msiPL: Non-linear Manifold and Peak Learning of Mass Spectrometry Imaging Data Using Artificial Neural Networks

Walid M. Abdelmoula¹, Begona Gimenez-Cassina Lopez¹, Elizabeth C. Randall², Tina Kapur², Jann N. Sarkaria³, Forest M. White⁴, Jeffrey N. Agar⁵, William M. Wells^{2,6}, and Nathalie Y.R. Agar^{*,1,2,7}

¹Surgical Molecular Imaging Laboratory, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

²Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

³Department of Radiation Oncology, Mayo Clinic, 200 First St SW, Rochester MN 55902, USA

⁴Department of Biological Engineering, Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, 02142, USA

⁵Department of Chemistry and Chemical Biology, Northeastern University, 412 TF (140 The Fenway), Boston, MA 02111, USA

⁶Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

⁷Department of Cancer Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115

Corresponding Author: Nathalie Y. R. Agar, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, 617-525-7374, Nathalie_Agar@dfci.harvard.edu

Abstract

Mass spectrometry imaging (MSI) is an emerging technology that holds potential for improving clinical diagnosis, biomarker discovery, metabolomics research and pharmaceutical applications. The large data size and high dimensional nature of MSI pose computational and memory complexities that hinder accurate identification of biologically-relevant molecular patterns. We propose msiPL, a robust and generic probabilistic generative model based on a fully-connected variational autoencoder for unsupervised analysis and peak learning of MSI data. The method can efficiently learn and visualize the underlying non-linear spectral manifold, reveal biologically-relevant clusters of tumor heterogeneity and identify underlying informative m/z peaks. The method provides a probabilistic parametric mapping to allow a trained model to rapidly analyze a new unseen MSI dataset in a few seconds. The computational model features a memory-efficient implementation using a minibatch processing strategy to enable the analyses of big MSI data (encompassing more than 1 million high-dimensional datapoints) with significantly less memory. We demonstrate the robustness and generic applicability of the application on MSI data of large size from different biological systems and acquired using different mass spectrometers at different centers, namely: 2D Matrix-Assisted Laser Desorption Ionization (MALDI) Fourier Transform Ion Cyclotron Resonance (FT ICR) MSI data of human prostate cancer, 3D MALDI Time-of-Flight (TOF) MSI data of human oral squamous cell carcinoma, 3D Desorption Electrospray Ionization (DESI) Orbitrap MSI data of human colorectal adenocarcinoma, 3D MALDI TOF MSI data of mouse kidney, and 3D MALDI FT ICR MSI data of a patient-derived xenograft (PDX) mouse brain model of glioblastoma.

Significance:

Mass spectrometry imaging (MSI) provides detailed molecular characterization of a tissue specimen while preserving spatial distributions. However, the complex nature of MSI data slows down the processing time and poses computational and memory challenges that hinder the analysis of multiple specimens required to extract biologically relevant patterns. Moreover, the subjectivity in the selection of parameters for conventional pre-processing approaches can lead to bias. Here, we present a generative probabilistic deep-learning model that can analyze and non-linearly visualize MSI data independent of the nature of the specimen and of the MSI platform. We demonstrate robustness of the method with application to different tissue types, and envision it as a new generation of rapid and robust analysis for mass spectrometry data.

Introduction

Mass spectrometry imaging (MSI) is a rapidly growing technology that holds high promise to impact the practice of anatomic pathology and drug development¹⁻⁶. MSI provides simultaneous mapping of hundreds to thousands of molecules directly from a tissue section in a label free manner⁷. Moreover, MSI can provide direct molecular imaging of multiple types of molecules such as proteins, peptides, lipids, metabolites and drug molecules with high sensitivity and molecular specificity. These molecular data can play a substantial role for improving clinical diagnosis and prognosis^{4,8}, pathway identification^{9,10}, biomarker discovery^{11,12}, and surgical guidance^{4,13}. Sample preparation, ionization techniques and instrumentation are determinants for the effectiveness of analyte detection. Matrix-assisted laser desorption ionization (MALDI) and derivatives of electrospray ionization (ESI) such as desorption (DESI) and continuous flow surface sampling are among the most common ionization techniques for MS towards clinical applications^{1,14}. For a mass spectrometry image acquisition, molecules are desorbed and ionized from the surface of a sample, and then separated in a mass analyzer based on their mass-to-charge ratio (m/z) and detected to measure their relative abundance forming a mass spectrum.

MSI data is large and complex. For instance, a raw high mass resolution MSI file size can reach up to a few terabytes of spectral information. Computational developments that would more efficiently and accurately mine MSI data to identify molecular signatures of clinical importance and enable new biomarker discovery have the potential to expand the applicability of MSI^{15–17}. However, the complex nature of MSI data hinders efficient data mining, clustering, visualization and classification using traditional machine learning techniques¹⁸. This data complexity poses memory and computational challenges, namely due to “the curse of dimensionality” in which original MSI data hold up to tens of thousands of spectra each of which has $10^4 - 10^6$ m/z spectral bins, and the non-linear separability of the underlying spectral manifold in the high dimensional space.

Peak picking is currently a fundamental data pre-processing step for the analysis of original mass spectral data at the basis of biomarker discovery^{19–22}. Peak picking is used to alleviate the sparsity and reduce the original spectral dimensionality while optimally increasing the signal-to-noise ratio through retaining as many informative m/z features as possible. In addition, peak picking is essential for identification, quantification, and discovery of molecular biomarkers²³. Despite the generally acceptable performance of peak picking algorithms, each applied parameter, e.g. baseline subtraction, peak width, signal-to-noise ratio (S/N), and smoothing introduces a level of subjectivity that influences the resulting peak list^{19,24}. The optimization of parameter selection largely relies on the user’s expertise and can therefore lead to significant discrepancy in overall biomarker identification²⁵. These limitations are exacerbated when applying these workflows to MSI data, which is large and has an added level of complexity with spatial information¹⁶. The existing data analysis workflows are then slow and might lack the required sensitivity needed to preserve low intensity spectral information present in small regions of the tissue but with high specificity^{16,26}.

Following peak picking the original dimensional complexity is reduced, however, MSI data is still of high dimensional nature as one 2D image is typically composed of thousands of high dimensional pixels (spectra) each of which has hundreds of peaks. High dimensional statistics for dimensionality reduction

are commonly used^{27–30}. Dimensionality reduction aims at projecting the high dimensional points into a smaller subspace to enable the capture and visualization of the underlying latent variables. Those latent variables reveal molecular patterns, reflecting clusters of similar spectra that might hold biological relevance^{12,31}. Linear dimensionality reduction methods of principal component analysis (PCA) and non-negative matrix factorization (NNMF) have been widely used for MSI data analysis^{31,32}. A major limitation of these methods is their inherited linearity constraints (e.g. the original data is linearly mapped based on a linear combination of lower-dimensional vectors) that prevent capturing the complex non-linear manifold of spectral structures, impacting accurate identification of latent variables. In contrast, non-linear dimensionality reduction methods such as t-distributed stochastic neighbor embedding (t-SNE) have gained popularity in the last few years for efficient omics data analysis^{12,33,34}. Nevertheless, t-SNE does not provide parametric mapping needed to project new unseen data into the already computed embedding. Despite recent progress on improving both the t-SNE computational and memory scalability^{35,36}, it still needs the full data to be loaded into the RAM which limits its application on data with large sizes such as 3D MSI³⁷. Loading the full data is necessary for the K-nearest neighbor graph creation which is instrumental in establishing spectral pairwise similarities to compute the final t-SNE embedding. Recent preliminary results, on a single pre-processed 2D MSI dataset that underwent peak picking, produced by using a neural network-based method of classical autoencoder have shown promise for efficient non-linear dimensionality reduction of MSI data compared to PCA and NNMF methods but with limited speed of analysis³⁸. We propose to extend this neural network approach by developing more rapid, generic and robust deep learning architecture that can analyze MSI data, without prior peak picking, independent of the nature of the specimen and of the mass spectrometer (ionization source and analyzer).

