobol sequences which have been implemented in simsMVA. Considering such correlation amongst neighbouring pixels/voxels, there must be a subsampling limit where the achieved result is as good as if one did the analysis using all pixels/voxels. Assuming that memory is not an issue, it is also important to take into account the gain in *computational time and error* for larger subsamples. A maximum value of the “*quality parameter*” will represent the ideal subsample size:

$$\text{quality parameter} = (\text{error} \times \text{time})^{-1}$$

Fig. 3 shows the error and quality parameter for NMF ($k = 3$) of the polymer blend dataset as a function of subsample size and after different number of iterations.

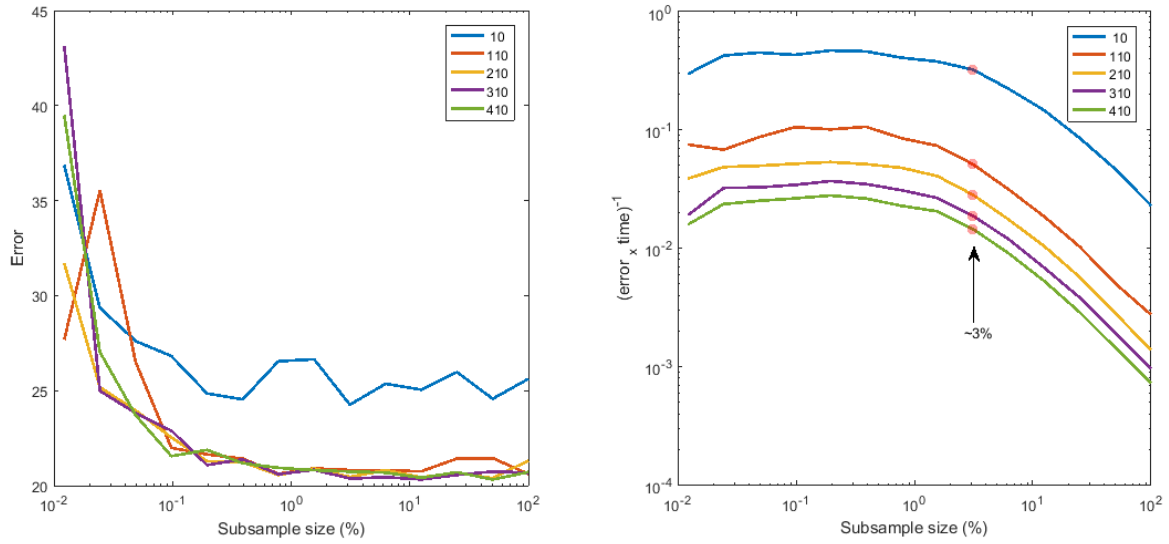


FIG.3. Error end quality parameter for NMF of one patch of the imaging ToF-SIMS dataset presented in Section 3. The different colours represent different numbers of iterations.

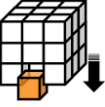
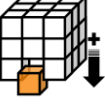



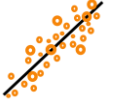
The results show that after around 3% subsample size, the results do not get much better as the subsample size and calculation time increase, resulting in a significant drop in the “quality parameter”. Using 3% subsample sizes as standard can be very useful since it speeds up the calculation time significantly.








3 Main features of the GUI

The main window of simsMVA contains a menu bar with four items: “New tab”, “Theme”, “Examples” and “Help”. The “New tab” menu contains five different options:

“Spectra”, “Images”, “Profiles”, “3D” and “Multi”. Every time one of these options is selected, a new tab is created and the user is prompted to give it a name. The “Examples” menu creates tabs loaded with example datasets acquired at The Surface Analysis Laboratory of the University of Surrey. The “Theme” menu contains a set of different colour schemes that can be applied to the GUI. A detailed description of each feature present in each kind of tab is given in Table I. Every time a dataset of any kind goes through multivariate analysis (PCA, NMF, k -means clustering or PLS), an “MVA results” tab is created. The next sections will describe the main features of each mode of simsMVA with some example datasets acquired at The Surface Analysis Laboratory of the University of Surrey.

TABLE I: Detailed description of main buttons present in each kind of tab.

Item	Description	Spectra	Images	Profiles	3D
New list (Load) 	Opens a file selection dialog that allows the user to select matrices from MATLAB’s workspace or one or more .txt (Spectra and Profiles) or .BIF6 (Images and 3D) files exported from the SurfaceLab software. Files for different samples must have the same peak list.	X	X	X	X
Add extra 	Opens a file dialog that allows the user to select one or more .txt files exported from the SurfaceLab software that will be added to the already loaded files.	X			
Set groups 	Creates a window with all sample names. The user can then give the same name to technical repeats and select which samples to be considered for analysis. Once the groups are set, the scatter plot in the bottom left figure will be updated with matching colours for samples within the same group.	X			
PCA 	Performs principal component analysis and creates an “MVA results” tab with the PCA results.	X	X	X	X
NMF 	Creates a window (NMF menu) that allows the user to select input parameters for non-negative matrix factorisation (algorithm, number of endmembers, number of iterations). The Run button creates an “MVA results” tab with the NMF results.	X	X	X	X
PLS 	<i>Under construction</i>	X			

k-means 	Creates a window (<i>k</i> -means menu) that allows the user to select input parameters of the MATLAB <i>kmeans</i> function (Distance, number of clusters, number of iterations and number of replicates). The Run button creates an “MVA results” tab with the k-means clustering results.		X	X
Save project 	Opens a save file dialog that allows the user to save the project as a .mat file.	X	X	
Load project 	Opens a load file dialog that allows the user to open any saved projects.	X	X	
Line scan 	Changes the mouse cursor to a cross. The user then has to click on two or more points on the left hand side map and press <i>enter</i> to display (in the figure at the bottom right hand side) the ion intensity across the line connecting the selected points (performed at an specific level for 3D data).		X	X
Overlay 	Creates a window that enables the overlay of different ion maps.		X	X
3D view  3D view	Creates a window that shows a 3D view of the intensities of a selected mass peak. The sliders on the left hand side allow the user to slice through the data cuboid.			X
3D overlay  3D overlay	Creates a window that enables a 3D overlay of selected peak intensities.			X

3.1 Spectra mode

The spectra mode is intended for the analysis of point ToF-SIMS spectra but can also be applied to any set of multivariate observations with no specific spatial or temporal inter-relations. Fig. 4 shows a screenshot of a Spectra mode tab loaded with ToF-SIMS data of different wood growth regions. Additionally to wood, samples of reference cellulose and organosolv lignin were also measured. More details about the experiment and results can be found elsewhere [21]. Several different regions were analysed for each sample at both early and late growth regions and a high spectral resolution peak list with areas of forty nine characteristic fragments of lignin and cellulose was created. A Spectra mode tab contains a

table with all loaded data: peak areas of different samples and their associated masses and labels. The first column contains tick boxes that allow the user to unselect specific variables prior to multivariate analysis. The plots on the right hand side show the average peak list for all samples. The upper plot will always show the original data and the bottom plot will show the pre-processed data. The panel on the bottom left will contain a list of all variables where the user can select which peak areas to plot against each other in a matrix. The main data pre-processing steps employed were Poisson scaling, normalisation by total counts, and mean centring.

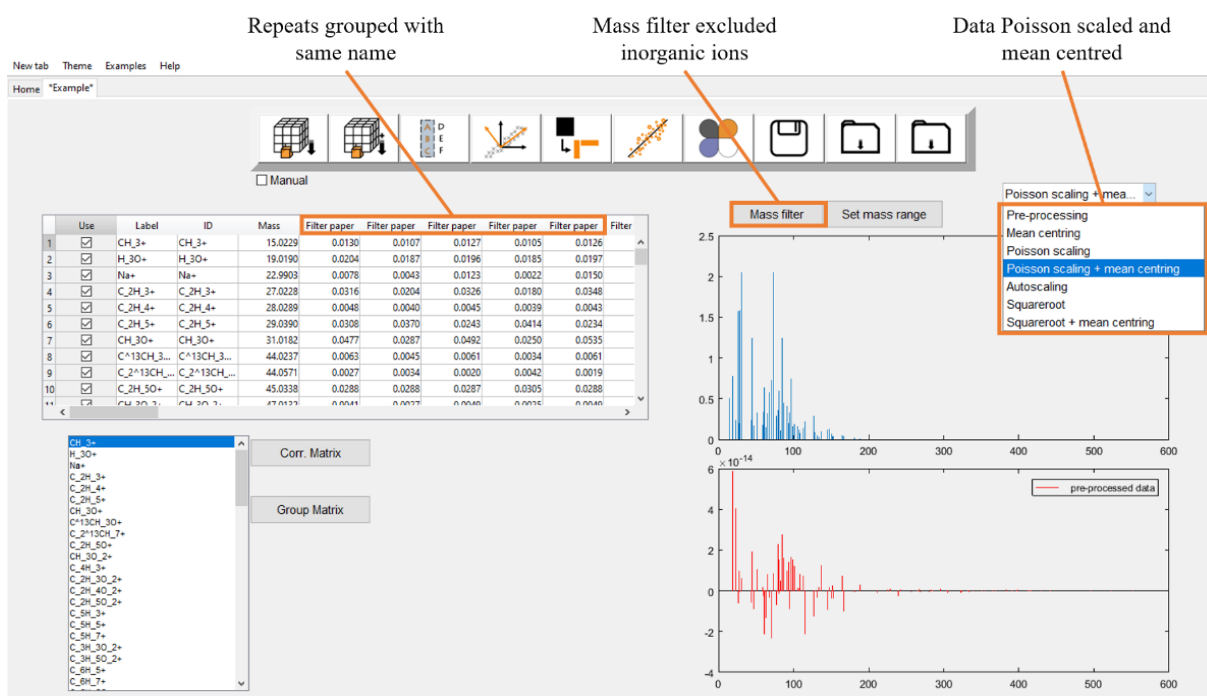


FIG. 4. Screenshot of a Spectra mode tab loaded with ToF-SIMS data wood, lignin and cellulose.

