Contents lists available at ScienceDirect

Analytica Chimica Acta

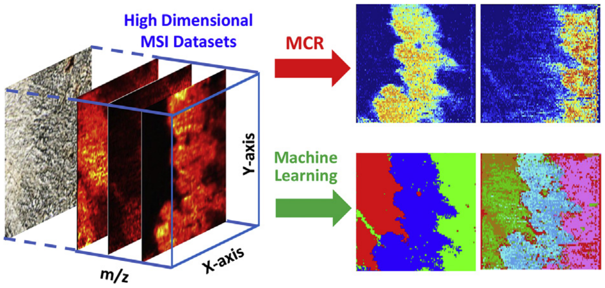
journal homepage: www.elsevier.com/locate/aca

Towards enhanced metabolomic data analysis of mass spectrometry image: Multivariate Curve Resolution and Machine Learning

Xiang Tian [1], Genwei Zhang [1], Yihan Shao [*], Zhibo Yang [*]

Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Parkway, Norman, OK, 73019, USA

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 1 November 2017
Received in revised form
8 February 2018
Accepted 10 February 2018
Available online xxx

Keywords:
Mass spectrometry imaging
Metabolomic analysis
Multivariate curve resolution
Machine learning
t-SNE (*t*-distributed stochastic neighbor embedding)

ABSTRACT

Large amounts of data are generally produced from mass spectrometry imaging (MSI) experiments in obtaining the molecular and spatial information of biological samples. Traditionally, MS images are constructed using manually selected ions, and it is very challenging to comprehensively analyze MSI results due to their large data sizes and highly complex data structures. To overcome these barriers, it is obligatory to develop advanced data analysis approaches to handle the increasingly large MSI data. In the current study, we focused on the method development of using Multivariate Curve Resolution (MCR) and Machine Learning (ML) approaches. We aimed to effectively extract the essential information present in the large and complex MSI data and enhance the metabolomic data analysis of biological tissues. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) algorithm was used to obtain major patterns of spatial distribution and grouped metabolites with the same spatial distribution patterns. In addition, both supervised and unsupervised ML methods were established to analyze the MSI data. In the supervised ML approach, Random Forest method was selected, and the model was trained using the selected datasets based on the distribution pattern obtained from MCR-ALS analyses. In the unsupervised ML approach, both DBSCAN (Density-based Spatial Clustering of Applications with Noise) and CLARA (Clustering Large Applications) were applied to cluster the MSI datasets. It is worth noting that similar patterns of spatial distribution were discovered through MSI data analysis using MCR-ALS, supervised ML, and unsupervised ML. Our protocols of data analysis can be applied to process the data acquired using many other types of MSI techniques, and to extract the overall features present in MSI results that are intractable using traditional data analysis approaches.

[*] Corresponding authors.
E-mail addresses: Yihan.Shao@ou.edu (Y. Shao), Zhibo.Yang@ou.edu (Z. Yang).

[1] Both authors equally contributed to this work.

https://doi.org/10.1016/j.aca.2018.02.031
0003-2670/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

MS imaging (MSI) is a powerful tool to construct the spatial distribution of wide ranges of molecules in biological tissue samples. Among all MSI methods that have been developed, matrix assisted laser desorption ionization (MALDI) and secondary ion MS (SIMS) are the most widely used sampling and ionization approaches in non-ambient methods [1]. MALDI has broad detection ranges of molecular weight (>500,000 Da) [2], whereas SIMS provides the highest spatial resolution (<700 nm) [3]. Ambient MSI methods, such as DESI (desorption electrospray ionization), require minimum or no sample-preparation, and they can be conveniently used in atmospheric pressure conditions [4]. MSI experiments provide rich chemical information of the biological sample surface, and the experimental datasets generally have very large sizes and complex high-dimensional data structures [5,6]. For example, depending on the spatial resolution (100 5 μm), a 1 mm^2 MADLI MS image is composed of pixels ranging from 100 to 4×10^4 [7]. Particularly, each pixel represents a complex chemical profile with the specific spatial information, and the size of a MS image can range from hundreds of megabytes to several gigabytes [8]. Conventional data analyses are typically carried out by focusing on the selected ions among all species detected from samples. However, it is extremely difficult to manually conduct a comprehensive analysis to extract overall features from MSI data through traditional approaches.

To effectively extract essential chemical and spatial information from large and complex MSI data, a number of statistical analysis methods, such as Principal Component Analysis (PCA) [9,10], clustering [8], and other multivariate analysis [11,12], have been utilized in previous studies. PCA can be used to determine the major molecular components contributing to the differences of spatial distributions. However, the applications of PCA to MSI data analysis are limited due to same drawbacks: the negative values of PCA score plots have no physical meaning (i.e., the ion intensity cannot be negative), incapability to define region of interest, and inconsistence with the corresponding loading plot [13,14]. To overcome these drawbacks, other methods such as probabilistic latent semantic analysis (pLSA) and non-negative parallel factors analysis (NN-PARAFAC), can be applied to achieve non-negative components decomposition; however, the number of components needs to be specified prior to analysis [1,15]. Unsupervised clustering methods, such as Hierarchical Clustering [16], *k*-means [17], fuzzy c-means [18], and Iterative Self-Organizing Data Analysis Technique (ISODATA) [19], share a common drawback: they cannot yield the spectra (i.e., molecular information) in each cluster [8]. In contrast, Multivariate Curve Resolution (MCR) algorithms, a family of methods for analyzing mixtures, have been proven as effective approaches to overcome this limitation. MCR techniques resolve mixed datasets by obtaining the number of components, and the signal profile and abundance of each component [11]. The first application of MCR was to investigate two-compound mixtures in UV spectroscopy [20]. Then, a number of MCR algorithms have been developed to analyze complex data. For example, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS), one of the most popular MCR methods utilizing ALS algorithm, has been used in many research fields [21] such as analyzing data from LC-MS [22], fluorescence [23], and MSI experiments [12]. Among all these applications, MCR-ALS has been successfully used to group major chemical components possessing similar spatial distributions from MSI studies [11]. This method can significantly improve the data analyzing efficiency, and potentially provide more information for a better understanding of biologically relevant metabolites in terms of their spatial distributions in tissues [24,25]. However, due to relatively intensive computation needed, MCR-ALS is unlikely to be the most efficient approach for the analyses of large sizes of MSI results [11]. For example, MCR needs to be separately performed to analyze individual sets of data obtained from multiples slices of the same tissue (i.e., 3D MSI), which share a large amount of similarities.