We introduce msiPL, a deep learning tool for the analysis and peak learning of MSI data, which is based on a fully-connected variational autoencoder neural-network³⁹. This is a probabilistic generative model that learns unsupervised and non-linear parametric mapping between high and low dimensional spaces

and has been efferently applied to other fields such as single cell omics⁴⁰ and medical image segmentation⁴¹. The low dimensional embedding learns a non-linear manifold that captures latent variables that we refer to as encoded features. The encoded features represent molecular patterns that are used to reconstruct the original data. Therefore, minimizing the error between original and reconstructed data would imply capturing accurate encoded features. Batch normalization is incorporated into the proposed neural network architecture to correct for covariate shift and improve both learning stability and convergence⁴². We also propose a method based on analyzing the neural network weight matrix to relate the encoded features to the original m/z features. Both the encoded features and their associated m/z ion features would support clustering and classification tasks required for biomarker identification. The performance of the proposed method was tested using various 2D and 3D MSI data of biological samples collected from different organs and acquired in different laboratories using different MSI platforms. Namely, MSI data from both human and animal tissue specimens were acquired using: MALDI TOF MSI, DESI MSI, and 9.4 Tesla MALDI FT ICR MSI with some of these datasets publicly available³⁷.

Results

Variation inference for manifold learning of MSI data

The computational architecture, shown in Figure 1, is based on a probabilistic generative model to establish efficient unsupervised learning, non-linear dimensionality reduction and stochastic variational inference using neural networks^{39,43}. The variational autoencoder (VAE) aims to jointly optimize two distinct models, namely: probabilistic encoder for variational inference and probabilistic decoder for unsupervised learning. The probabilistic encoder is used as a recognition model $q_{\phi}(z|x)$ to infer approximate estimate of the true but intractable distribution $p_{\theta}(z|x)$ of the k –dimensional latent variable z (i.e. encoded features) underlying the complex high dimensional MSI data $X = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, where N represents the total number of spectra and each spectrum $x^{(i)} \in \mathbb{R}^d$ is of d –dimensions, and $d \gg k$. The probabilistic encoder model assumes sampling of the latent variable z

from a multivariate normal distribution with parameters $\mu_{\phi}(x)$ and $\sigma_{\phi}(x)$; both are functions of x and computed by a fully connected neural network. The probabilistic decoder is used as a generative model $p_{\theta}(x|z)$ for data reconstruction given solely the encoded features. Both the recognition and generative models' parameters, ϕ and θ , are jointly optimized and computed from the neural network parameters. More details are provided in the Methods section and we refer to³⁹ for more information on VAE.

Hyperparameters and implementation details

Original MSI data were analyzed using msiPL, a VAE architecture of a fully connected neural network given in Figure 1.b. The proposed neural network architecture consists of 5 layers, namely: input layer (L_1), three hidden layers (h_1, h_2 , and h_3), and output layer (L_5). The number of artificial neurons for both L_1 and L_2 is equivalent to the number of m/z bins, whereas for the hidden layers h_1, h_2 , and h_3 it is 512, 5, 512 neurons, respectively. The hidden layer h_2 captures the encoded features which represent non-linear dimensionality reduction of original MSI data and compressed in a 5-dimensional space. Batch normalization was applied on each layer's input before any neuron activation to correct for co-variate shift that would degrade the learning process⁴². Following the batch normalization, the rectified linear unit (ReLU) function⁴⁴ was used for neuron activation in all layers except the neurons of the output layer L_5 which were activated using the sigmoid function⁴⁵. The unsupervised learning occurs through minimizing the reconstruction loss between original and reconstructed data mainly by optimizing the VAE cost function which consists of Kullback-Leibler divergence (KL- divergence) and marginal likelihood modeled as a categorical cross-entropy³⁹. The Adam stochastic gradient optimizer with learning rate of 0.001 was used to train the network on minibatches of 128 spectra for 100 epochs⁴⁶. This network was trained on *total-ion-count* (TIC) normalized spectra and implemented using the open source deep learning library of Keras⁴⁷ and running on Tensorflow⁴⁸.

Linking encoded features to observed m/z variables

The encoded features represent the learned non-linear manifold in the lower dimensional space, and enabled capturing spatial patterns of molecules from the original high dimensional space. These patterns

were formed based on a smaller subset of m/z features and it is therefore of interest to identify those underlying m/z features that are expected to hold biological relevance. We propose a backpropagated-based threshold analysis on the weight parameter $W^{(L)}$ of the neural network at layer L , as visually illustrated in Figure 1.d and demonstrated in equation (4). Since the original MSI data was analyzed without prior pre-processing for peak picking, the identified m/z features can then be assigned to a peak. As such, a peak is identified on the average spectrum as the nearest local maximum of a given m/z feature. More details are given in the Methods section.

Analysis of FT ICR MSI data from human prostate cancer tissue specimen

Ultra-high spectral resolution 2D FT ICR MSI data from a human prostate cancer specimen was computationally analyzed using msiPL. The original MSI data encompassed 12,716 pixels (spectra) and each pixel is a high dimensional data point that contained 730,403 m/z values (dimensions) for the mass range m/z 250-1000. The data were exported in the standardized format imzML⁴⁹ and converted to HDF5 format⁵⁰ (using the python package “h5py”) for variational autoencoder analysis. The original dataset was highly sparse and with enormous dimensionality and comprised 730,403 m/z values causing computational and memory challenges to optimize the 748 million parameters of the proposed neural network. To reduce the sparsity and the large number of m/z values and avoid allocating unnecessary memory, local maxima were identified on the average spectrum using the python function “argrelextrema” in which a local maximum is defined as a datapoint of intensity higher than its two neighbors. This significantly reduces the spectral dimensionality from approximately 730,403 to 61,343 m/z values, and this has significantly impacted the spectral sparsity but not the spectral representation (See Supplementary Figure1). The neural network performs unsupervised learning in an iterative manner to minimize the reconstruction loss and as shown in Figure 2.a, the optimizer converged after less than 100 iterations with a total running time of about 40 minutes on a PC workstation (Intel Xenon 3.3GHz,

512 GB RAM, 64-bit Windows, 2 GPUs NVIDIA TITAN Xp). The distributions of *TIC* normalized average spectra of both original and reconstructed data are given in Figure 2.b and their overlay reflects high estimation quality. The encoded features, shown in Figure 2.c, serve as a non-linear embedding that enables visualization and reveals molecular patterns embedded in a compact representation of the original high dimensional data solely in the latent space of 5-dimensions. The generative model estimated the observed spectral data given only these 5-dimensional encoded features, with an overall reconstruction loss of mean squared error (MSE) of 2.42×10^{-5} between *TIC* normalized spectra of original and predicted data. To visually communicate the reconstruction quality of the MSI data, spatial distributions of a few selected *m/z* ions of both original and reconstructed data are presented in Figure 2.d. The encoded features were then linked to the original *m/z* variables, using equation (4) with setting parameter β to 2.5. A reduced list of 244 *m/z* values revealed the main determinants of molecular patterns captured in the latent space (Supplementary File PeakList_Result.xlsx).