If the user clicks the “PCA” button, for example, a new “MVA results” tab is created. The tab will contain the loadings on the top panel, the scores on the bottom right panel and the principal components captured variance on the bottom left panel. Fig. 5 shows a screenshot of a PCA results tab of the wood dataset. The Loadings panel has a slider that controls a threshold of which variable labels to be shown in the plot for each principal component. Additional tools

of the loadings panel include the splitting of positive and negative loadings, the option to switch between mass number and labels (usually chemical assignment), a selective magnification tool and a grid plot containing an overview of all or a given number of components. The PC 1 loadings shown in Fig. 5 have lignin characteristic peaks on the positive side and cellulose characteristic peaks on the negative side.

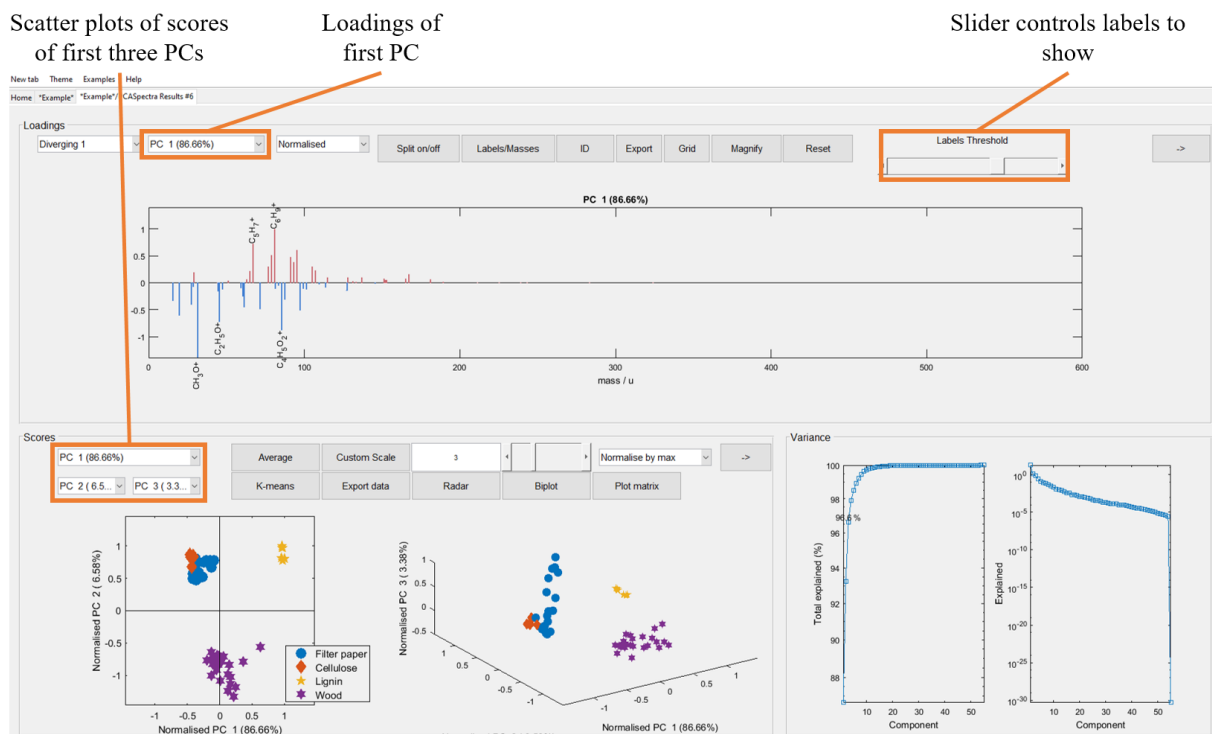


FIG.5. Screenshot of an MVA results tab for Spectra data.

The Scores panel has three drop down menus to select up to three principal components scores to be shown as 2D and 3D scatter plots. The colours of the symbols correspond to previously assigned groups in the Spectra mode tab. There are a few different functions for data visualisation such as averaging groups, radar plots and creation of a custom scale to plot the Scores against. For the wood data set, PC 1 separates lignin from cellulose + wood and PC 2 separates lignin + cellulose from wood. Two useful functions to visualise these groups are the use of biplots combined with the creation of Voronoi cells [22] as shown in Fig. 6.

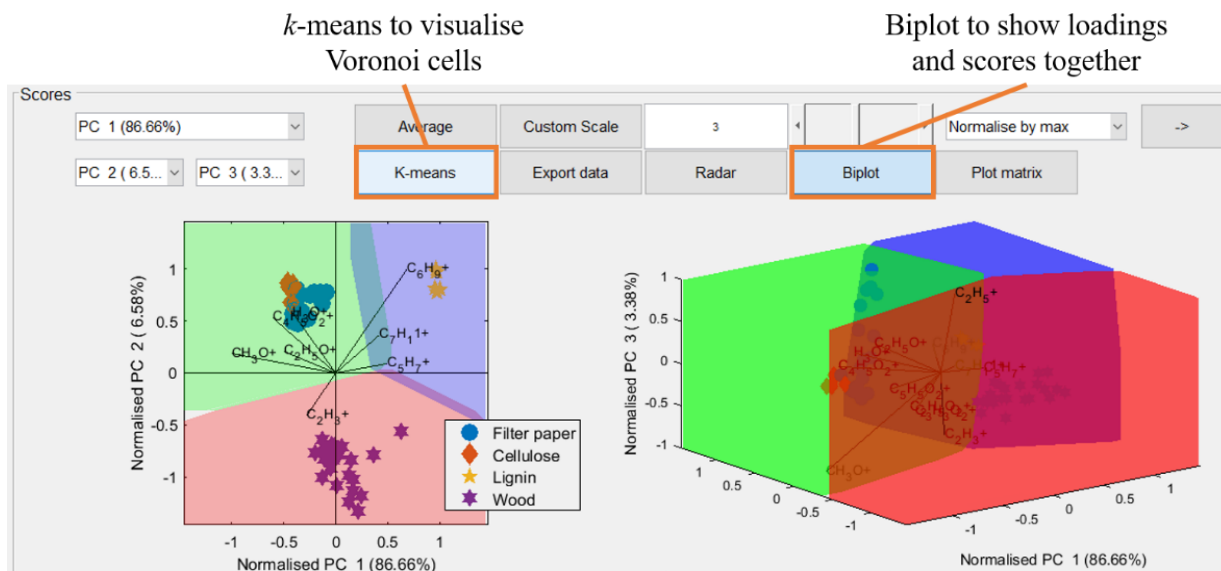


FIG.6. Screenshot of a Scores panel with a combination of biplot and Voronoi cells created on the scatter plots.

For the creation of Voronoi cells, a *k*-means clustering algorithm was applied to the scores of the first three principal components. Once the cluster centres are determined, all points in the visible space PC1 x PC2 x PC3 are tested to check which centre is the closest and coloured accordingly. With the Voronoi cells and biplot on the left hand side of Fig. 6 (scatter plot of scores of PC 1 against PC 2) it is clear that there are three groups comprising of the pure lignin samples, the pure cellulose samples and the wood samples. The biplot shows the most characteristic peaks of each group.

3.2 Profiles mode

The profiles mode is applicable to sequential point measurements typically of the same sample. These can be line scans, temporal profiles or, more commonly for ToF-SIMS data, depth profiles. Fig. 7 shows a screenshot of a profiles mode tab loaded with ToF-SIMS depth profiling data of a metallic layered sample. The depth profiles were acquired using the dual-beam depth profiling mode of the TOF.SIMS 5 (IONTOF GmbH) with a 25 keV Bi_3^+ primary ion beam delivering 0.18 pA of current and raster scanned over a $50 \times 50 \mu\text{m}^2$ area at

the centre of the etch crater formed using a 3 keV Cs⁺ beam raster scanned over an area of 400 × 400 μm². The depth profiling analysis was performed in the ‘interlaced’ mode, where the sources can operate in a simultaneous and continuous fashion. A simsMVA profiles mode tab contains a table on the right hand side that allows the selection of which peak intensity profile to plot on the large set of axes on the left hand side. The small axes on the right hand side show the total spectrum of all levels. It is possible to select specific ranges of masses and levels to be processed. This is useful for example in the presence of artefacts on the first or last few levels or when there is implantation of ions of the sputter beam. The slider on the top left applies a moving average filter to the profiles of all ions and there are other functions such as normalisation and data pre-processing.