Due to a rapid growth of applications of AI (Artificial Intelligence) in a broad range of areas, using ML approaches to analyze MSI data has become an emerging trend [26–29]. ML uses computational approaches to learn from complex instance data, identify patterns and relationships present within the instance data, and achieve predictive data mining [30]. ML has been applied in many branches of biomedical research, including cancer prediction and prognosis [31], genetics and genomics [32], proteomics [33], and medical imaging [34]. However, due to the large size and high complexity of MSI data as well as the infancy of ML applications, this promising data analysis method has only been utilized in a very few MSI studies [26–29].

In general, ML methods can be divided into two categories: supervised ML and unsupervised ML. Supervised ML predicts features of a dataset after the model is trained using the labelled training data (i.e., datasets of ions with assigned tissue types or regions). Data analysis using supervised ML usually involves three steps: training data selection, model optimization/validation, and prediction of new dataset. Supervised ML has been used in MSI data analysis such as in DESI MSI studies to define the histological subtypes and estimate tumor cell abundances in a gliomas tissue [26]. Most importantly, a trained ML model can be employed to rapidly process 3D MSI data collected from a series of sections of the same tissue, eliminating time-consuming procedures needed in the multivariate analysis of individual MS images. Multiple supervised ML methods, such as support vector machines (SVM) [35] and Random Forest [36], have been developed for data analysis. In our study, the Random Forest algorithm was selected as the supervised ML tool because of its robustness of classification without overfitting issues [37], high prediction accuracy, and capability of handling a large number of input variable [38]. Compared with supervised ML, unsupervised ML (i.e., unsupervised clustering) methods can be used to directly cluster data into major components without known sample labels in advance, suggesting they are appropriate approaches to analyzing MSI experimental results without labels. Common unsupervised ML methods include k-means, PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications), and DBSCAN (Density-based Spatial Clustering of Applications with Noise). k-means is one of the most widely used methods, but it is sensitive to anomalous data points and outliers (i.e., data points that are distant from the mass of data) [39]. Compared with k-means, PAM is a more robust method to process data containing outliers. As an extension of PAM, CLARA has been developed to effectively analyze big data (e.g., more than several thousands of observations), and it is superior to other methods for its capability of self-optimizing the number of major components [40]. With its superior capability of picking up system outliers, DBSCAN has also become one of the most cited data clustering algorithms [41].

Generally, unsupervised ML approaches are more likely to be suitable to analyze large sizes of MSI data. However, due to the intrinsic high-dimensionality of MSI datasets, which contain both spatial (i.e., x, y-coordinates) and molecular (i.e., m/z values with their intensities) information of each pixel, these methods cannot be directly applied to process the raw MSI data. Instead, the dimensionality reduction of MSI data is needed before unsupervised ML can be conducted. The common dimensionality reduction techniques include the aforementioned PCA, SOM (Self-Organized Maps), and *t*-SNE (*t*-distributed Stochastic Neighbor Embedding). Among these techniques, PCA, as a common tool for linear dimensionality reduction, is widely used to extract key information from complex data, and to visualize them in low-dimensional score plots of top major components. However, the drawbacks of PCA score plot as mentioned above indicate that it is an inadequate tool for the dimensionality reduction of MSI data. SOM reduces the dimensionality of imaging data by projecting each intensity imaging to a fixed grid (hexagonal or rectangular topology), resulting in a loss of the variants relationship (i.e., spatial information of clusters) [28,42]. As a nonlinear dimensionality reduction technique, *t*-SNE is particularly suitable for clustering and visualizing highly complex multi-dimensional data [43]. With crucial information in the high-dimensional datasets retained in the low-dimensional datasets upon dimensionality reduction [28,44], *t*-SNE shows a more desirable dimension-reducing ability.

The potential applications of MCR and ML methods to a variety of areas, such as pharmacometabolomics, biomarker discovery, disease diagnosis, and clinic efficacy monitoring, have been reported [45–49]. Particularly, MCR-ALS [11] and ML [27,29] techniques have been recently applied to MSI data analysis. In the current study, we used MCR-ALS and ML (both supervised and unsupervised ML) approaches to carry out enhanced metabolomics analysis of high-spatial resolution (8 ± 2 μm) MSI data of mouse kidney sample. The datasets analyzed in this study were adapted from our previously published work using the Single-probe MSI method [50], which is a versatile technique that has been used in a variety of studies such as MS imaging [50,51], live single cell analysis [52,53], and extracellular metabolites measurement in live multicellular spheroids [54]. In the current study MCR-ALS was first utilized to group the molecular species possessing similar spatial patterns within MS images. Then, to establish an efficient platform for more comprehensive analysis, we performed both supervised and unsupervised ML studies. The Random Forest method was used in the supervised ML approach after MCR-ALS analysis, whereas DBSCAN and CLARA were utilized to cluster datasets upon dimensionality reduction using *t*-SNE. Our application of machine learning methods can be greatly helpful to reveal the hidden information present in large amounts of MSI data, which can potentially benefit biochemical and medical studies.

2. Data pre-processing and analysis

The experimental data used in the current study were obtained from our previous work [50], which describes detailed experimental protocols of the Single-probe fabrication, tissue sample preparation, and MSI data collection. A brief description of experimental protocols is provided in the Supporting Information.

2.1. Flow of MSI data analysis

As summarized in Fig. 1, the complete procedures of our MSI data analysis include four major steps: data pre-processing, MCR-ALS analysis, supervised ML, and unsupervised ML. 1) Data Pre-processing. We carried out the pre-processing of MSI raw data to obtain datasets that are suitable for the subsequent analyses. 2) MCR-ALS Analysis. MCR-ALS algorithm was used to analyze datasets upon accomplishing the pre-processing step. 3) Supervised ML. We labelled partial MCR-ALS results, and used them as the training data (Training Data Selection) for the optimization of the supervised ML model (Random Forest). The trained model was then used to classify the entire histological regions of MSI data (Tissue label prediction). The spatial distribution of clusters was constructed using the classified datasets according to their tissue labels. Supervised ML results were compared with manually assigned tissue labels under the assistance of parametric *t*-SNE (henceforth referred visualization). 4) Unsupervised ML. We used *t*-SNE algorithm to reduce the dimensionality of the high-dimensional datasets obtained from the pre-processing step, and then used unsupervised ML (CLARA and DBSCAN clustering) to analyze the lower-dimensional datasets. Similarly, the spatial distribution of clustered datasets was constructed according to their tissue labels.