Identification of molecular patterns associated with tumor regions in human prostate tissue

The histopathological annotation of the prostate tumor regions revealed a Gleason score (GS) of (3+4)=7 (with cribriform cell morphology) (Figure 2.e)⁵¹. The understanding of molecular patterns underlying the annotated histopathological tumor region could contribute to the development of molecular diagnostics. The encoded features were clustered by the Gaussian-mixture model (GMM) with *k*-clusters (*k*=6) (Figure 2.f) where the light-blue structure (cluster#1) represents a molecular-based tumor pattern with concordance to the histologically-annotated tumor regions. This molecular-based tumor cluster was segmented (Figure 2.g) and correlated with the reduced MSI data of 244 *m/z* values. For example, the ion feature *m/z* 739.4664 ± 0.001 with a Pearson correlation coefficient of 0.7 was tentatively assigned to $C_{39}H_{73}O_8P$ by searching the Human Metabolome Database (HMDB)⁵² based on the accurate mass and with a tolerance window of 1.44 ppm, *m/z* 985.5567 ± 0.001 with a correlation coefficient of 0.65 was tentatively identified as PIP(P-42:6) with an error of -0.14 ppm, and *m/z* 738.4548 ± 0.001 with a correlation coefficient of 0.64 was tentatively identified as PI-Cer(t30:2) with an error of 0.53 ppm. A list

of the top determinant m/z values for this tumor cluster with tentative molecular assignments are presented in Supplementary Table S1.

Tumor-specific metabolic signatures identified in a PDX mouse brain model of glioblastoma

Four consecutive tissue sections of 12 μm thickness were sampled from a patient-derived xenograft (PDX) mouse brain model of glioblastoma (GBM 12) and analyzed by MALDI FT ICR MSI. The original MSI data was highly sparse and constituted of 3,570 spectra each of 661,402 m/z bins, which was reduced to 21,241 m/z values as presented above. The unsupervised learning process reached stable convergence within less than 100 iterations with a computational time of about 3.6 minutes (Figure 3.a). The original data was reconstructed with an overall mean squared error of 4.5×10^{-4} (Figures 3.b and 3.e). The five-dimensional encoded features shown in Figure 3.c capture molecular structures located at a non-linear manifold in the original high dimensional space. These encoded features were clustered using GMM ($k=8$) and the clustered image (Figure 3.d) reveals molecularly distinct tissue regions such as heterogeneous tumor regions (cluster#2 and cluster#8) and a tumor rim (cluster#4). Figure 3.e shows spatial distribution of some m/z values that were determinant of some molecular clusters. The EGFR inhibitor erlotinib (m/z 394.1757 ± 0.001) and m/z 529.9846 ± 0.001 (tentatively identified as ATP/dGTP with an error of 0.69 ppm) show colocalization with tumor cluster#2, whereas m/z 558.2953 ± 0.001 (tentatively identified as lysoPC(18:2) with an error of 0.62 ppm) is found as part of a second tumor cluster#8. The tumor rim cluster#4 was defined in part by m/z 438.2978 ± 0.001 (tentatively assigned to palmitoylcarnitine with an error of 0.27 ppm). There is a noticeable inverse relationship between the intensity distributions of ATP/dGTP and palmitoylcarnitine within the tumor region. The increased distribution of the palmitoylcarnitine at the interface between normal and tumor tissues was discussed in more details by Randall et al⁵³. Analysis of the neural network weight variable using equation (4) and setting parameter β to 2.5 enabled identification of 186 m/z values from the original 21,241 m/z values. Interestingly, m/z 394.175 ± 0.001 which corresponds to the EGFR inhibitor erlotinib was identified despite a mean peak intensity of less than 1% the value of the largest peak from the mean spectrum. For

each of the identified tumor clusters, a list of the top determinant m/z values with tentative molecular assignments are presented in Supplementary Tables S2-S4.

Rapid analysis of 3D MSI data with ultra-high spectral resolution from a PDX mouse brain model of glioblastoma

The trained VAE model presented in Figure 3 was used to analyze an unseen 3D FT ICR MSI dataset acquired from 3 consecutive tissue sections (12 μm thickness) with a separation distance of 160 μm of a PDX mouse brain model (GBM12). The dataset was constituted of 11,263 spectra each of 661,402 m/z bins, which was reduced to 21,241 m/z values as presented above and was analyzed using the trained VAE model with a total running time of 6 seconds for the probabilistic encoder to capture latent variables (encoded features) and about 8 seconds for the generative model to reconstruct the original spectral data. As shown in Figure 4.a, the original spectral data was reconstructed with an overall mean squared error of less than 4.12×10^{-4} . The captured encoded features act as a non-linear embedding of the high dimensional data and reveal distinct molecular patterns in the lower dimensional space of 5 dimensions (Figure 4.b). Unsupervised clustering of these encoded features using GMM (k=11) (Figure 4.c) revealed biologically-relevant clusters of two heterogenous regions in the tumor core (clusters# 4 and 11) and a tumor rim (cluster#8). Analysis of the neural network weight variable, using equation (4) with setting parameter β to 2.5, enabled identification of 198 m/z values from the original 21,241 m/z values. The reduced peak list was correlated with the identified biologically-relevant clusters and ion features at m/z 438.2978 ± 0.001 , 558.2953 ± 0.001 and 529.9846 ± 0.001 which were found to be distributed in the tumor rim region and each of the two heterogenous regions within the tumor core, respectively. Tentative molecular assignment of these m/z values is presented in Supplementary Tables S2-S4.

Scalability on large 3D MALDI MSI dataset with more than one million spectra

A 79 gigabytes volumetric MSI sample of a mouse kidney that encompasses 73 consecutive tissue sections, each with a thickness of 3.5 μm , was acquired by MALDI MSI in the mass range m/z 2,000-20,000 and yielded a 3D spectral image that encompasses 1,362,830 spectra each of 7671 dimensions (i.e. m/z bins)³⁷. Of note, this public dataset is available without prior peak picking but underwent some

conventional spectral pre-processing such as Gaussian spectral smoothing and baseline subtraction³⁷. The computational model was first trained on 18,536 spectra from a 2D MALDI MSI acquisition (section #1), and then tested on the 3D MSI data of the withheld 72 tissue sections (1,342,294 spectra). The unsupervised learning process of the training model reached convergence after less than 100 iterations with a total running time of approximately 8.6 minutes (Figure 5.a) and provided a reconstruction of the original TIC normalized spectra with an overall mean squared error of 5.5×10^{-3} (Figure 5.b-d). The five-dimensional encoded features revealed structural information with distinct molecular patterns (Figure 5.e) that were clustered using GMM (k=7) as shown in Figure 5.f. The molecularly clustered image revealed anatomical structures of the kidney such as the renal cortex (clusters# 2&6), renal medulla (cluster# 1), renal pelvis (cluster# 7), renal artery & vein (cluster# 4), and ureter (cluster# 5) in accordance with histology and previous studies^{15,54}. Figure 5.h shows the spatial distribution of selected m/z values found as part of these molecular patterns. The highly weighted spectral features that constitute the molecular patterns captured by encoded features were identified using equation (4) with setting parameter β to 2.5 and highlighted in the mean spectrum with m/z bins highlighted in red and resulting peaks highlighted in green (Figure 5.g). The complete peak list with 124 m/z values is provided in the Supplementary File PeakList_Result.xlsx.

The trained model was then applied on the 3D MSI data of the withheld 72 tissue sections with spectra corresponding to each 2D MSI tissue section independently loaded into the RAM and analyzed by the trained probabilistic model with an overall running time of about 10 seconds. The approach was 4 times faster with 20 times less memory requirement compared to the state-of-the-art computational development used to analyze the same dataset¹⁵. Figure 6 shows results from selected test samples at different volumetric depth within the tissue volume, and results from the analysis of the complete test data are presented in Supplementary Figure S2. The low-dimensional latent space captured molecular patterns from the high dimensional space (Figure 6.b) and the encoded features of the entire 3D MSI data were clustered using the GMM (k=8) revealing molecular patterns that highlight anatomical structures of the

kidney (Figure 6.c and Supplementary Figure S2). The original TIC normalized MSI data was reconstructed with an overall mean squared error of 3.11×10^{-3} .

