# ROBUST SIMULATION OF IMAGING MASS SPECTROMETRY DATA

Anastasia Sarycheva
Skolkovo Institute of Science and
Technology
Bolshoy Boulevard 30, bld. 1,
Moscow 121205, Russia
E-mail: sarycheva.anastasia@gmail.com

Anton Grigoryev
Institute for Information
Transmission Problems, RAS
Bolshoy Karetny per. 19,
Moscow, 127051, Russia
E-mail: me@ansgri.com

Evgeny N. Nikolaev
Skolkovo Institute of Science and
Technology
Bolshoy Boulevard 30, bld. 1,
Moscow 121205, Russia
E-mail: e.nikolaev@skoltech.ru

Yury Kostyukevich
Skolkovo Institute of Science and
Technology
Bolshoy Boulevard 30, bld. 1,
Moscow 121205, Russia
E-mail: y.kostyukevich@skoltech.ru

## KEYWORDS

Mass spectrometry imaging, simulation, instrument response function.

## ABSTRACT

Mass spectrometry imaging (MSI) with high resolution in mass and space is an analytical method that produces distributions of ions on a sample surface. The algorithms for preprocessing and analysis of the raw data acquired from a mass spectrometer should be evaluated. To do that, the ion composition at every point of the sample should be known. This is possible via the employment of a simulated MSI dataset. In this work, we suggest a pipeline for a robust simulation of MSI datasets that resemble real data with an option to simulate the spectra acquired from any mass spectrometry instrument through the use of the experimental MSI datasets to extract simulation parameters.

## INTRODUCTION

High-resolution mass spectrometry is an analytical technique based on the precise measurement of mass-to-charge ratio ($m/z$) of ionized molecules found in a sample and their relative amount. The mass spectrometry (MS) experiment includes the following main steps: sample preparation, ionization, ion separation (employing electric and magnetic fields), ion detection and signal processing. The result of such an analysis is represented as the so-called mass spectrum (see an example of a profile mass-spectrum in Figure 1: in profile mode, a peak is represented by a collection of signals over several MS experiments) where the signal intensities (i.e. the relative number of ions with certain $m/z$) are plotted as y-axis versus corresponding mass-to-charge ratios along x-axis. Mass spectrum is used to determine the compounds of the sample. For each compound information on molecular mass, composition and structure can be derived through the analysis of experimental spectra. This makes mass spectrometry an essential technique utilized in many applied and basic sciences such as Chemistry, Biology, Medicine, Ecology, Forensic science, etc. (De Hoffmann, 2000)
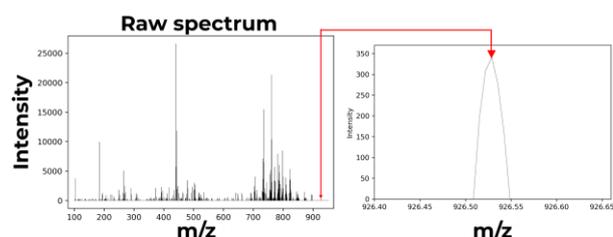