2.1.1. MSI data pre-processing

In data pre-processing step, the original MSI data were converted into a format that is suitable for subsequent processing. Technical details of the data pre-processing were provided in the Supporting Information. To import the MSI data into the pre-processing platform in MATLAB, the original MSI data (.raw) was converted to the imzML format using imzMLconverter software [55]. Data pre-processing (Fig. 1), including smoothing, peak identification, noise removal, peak alignment and normalization, was carried out using the built-in functions of MATLAB Bioinformatics Toolbox by following the general procedures described by Robert C. Glen et al. [28]. Specifically, Savitzky-Golay filtering was used to perform the smoothing of spectral profiles. Peak picking was conducted using the 'mspeaks' function, and the peak assignment was determined by the sign changes of the first derivatives of the spectral profiles. Noise signal was labelled through median absolute deviation (MAD) estimation and removed by setting the threshold of noise as the signal-to-noise ratio (S/N) less than five ($S/N < 5$). For peak alignment, the 'mspalign' command was applied to the spectra obtained from the previous steps. The default estimation method, histogram (kernel density function), was used to determine the peak locations (i.e., *m/z* values). For the comparison of relative ion intensities among different MS scans, peak intensities were normalized to the total ion current (TIC). Upon accomplishing the data pre-processing, an aligned data matrix (14,100 × 182) was obtained for subsequent MCR-ALS analysis: one dimension of the matrix is the number of aligned *m/z* values (182), and the other is the number of pixels (188 pixel × 75 line = 14100).

2.1.2. MCR-ALS analysis

To conduct MCR-ALS analysis, Multivariate Curve Resolution Toolbox, developed by Tauler et al. [19], running under MATLAB environment was used to analyze the data matrix obtained from the pre-processing step. To better differentiate low intensity ions in the MCR plots, a logarithmic transformation ($\log_2(x)$) of ion intensities in the data matrix (182 × 14100) was conducted prior to the MCR-ALS analysis. The Singular Value Decomposition (SVD) approach was used to estimate the appropriate number of eigenvalues (i.e., components), and each component represents a unique pattern of the spatial distribution of a group of ions. Ions with similar spatial distribution were assigned to the same group. Five major components (explained variance of 97.7%) were found to be an optimal number of coexisted patterns. The spatial distribution of each component was exported from the MCR-ALS through a user-friendly interface, and the corresponding ions were shown as mass spectra using R language (details are provided in the SI).

2.1.3. Supervised machine learning (ML): Random Forest

Training data and testing data (Training Data Selection; Table S1) were manually selected from three classified histological regions of mouse kidney (Fig. 3A) determined by MCR-ALS analysis to provide input features (intensities of ions) and output labels (inner medulla, outer medulla, or cortex). The Random Forest approach provided in R language was trained using the selected training data. To validate this supervised ML model, testing data were used to test the accuracy of prediction. Three trails have been executed, and the average prediction accuracy is >99% (Table S2). Then, the optimized model (Optimized Decision Forest) was used to process the rest of the datasets for tissue label prediction, i.e., to

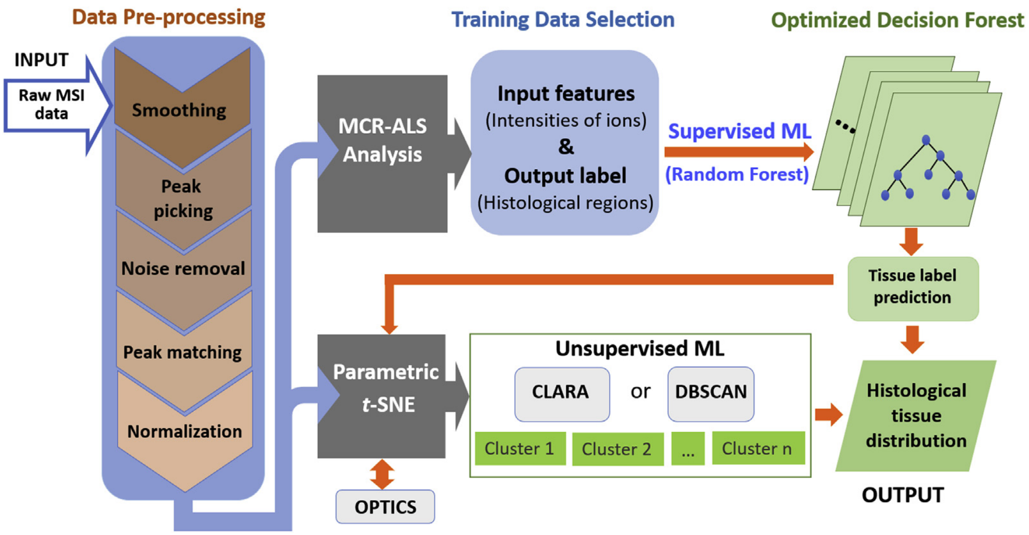


Fig. 1. Flowchart of MSI data analysis.

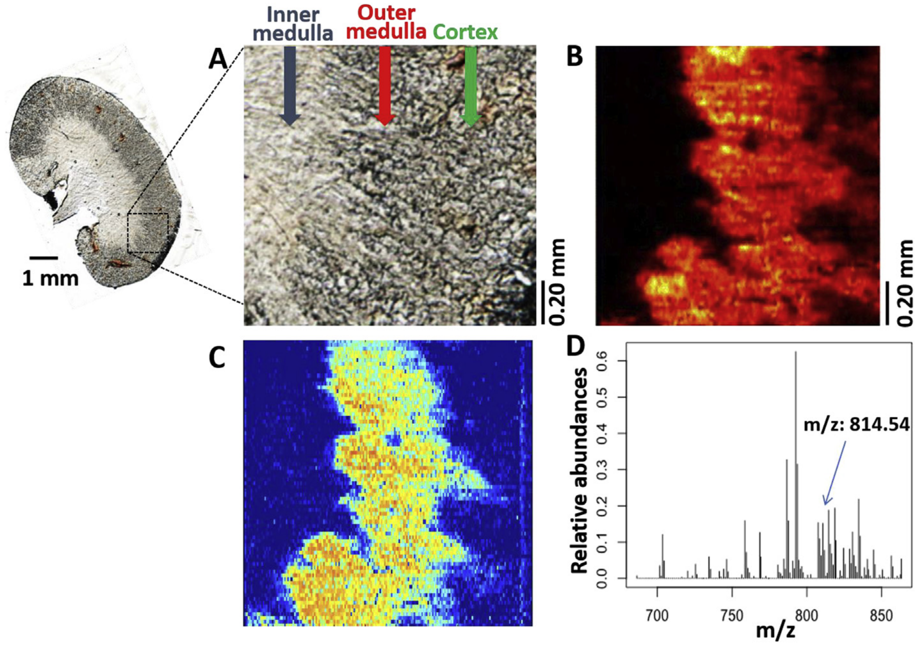


Fig. 2. MSI data analysis using selected ion or MCR methods. (A) Optical image of mouse kidney illustrating three histological regions: inner medulla, outer medulla, and cortex. (B) An example of MS image constructed using the selected ion ([PC (38:5) + Na]$^+$; *m/z* 814.5726). (C) An example of spatial distribution pattern obtained from MCR-ALS analysis. (D) The mass spectrum of the grouped molecules possessing the same pattern of spatial distribution shown in C. Note: Figs. A and B are adapted from "High Resolution Tissue Imaging Using the Single-probe Mass Spectrometry under Ambient Conditions," by Wei Rao, Ning Pan, and Zhibo Yang, 2015, *Journal of the American Society of Mass Spectrometry*, 26, 986-993 [50]. Copyright 2015 by *The American Society of Mass Spectrometry*.

classify ions in the MSI data into each of three regions. To provide clear physical meanings of the overall results obtained from the supervised ML approach, the predicted tissue labels (Fig. 3B) were transformed into histological tissue distribution using the R language.