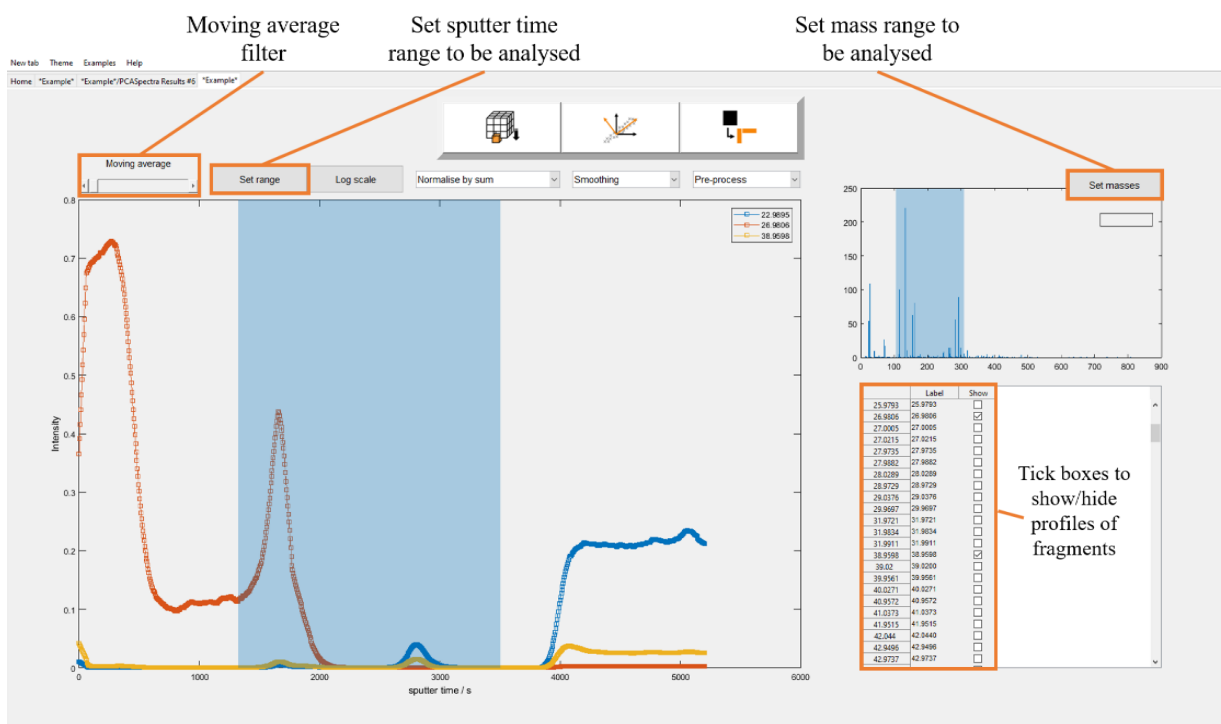


FIG.7. Screenshot of a profiles mode tab loaded with ToF-SIMS data of a layered metallic sample.

If the user clicks the “NMF” button, a small window is created with a number of options for the factorisation. The available options are “algorithm”, “number of endmembers”, “number of iterations”, “number of repeats”, “live visual output”, “calculate

lack-of-fit” and “*use sparse matrices*”. Once the factorisation is done, similarly to PCA, a new “MVA results” tab is created. The tab will contain the endmembers spectra on the top panel, the endmembers intensities on the bottom right panel and the error per iteration on the bottom left panel. NMF was performed using 6 endmembers, a multiplicative update-based algorithm and 500 iterations. The “*Overview*” button creates a window containing an overlay of the profiles of all endmembers together with their characteristic spectra, as shown in Fig. 8 for the metallic layered sample.

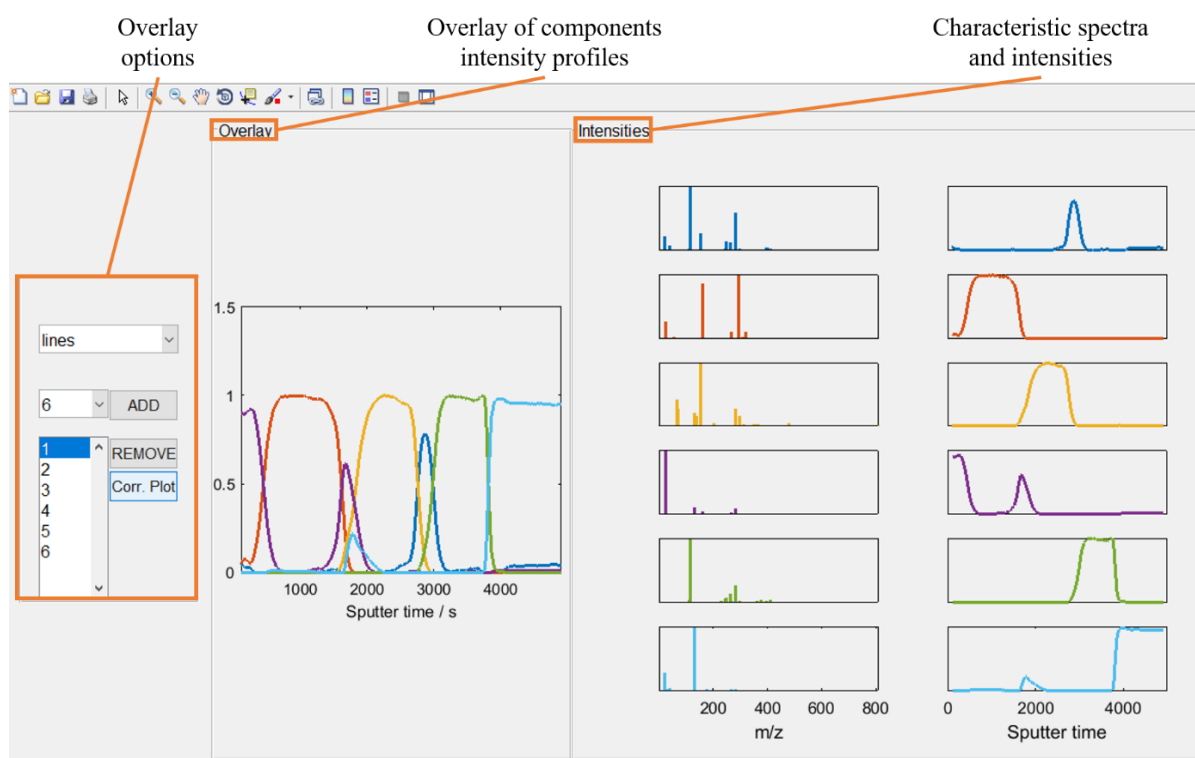


FIG. 8. Screenshot of an overlay window with NMF results of depth profiling data.

3.3 Images mode

Even though it is optimised for ToF-SIMS imaging datasets, the processing and visualisation tools in the images mode of simsMVA can be applied to any chemical mapping or hyperspectral mapping dataset. To demonstrate these tools, it was chosen an example ToF-SIMS imaging dataset from a low-angle taper-section of an organic coating painted onto a

metallic substrate. The method used to analyse the polymer/polymer and polymer/metal buried interfaces has been developed at University of Surrey and extensively explored [23,24]. The sample is cut using ultra-low-angle microtomy, exposing the interfacial regions. For the ToF-SIMS analysis, the sample is then tilted to ensure the exposed region is normal to the extraction optics of the time-of-flight analyser. More details of the experiment and results can be found elsewhere [25]. The area covering the whole taper is larger than the raster range of the primary ion beam of the equipment, therefore in order to analyse the whole region, several *patches* must be acquired. Instruments such as the TOF.SIMS 5 offer an automated way of analysing large areas by rastering the stage, however, when the sample is formed of regions with very different chemistry and conductivity, such automated modes will only be optimised for one end of the analysed area and the solution for getting good quality data is the separate acquisition of several $500 \times 500 \mu\text{m}^2$ patches. The images mode of simsMVA offers a tool for *stitching* several hyperspectral patches and transform them into one dataset. Fig. 9 illustrates this process for a grid of 3×2 hyperspectral patches with the same pixel size, however, it is also possible to combine patches of different sizes arranged in different grids with no total size limit. The only requirement is that all patches have the same number of variables. Fig. 10 shows a screenshot of an images mode tab loaded with the resulting *stitched* dataset of the organic coating.. An images mode tab will have at the top left hand side the intensity maps for a selected ion that can be normalised by the total ion intensity or any other ion map. The two plots on the right hand side will show the intensity distribution of the peak list (after pre-processing) outside (top) and inside (bottom) of the region of interest determined by a red resizable polygon on the left hand side map. The user can choose to process only the region within the polygon or perform subsampling of the data using low discrepancy which have been shown to generate, in much less time, results as good as if the whole dataset was processed [17,18,26]. The image at the bottom left will contain an RGB overlay of different ion maps selected by the user.

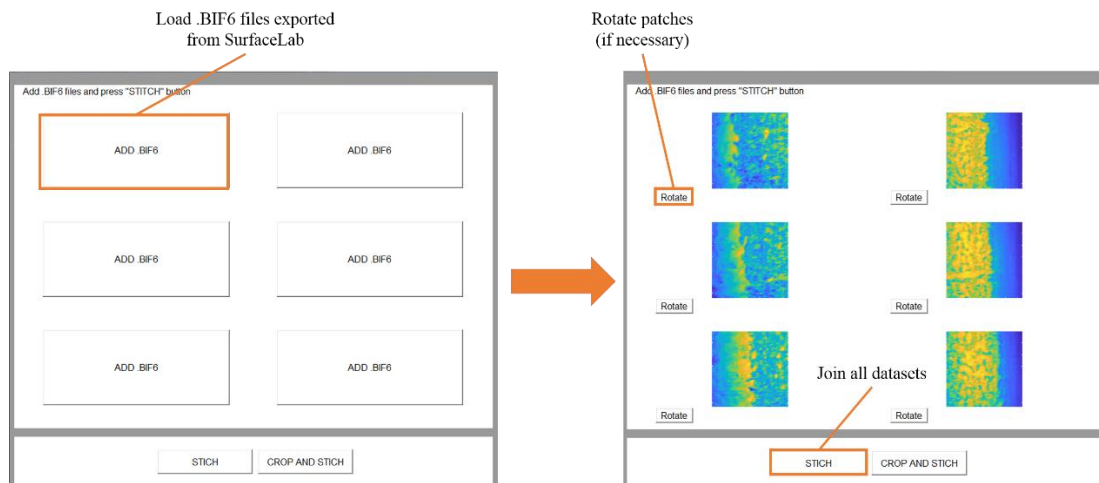


FIG. 9. Dataset “stitching” process for a grid formed of 3 x 2 patches of hyperspectral imaging datasets of the same pixel size.