Figure 1: A raw (not preprocessed) profile mass spectrum acquired via MALDI-Orbitrap mass spectrometer (Thermo Scientific Q Exactive Orbitrap) collected from a single region ($35 \times 35 \ \mu m^2$) of a mouse full body section. This mass spectrum includes 4934 individual $m/z$ with corresponding intensities. Zoomed peak 926.529 $m/z$ illustrates a typical peak shape — Gaussian.
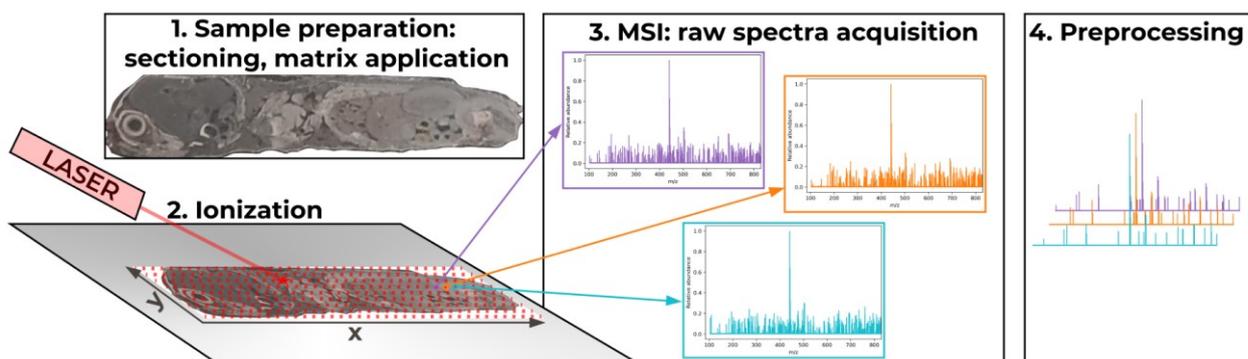
Figure 2: MALDI MSI experiment (sample: mouse full body section).

An MS instrument, mass spectrometer, includes an ion source, a mass analyzer, and a detector (De Hoffmann, 2000). The ion source produces ions of the analyzed sample, the mass analyzer separates these ions according to their $m/z$, and the detector counts the number of ions for every $m/z$ bin (the detector aggregates the continuous $m/z$ values produced by the analyzer into discrete $m/z$ bins, the binning depends on the instrument and its settings).

Mass spectrometry imaging (MSI) is the sequential mass spectrometry analysis of the regions on the surface of the sample. Based on this information, the spatial distributions of the detected ions on the sample's surface are generated. MSI is commonly used in diagnostic applications in the medical and biomedical field (e.g. abnormal regions detection such as tumors, biomarkers search), in medicinal chemistry (medicinal drugs development, research of drugs and their metabolites localization in tissues) (Römpp and Spengler 2013).

There are many ionization techniques developed for MSI, but in this work, we will briefly describe the most popular one: Matrix-Assisted Laser Desorption/Ionization (MALDI) (Baker et al. 2017). The matrix (typically an organic acid) is chosen by the researcher based on the analytes (the compounds of interest to be ionized and detected). The matrix co-crystallizes the analytes, fixing them in place, and facilitates the ionization process. During the desorption/ionization stage, the laser simultaneously vaporizes and ionizes the region it is directed towards, covering the surface of the sample with the given step (raster step). Thus, for each raster (which represents a pixel on the resulting spatial ion distributions) of the sample surface, a mass spectrum (which is a set of detected ions with corresponding signal intensities) is acquired. The MALDI MSI experiment workflow is illustrated in Figure 2. Preprocessing of raw MSI spectra is necessary as the amount of the detected ions for each region is too large for high resolution in mass and space MS instruments (Römpp and Spengler 2013). Preprocessing algorithms should reduce the size of raw MSI dataset, remove noise, eliminate inaccuracies, and make mass spectra from different regions comparable. These algorithms (and/or parameters for them) have to be evaluated, which poses a question of ground truth data in MSI. Due to the complicated nature of the MSI data acquisition, there is no way to get ground truth data for

the samples, i.e. it is not possible to know the exact ion composition at each region of the sample in order to compare it to the ion composition revealed by preprocessed raw spectra. Thus, in order to evaluate the preprocessing algorithms, these algorithms are applied to simulated MSI datasets (Palmer 2014; Verbeek 2014; Wijetunge et al. 2015; Guo et al. 2019; Lieb et al. 2020; Booij 2021).