The supervised ML results were verified through *t*-SNE visualization. *t*-SNE has become a popular method to transform the high-dimensional structures of data into two- or three-dimensional formats that can be conveniently visualized. In our studies, we used *t*-SNE, which is provided as a built-in function in Statistics and Machine Learning Toolbox of MATLAB, to visualize the results from both manually assigned tissues labels (Fig. 3A) and ML-generated (Fig. 3B) upon finishing the dimensionality reduction (Fig. 3C and D). As a validation of *t*-SNE results, the OPTICS (Ordering Points To Identify the Clustering Structure) approach was also used to index the data points in order to identify the clustering structure (Fig. S2B). The optimized parameters of *t*-SNE and OPTICS were detailed in the Supporting Information.

2.1.4. Unsupervised machine learning (ML): CLARA and DBSCAN

We performed CLARA analysis using R, provided within package 'cluster', with six optimal clusters and the sample size equals to 50. CLARA results were visualized using the R language (Fig. 4A). DBSCAN, another popular unsupervised ML method with an

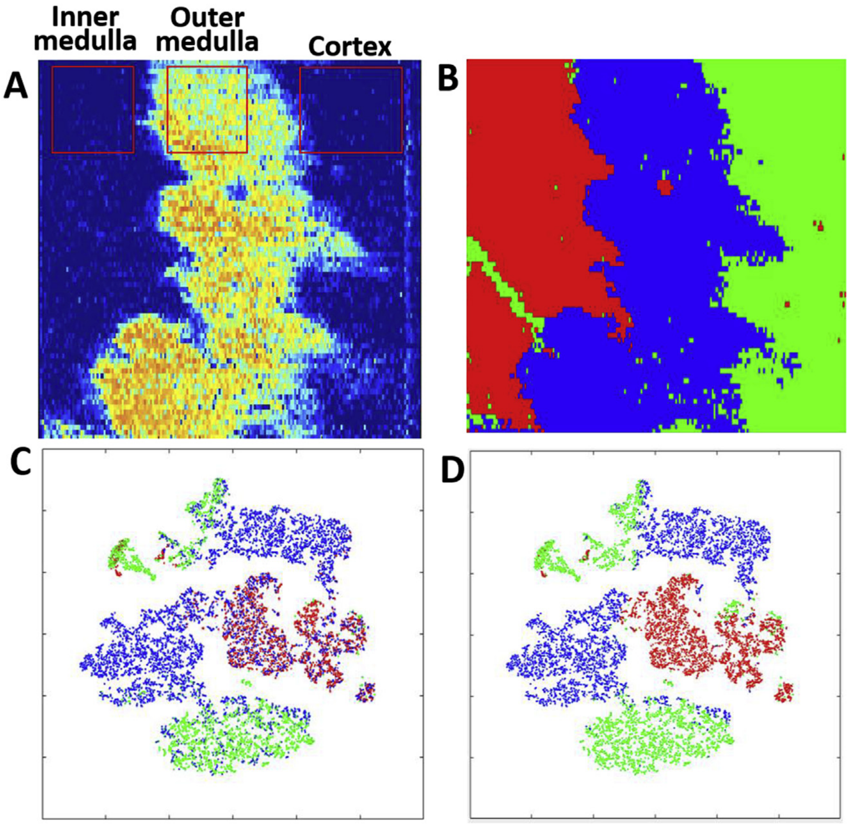


Fig. 3. MSI data analysis using supervised ML method. (A) Training data for supervised ML were manually selected from three regions (inner medulla, outer medulla, and cortex) based on the MCR-ALS results. (B) Histological tissue distribution constructed using supervised ML (Random Forest) results. (C) The clustering capability of MCR-ALS method was evaluated using *t*-SNE. (D) The clustering capability of supervised ML method was evaluated using *t*-SNE.

excellent capability of picking up system outliers, was used to perform a comparison with CLARA. Similarly, DBSCAN results were visualized using the R (Fig. 4C). Finally, to interpret the physical meanings of all the clustered data from CLARA and DBSCAN analyses, the classification maps of mouse kidney tissue were constructed using the R (all packages can be found at https://cran.r-project.org/web/packages) shown in Figs. 4B and 4C. As a comparison of *t*-SNE results, PCA was also used to reduce the dimensionality of the MSI data; however, the boundaries among clusters (Fig. S7) are not as clear as those from *t*-SNE results (Fig. 3C), indicating that PCA is inadequate to visualize high dimensional MSI dataset.