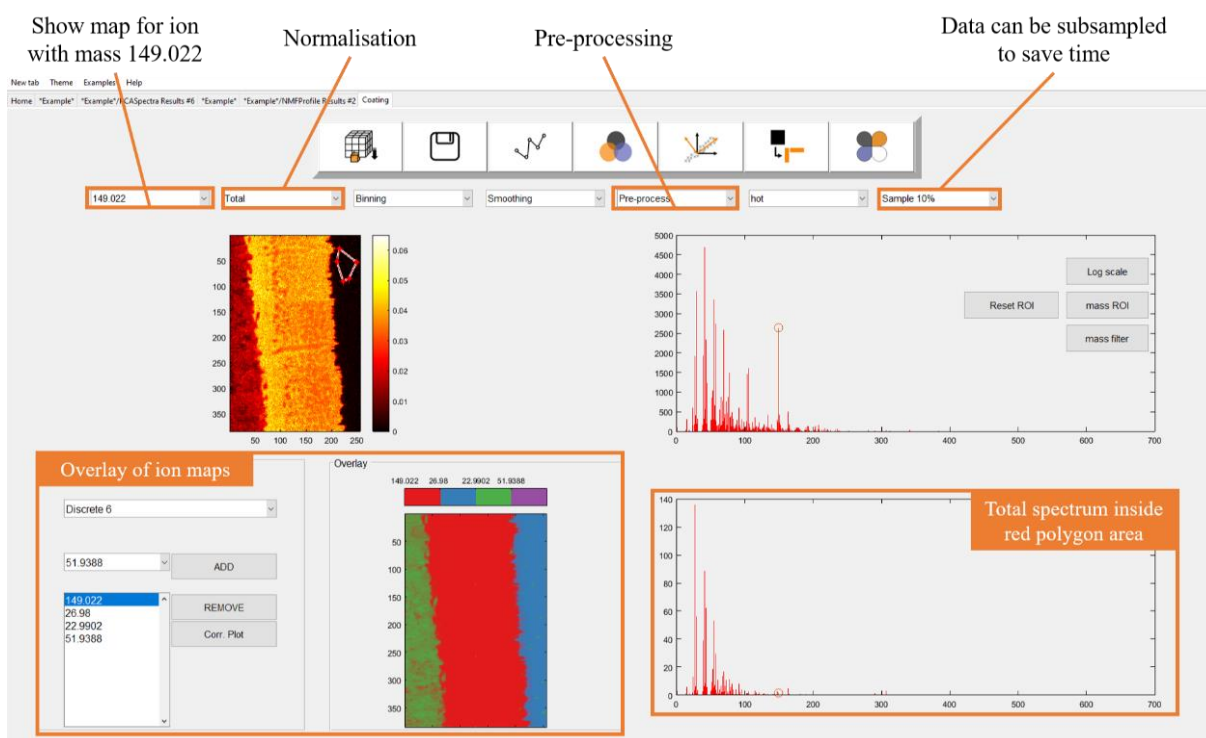


FIG. 10. Screenshot of an Images mode tab loaded with ToF-SIMS data of a cross-sectional taper of an organic coating.

Fig. 11 shows a screenshot of a PCA results tab for the example dataset. Prior to PCA, the data was pre-processed in the following order: Poisson scaled, normalised by total ion counts per pixel and mean centred. The Loadings panel looks the same for all previously described modes of analysis with the advantage of using different colour maps for the plots. A useful practice is to match the colour maps of the loadings and scores plots. This is useful for PCA results where the Scores are shown using a diverging colour map. In both Loadings and Scores panels, positive values are scaled to red and negative to blue, while values closer to zero will be coloured black. The scatter plot on the Scores panel represent the scores of two chosen principal components for all pixels. The “Brush” button enables the user to select pixels on the scatter plot and those will be highlighted on the intensity map on the left-hand side.

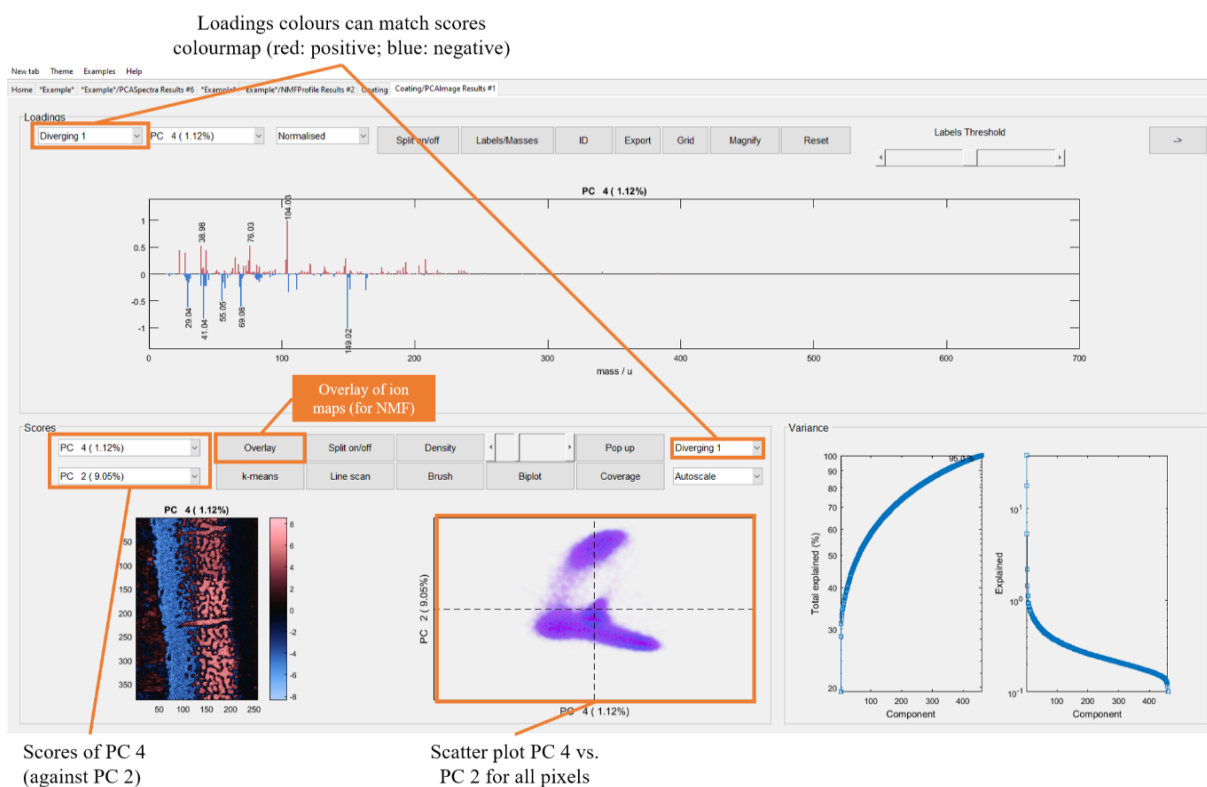


FIG.11. Screenshot of a MVA results tab with PCA results of a cross-sectional taper of an organic coating...

The PCA results of the taper showed that there are two phases on the bulk of the organic coating, with a different composition on the top surface. It also showed that the metal

substrate is covered with a thin chromium layer. If instead of PCA, NMF with 5 endmembers is done, a useful feature is the “*Overlay*” button, that creates a window enabling to overlay all endmembers intensities in a single map, according to a discrete colour scale, as shown in Fig. 12. For this dataset specifically, the NMF results show clearly the top surface (purple), polymer 1 (blue), polymer 2 (green), chromium layer (yellow) and metal substrate (orange).

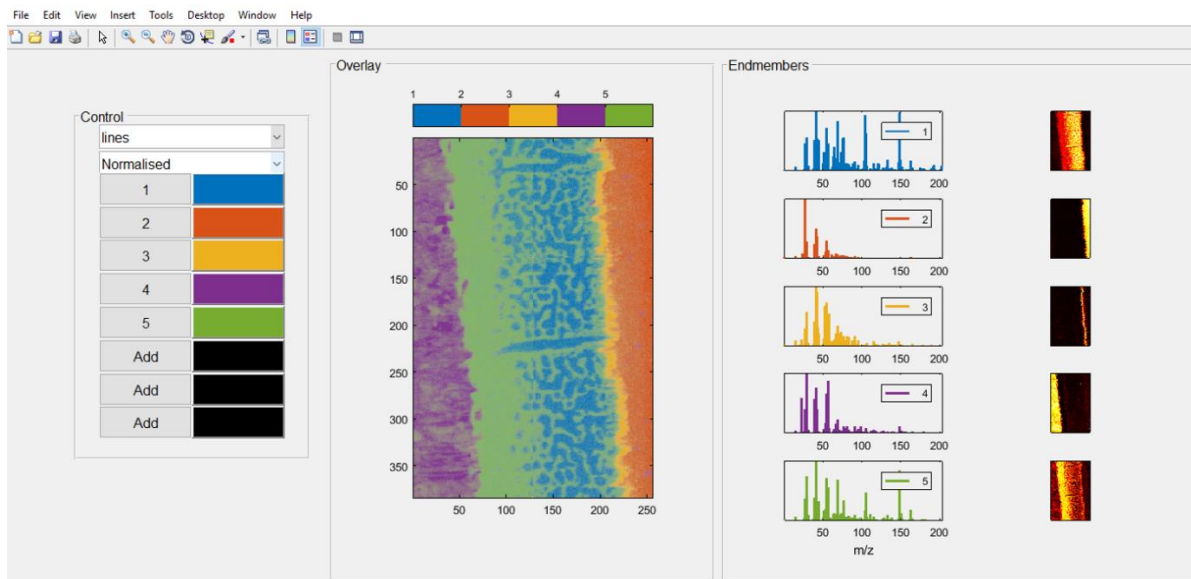


FIG.12. Screenshot of an intensities overview window.