But even after successful preprocessing steps, the number of individual ion distributions is large. Preprocessed MSI data can be treated as a multichannel image (similarly to an optical image taken with a usual camera; such an image has three channels: red, green, and blue, each channel representing the corresponding light wavelength and its intensity in each part of an image), where each channel represents a certain mass-to-charge ratio ($m/z$) and its intensity at each raster of the surface, i.e. each channel is a visualization of a single ion distribution. So preprocessed MSI data are organized in so-called data cubes (see Figure 3).
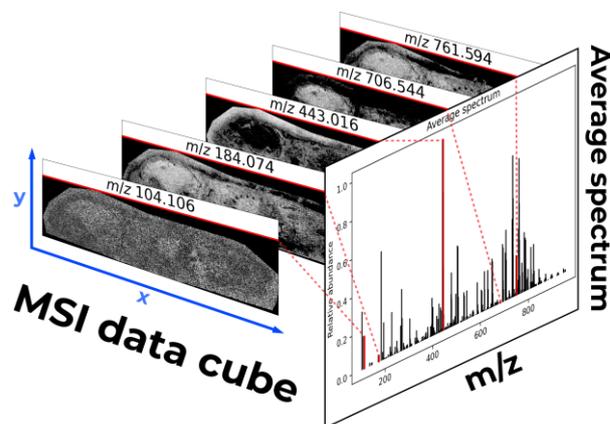


Figure 3: MSI data cube (sample: mouse full body section) which includes spatial distributions of 417 individual ions ($m/z$). Average spectrum is centroid: the signals are displayed as discrete $m/z$ with corresponding intensities with zero line widths.

Various feature selection, clustering and visualization algorithms have to be applied to MSI data cubes, in order to perform a sufficient analysis of such data. These algorithms also have to be evaluated, which again requires realistically simulated data (Buchberger et al. 2018; Verbeek et al. 2020, Sarycheva et al., 2020).

While there is a publically available MSI simulation functions for R language (Bemis et al. 2015; Bemis and Harry 2017; Bemis 2020), which will be discussed in detail in the section dedicated to Simulation of imaging mass spectrometry data, they do not offer as much flexibility as the proposed pipeline.

The important feature of the proposed approach is that it allows for simulation of MSI spectra acquired from an individual instrument. This reduces the gap between evaluation of preprocessing and analysis algorithms performances for the experimental MSI dataset acquired by a certain instrument and the simulated dataset for the same instrument.

## SIMULATION IN MASS SPECTROMETRY

Mass spectrometry data processing and analysis are not possible without computational methods. These methods should be validated, evaluated and compared to guarantee the credibility of the acquired results as the chosen algorithms and/or parameters have been applied to a MS dataset. To gauge the performance of the algorithms (and/or tune the parameters), the reference results (benchmark) should be determined. While it could be argued that the output of a trusted method (or results amalgamation of multiple trusted methods) might be picked as a reference, such an approach is incorrect for the following reasons. First, the result of the existing methods application might be inaccurate (especially, due to the nature of MS data — there are no flawless tools or data processing pipelines). Moreover, a new method might disagree with the results of the existing ones because it might correct the underlying bias of the latter. Thus, it is not correct to expect the result of a new method to replicate the result of the state-of-the-art methods. Second, such an evaluation of a new method would depend on the set of trusted methods chosen to generate reference. So if the set of trusted methods would be changed, even if the input raw data remains the same, the evaluation conclusions might change, too. Thus, the benchmark should not depend on the compared methods.

Realistic data simulation, where the ground truth is defined, serves as a way to create the reference. For the evaluation of raw data processing, e.g. peak picking, feature detection, raw MS data are simulated (e.g., Schulz-Trieglaff et al. 2008; Bielow et al. 2011; Wijetunge et al. 2015). To evaluate the algorithms for the analysis, the processed MS data are simulated (e.g., Awan and Saeed 2018). The main advantages of simulated data: 1) the ground truth is defined, which is often cannot be achieved in MS experiment due to the competitive ionization processes during the ionization stage, instrument noise, and etc; 2) a lot of datasets can be simulated, represented data from different instruments and from different samples; the acquisition of various experimental MS datasets might be expensive or difficult (Gatto et al. 2016).