3. Results and discussion

3.1. MCR-ALS approach grouped molecules with similar spatial distribution pattern from MS images

Fig. 2A and B, which are adapted from our previous publication [50], illustrate the optical image of kidney slice (with the analyzed area in MSI experiments) and the MS image of a selected ion ([PC(38:5) + Na]$^+$, m/z 814.5726), respectively. Three different histological regions (i.e., inner medulla, outer medulla, and cortex) can be generally observed from the optical image. Using the high-spatial resolution (8.5 μm) of our Single-probe MSI techniques, 14100 pixels (equivalent to 14,100 mass spectra) were obtained from the MS image with an area of ~1.0 × 1.0 mm^2 (Fig. 2A). Data analysis was carried out using a Dell Precision T5500 work station (processor: dual Intel (R) Xeon(R) CPU ×5650 2.66 GHz; memory (RAM): 72.0 GB). 182 common ions were generated from pretreatment for the subsequent analysis as outlined in Method 2.1.2. A complete MCR analysis of the pretreated MSI data took about 20 min to accomplish. MCR-ALS approach was used to determine the number of major components present in the data matrix. Using the Singular Value Decomposition (SVD) method (Fig. S5), we concluded that five eigenvalues (i.e., components) were sufficient to represent the majority information (explained variance of 97.7%, Fig, S6 and Table S7) of the entire MSI dataset. Species in each component (i.e., m/z values with their corresponding relative intensities) were extracted from the data matrix. For example, a component obtained from MCR-ALS analysis (Fig. 2C) exhibits very similar spatial features as a MS image constructed using a selected ion (m/z 814.5726) within outer medulla (Fig. 2B) [50]. In fact, this selected ion is among many others with very similar spatial features as summarized in Fig. 2D. It is very likely that compounds present similar spatial distribution have correlated metabolomic functions. For example, previous studies indicate that PC (40:6) and PC (36:4) have high abundances in the kidney outer medulla, and they both are upregulated in response to cisplatin treatment by shaping membrane-protein function [56,57]. The MCR-ALS classification can potentially benefit the future discovery of the metabolomic pathways among species possessing similar patterns of spatial distribution. We summarized the top 15 most abundant ions grouped in the component 1 in Table 1, which include *m/z* values from MCR analysis, experimental *m/z* values, and tentatively assigned metabolites acquired from METLIN (https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage) [58]. In addition, the loading scores were provided to illustrate the relative contribution of each metabolite to component 1. Using similar approaches, other four major components representing the

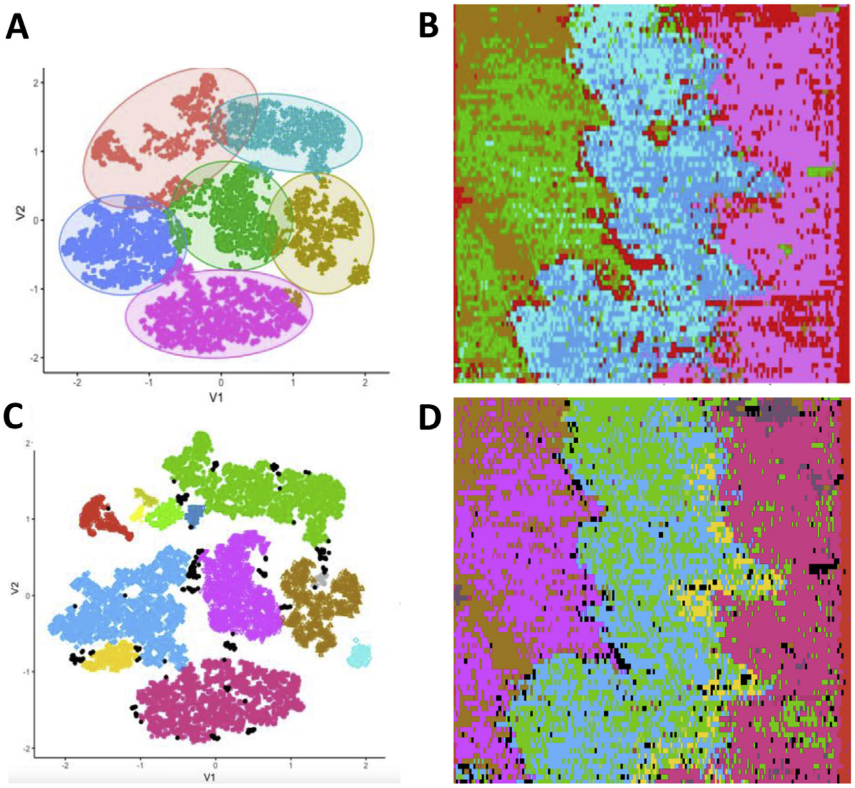


Fig. 4. MSI data analysis using unsupervised ML methods. (A) CLARA was used to cluster the MSI data into six major groups upon optimization. (B) Histological tissue distribution constructed using CLARA results. (C) DBSCAN was applied to cluster the MSI data into seven major groups. (D) Histological tissue distribution constructed using DBSCAN results.

Table 1
Top 15 most abundant ions grouped in component 1 (Fig. 2D).

MCR (m/z)	Experiment (m/z)	Tentative Assignment[a]	Exact (m/z)	ppm	Loading Score
792.5428	792.5911	PC(18:1/20:4)	792.5902	1	0.6251
786.6442	786.6018	PC(18:1/18:1)	786.6007	1	0.3268
793.6425	793.5951	PG (22:0/15:0)	793.5953	0	0.3146
834.6404	834.6015	PC(20:4/20:2)	834.6007	0	0.2185
818.6404	818.6063	PC(18:1/22:5)	818.6058	0	0.1934
814.5414	814.5729	PC(18:1/20:4)	814.5721	0	0.1878
758.5451	758.5705	PE (22:2/15:0)	758.5694	1	0.1588
810.6423	810.6019	PC(20:3/18:1)	810.6007	1	0.1511
830.5394	830.5464	PC(22:5/16:0)	830.5460	0	0.1276
768.5447	768.5550	PE (22:4/16:0)	798.5538	1	0.1262
703.5506	703.5757	SM(18:1/16:0)	703.5749	1	0.1206
808.5428	808.5864	PC(20:4/18:1)	808.5851	1	0.1086
819.3422	819.6100	PG (21:0/18:1)	819.6110	1	0.1037
815.5661	815.5763	PG (18:0/19:0)	815.5773	1	0.0939
824.5409	824.5578	PC(18:2/18:0)	824.5566	1	0.0824

[a] PC: Phosphatidylcholines, PE: Phosphatidylethanolamine, SM: Sphingomyelin, PG: Phosphatidylglycerol.

spatial and molecular features of the MS images for the inner medulla and cortex regions were provided in the Supporting Information (Fig. S1 and Table S3-S6).

3.2. Supervised ML utilized MCR-ALS results and improved discovery of subtle features from MSI data

Since the quality of training data can significantly affect the ML prediction capability, it is crucial to carefully select the appropriate training datasets representing characteristics of the overall data. We selected the datasets according to results obtained from the MCR-ALS analysis, and divided them into training and testing datasets (Table S1), which are grouped *m/z* values of species present on each of those three regions on the tissue slice (i.e., inner medulla, outer medulla, and cortex) (Fig. 3A). A Random Forest classification model was trained using the training datasets. We evaluated the accuracy of this trained model using the testing datasets and achieved a high predictive accuracy (>99%; details are provided in Table S2). This trained Random Forest model was then applied to the rest of the MSI datasets to predict their tissue labels, i.e., to classify the detected molecules to each of those three physiological regions on tissue slice. To interpret the physical meanings of the supervised ML results, classified data were used to generate the tissue classification map, in which three different colors represent three physiological regions on tissue (Fig. 3B). Because the training datasets were selected based on MCR-ALS results, the supervised ML generally reproduced the overall features of three histological regions. Particularly, the spatial distribution of molecules in the inner medulla exhibited similar features that can be observed in a spatial pattern obtained from MCR-ALS analysis (Fig. 3A). Moreover, our supervised ML analysis enhanced some subtle features in MS images, and generated clearer boundaries between different regions.