3.4 3D mode

Fig. 13 shows a screenshot of a 3D mode tab loaded with ToF-SIMS data of an automotive grade polypropylene sample. This kind of material undergoes flame or plasma treatment prior to paint application and therefore it is important to understand the surface and bulk properties of the as-received form. More details of the samples and effects of plasma and flame treatment were published elsewhere [27,28]. The 3D ToF-SIMS data were acquired using the dual-beam depth profiling set-up of the TOF.SIMS 5 (IONTOF GmbH) with a 25 keV Bi_3^+ primary ion beam delivering 0.18 pA of current and raster scanned over a $400 \times 400 \mu\text{m}^2$ area at the centre of the etch crater formed using a 1 keV C_{60}^+ beam raster scanned over an area of $600 \times 600 \mu\text{m}^2$. The analysis was performed in the ‘non-interlaced’ mode with an

electron flood to neutralise charge build up. The sputter time and pause time per level were set respectively as 0.5 s and 1 s. A simsMVA 3D mode tab has a map on the top left side that will show the intensity maps for the selected ion at the selected level. Similarly to an images mode tab, the two plots on the right hand side will show the intensity distribution of the peak list (after pre-processing) outside (top) and inside (bottom) of the region of interest determined by a red polygon on the left hand side map. The axes on the bottom left will have a 3D map of intensities with a slice on the currently selected level.

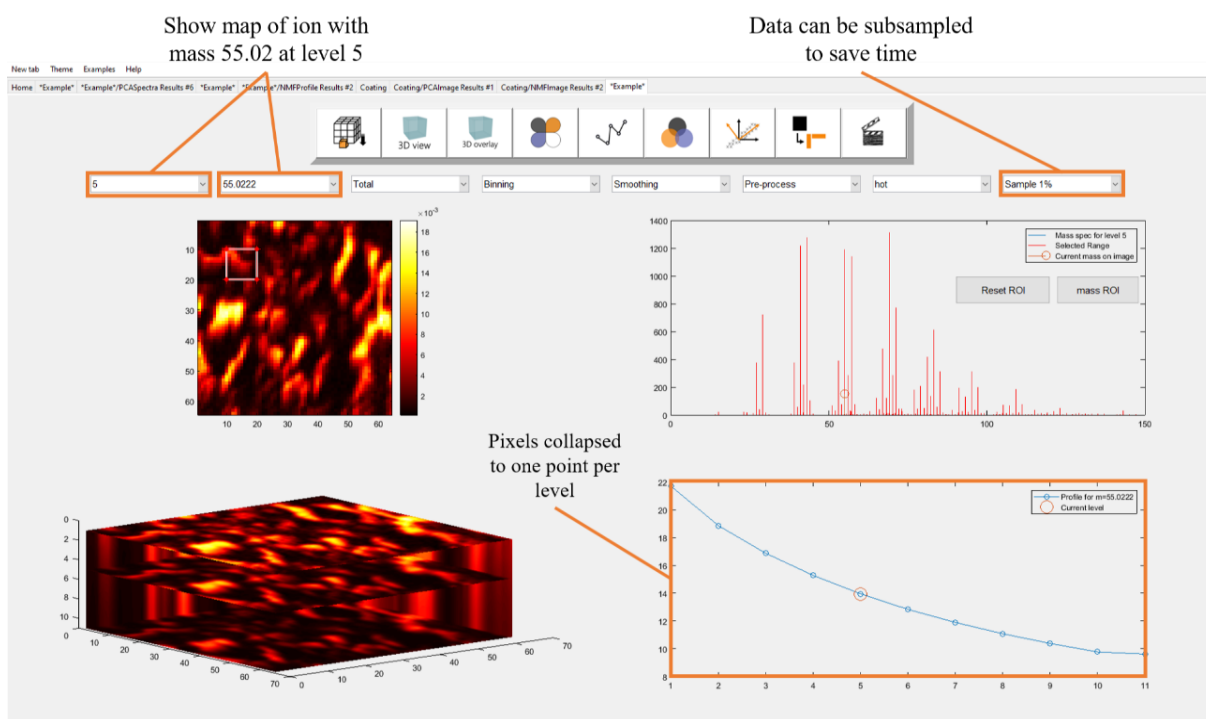


FIG.13. Screenshot of a 3D mode tab loaded with ToF-SIMS data of an automotive grade polypropylene sample.

NMF was performed in the polypropylene data in a subsample of 1 % of the voxels, using 3 endmembers, a multiplicative update-based algorithm and 500 iterations. The tab containing the results is very similar to the ones for spectra, profiles, and images. The “Profile view” button collapses all pixels of the NMF endmembers of each level to one data point per level and plots the results on the right-hand side axes.

When the user clicks the “3D view” or “3D overlay” buttons, they are prompted with an option to whether or not perform z-correction using the XZ or the YZ planes of a specific NMF endmember. In some cases, the correction of the z-coordinates of the voxels will aid visualisation and interpretation of 3D MVA results [29,30]. The correction is done based on a threshold that is visually set by the user in a window that allows the exploration of different XZ or YZ planes. Fig. 14 shows both “3D View” (for one endmember) and “3D overlay” (for all three endmembers) of the NMF results of the polypropylene dataset after z-correction. There are three possible styles of 3D visualisation (“Scatter”, “Slices” or “Rendered Isosurfaces”) and parameters such as colour map, transparency, normalisation, smoothing, marker size and aspect ratio can be controlled.

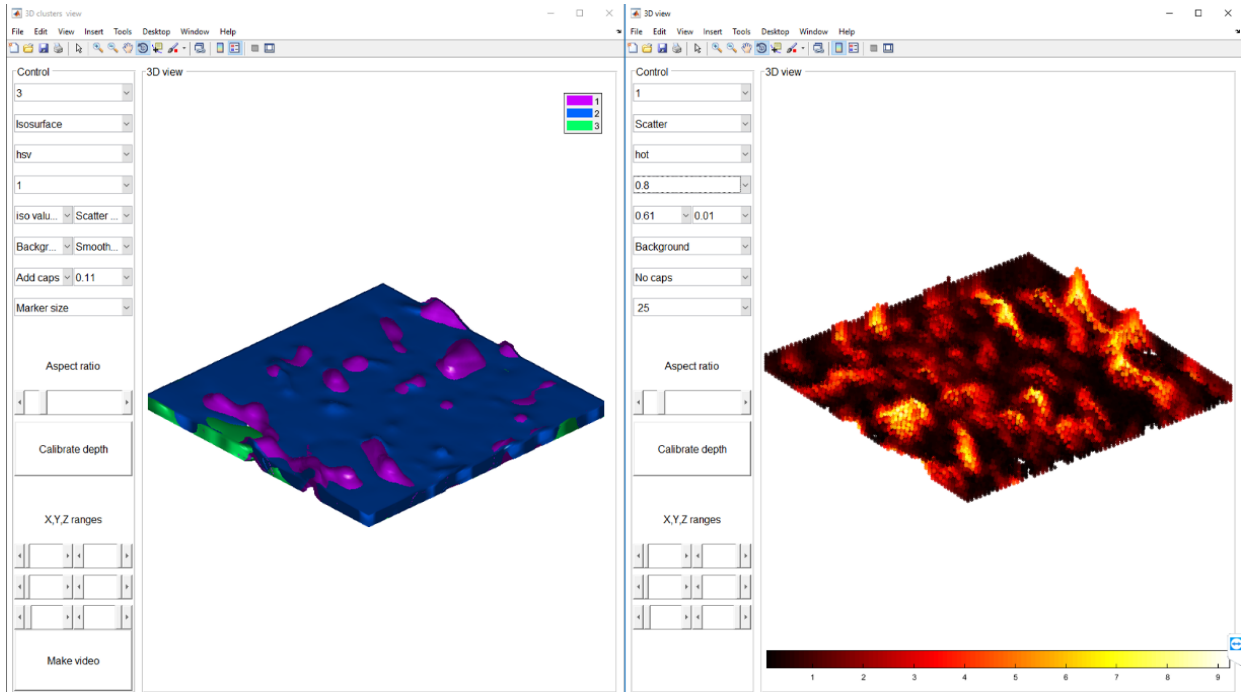


FIG.14. 3D visualisation of the NMF endmember intensities of the polypropylene dataset. Left: Screenshot of a 3D overlay window in the isosurfaces mode. Right: Screenshot of a 3D view window in the scatter plot mode for endmember 1.

3.5 Multi mode

The multi mode is a combination of all previously described modes and is intended as a tool to process various differently structured datasets as a single matrix. In the current version, the main tab contains four panels that allow the user to load four different datasets (one of each kind). The only restriction is that they have the same number of variables. Fig. 15 shows a schematic of how the data matrix is arranged in the multi mode.