Identification of tumor and connective tissue types in 3D DESI MSI of colorectal adenocarcinoma dataset

DESI MSI data was acquired from 26 consecutive (acquired at every 100 μm) 10 μm thickness tissue sections to reconstruct a 3D MSI volume from a human colorectal adenocarcinoma specimen in the mass range m/z 200-1050³⁷. The 148,044 spectra each of 8,073 dimensions constituting the 3D DESI MSI data volume was analyzed using msiPL. Data from a single tissue section was arbitrary selected (section#1) to train the model with 5,694 spectra, and the model converged after tens of iterations (Supplementary Figure S3.a) with a total running time of approximately 3.2 minutes. Based on the learned latent variables, the original spectral data was reconstructed with an overall mean squared error of 2.02×10^{-4} (Supplementary Figure S3) and the encoded features were then linked to the original m/z values using equation (4) with parameter β set to 2.5 resulting in a reduced peak list with 24 m/z values. The encoded features were then clustered by GMM ($k=5$), revealing a tumor region (red cluster#5) and connective tissue (light-blue cluster#2) in agreement with histological evaluation (Supplementary Figure S3). These two clusters were correlated with the reduced peak list revealed ions at m/z 279.2 ± 0.1 and m/z 421.3 ± 0.1 that were found to be elevated in the tumor and connective tissue clusters with Pearson correlation values of 0.773 and 0.574, respectively.

The trained model was then applied to analyze the 3D MSI data volume from the remaining 25 tissue sections with 142,350 spectra. To gain memory advantages, the spectra corresponding to each 2D MSI dataset were independently loaded into the memory and analyzed by the VAE model with a running time of approximately 2 seconds for the probabilistic encoder to predict the encoded features and another 2 seconds for the generative model to reconstruct the data. The 3D MSI data was reconstructed with an overall mean squared error of 1.77×10^{-4} (Supplementary Figures S3 and S5) and the 3D encoded features are presented in Supplementary Figure S4, representing a 5-dimensional non-linear embedding of

the original high dimensional data. Further GMM clustering ($k=5$) of these encoded features revealed two distinct molecular clusters associated with tumor and connective tissues in accordance with previous studies^{15,37} and images of the H&E stained tissue sections (Supplementary Figure S5).

Identification of α -defensins in human oral squamous cell carcinoma by 3D MALDI MSI data analysis

The deep learning data analysis strategy was further applied to a 3D MALDI MSI data volume acquired from 58 consecutive tissue sections (10 μm thickness) of a human oral squamous cell carcinoma (OSCC) specimen. Data was acquired in the mass range m/z 2,000-20,000 and resulted in a total of 825,558 of pre-processed spectra (Gaussian spectral smoothing and baseline subtraction) each of 7,665 dimensions³⁷. The computational model was first trained on a single 2D MALDI MSI dataset (12,875 spectra) of an arbitrary chosen tissue section (section #1), and then tested on MSI data from the withheld 57 tissue sections (815,683 spectra). The training phase reached a stable convergence after less than 100 iterations (Supplementary Figure 6S.a), with a total running time of approximately 6.1 minutes. The original spectral data was reconstructed with an overall mean squared error of 3.7×10^{-3} (Supplementary Figure S6). The captured encoded features were clustered by GMM ($K=7$) and revealed clusters (Supplementary Figure S6) with underlying distribution of peptide ions at m/z 3,445 and 3,488 found to be elevated (Pearson correlation coefficient of 0.713) with cluster#3. These peptides have previously been proposed to be putative defensins HNP1-3 produced by neutrophils which can induce tumor angiogenesis^{37,55}. Analysis of the neural network weight variable using equation (4) with parameter β set to 2.5 enabled the identification of 177 m/z bins (red markers in Figure S6.f) that contributed in forming the molecular patterns captured by the encoded features, leading to a reduced peak list of 44 m/z values (green markers in Figure S6.f; Supplementary File PeakList_Result.xlsx).

Spectra from each of the 2D MSI datasets from the remaining 57 tissue sections were independently loaded into the RAM and analyzed by the VAE with an average running time of approximately 8 seconds. The proposed approach was 4 times faster and used 30 times less memory compared to reported analysis

of the same dataset¹⁵. Results from a subset 2D MSI datasets taken at different volumetric depth within the tissue volume are presented in Figure S7 with results from the complete data volume (Supplementary Figure S8). The original TIC normalized MSI data was reconstructed with an overall mean squared error of 3.03×10^{-3} (Figure S7.a) and the encoded features captured molecular patterns that were clustered by GMM (k=7). Cluster#2 (Figure S7.e) highlighted a molecular structure in which lipid ions at m/z 3,373, 3,445 and 3,488 were co-localized and elevated (Figure S7.f) as described above for the 2D MSI data analysis and in agreement with a previous study¹⁵. A reduced list of 57 m/z values strongly correlated to the identified patterns (Figure S7) and results from the complete 57 tissue section volume are shown in Supplementary Figure S8.

Discussion

We have proposed msiPL, a generic neural network-based method for the analysis and peak learning of MSI data from different types of mass spectrometer and tissue type. The neural network showed stability, robustness, and provided time and computation efficient analysis of various types of MSI data. The regularization provided by both the KL-divergence and batch normalization resulted in the stability of the neural network and minimized its reliance on optimization of the hyper-parameter initialization⁴². The KL-divergence is embedded in the VAE loss function³⁹, and the batch normalization was incorporated into the proposed network to normalize each layer's input right before any neuron activation to correct for the co-variate shift⁴². Of note, the number of latent space dimensions is a tunable parameter which was here empirically set to 5 dimensions. Generally, a main consideration for choosing the number of latent space dimensions is to minimize the reconstruction error of the generative model through optimizing the VAE cost function shown in Equation (3). For example, supplementary Table S5 shows the effect of different model parameters on the model complexity and the quality of data reconstruction. Those 5-dimensional encoded features are a compressed representation capturing molecular patterns from original high dimensional data, with each encoded feature combining a set of spectral features rather than a single spectral feature. The link between the encoded features and the original m/z variables was established

through backpropagated threshold-based analysis on the neural network weight matrix. The same threshold weight was given to each encoded feature as indicated by the parameter β as defined in equation (4). Since some encoded features could capture more significant molecular patterns than other encoded features, the β parameter can therefore be optimized for each of those encoded features, which would require further investigation.

The robustness of the network has been shown from 5 different MSI datasets including 2D MALDI FT ICR MSI data from human prostate cancer (exported in an imzML file of 9.4 gigabytes of spectral data), 3D DESI MSI data from human colorectal adenocarcinoma (8.9 gigabytes of spectral data), 3D MALDI MSI data from human oral squamous cell carcinoma (47.3 gigabytes of spectral data), and 3D MALDI MSI data from a mouse kidney (78.9 gigabytes of spectral data), and 3D MALDI FT ICR MSI data from PDX mouse brain model of glioblastoma (3.3 gigabytes of spectral data).

Of note, prior to utilizing the msiPL processing the MSI data was normalized based on the common method of total-ion-count (TIC). However, there are different normalization strategies for MSI data, for example such as those covered in^{56,57}. It is up to the end users to choose their best MSI normalization strategy, however the msiPL expects the normalized input data to be bounded within the interval [0,1]. That is because the output layer of the msiPL is based on a sigmoid activation function which yields values within the range [0,1]. The consistency in the dynamic range at both the input and output layers is crucial to properly optimize the VAE loss function shown in Equation (3) and eventually minimizing the reconstruction error.