To verify that the supervised ML method provides more capability of classification than MCR-ALS approach, *t*-SNE has been employed to process the MCR-ALS results. Since *t*-SNE requires tissue labels to be known prior to plotting the *t*-SNE clustering results, three types of tissue labels (i.e., inner medulla, outer medulla, and cortex) were assigned to the 14,100 scans based on the MCR-ALS results (Fig. 3A). Using the *t*-SNE to cluster these labels, overlapped color dots were observed between outer medulla and inner medulla or cortex (Fig. 3C), indicating unclear assignments of boundary labels can arise from manually assigned tissue labels solely based on the MCR-ALS results. In contrast, for tissues labels predicted from the supervised ML analysis, an improved *t*-SNE cluster map was generated (Fig. 3D), implying less biased determination of boundary pixels was achieved using supervised ML rather than human intuition. More importantly, the established ML models can be directly used for efficient analyses of many other MS images obtained from the same tissues such as slices used in 3D MSI studies of a given tissue.

3.3. Unsupervised ML extracted more molecular and spatial information from MSI data

Supervised ML requires training data to optimize the model prior to any applications. However, additional efforts, such as MCR-ALS or H&E staining, are needed to provide histological information for selecting the training data. In contrast, such training data selection is not required for unsupervised ML methods. To generate datasets with suitable size and structure, *t*-SNE algorithm was utilized to reduce the high-dimensional MSI data. Notably, directly applying unsupervised ML (without *t*-SNE dimensionality reduction) to analyze our original MSI dataset cannot produce any optimal clusters through parameter-changing attempts (Fig. S4), indicating that using *t*-SNE for dimensionality reduction is a key step to obtain the classification of high-dimensional MSI dataset. To verify the effectiveness of *t*-SNE for dimensionality reduction, OPTICS algorithms were used to generate the point reachability plot of the lower-dimensional datasets (Fig. S3 B), in which the reachability indicates the extent of separation in *t*-SNE plot (Fig. 3) (details of OPTICS are in Fig. S2B).

Two unsupervised ML approaches, CLARA and DBSCAN, were chosen to perform the classification. CLARA is suitable to deal with big dataset and capable of optimizing the number of clusters, whereas DBSCAN has the advantage of picking up system outliers. The results of CLARA and DBSCAN were shown in Fig. 4A and C, respectively. The physical meanings of the unsupervised ML-generated data clusters were further investigated through reconstructing their spatial distributions (Fig. 4B and D). Interestingly, without given information of tissue distribution patterns, CLARA approach generated an optimal number (six) of clusters possessing spatial distributions that are relevant to the histological regions shown on the optical image (Fig. 2A). These six optimal clusters coexist in our dataset, and two sub-regions were discovered in both the inner and outer medulla regions (Fig. 4A and B). DBSCAN analysis also discovered similar cluster patterns as well as several minor sub-regions that are comparable to CLARA results. In addition, another advantage of using DBSCAN was to identify the system outliers, labelled as the black dots in Fig. 4C and D, that were absent from the CLARA analysis. Noteworthy, clusters assigned with red color from both CLARA and DBSCAN results are basically classified as margins of the tissue slice (Fig. 4).

The reconstructed histological tissue distributions were obtained from both supervised (Fig. 3B) and unsupervised (Fig. 4B and D) ML approaches, and similar features of spatial distributions were obtained. In supervised ML method, a portion of MCR-ALS results was used as the training data to optimize the model, which was further used to reproduce all three clusters. In unsupervised ML methods, including CLARA and DBSCAN, the *t*-SNE results were used to perform the tissue classification (i.e., to allocate the spatial locations of detected molecule). Encouragingly, two different approaches led to similar results, indicating that the clusters we identified and the analysis protocols in use were reliable.

Although both supervised and unsupervised ML methods have been successfully applied to our MSI data analyses, they have their own inherent advantages and disadvantages. Previous studies indicate that supervised ML is a suitable tool to identify different tissue features or potentially distinguish pathological cells (e.g., cancer cells) from normal cells in MSI data analysis [59]. Although supervised ML algorithm shows an adaptability for rapid prediction of unknown MS images, careful selection of training data, which requires additional information, is critical for the optimization of models. This type of training process is obligatory before supervised ML can be applied for any MSI data analysis. In contrast, unsupervised ML can be conveniently utilized to analyze MSI data without model training. However, the results need to be validated by comparing with those obtained from other labeling studies or techniques such as tissue staining [28,60]. Nevertheless, it is still possible that some subtle physical or chemical features can be overlooked, or experiments cannot be performed due to limitations of available techniques or samples. Therefore, to significantly increase the data-analyzing efficiency while minimizing uncertainties during the MSI data analyses, we suggest using a combined method including both supervised and unsupervised ML approaches.

4. Conclusion

As an emerging molecular imaging technique used for biological tissue analysis in fundamental research and biomedical applications, MSI experiments usually generate huge amounts of data. Traditionally, MSI data analysis is generally carried out for manually selected ions. Due to the large size and high-dimensionality of MSI datasets, it has been very challenging to conduct comprehensive data analysis to extract overall features representing essential molecular and spatial information present in biological tissues [14]. Therefore, advanced data analysis methods are needed to perform more efficient analysis of large sizes of MSI data [61].

MCR-ALS is a multivariate analysis method that can decompose the complex MSI dataset into major components with spatial distribution patterns and grouped ions [11,12]. The application of MCR-ALS can enhance the MSI data analysis without tissue-histological knowledge [62]. To further increase the efficiency of data mining from larger sizes of MSI results, machine learning (ML) methods are likely to be more efficient approaches. In the current study, both supervised and unsupervised ML methods have been utilized to distinguish histological regions from the high-spatial resolution MS images of mouse kidney slice. Two programming languages, R and MATLAB, were cooperatively used to implement the methods for MSI data processing. Regardless of the type of ML approach used in the analysis of MSI data, reducing the dimensionality of target datasets was an obligatory step. As a prevailing dimensionality reduction tool, *t*-SNE was employed in both supervised and unsupervised ML studies. In supervised ML (i.e., Random Forest) studies, the histological regions determined by

MCR-ALS analysis were used as a guide to select the defined training datasets, and *t*-SNE algorithm was utilized to reduce the high-dimensional datasets allowing us to visualize the results obtained from Random Forest analyses. In unsupervised ML studies (CLARA and DBSCAN), *t*-SNE has been proven as an effective approach to process the pre-treated MSI datasets and ensure them to be suitable for subsequent unsupervised ML processes. Both supervised and unsupervised ML approaches are effective for MSI data analysis. However, combined supervised and unsupervised ML studies are likely to be more effective to extract the overall chemical and spatial features from complex MSI data with minimum overlooked information. Our studies indicate that advanced data analysis methods, including MCR-ALS and ML approaches, are efficient tools for comprehensive analysis of large amounts of high-spatial resolution MS images obtained using our Single-probe MSI techniques. These emerging methods can be broadly utilized for many other MSI studies conducted using different techniques, and to promote the growth of data-analyzing tools needed for big data science.