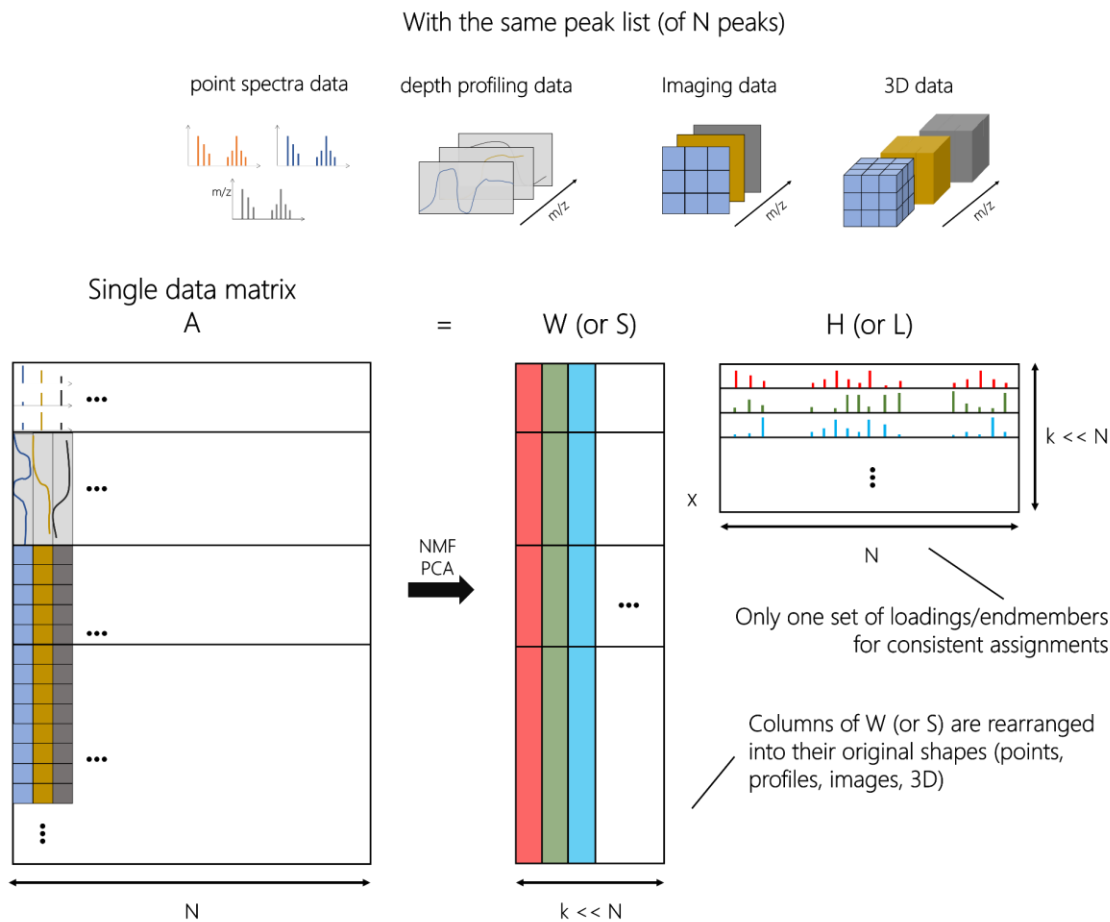


FIG.15. *simsMVA* multi mode. Datasets with different structures are combined into a single matrix.

Once loaded, a multi mode dataset can be normalised by total variables intensities and pre-processed in the same manner as in the individual modes. *simsMVA* can then perform PCA or NMF and the results are shown in a multipanel tab with separate views for scores (or

intensities) and all data visualisation functionalities previously described. The difference is that all datasets share the same loadings (or characteristic NMF spectra). In order to illustrate simsMVA multi mode, example datasets were created from the 3D dataset presented in Section 3.4: a spectra dataset containing 20 individual voxels; a profile dataset containing 11 observations where each of them is a sum of all pixels per depth level; an imaging dataset created from the map of a specific level cropped in an area containing a particle (and upscaled to 55 x 55 pixels). The original 3D dataset contains 11 levels of 64 x 64 pixels each. Fig. 16 shows PCA results of the example multi dataset. It can be seen that PC 3 separates the bulk material from the particles and this is consistent across all datasets. An example of a situation where the simsMVA multi mode can be useful is when one intends to identify the presence and distribution of standard materials in a mixed samples. A 3D dataset of the samples could be loaded together with point spectra data of the standard materials and MVA would identify, in an unsupervised fashion, the distribution of such standard materials within the sample.

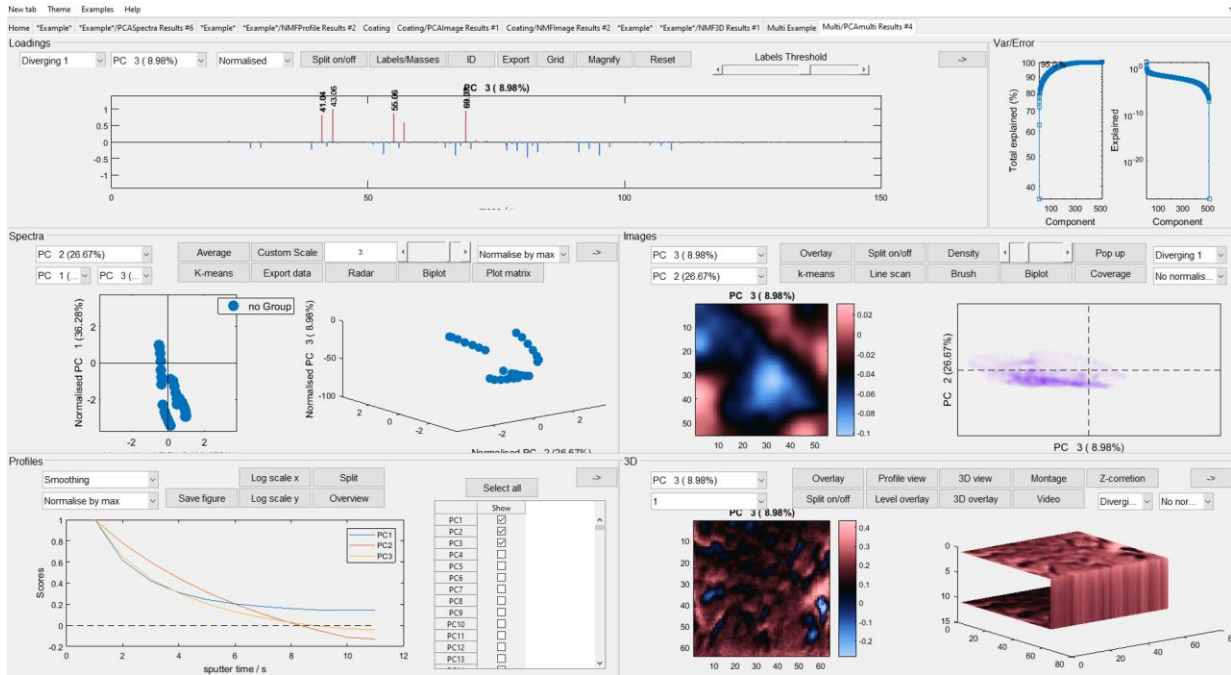


FIG.16. PCA results of an example multi mode dataset. All visualisation tools for the different data structures are kept. The difference is that all datasets share the same principal components.

4 System requirements and availability

simsMVA is available both as an application inside MATLAB and as a standalone version. The MATLAB version runs on any operational system and can load any data matrix from MATLAB's workspace. The disadvantage is that it requires an installation of MATLAB 2015b or newer with both the *Statistics and Machine Learning* and the *Image Processing* toolboxes. The standalone version runs on 64 bit Windows PCs and does not need a MATLAB installation nor any of the toolboxes, however it is currently limited to loading data exported from IONTOF's SurfaceLab software. Additionally to these requirements, for large area imaging or 3D datasets, a minimum of 8GB of RAM is recommended.

simsMVA is in constant development and the current version is freely available for non-commercial use. Copies can be requested via the website <http://www.mvtools.com> or by contacting the corresponding author. The website is also regularly updated with tutorials and news.

5 Declaration of independent implementation

Independently tested by:

Kristof Marcoen

PhD researcher, Vrije Universiteit Brussel

Research group Electrochemical and surface engineering (SURF)

Report:

I hereby declare that that I successfully employed simsMVA to process ToF-SIMS data on my own computer at Vrije Universiteit Brussel (Brussels, Belgium).

I used it in the study of the formation of a lithium-based corrosion protection layer in coating defects on aluminium alloys. The formation of this protective layer occurs in different stages, with competitive growth between two layered structures with chemical compositions that are characterised by similar mass fragments. A clear distinction between both compositions from mass spectra could be made by considering the differences in ion intensity ratios. Large area (2 mm x 1 mm) ToF-SIMS images were acquired to study the formation and

spread of the protective layer in artificial coating defects. PCA was applied on the image spectra files and confirmed that two compositions were present in the protective layer. NMF scores obtained from NMF analysis on the same spectra files could be interpreted as relative concentrations for both components. These MVA results enabled us to unravel the formation mechanism of the protective layer ^[1]. In a later stage NMF was successfully applied on the image files as well, to visualise the spread of both components in large coating defects.

I experienced simsMVA as a userfriendly tool to process complex ToF-SIMS datasets. It provides several options for a nice visualisation of the data. simsMVA has proven its great potential for application in large area ToF-SIMS imaging and offers a wide range of possibilities for processing spectra, maps, depth profiles or 3D ToF-SIMS datasets.

Yours faithfully

Kristof Marcoen

Phd researcher, Vrije Universiteit Brussel

Research group Electrochemical and surface engineering (SURF)

[1] K. Marcoen et al., Compositional study of a corrosion protective layer formed by leachable lithium salts in a coating defect on AA2024-T3 aluminium alloys, Prog. Org. Coatings. (2018) 65-75. doi: 10.1016/j.porgcoat.2018.02.011

6 Exporting data from SurfaceLab 6

The current version of simsMVA is optimised to import files generated and exported by the TOF.SIMS 5 spectrometer. Figures 17 to 19 are intended as a guideline on how to export files using SurfaceLab software provided by IONTOF alongside the instrument. A pre-requisite for generating new datasets from the same measurement is that it was recorded in an ITM raw data file format.