The simulation of individual spectra can be based on a mathematical model which describes physics of an instrument, if there are analytical expressions describing the latter. For example, the realistic mass spectra can be simulated for Time of Flight (TOF) mass spectrometers (Coombes et al. 2005): the flight time of a given ion of mass $m$ and charge $z$ (known for a range of proteins found in biological tissue or fluids) in an electric field is simulated given the parameters describing the virtual MS instrument (Morris et al. 2005) and the amount of detected ions. This simulation model accounts for two factors affecting the mass resolution (the ability of the instrument to provide a mass spectrum where two slightly different masses are distinguishable): the acquisition time resolution of the detector and the distribution of the initial velocities of the ions. The isotope distributions of individual proteins are included in simulation, since proteins mostly consist of the atoms of carbon, oxygen, and nitrogen. To sum up, this simulation approach employs Instrument Response Function (IRF) calculated from physical laws which result in an approximation for a virtual TOF mass spectrometer.

However, a creation of a detailed physical model of mass spectra generation is not possible for every MS instrument. In a simulation tool LC-MSsim (implemented in C++ programing language) for liquid chromatography mass spectrometry (LC–MS) data (Schulz-Trieglaff et al. 2008), the peak shape of an input $m/z$ is modelled using a Gaussian distribution. In this simulation, the peak width is chosen by a user in terms of the Full-Width-At-Half-Maximum (FWHM) of a peak in mass spectrum, which is defined as the difference between $m/z$ at which the intensity equals half of the maximum intensity of this peak. Since peak shape is modelled as a Gaussian, FWHM of a Gaussian is defined as follows:

$$FWHM_G = 2\sqrt{2ln2}\sigma, \qquad (1)$$

where $\sigma$ is the standard deviation of the Gaussian.

Any real MS dataset includes not only signals caused by the ionized compounds present in the sample, but also noise, which should be accounted for in a simulated spectrum. In LC-MSsim, the FWHM of peaks is used to simulate MS instruments with different mass accuracies (mass accuracy is the difference between measured and actual mass) and resolutions. The inaccuracies in measured peak intensities are simulated by adding Gaussian-distributed noise to peaks. The statistical fluctuations found in MS spectra if the measured intensity of ions with certain $m/z$ is very low (i.e. high-frequency noise of low intensity in a mass spectrum), so-called shot noise, is not well defined in MS, yet for Q-TOF and Ion Trap instruments it can be modeled by Poisson distribution (Du et al. 2008). So in LC-MSsim, the number of shot noise signals is sampled from a Poisson distribution, while $m/z$ are sampled from Gaussian distribution and intensities of these signals are sampled from Exponential distributions (it was approximated based on the experimental mass spectra). The baseline signal in mass spectra (especially prominent within MALDI MS instruments), which decays with increasing $m/z$, in LC-MSsim is simulated by adding an

exponentially-decaying baseline to a simulated mass spectrum.

A simulation tool MSSimulator (implemented in C++ programing language) for LC–MS and LC–MS/MS (MS/MS is tandem mass spectrometry, where the selected ions, separated by their $m/z$ in MS experiment, are split into smaller fragment ions, and then these fragments are also separated and detected by MS experiment) data (Bielow et al. 2011), uses either a truncated Gaussian or Lorentzian distribution for peaks modeling, the width of peaks can be controlled by a user based on the resolution. MSSimulator also provides three models of resolution models in common instruments: resolution is constant in TOF; resolution is degrading linearly with $m/z$ in Fourier transform ion cyclotron resonance (FTICR) instruments; resolution is degrading linearly with the square root of $m/z$ in Orbitrap mass spectrometers (Makarov et al. 2006).