Acknowledgements

We thank Paolo Inglese and Dr. Robert Glen (Imperial College London) for sharing their MATLAB scripts for data pre-processing. We thank Dr. Joaquim Jaumot and Dr. Roma Tauler for suggestions of using Multivariate Curve Resolution Toolbox. We also thank two anonymous reviewers for providing helpful comments. This research was supported by grants from National Institutes of Health (R01GM116116 and R21CA204706).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2018.02.031.

References

[1] E.R. Amstalden van Hove, D.F. Smith, R.M. Heeren, A concise review of mass spectrometry imaging, J. Chromatogr. A 1217 (2010) 3946–3954.
[2] A. Cohen, Mass spectrometry, review of the basics: electrospray, MALDI and commonly used mass analyzers, Appl. Spectrosc. Rev. 44 (2009) 210–230.
[3] J. Brison, M.A. Robinson, D.S.W. Benoit, S. Muramoto, P.S. Stayton, D.G. Castner, TOF-SIMS 3D imaging of native and non-native species within HeLa cells, Anal. Chem. 85 (2013) 10869–10877.
[4] D.R. Ifa, J.M. Wiseman, Q.Y. Song, R.G. Cooks, Development of capabilities for imaging mass spectrometry under ambient conditions with desorption electrospray ionization (DESI), Int. J. Mass Spectrom. 259 (2007) 8–15.
[5] L.A. McDonnell, R.M. Heeren, Imaging mass spectrometry, Mass Spectrom. Rev. 26 (2007) 606–643.
[6] C. Wu, A.L. Dill, L.S. Eberlin, R.G. Cooks, D.R. Ifa, Mass spectrometry imaging under ambient conditions, Mass Spectrom. Rev. 32 (2013) 218–243.
[7] J.M. Spraggins, R. Caprioli, High-speed MALDI-TOF imaging mass spectrometry: rapid ion image acquisition and considerations for next generation instrumentation, J. Am. Soc. Mass Spectrom. 22 (2011) 1022–1031.
[8] G. McCombie, D. Staab, M. Stoeckli, R. Knochenmuss, Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis, Anal. Chem. 77 (2005) 6118–6124.
[9] I. Yao, Y. Sugiura, M. Matsumoto, M. Setou, In situ proteomics with imaging mass spectrometry and principal component analysis in the Scrapper-knockout mouse brain, Proteomics 8 (2008) 3692–3701.
[10] A.L. Dill, L.S. Eberlin, A.B. Costa, C. Zheng, D.R. Ifa, L.A. Cheng, T.A. Masterson, M.O. Koch, O. Vitek, R.G. Cooks, Multivariate statistical identification of human bladder carcinomas using ambient ionization imaging mass spectrometry, Chem. Eur J. 17 (2011) 2897–2902.
[11] J. Jaumot, R. Tauler, Potential use of multivariate curve resolution for the analysis of mass spectrometry images, Analyst 140 (2015) 837–846.
[12] W. Rao, D.J. Scurr, J. Burston, M.R. Alexander, D.A. Barrett, Use of imaging multivariate analysis to improve biochemical and anatomical discrimination in desorption electrospray ionisation mass spectrometry imaging, Analyst 137 (2012) 3946–3953.
[13] T. Alexandrov, M. Becker, S.O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, P. Maass, Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering, J. Proteome Res. 9 (2010) 6535–6546.
[14] T. Alexandrov, MALDI imaging mass spectrometry: statistical data analysis and current computational challenges, BMC Bioinf. 13 (16) (2012) S11.
[15] M. Hanselmann, M. Kirchner, B.Y. Renard, E.R. Amstalden, K. Glunde, R.M. Heeren, F.A. Hamprecht, Concise representation of mass spectrometry images by probabilistic latent semantic analysis, Anal. Chem. 80 (2008) 9649–9658.
[16] S.O. Deininger, M.P. Ebert, A. Futterer, M. Gerhard, C. Rocken, MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers, J. Proteome Res. 7 (2008) 5230–5236.
[17] D. Trede, S. Schiffler, M. Becker, S. Wirtz, K. Steinhorst, J. Strehlow, M. Aichler, J.H. Kobarg, J. Oetjen, A. Dyatloy, S. Heldmann, A. Walch, H. Thiele, P. Maass, T. Alexandrov, Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: three-dimensional spatial segmentation of mouse kidney, Anal. Chem. 84 (2012) 6079–6087.
[18] K.S. Chuang, H.L. Tzeng, S. Chen, J. Wu, T.J. Chen, Fuzzy c-means clustering with spatial information for image segmentation, Comput. Med. Imaging Graph. 30 (2006) 9–15.
[19] I. Klinkert, K. Chughtai, S.R. Ellis, R.M.A. Heeren, Methods for full resolution data exploration and visualization for large 2D and 3D mass spectrometry imaging datasets, Int. J. Mass Spectrom. 362 (2014) 40–47.
[20] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36.
[21] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12.
[22] I.S. Perez, M.J. Culzoni, G.G. Siano, M.D.G. Garcia, H.C. Goicoechea, M.M. Galera, Detection of unintended stress effects based on a metabonomic study in tomato fruits after treatment with carbofuran pesticide. Capabilities of MCR-ALS applied to LC-MS three-way data arrays, Anal. Chem. 81 (2009) 8335–8346.
[23] A.C. Neves, R. Tauler, K.M. de Lima, Area correlation constraint for the MCR-ALS quantification of cholesterol using EEM fluorescence data: a new approach, Anal. Chim. Acta 937 (2016) 21–28.
[24] C. Bedia, R. Tauler, J. Jaumot, Analysis of multiple mass spectrometry images from different Phaseolus vulgaris samples by multivariate curve resolution, Talanta 175 (2017) 557–565.
[25] H.D. Bean, J.J. Zhu, J.E. Hill, Characterizing bacterial volatiles using secondary electrospray ionization mass spectrometry (SESI-MS), Jove-Journal of Visualized Experiments (2011).
[26] L.S. Eberlin, I. Norton, A.L. Dill, A.J. Golby, K.L. Ligon, S. Santagata, R.G. Cooks, N.Y. Agar, Classifying human brain tumors by lipid imaging with mass spectrometry, Cancer Res. 72 (2012) 645–654.
[27] M. Galli, I. Zoppis, A. Smith, F. Magni, G. Mauri, Machine learning approaches in MALDI-MSI: clinical applications, Expert. Rev. Proteomic. 13 (2016) 685–696.
[28] P. Inglese, J.S. McKenzie, A. Mroz, J. Kinross, K. Veselkov, E. Holmes, Z. Takats, J.K. Nicholson, R.C. Glen, Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer, Chem. Sci. 8 (2017) 3500–3511.
[29] Z.P. Zhou, R.N. Zare, Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning, Anal. Chem. 89 (2017) 1369–1372.
[30] S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas, Machine learning: a review of classification and combining techniques, Artif. Intell. Rev. 26 (2006) 159–190.
[31] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.
[32] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (2015) 321–332.
[33] A.L. Swan, A. Mobasheri, D. Allaway, S. Liddell, J. Bacardit, Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology, OMICS 17 (2013) 595–610.
[34] M.N. Wernick, Y.Y. Yang, J.G. Brankov, G. Yourganov, S.C. Strother, Machine learning in medical imaging, IEEE Signal Process. Mag. 27 (2010) 25–38.
[35] B. Schölkopf, A.J. Smola, Learning with Kernels : Support Vector Machines, Regularization, Optimization, and beyond, MIT Press, Cambridge, Mass, 2002.
[36] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[37] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.
[38] M. Hanselmann, U. Kothe, M. Kirchner, B.Y. Renard, E.R. Amstalden, K. Glunde, R.M.A. Heeren, F.A. Hamprecht, Toward digital staining using imaging mass spectrometry and random forests, J. Proteome Res. 8 (2009) 3558–3567.
[39] L. Ertoz, M. Steinbach, V. Kumar, Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, in: Proceedings of the Third Siam International Conference on Data Mining, 2003, pp. 47–58.
[40] A. Nagpal, A. Jatain, D. Gaur, Review based on data clustering algorithms, in: 2013 Ieee Conference on Information and Communication Technologies (Ict 2013), 2013, pp. 298–303.
[41] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd (1996) 226–231.
[42] P. Franceschi, R. Wehrens, Self-organizing maps: a versatile tool for the automatic analysis of untargeted imaging datasets, Proteomics 14 (2014) 853–861.
[43] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res.