In SurfaceLab's Spectra program, once the spectra of all samples are calibrated, the user can then apply a peak list to one of them and follow the steps shown in Figure 5.14: 1) On the left-hand side menu, select all samples desired to be on the peak list. 2) Click on

“statistics” button, a window with the peak list will pop up. 3) On the top icons bar, unselect all extra statistics metrics. 4) Click “options” and change “description” to “Mass”. 5) Click on save button and save peak list as a .txt file.

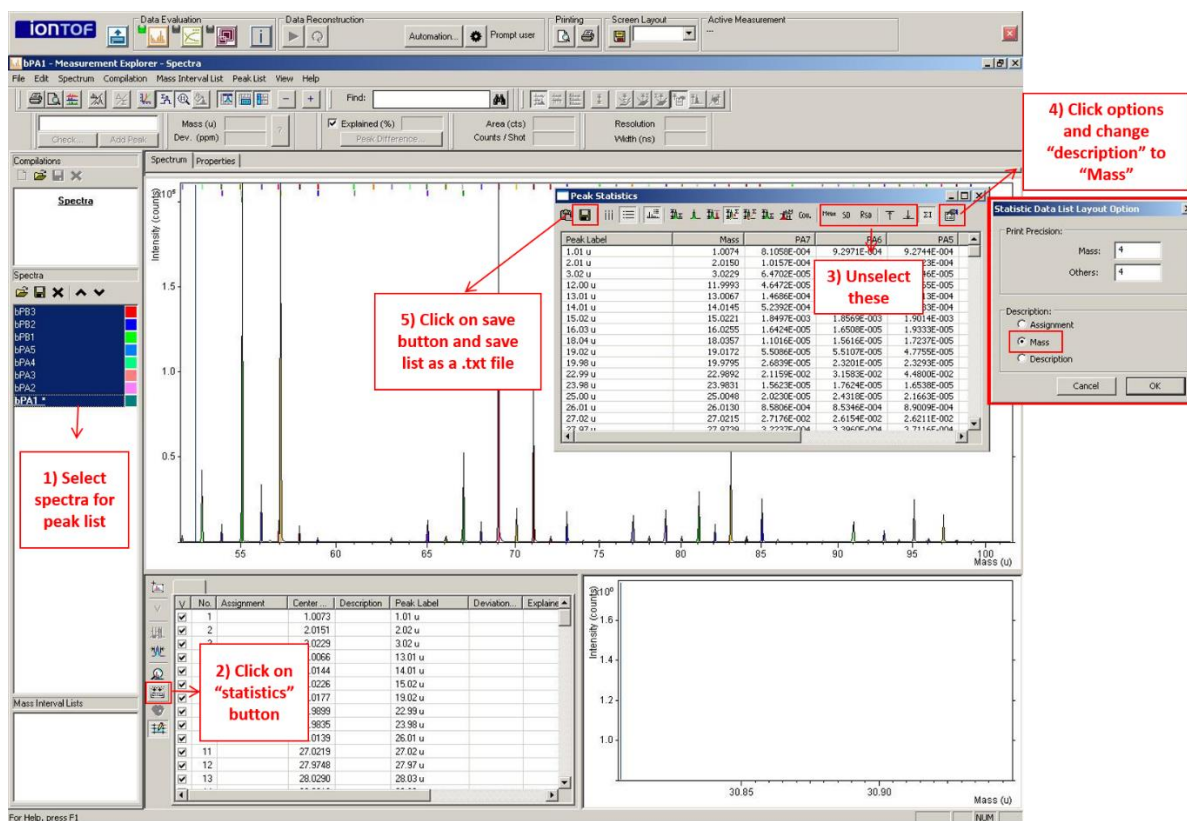


Figure 17: Guideline of how to export a peak list from SurfaceLab's Spectra program.

In SurfaceLab's Profiles program, the user has to reconstruct the data from an .ITM file using a desired peak list and then follow the steps shown in Figure 5.15: 1) On the top menu bar, click "File->Export", an options window will pop up. 2) Select "Data Point" and "Sputter Time" for X-Axis and one of the other options for Y-axis, press OK and save the profile data as a .txt file.

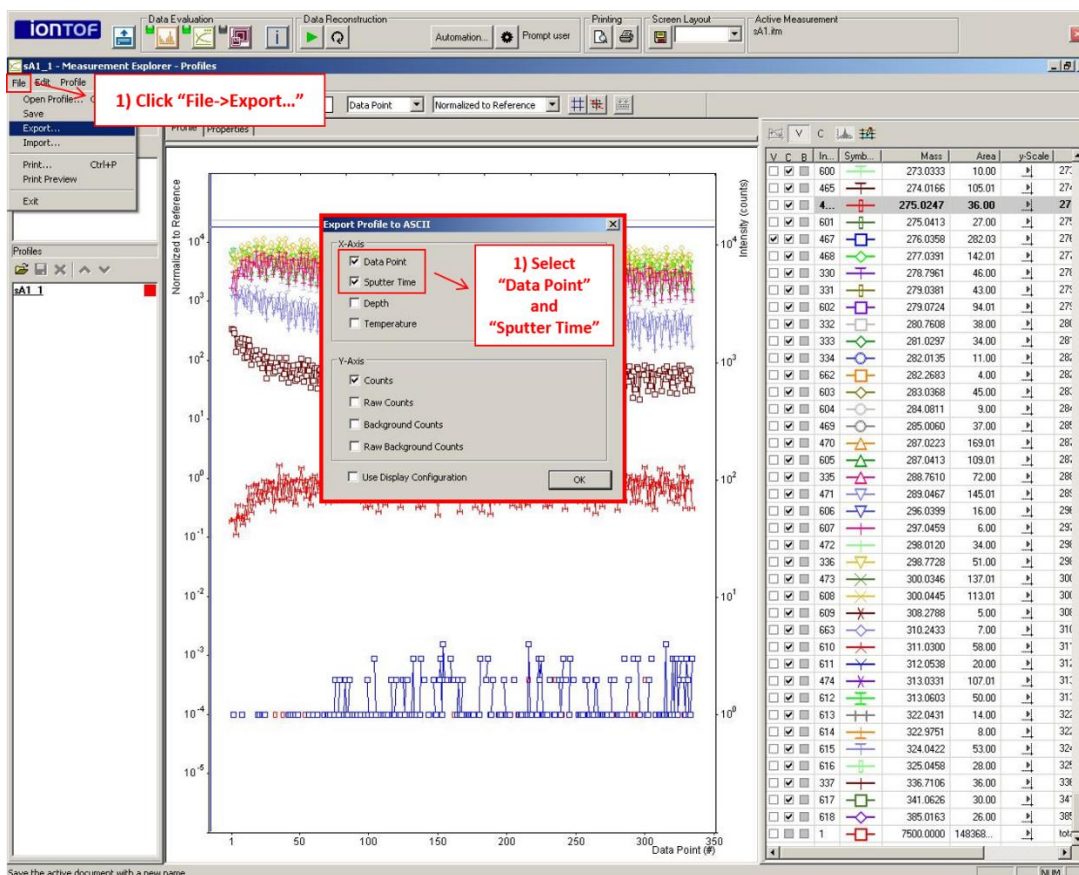


Figure 18: Guideline of how to export a peak list from SurfaceLab's Profiles program

In SurfaceLab's Images program, the user also has to reconstruct the data from an .ITM file using a desired peak list and then follow the steps shown in Figure 5.16: 1) On the top menu bar, click "File->Export", an options window will pop up. 2) On the left-hand side menu, select the desired mass range. 3) Select "Export Summed Image" for Images data or "Export Scan Resolved Images" for 3D data. 4) In "Exported Data Format" Choose "Binary (BIF6)" and press the "Export" button to create the BIF6 files that can then be loaded into simsMVA.

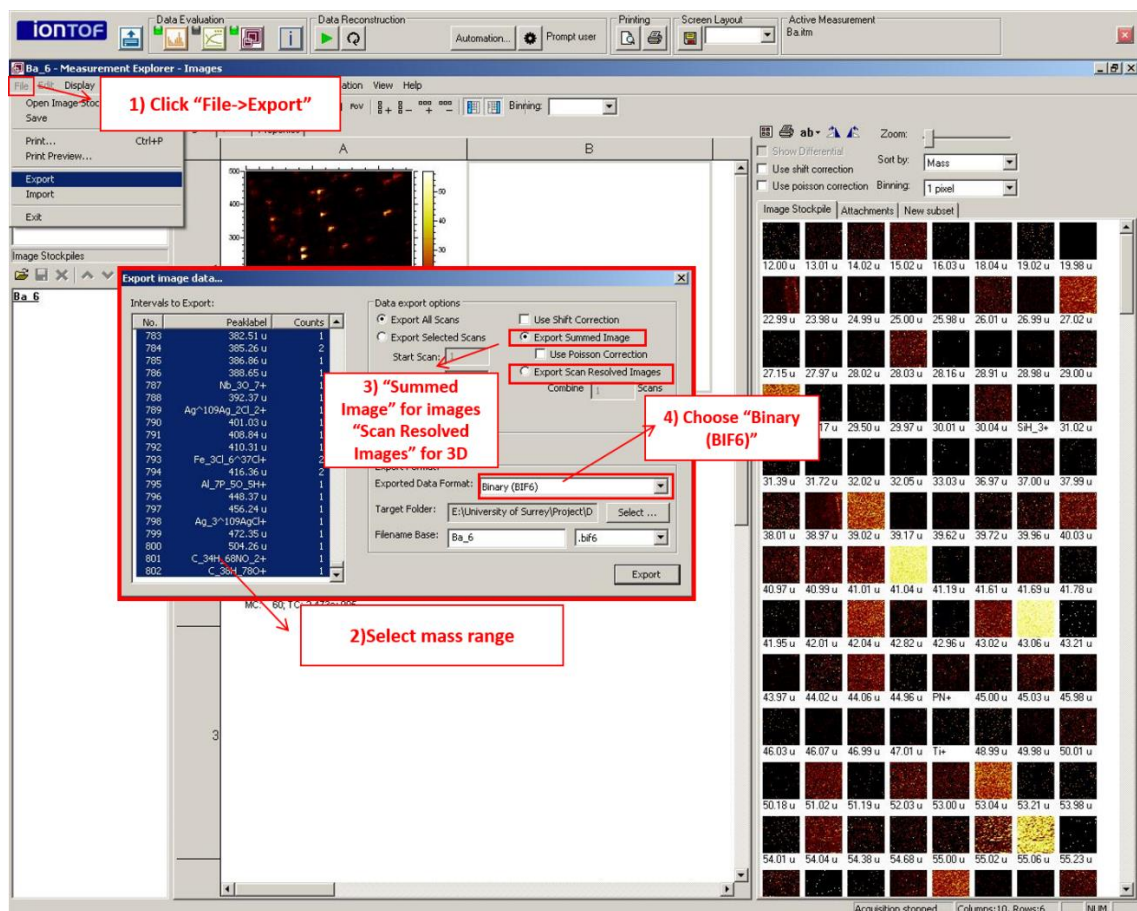


Figure 19: Guideline of how to export a peak list from SurfaceLab's Profiles program