A simulation tool Mspire-Simulator (implemented in Ruby programing language) for LC–MS data (Noyce et al. 2013) employs IRFs calculated from experimental data for three different instruments, acquired from LTQ-Orbitrap, Orbitrap-Velos, Bruker MicrOTOF-Q mass spectrometers. These default models can be replaced by models provided by the user which would mimic other settings and/or instruments: the simulation parameters can be acquired from LC–MS files using a genetic curve fitting algorithm.

To summarize, simulation is used in MS field as benchmark data to assess various algorithms, since the creation of annotated MS datasets acquired by various instruments with various settings is complicated and expensive, and publically available experimental datasets are scarce (Wijetunge et al. 2015; Gatto et al. 2016; Awan and Saeed 2018).

## SIMULATION OF IMAGING MASS SPECTROMETRY DATA

MSI experiments, being a compilation of multiple MS experiments for various points of a sample surface, take more time and are more expensive than the routine MS experiments. Due to the complicated nature of an MSI dataset, the amount of publically available comprehensibility annotated testing datasets is often insufficient (Palmer 2014). For certain methods, the testing datasets might not be available at all. Thus, the simulated MSI data are used for validation and evaluation of MSI data processing and analysis algorithms (Palmer 2014; Verbeek 2014; Guo et al. 2019; Lieb et al. 2020; Booij 2021).

Input data for the MSI simulation is usually the list of ions (it is used if the ionization process is hard to model; otherwise, the list of ions can be predicted from the list of input compounds) with corresponding spatial distributions, and an output MSI dataset is formed according to a statistical model which corresponds to a desired MS instrument.

Verbeek in (Verbeek 2014) describes the simulation approach to datasets creation employed to benchmark MSI analysis algorithms (Booij 2021). An artificial

dataset includes areas representing different tissue regions (and thus having distinct spectral composition) which might overlap: the corresponding characteristic spectra are mixed. Each pixel contains $N$ $m/z$ bins in a certain mass range $m/z_{min}$ to $m/z_{max}$. Characteristic spectra contain certain amounts of peaks with various intensities within mass range. Gaussian noise is added to the mass spectrum of each pixel.

Palmer in (Palmer 2014) describes an IRF modeling approach to MSI spectra generation using QqTOF (Quadrupole-time-of-flight mass spectrometer) instrument as the example. IRFs are approximated via fitting mathematical functions to experimental data, which allows for an approximation of any instrument. The continuous $m/z$ values are aggregated into discrete mass bins. Binning is defined by the instrument and settings. It is simulated according to the resolution of the desired virtual instrument, and the input list of ions is mapped to the corresponding bins (and their intensities are summed). The binned $m/z$ values with corresponding intensities are worked up by IRFs. The latter mimic signal blurring in mass analysers: intensities of neighboring bins affect the input bin intensity (simulated by Gaussian filter moved along $m/z$ axis). IRFs also add detection noise to each input bin (e.g., baseline noise for TOF instruments, electronic noise due to detection circuitry's thermal electron motion sampled from Gaussian distribution, shot noise for all counting detectors, chemical noise which adds the detection of randomly distributed ions on the sample surface).

In (Guo et al. 2019), ion distributions with complex morphology are simulated. The ion spatial variation was simulated as follows. The intensity of an ion at each pixel is generated as a sum of the following terms: the mean intensity of morphological component (i.e. a distinct tissue region) of this pixel, the spatial auto-correlation (simulated via the intrinsic conditional auto-regression (ICAR) model: spatial effect is varying around mean spatial effects at neighboring locations drawn from Normal distribution) which reflects similarity or disagreement in ion composition of neighboring pixels, and the random noise (i.e. measurement error).

Dexter in (Dexter 2018) uses multivariate normal distribution for statistical modeling using experimental MSI data, since he demonstrated that the clustered MSI data from the coronal mouse brain (data acquired via MALDI QqTOF instrument) converted to polar coordinates can be approximated by a multivariate normal distribution. Normality testing (the chi squared quantile plots) is performed for the experimental dataset, and if the latter is close to normally distributed, it is used as a reference for simulation. Simulated data are sampled probabilistically from a multivariate normal distribution.