The proposed computational model is set to be trained on the spectral level without considering spatial information with each pixel providing a spectrum as part of a training sample. Individual spectra should be aligned to a common reference, independently of their spatial organization with a 2D or 3D image. The neural network also features mini-batch implementation for parallel processing and for improving memory complexity as it only needs to load a small subset of spectral data into the memory allowing to efficiently process large and complex data such as FT ICR MSI or 3D MSI. This provides significant

improvement over existing computational approaches that require the full data loaded into memory to calculate pairwise similarities between spectra^{15,54}. Therefore, the msiPL could be trained on a typical workstation with a RAM capacity of 32-64 GB. However, and due to the complexity of the neural network that involves optimizing millions of parameters, training the msiPL on the GPU would significantly improve the running time compared to utilizing only the CPU as demonstrated in Supplementary Table S6. The proposed development provides unprecedented computational boost and memory advantages that could support a wide variety of MSI applications.

The proposed strategy for 3D MSI data analysis is based on a training/testing framework. The training phase runs on a subset of spectra from MSI data from either a single or several tissue sections. The training constitutes of unsupervised learning of the underlying spectral manifold that can subsequently be clustered to identify molecularly distinct regions. The testing phase is applied on the withheld data that do not necessarily need to be fully loaded into the memory at once since the computational model can analyze a subset and capture its underlying molecular patterns in few seconds. This new data analysis approach enabled the overall analysis of 3D MSI data with 20 times less memory and was 4 times faster than a recently reported application of Hierarchical Stochastic Neighbor Embedding (HSNE)¹⁵. Once identified, the molecular patterns from distinct tissue sections in the volumetric specimen can be reconstructed to form a 3D volume representing the specimen. As an interesting observation, the neural network was able to identify, on test data, spectral patterns of different morphological appearance but similar molecular phenotypes to those were encompassed by the training data (e.g. see artery and vein structures in the encoded features of Figures 5 and 6). It is natural to observe such variations in the morphological phenotype of the tissue anatomy at different locations within the volumetric specimen, but their spectral phenotype should be similar. Therefore, if a certain spectral phenotype was not presented in the training data it would probably not be detected during testing. It should be taken into consideration to balance the spectral phenotypes held in both training and testing datasets, for instance through cross-

validation⁵⁸ as shown in Figure 7 and Supplementary Figures S9-S10, especially in 3D MSI data that may expose molecular heterogeneity within the volumetric specimen.

The stability performance of the computational model was investigated using 5-fold cross-validation on the 3D MALDI MSI dataset of mouse kidney as shown in Figure 7 and Supplementary Figure S9. The full 3D MSI dataset was randomly shuffled and split into a 20% training set (spanning spectra from 14 tissue sections) and 80% testing set (spanning spectra from 58 tissue sections), see Figure 7.a. For each of the five cross-validation iterations, the msiPL method was applied on the training set to optimize the artificial neural network and the optimized model was then applied on the unseen test set. These analyses revealed comparable performance on both the training and testing sets, as shown in Supplementary Figure S9. The best cross-validation model was able to reconstruct the original MSI dataset with minimal mean squared error of 6.18×10^{-3} and 5.17×10^{-3} for the associated training and testing sets, respectively as given in Figures 7.b-c. The stability of peak learning was investigated across the different cross-validation models. Figure 7.d. shows the frequency distribution of all m/z peaks identified in the 5-fold cross-validation analyses and the peak count for each frequency is given in Figure 7.e. Overall, 69.6% of the m/z peaks were found stable as they were consistently identified in 80% of the cross-validation analyses. Figure 7.f shows 3D spatial distribution of some of those stable m/z peaks and each of which reveals high co-localization to a specific structure that reconciles the kidney's anatomy—reflecting quality of the learned peaks. The full peak list from the different cross-validation models is provided in the Supplementary File PeaksList_CrossValidation.xlsx. Similarly, another cross-validation analysis was applied on the 3D DESI MSI dataset of colorectal carcinoma and it revealed stable peaks such as m/z 279.2 and m/z 766.5 which are localized and elevated in the tumor and normal tissue clusters, respectively as shown in the Supplementary Figure S10.

While the neuron activation function of rectified linear unit (ReLU) was used in all layers of the proposed deep learning network, it was not used in the output decoder layer in which the sigmoid function was

employed instead. The main reason is that the sigmoid function at the output layer is more adequate for the underlying VAE loss function. The VAE loss function, as illustrated in Equation (3) consists of summation of two terms of KL-divergence and the marginal likelihood estimate that was modeled using categorical cross-entropy. The KL-divergence acts as a regularizer for the probabilistic encoder to measure the similarity between the approximate estimate distribution $q_{\phi}(z|x)$ and the true but intractable distribution $p_{\theta}(z|x)$. The categorical cross-entropy was used to measure the reconstruction loss between two probability distributions of original input and the estimated marginal likelihood represented by the decoded output. Unlike the ReLU activation function, the sigmoid function output can be interpreted as a probability since its range varies between 0 and 1⁵⁹.

Analysis of mass spectrometry imaging data with bypassing conventional spectral pre-processing would enrich the biomarker discovery process by increasing identification sensitivity and specificity for molecular ions with biological relevance. For instance, this would minimize the effect of optimization in pre-processing steps required for feature extraction such as peak picking that have an impact on overall biomarker identification²¹. Despite manufacturers currently offering default parameter settings for the peak picking process (e.g. spectral smoothing, signal-to-noise ratio (S/N), peak shape, and peak threshold), these default values are general and lack the sensitivity needed to reliably identify peaks⁶⁰ (e.g. missing the erlotinib drug peak as shown in Supplementary Figure S11). Additionally, the quadratic computational complexity of the analysis slows down the peak picking process with an increase in the number of spectra¹⁶, see Supplementary Table S7. Of note, the local maxima detection applied here only to reduce the original FT ICR MSI data complexity conceptually differs from peak picking. The former simply identifies local maxima between 3 consecutive m/z variables^{16,20}, while the latter seeks to identify relevant peaks induced by molecules which are spectrally characterized by a structure that includes (peak height, width, slope and baseline) and minimize the contribution of noise-to-signal ratio. However, the local maxima approximation applied to the FT ICR MSI data could be avoided in future developments through, for example, investigating an integrated multi-spectral scale neural network architecture. Such

future work would allow to cope with even more complex spectral data such as those acquired by ultrahigh resolution FT ICR MSI instruments⁶¹. Future developments may also consider another level of spectral complexity such that provided by the collision cross section property of the trapped ion mobility spectrometry (TIMS) based technology which holds promise for higher accurate molecular identification^{62,63}. Classical machine learning approaches are of limited capabilities to analyze original mass spectrometry data at full spectral dimensions. Mainly, because those approaches suffer a common issue known as curse-of-dimensionality that deteriorates the clustering/classification accuracy on high dimensional data⁶⁴. In contrast, deep learning-based approaches have shown the ability to avoid the curse-of-dimensionality and to establish self-learning of relevant features that increase classification accuracy⁶⁵. This expansion in data analysis could impact complex applications such multi-class classification required to resolve molecular intratumor heterogeneity⁶⁶.

The GMM-based clustering method on the encoded features has been found computationally fast (i.e. a few seconds) as well as efficient in identifying spatial clusters of biological relevance such as distinguishing normal and tumor clusters. The number of K-clusters is a user tunable parameter that could be set either manually or automatically. In our strategy, an iterative approach was followed in which the GMM clustering process was applied on a different number of K-clusters within an expected range. The best model was then manually chosen as such it exhibited a clustering balance that avoids under-/over-estimation. Contrary, an information theory-based optimization process can be utilized for automated model selection, see Supplementary Figure S12. That is the Bayesian information criterion (BIC) algorithm in which the best model is theoretically expected to achieve the minimal BIC score⁶⁷. Since the distribution of the BIC scores is gradually decreasing within the searchable range, as shown in Supplementary Figure S12, the Kneedle algorithm⁶⁸ was then applied on the BIC scores to detect the critical point of maximum curvature (also known as elbow/knee point) at which the best model was selected. Compared to the manual approach, there was a noticeable under-estimation for the PDX GBM

dataset in which intra-tumor heterogeneous clusters were missed whereas an over-estimation was observed for the colorectal cancer dataset.