9 (2008) 2579–2605.

[44] W.M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M.J.T. Reinders, A. Walch, L.A. McDonnell, B.P.F. Lelieveldt, Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 12244–12249.

[45] J.P. Connor, M. Symons, G.F. Feeney, R.M. Young, J. Wiles, The application of machine learning techniques as an adjunct to clinical decision making in alcohol dependence treatment, Subst. Use Misuse 42 (2007) 2193–2206.

[46] Y.K. Kim, K.S. Na, Application of machine learning classification for structural brain MRI in mood disorders: critical review from a clinical perspective, Prog. Neuro-Psychopharmacol. Biol. Psychiatry, 80 (Pt B) (2017) 71–80.

[47] P. Thottakkara, T. Ozrazgat-Baslanti, B.B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, A. Bihorac, Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications, PLoS One 11 (2016) e0155705.

[48] Z. Zhang, When doctors meet with AlphaGo: potential application of machine learning to clinical medicine, Ann. Transl. Med. 4 (2016) 125.

[49] G. Lin, Y.L. Chung, Current opportunities and challenges of magnetic resonance spectroscopy, positron emission tomography, and mass spectrometry imaging for mapping cancer metabolism in vivo, BioMed Res. Int. 2014 (2014) 625095.

[50] W. Rao, N. Pan, Z. Yang, High resolution tissue imaging using the single-probe mass spectrometry under ambient conditions, J. Am. Soc. Mass Spectrom. 26 (2015) 986–993.

[51] W. Rao, N. Pan, X. Tian, Z. Yang, High-resolution ambient MS imaging of negative ions in positive ion mode: using dicationic reagents with the single-probe, J. Am. Soc. Mass Spectrom. 27 (2016) 124–134.

[52] N. Pan, W. Rao, N.R. Kothapalli, R. Liu, A.W. Burgett, Z. Yang, The single-probe: a miniaturized multifunctional device for single cell mass spectrometry analysis, Anal. Chem. 86 (2014) 9376–9380.

[53] N. Pan, W. Rao, S.J. Standke, Z. Yang, Using dicationic ion-pairing compounds to enhance the single cell mass spectrometry analysis using the single-probe: a microscale sampling and ionization device, Anal. Chem. 88 (2016) 6812–6819.

[54] M. Sun, X. Tian, Z. Yang, Microscale mass spectrometry analysis of extracellular metabolites in live multicellular tumor spheroids, Anal. Chem. 89 (2017) 9069–9076.

[55] A.M. Race, I.B. Styles, J. Bunch, Inclusive sharing of mass spectrometry imaging data requires a converter for all, J. Proteomics 75 (2012) 5111–5112.

[56] E. Moreno-Gordaliza, D. Esteban-Fernandez, A. Lazaro, B. Humanes, S. Aboulmagd, A. Tejedor, M.W. Linscheid, M.M. Gomez-Gomez, MALDI-LTQ-Orbitrap mass spectrometry imaging for lipidomic analysis in kidney under cisplatin chemotherapy, Talanta 164 (2017) 16–26.

[57] R. Phillips, T. Ursell, P. Wiggins, P. Sens, Emerging roles for lipids in shaping membrane-protein function, Nature 459 (2009) 379–385.

[58] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: a metabolite mass spectral database, Ther. Drug Monit. 27 (2005) 747–751.

[59] C. Heylman, R. Datta, A. Sobrino, S. George, E. Gratton, Supervised machine learning for classification of the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes, PLoS One 10 (2015) e0144572.

[60] T.J. Fuchs, J.M. Buhmann, Computational pathology: challenges and promises for tissue analysis, Comput. Med. Imaging Graph. 35 (2011) 515–530.

[61] E.A. Jones, S.O. Deininger, P.C. Hogendoorn, A.M. Deelder, L.A. McDonnell, Imaging mass spectrometry statistical analysis, J. Proteomics 75 (2012) 4962–4989.

[62] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, J.P. Huvenne, Multivariate curve resolution methods in imaging spectroscopy: influence of extraction methods and instrumental perturbations, J. Chem. Inf. Comput. Sci. 43 (2003) 2057–2067.