Acknowledgements

The authors wish to thank the Coordination for the Improvement of Higher Education Personnel – CAPES (project 11995-13-0) for funding this work. The authors are also thankful to Dr. Rene Tshulu, Dr. Jorge Banuls Ciscar, Dr. Taraneh Bozorgzad Moghim, Mr Kristof Marcoen and Ms Annelies Voorhaar, for their constant feedback on simsMVA since early versions and to Mrs Helen Sampaio Ferraz for her creative insights and help with the design of the icons and GUI.

References

- [1] J.C. Vickerman, D. Briggs, TOF-SIMS: Materials Analysis by Mass Spectrometry, IM Publications, 2013.

- [2] J.C. Vickerman, N. Winograd, International Journal of Mass Spectrometry SIMS — A precursor and partner to contemporary mass spectrometry, *Int. J. Mass Spectrom.* 377 (2015) 568–579. doi:10.1016/j.ijms.2014.06.021.
- [3] J.L.S. Lee, I.S. Gilmore, M.P. Seah, Quantification and methodology issues in multivariate analysis of ToF-SIMS data for mixed organic systems, *Surf. Interface Anal.* 40 (2008) 1–14. doi:10.1002/sia.2713.
- [4] J.L.S. Lee, I.S. Gilmore, I.W. Fletcher, M.P. Seah, Multivariate image analysis strategies for ToF-SIMS images with topography, *Surf. Interface Anal.* 41 (2009) 653–665. doi:10.1002/sia.3070.
- [5] J.L.S. Lee, B.J. Tyler, M.S. Wagner, I.S. Gilmore, M.P. Seah, The development of standards and guides for multivariate analysis in surface chemical analysis, *Surf. Interface Anal.* 41 (2009) 76–78. doi:10.1002/sia.2935.
- [6] M.P. Seah, J.L.S. Lee, B.J. Tyler, M.S. Wagner, I.S. Gilmore, G. Term, Proposed terminology for Multivariate Analysis in Surface Chemical Analysis – Vocabulary – Part 1 : General Terms and Terms for the Spectroscopies, *Chem. Anal.* (2008) 1–10.
- [7] B.J. Tyler, G. Rayal, D.G. Castner, Multivariate analysis strategies for processing ToF-SIMS images of biomaterials, *Biomaterials.* 28 (2007) 2412–2423. doi:10.1016/j.biomaterials.2007.02.002.
- [8] B.J. Tyler, The accuracy and precision of the advanced Poisson dead-time correction and its importance for multivariate analysis of high mass resolution ToF-SIMS data, *Surf. Interface Anal.* (2014) 581–590. doi:10.1002/sia.5543.
- [9] D.J. Graham, D.G. Castner, Multivariate analysis of ToF-SIMS data from multicomponent systems: The why, when, and how, *Biointerphases.* 7 (2012) 1–12. doi:10.1007/s13758-012-0049-3.
- [10] M.R. Keenan, P.G. Kotula, Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images, *Surf. Interface Anal.* 36 (2004) 203–212. doi:10.1002/sia.1657.
- [11] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, D. O’Sullivan, *PLS_Toolbox 7.9 for use with MATLAB*, (2014).
- [12] D.J. Graham, NBtoolbox (<https://www.nb.uw.edu/mvsa/nbtoolbox>), <https://www.nb.uw.edu/mvsa/nbtoolbox>. (n.d.). <https://www.nb.uw.edu/mvsa/nbtoolbox> (accessed November 22, 2017).
- [13] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, 2005. doi:10.1016/j.chemolab.2004.12.007.
- [14] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization., *Nature.* 401 (1999) 788–791. doi:10.1038/44565.
- [15] D. Lee, H. Seung, Algorithms for non-negative matrix factorization, *Adv. Neural Inf. Process. Syst.* (2001) 556–562. doi:10.1109/IJCNN.2008.4634046.
- [16] R. Bro, N.D. Sidiropoulos, Least squares algorithms under unimodality and non-negativity constraints, *J. Chemom.* 12 (1998) 223–247. doi:10.1002/(SICI)1099-128X(199807/08)12:4<223::AID-CEM511>3.0.CO;2-2.
- [17] G.F. Trindade, M. Abel, J.F. Watts, Non-negative matrix factorisation of large mass spectrometry datasets, *Chemom. Intell. Lab. Syst.* 163 (2017) 76–85. doi:10.1016/j.chemolab.2017.02.012.

- [18] S. Van Nuffel, C. Parmenter, D.J. Scurr, N.A. Russell, M. Zelzer, Multivariate analysis of 3D ToF-SIMS images: method validation and application to cultured neuronal networks, *Analyst*. 141 (2016) 90–95. doi:10.1039/C5AN01743B.
- [19] S. Van Nuffel, Three-dimensional Time-of-Flight Secondary Ion Mass Spectrometry Imaging of Primary Neuronal Cell Cultures, University of Nottingham, 2017.
- [20] P.J. Cumpson, I.W. Fletcher, N. Sano, A.J. Barlow, Rapid multivariate analysis of 3D ToF-SIMS data: graphical processor units (GPUs) and low-discrepancy subsampling for large-scale principal component analysis, *Surf. Interface Anal.* 48 (2016) 1328–1336. doi:10.1002/sia.6042.
- [21] G.F. Trindade, J. Bañuls-Ciscar, C.K. Ezech, M.L. Abel, J.F. Watts, Characterisation of wood growth regions by multivariate analysis of ToF-SIMS data, in: *Surf. Interface Anal.*, 2016: pp. 584–588. doi:10.1002/sia.5915.
- [22] F. Aurenhammer, Voronoi Diagrams — A Survey of a Fundamental Data Structure, *ACM Comput. Surv.* 23 (1991) 345–405. doi:10.1145/116873.116880.
- [23] S.J. Hinder, C. Lowe, J.T. Maxted, J.F. Watts, A ToF-SIMS investigation of a buried polymer/polymer interface exposed by ultra-low-angle microtomy, *Surf. Interface Anal.* 36 (2004) 1575–1581. doi:10.1002/sia.1985.
- [24] S.J. Hinder, C. Lowe, J.T. Maxted, J.F. Watts, The morphology and topography of polymer surfaces and interfaces exposed by ultra-low-angle microtomy, *J. Mater. Sci.* 40 (2005) 285–293. doi:10.1007/s10853-005-6081-7.
- [25] G.F. Trindade, M.L. Abel, C. Lowe, R. Tshulu, J.F. Watts, A Time-of-Flight Secondary Ion Mass Spectrometry/Multivariate Analysis (ToF-SIMS/MVA) Approach to Identify Phase Segregation in Blends of Incompatible but Extremely Similar Resins, *Anal. Chem.* 90 (2018) 3936–3941. doi:10.1021/acs.analchem.7b04877.
- [26] P.J. Cumpson, I.W. Fletcher, N. Sano, A.J. Barlow, Rapid multivariate analysis of 3D ToF-SIMS data: graphical processor units (GPUs) and low-discrepancy subsampling for large-scale principal component analysis, *Surf. Interface Anal.* 48 (2016) 1328–1336. doi:10.1002/sia.6042.
- [27] D.F. Williams, M.L. Abel, E. Grant, J. Hrachova, J.F. Watts, Flame treatment of polypropylene: A study by electron and ion spectroscopies, *Int. J. Adhes. Adhes.* 63 (2015) 26–33. doi:10.1016/j.ijadhadh.2015.07.009.
- [28] G.F. Trindade, D.F. Williams, M.L. Abel, J.F. Watts, Analysis of atmospheric plasma-treated polypropylene by large area ToF-SIMS imaging and NMF, *Surf. Interface Anal.* (2018). doi:10.1002/sia.6378.
- [29] D. Breitenstein, C.E. Rommel, R. Möllers, J. Wegener, B. Hagenhoff, The chemical composition of animal cells and their intracellular compartments reconstructed from 3D mass spectrometry, *Angew. Chemie - Int. Ed.* 46 (2007) 5332–5335. doi:10.1002/anie.200604468.
- [30] M.A. Robinson, D.J. Graham, D.G. Castner, ToF-SIMS depth profiling of cells: Z-correction, 3D imaging, and sputter rate of individual NIH/3T3 fibroblasts, *Anal. Chem.* 84 (2012) 4880–4885. doi:10.1021/ac300480g.