Taken together, the deep learning data analysis strategy presented here provides the ability to learn the underlying non-linear manifold required to accurately and effectively identify and visualize molecular patterns from original high dimensional data, avoiding conventional time-consuming pre-processing computation. The resulting workflow provides unsurpassed data analysis time of large and complex new data, while delivering an enriched biomarker discovery process through unsupervised identification of complex molecular patterns with identification of their determinant m/z values.

Materials and Methods:

Experimental Datasets:

MSI datasets from five different tissue types were analyzed and their description is given in the Supplementary Materials and Methods. Briefly, three of these MSI datasets are publicly available³⁷, which include: 3D DESI MSI dataset of human colorectal adenocarcinoma, 3D MALDI MSI dataset of human oral squamous cell carcinoma, and 3D MALDI MSI data of mouse kidney. The other two MSI datasets were collected and acquired at our institution, which include: 2D MALDI FT-ICR MSI dataset of human prostate cancer⁶⁹, and 3D MALDI FT-ICR MSI dataset of PDX mouse brain model of glioblastoma³. The MSI datasets, without prior peak picking, were exported in the standardized format imzML⁴⁹ and converted to HDF5 format⁵⁰ for variational autoencoder analysis.

Variational Autoencoder:

Mathematically, spectra of MSI data can be represented as a set of high-dimensional vectors $X = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$, where N represents the total number of spectra (or pixels) and each spectrum $x^{(i)} \in \mathbb{R}^d$ is of d -dimensions. We assume these d -dimensional i.i.d. vectors were generated by a random process that involves an unobserved lower-dimensional latent variable $z \in \mathbb{R}^k$, where $d \gg k$. The latent variable is sampled from a prior distribution $p_{\theta^*}(z)$ and the datapoint $x^{(i)}$ is generated from the conditional distribution $p_{\theta^*}(x|z)$. The true posterior distribution $p_{\theta^*}(z|x)$ would provide a compressed representation (we referred to it as encoded features) of the observed high-dimensional data. The high dimensional nature of $x^{(i)}$ makes the posterior distribution computationally intractable. The variational inference aims therefore at introducing a recognition model $q_{\phi}(z|x)$ that approximates the true intractable posterior $p_{\theta}(z|x)$. The recognition model is assumed to be sampled from a normal distribution parameterized by $\mu_{\phi}(x)$ and $\sigma_{\phi}(x)$ as shown in equation (1). The inference of the latent variable z would enable the generative model of marginal likelihood estimator $p_{\theta}(x|z)$ to reconstruct datapoint $x^{(i)}$.

$$q_{\phi}(z|x) \sim \mathcal{N}(\mu_{\phi}(x), \sigma_{\phi}(x)\mathbf{I}). \quad (1)$$

The variational parameter ϕ needs to be estimated as such it makes $q_{\phi}(z|x)$ as close as possible to the true posterior $p_{\theta}(z|x)$. The Kullback-Leibler (KL) divergence, given in equation (2), can assess the closeness between these two distributions. Because the KL-divergence is always non-negative, the term $(E_{q_{\phi}(z|x)}[\log q_{\phi}(z|x)] - E_{q_{\phi}(z|x)}[p_{\theta}(z, x)] = \mathcal{L})$ represents the variational lower bound on the marginal distribution, where $\mathcal{L} \leq \log p_{\theta}(x)$. The optimum estimate of the parameters ϕ and θ would maximize the variational lower bound that can be re-written as shown in equation (3).

$$\begin{aligned} KL(q_{\phi}(z|x) \parallel p_{\theta}(z|x)) &= \int q_{\phi}(z|x) \log \frac{q_{\phi}(z|x)}{p_{\theta}(z|x)} dz \\ &= E_{q_{\phi}(z|x)}[\log q_{\phi}(z|x)] - E_{q_{\phi}(z|x)}[p_{\theta}(z, x)] + \log p_{\theta}(x). \quad (2) \end{aligned}$$

In variational autoencoder (VAE), the recognition model represents a probabilistic encoder and the generative model represents a probabilistic decoder. The recognition and generative model parameters ϕ

and θ are computed from the neural network parameters and jointly optimized by maximizing the cost function of the variational lower bound $\mathcal{L}(\phi, \theta; x^{(i)})$, given in equation (3), which eventually would minimize the overall VAE loss. The first term of this cost function acts as a regularizer for the encoder and it measures the closeness between the approximated posterior and the prior. The second term represents the expected value of the reconstructed error which we modeled as a cross-entropy.

$$\mathcal{L}(\phi, \theta; x^{(i)}) = -KL(q_{\phi}(z|x^{(i)}) \parallel p_{\theta}(z)) + E_{q_{\phi}(z|x^{(i)})}[\log p_{\theta}(x^{(i)}|z)]. \quad (3)$$

Kingma *et al.* introduced a reparameterization trick to make $\mathcal{L}(\phi, \theta; x^{(i)})$ differentiable with respect to ϕ and θ . That is, incorporating first an auxiliary variable $\varepsilon \sim \mathcal{N}(0,1)$ with an input datapoint $x^{(i)}$ to form a continuous function $g_{\phi}(\varepsilon, x^{(i)})$ that can then be used to sample the latent variable from the approximated posterior $z^{(i)} \sim q_{\phi}(z|x^{(i)})$ as such $z^{(i)} = g_{\phi}(\varepsilon, x^{(i)}) = \mu_{\phi}(x^{(i)}) + \text{diag}(\sigma_{\phi}(x^{(i)})) \cdot \varepsilon$. For more information on variational autoencoder, we refer to³⁹.

Identification of Informative m/z Peaks:

The inferred multi-variate latent variable z represents encoded features that capture molecular patterns in the original MSI data. It is therefore of interest to identify m/z features underlying those learnt patterns. We propose a threshold analysis on the weight parameter $W^{(L)}$ of the neural network at layer L , as depicted in Figure 1.d. Briefly, for each encoded feature represented by the i^{th} neuron at layer h_2 we first identify the j^{th} neuron at the previous hidden layer h_1 with maximum scalar weight value $w_{ij}^{(2)} \in W^{(2)}$, see red line in Figure 1.d. Then, a threshold T was computed, equation (4), using the weight vector $w_{dj}^{(1)} \in W^{(1)}$ which is a one-dimensional vector holds the weights between all d neurons of the input layer L_1 and the identified j^{th} neuron at h_1 . Eventually, a set of p neurons at the input layer L_1 whose weights are larger than T (*i.e.* $w_{dj}^{(1)} \geq T$) were retrieved and each of which represents an m/z variable, for schematic illustration see blue highlighted lines in Figure 1.d. Since the original MSI data were analyzed

without prior pre-processing for peak picking, the retrieved observed variables represent m/z bins that need then to be assigned to its associated m/z peak. As such, an m/z peak has been identified on the average spectrum as the nearest local maximum to a given m/z bin.

$$T = \text{mean}(w_{dj}^{(1)}) + \beta * \text{std}(w_{dj}^{(1)}); \text{ where } \beta \in [1, 2.5]. \quad (4)$$

Data Clustering using Gaussian Mixture Model (GMM):

The encoded features reduce the original dimensional complexity and enable application of a simple clustering algorithm such as Gaussian mixture model (GMM)⁷⁰. The encoded features are expected to learn a non-linear manifold to allow capturing and visualizing molecular patterns from original high dimensional data. The clustering algorithm would cluster those patterns to form one image in which distinct clusters represent molecularly distinct regions. The number of clusters (k) for the GMM clustering algorithm is a user tunable parameter. A cluster of interest is then correlated with the MSI data of reduced peak list to identify co-localized m/z peaks with the highest Pearson correlation coefficient⁷¹.

Data availability:

These three MSI datasets (3D MALDI TOF MSI data of human oral squamous cell carcinoma, 3D MALDI TOF MSI data of mouse kidney and 3D DESI Orbitrap MSI data of human colorectal adenocarcinoma) were previously published and publicly available by Oetjen *et al.*³⁷. The 2D MALDI FT ICR MSI data of a human prostate tissue, and the 3D MALDI FT ICR MSI data of a PDX mouse brain model of glioblastoma will be made publicly available at the time of publication.

Code availability:

Source code is publicly available on GitHub: <https://github.com/wabdelmoula/msiPL.git>

References:

1. Aichler, M. & Walch, A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab Invest* **95**, 422–431 (2015).
2. Norris, J. L. & Caprioli, R. M. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chem. Rev.* **113**, 2309–2342 (2013).
3. Randall, E. C. *et al.* Integrated mapping of pharmacokinetics and pharmacodynamics in a patient-derived xenograft model of glioblastoma. *Nat. Commun.* **9**, (2018).
4. Santagata, S. *et al.* Intraoperative mass spectrometry mapping of an onco-metabolite to guide brain tumor surgery. *Proc. Natl. Acad. Sci.* (2014). doi:10.1073/pnas.1404724111
5. Schulz, S., Becker, M., Groseclose, M. R., Schadt, S. & Hopf, C. Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development. *Curr. Opin. Biotechnol.* **55**, 51–59 (2019).
6. Basu, S. S. *et al.* Rapid MALDI mass spectrometry imaging for surgical pathology. *npj Precis. Oncol.* **3**, 17 (2019).
7. McDonnell, L. A. & Heeren, R. M. Imaging mass spectrometry. *Mass Spectrom Rev* **26**, 606–643 (2007).
8. Calligaris, D. *et al.* Application of desorption electrospray ionization mass spectrometry imaging in breast cancer margin analysis. *Proc. Natl. Acad. Sci.* **111**, 15184–9 (2014).
9. Pandey, A. *et al.* Analysis of receptor signaling pathways by mass spectrometry: identification of vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 179–184 (2000).

10. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
11. Bauer, J. A. *et al.* Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. *Clin. Cancer Res.* **16**, 681–690 (2010).
12. Abdelmoula, W. M. *et al.* 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).
13. Calligaris, D. *et al.* Mass spectrometry imaging as a tool for surgical decision-making. *J. Mass Spectrom.* **48**, 1178–1187 (2013).
14. Eberlin, L. S. *et al.* Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1611–6 (2013).
15. Abdelmoula, W. M. *et al.* Interactive Visual Exploration of 3D Mass Spectrometry Imaging Data Using Hierarchical Stochastic Neighbor Embedding Reveals Spatiomolecular Structures at Full Data Resolution. *J. Proteome Res.* (2018). doi:10.1021/acs.jproteome.7b00725
16. Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics* (2012). doi:10.1186/1471-2105-13-S16-S11
17. Veselkov, K. A. *et al.* Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer. *Proc. Natl. Acad. Sci.* **111**, 1216–1221 (2014).
18. Behrmann, J. *et al.* Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics* (2018). doi:10.1093/bioinformatics/btx724
19. Yang, C., He, Z. & Yu, W. Comparison of public peak detection algorithms for MALDI mass

- spectrometry data analysis. *BMC Bioinformatics* **10**, (2009).
20. Kilgour, D. P. A. *et al.* Autopicker - a Robust and Reliable Peak Detection Algorithm for Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **28**, 253–262 (2017).
21. McDonnell, L. A., van Remoortere, A., de Velde, N., van Zeijl, R. J. & Deelder, A. M. Imaging mass spectrometry data reduction: automated feature identification and extraction. *J Am Soc Mass Spectrom* **21**, 1969–1978 (2010).
22. Wijetunge, C. D., Saeed, I., Boughton, B. A., Roessner, U. & Halgamuge, S. K. A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model. *BMC Genomics* **16**, (2015).
23. Adam, B. L. *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614 (2002).
24. House, L. L., Clyde, M. A. & Wolpert, R. L. Bayesian nonparametric models for peak identification in maldi-tof mass spectroscopy. *Ann. Appl. Stat.* **5**, 1488–1511 (2011).
25. Bauer, C., Cramer, R. & Schuchhardt, J. Evaluation of peak-picking algorithms for protein mass spectrometry. *Data Min. Proteomics. Humana Press* **696**, 341–352 (2011).
26. Coombes, K. R. *et al.* Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* **5**, 4107–4117 (2005).
27. Van Der Maaten, L. J. P., Postma, E. O. & Van Den Herik, H. J. Dimensionality Reduction: A Comparative Review. *J. Mach. Learn. Res.* **10**, 1–41 (2009).
28. Jones, E. A., Deininger, S. O., Hogendoorn, P. C., Deelder, A. M. & McDonnell, L. A. Imaging mass spectrometry statistical analysis. *J Proteomics* **75**, 4962–4989 (2012).

29. Abdelmoula, W. M. *et al.* Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding. *Anal Chem* **86**, 9204–9211 (2014).
30. Verbeeck, N., Caprioli, R. M. & Van de Plas, R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews* (2019).
doi:10.1002/mas.21602
31. Jones, E. A. *et al.* Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PLoS One* **6**, (2011).
32. Race, A. M., Steven, R. T., Palmer, A. D., Styles, I. B. & Bunch, J. Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets. *Anal. Chem.* **85**, 3071–3078 (2013).
33. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
34. Shekhar, K., Brodin, P., Davis, M. M. & Chakraborty, A. K. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc. Natl. Acad. Sci. U. S. A.* **111**, 202–207 (2014).
35. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
36. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical Stochastic Neighbor Embedding. *Comput. Graph. Forum* **35**, 21–30 (2016).
37. Oetjen, J. *et al.* Benchmark datasets for 3D MALDI- and DESI-imaging mass spectrometry. *Gigascience* **4**, 20 (2015).
38. Thomas, S. A., Race, A. M., Steven, R. T., Gilmore, I. S. & Bunch, J. Dimensionality Reduction of Mass Spectrometry Imaging Data using Autoencoders. *IEEE Symp. Ser. Comput. Intell.* 1–7

- (2016). doi:10.1109/SSCI.2016.7849863
39. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv Prepr. arXiv1312.6114* (2013).
 40. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).
 41. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
 42. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv Prepr. arXiv1502.03167* (2015).
 43. Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
 44. Nair, V. & Hinton, G. Rectified Linear Units Improve Restricted Boltzmann Machines. in *Proceedings of the 27th International Conference on Machine Learning* (2010).
doi:10.1.1.165.6419
 45. Nielsen, M. *Neural Networks and Deep Learning*. (USA: Determination press, 2015).
 46. Kingma, D. P. & Lei Ba, J. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. *arXiv* (2015).
 47. Chollet, F. Keras (2015). URL <http://keras.io> (2017).
 48. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. *12th {USENIX} Symp. Oper. Syst. Des. Implement. ({OSDI})* **16**, 265–283 (2016).
 49. Race, A. M., Styles, I. B. & Bunch, J. Inclusive sharing of mass spectrometry imaging data requires a converter for all. *J. Proteomics* **75**, 5111–5112 (2012).

50. Folk, M., Heber, G., Koziol, Q., Pourmal, E. & Robinson, D. An overview of the HDF5 technology suite and its applications. in *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases. ACM.* 36–47 (2011). doi:10.1145/1966895.1966900
51. Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urol.* **111**, 58–64 (1974).
52. Wishart, D. S. *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkl923
53. Randall, E. C. *et al.* Localized metabolomic gradients in patient-derived xenograft models of glioblastoma. *Cancer Res.* **80**, 1258–1267 (2020).
54. Trede, D. *et al.* Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: Three-dimensional spatial segmentation of mouse kidney. *Anal. Chem.* **84**, 6079–6087 (2012).
55. Magalhaes, M. A. O., Glogauer, J. E. & Glogauer, M. Neutrophils and oral squamous cell carcinoma: lessons learned and future directions. *J. Leukoc. Biol.* **96**, 695–702 (2014).
56. Deininger, S. O. *et al.* Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Anal Bioanal Chem* **401**, 167–181 (2011).
57. Fonville, J. M. *et al.* Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Anal. Chem.* **84**, 1310–1319 (2012).
58. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Proceedings of the 14th international joint conference on Artificial intelligence - 14(2)* 1137–1145 (1995).
59. Nielsen, M. A. *Neural networks and deep learning.* **25**, (Determination press San Francisco, CA, USA:, 2015).

60. Donnelly, D. P. *et al.* Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **16**, 587–594 (2019).
61. Stopka, S. A. *et al.* Ambient Metabolic Profiling and Imaging of Biological Samples with Ultrahigh Molecular Resolution Using Laser Ablation Electrospray Ionization 21 Tesla FTICR Mass Spectrometry. *Anal. Chem.* **91**, 5028–5035 (2019).
62. Lanucara, F., Holman, S. W., Gray, C. J. & Eyers, C. E. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nature Chemistry* 281–294 (2014). doi:10.1038/nchem.1889
63. Spraggins, J. M. *et al.* High-Performance Molecular Imaging with MALDI Trapped Ion-Mobility Time-of-Flight (timsTOF) Mass Spectrometry. *Anal. Chem.* **91**, 14552–14560 (2019).
64. Donoho, D. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Conf. Math Challenges 21st Century* (2000).
65. Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B. & Liao, Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing* (2017). doi:10.1007/s11633-017-1054-2
66. Balluff, B. *et al.* De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *J. Pathol.* **235**, 3–13 (2015).
67. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).
68. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior. in *Proceedings - International Conference on Distributed Computing Systems* 166–171 (2011). doi:10.1109/ICDCSW.2011.20
69. Randall, E. C. *et al.* Molecular characterization of prostate cancer with associated Gleason score using mass spectrometry imaging. *Mol. Cancer Res.* (2019). doi:10.1158/1541-7786.MCR-18-

1057

70. Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process. A Rev. J.* (2000). doi:10.1006/dspr.1999.0361
71. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* (2012).

Figure 1. The proposed neural network architecture of variational autoencoder for mass spectrometry imaging data analysis and peak learning. The network consists of 5 layers and trains on TIC-normalized spectra without considering their spatial relationships. The neural network is regularized using batch normalization, and informative m/z values were identified using statistical analysis on the neural network weight parameter.

Figure 2. Deep-learning based analysis of an ultra-high spectral resolution of 2D FT-ICR MSI of prostate cancer tissue: **a**. distribution of the optimization convergence with the number of iterations (epochs). **b**. The TIC-normalized distribution of original and predicted average spectra. **c**. Five dimensional encoded features (latent variable z) represent the learned non-linear manifold that enabled visualization and captured molecular patterns of the original high dimensional MSI data of 61343 dimensions. These encoded features are of high quality as it were used to predict the original data with an overall mean squared error of 2.42×10^{-5} . **d**. spatial distribution of a few arbitrary m/z peaks for original and predicted MSI data and it reveals high estimation quality of the original observed data. Comparison between H&E stained histology and clustered molecular patterns reveals molecular-based tumor region: **e**. Histopathological annotation of the tumor regions, **f**. Gaussian mixture model (with $k=6$) was applied to cluster the encoded features, and the tumor-associated pattern is represented by the light-blue structure (cluster#1) that was extracted (**g**) and correlated with the reduced MSI data. The highest Pearson correlation value was with the ion feature at m/z 739.4664 $\pm$ 0.001 and it reveals elevation in the tumor region (**h**).

Figure 3. Analysis of 2D MALDI FT-ICR MSI dataset of PDX mouse brain model of glioblastoma: **a**. Distribution of optimization convergence, and **b**. Overlay of the mean spectrum of both TIC normalized original (green) and reconstructed (red) data with an overall mean squared error of 4.5×10^{-4} . **c**. The clustered image of the encoded features (**d**) using GMM ($k=8$) reveals biologically interesting tissue types, such as: normal tissue (cluster#1), tumor heterogeneity (cluster#2 and cluster#8) and a rim around the tumor (cluster#4). **e**. Spatial distribution of a few biologically interesting m/z ions that were found highly co-localized within the clusters of interest, and there is a close similarity between the reconstructed and measured m/z ions.

Figure 4. Analysis on test dataset of MALDI FT-ICR of three consecutive section from a PDX mouse brain of glioblastoma based on the trained model shown in Figure 3. The analysis was performed independently on each tissue section to reveal: **a**. Overlay of the overall mean spectrum of both TIC normalized original (green) and reconstructed (red) data, **b**. Five-dimensional encoded features capture molecular structures located on a non-linear manifold in the original high-dimensional data, **c**. GMM based clustering ($k=11$) of the encoded features reveal biologically interesting clusters such as: tumor heterogeneous regions (cluster#4 and cluster#11) and a rim around the tumor (cluster#8), and **d**. spatial distribution of a few learned m/z peaks that were found highly co-localized within distinct tumor clusters.

Figure 5. Analysis of a training 2D MALDI MSI dataset of a mouse Kidney: **a.** convergence distribution, **b.** Overlay of the TIC normalized original mean spectrum and the estimated average spectrum (**c-d**) with an overall mean squared error of 5.5×10^{-3} . **e.** The low dimensional encoded features capture molecular patterns that were segmented using GMM ($k=7$) as shown in (**f**). The set of learned m/z variables that were strongly linked to the identified patterns are highlighted in the mean spectrum as such: m/z bins (red points) and their associated m/z peaks (green points). **h.** Spatial distribution of a few learned m/z features co-localized with some identified spatial patterns, and the close similarity of their distributions for original (top row) and reconstructed data (bottom row) reflects high unsupervised learning quality.

Figure 6. Analysis of 3D MALDI MSI test dataset of mouse kidney of total 72 tissue sections in which spectra of each 2D MSI dataset were independently analyzed: **a**. distribution of average spectrum of both TIC normalized original and reconstructed data of six datasets samples at different volumetric tissue depth (z-direction). **b**. Low dimensional encoded features capture molecular structures from original high dimensional data. **c**. 3D distribution of distinct clusters in the entire dataset that were identified by clustering the encoded features of the entire dataset using GMM ($k=8$) and each cluster represents a molecular pattern that reconciles the kidney's anatomy.

Figure 7. Cross-validation analysis for the 3D MALDI MSI data of mouse kidney: **a**. The full MSI dataset (73 consecutive sections) was randomly shuffled and split into a 20% training set and an 80% testing set, and this process was repeated 5 times as such for each time the *msiPL* model was applied on the training set to optimize the artificial neural network and the trained model was then applied on the unseen test set. **b**. The best cross-validation model was able to reconstruct the original associated training dataset with minimal mean squared error of 6.18×10^{-3} , and showing close distribution of their average TIC-normalized spectra. **c**. The trained model was applied on the unseen test set and revealed comparable performance. The stability of peak learning across different cross-validation models is with the frequency distribution of all m/z peaks identified in the 5-fold cross-validation analyses (**d**), and the peaks count for each frequency (**e**). Overall, 69.6% of the peaks were found stable as they were consistently identified in 80% of the cross-validation analyses. (**f**) 3D Spatial distribution of selected stable m/z values and each of which reveals high localization to a specific structure that reconciles with the kidney's anatomy, thereby reflecting relevance of the learned peaks.