A simulation of HR imaging mass spectrometry data (ims-simulator) scripts for python 2.7 are available at Github(https://github.com/metaspace2020/ims-simulator) as part of Metaspace project (Alexandrov et al. 2019). The input experimental centroided MSI dataset in imzML format (Schramm et al. 2012) is used as a template for the simulation. Other input data include: the

instrument type (two options: FTICR or Orbitrap); the resolving power (instrument's ability to distinguish between two adjacent ions of equal intensity) at $m/z = 200$; database with the list of metabolites (molecules) as well as the list of possible adducts (the adduct ions are formed during ionization process and contain a certain ion along with analyte molecule (M), e.g. hydrogen ion adducts $[M + H]^+$, sodium ion adducts $[M + Na]^+$, potassium ion adducts $[M + K]^+$, etc.) which might be found in the experiment. This information is used to provide false discovery rate (FDR)-controlled metabolite annotation (Palmer et al. 2016) of the input MSI dataset. This annotation (a list of adducts and molecules) is used to simulate a clean (without noise) dataset. Then the basic statistics for the experimental dataset are calculated (sparsity: histogram of $m/z$ differences between neighboring $m/z$ in each spectrum (i.e. in each pixel); histogram of intensities; minimum intensities for each spectrum). The input dataset's dimensionality is reduced via non negative matrix factorization (NMF): the amount of components is the input amount of desired layers for the simulated dataset (each layer, a simulated tissue, with the spatial distribution represented by an NMF component and with spectral composition represented by a pseudo-spectrum — the loadings of the corresponding NMF component). Noise parameters (median, standard deviation) for each $m/z$ value are calculated from the difference between the experimental data cube's ion intensity distribution and the distribution reconstructed by NMF.

The Cardinal, an R package for MSI data processing and analysis (Bemis et al. 2015), provides functions for the simulation of MS and MSI datasets and which were employed for MSI dataset simulation used for the evaluation of peak picking algorithm in (Lieb et al. 2020), based on the documentation (Bemis and Harry 2017) and the corresponding functions in the package. However, some of these functions were deprecated or changed in the newer version of the package, Cardinal 2 (Bemis 2020). It features a function simulateImage() which relies on simulateSpectrum() for the MSI data simulation. The simulation function input: spatial data (Pixel data) features coordinates x, coordinates y, and boolean columns for each spatial region (morphological substructure, reference image masks) specifying whether or not this region is present in (x,y); ions and intensities (Feature data) featuring $m/z$ (ions), and columns for each spatial region (morphological substructure) specifying intensities of ions describing the spectral composition of corresponding spatial regions. Minimum and maximum $m/z$ values for simulation mass range, as well as step-size for the observed $m/z$ values of the profile spectrum can be specified. There are additional parameters introducing noise and variation (virtual instrument parameters):

spatial autocorrelation (for spatial covariance calculation), standard deviation giving the run-to-run and/or pixel-to-pixel variance (sampled from Normal distribution), standard deviation for the distribution of the observed peaks, a multiplier for multiplicative variance, standard deviation of the random noise introduced in the spectrum, standard deviation of the mass error in the observed $m/z$ values of peaks, mass resolution, maximum intensity of the baseline and its exponential decay, whether output spectra will be in profile or centroided.

While R language is relatively easy to use, the description of Cardinal 2 simulation functions explicitly reads that they are designed for small proof-of-concept examples, and may not scale well to simulating larger datasets(https://rdrr.io/bioc/Cardinal/man/simulateSpectrum.html).

A perfect simulator should be able to produce datasets of any size with defined ground truth, with the option to mimic various instruments. Ideally — any instrument, if the user provides MSI data. This is achievable through the augmentation of simulation with parameters and noise distribution extracted from the experimental datasets.

## THE PROPOSED MSI DATA SIMULATION PIPELINE

If the experimental datasets are provided by the user (in imzML format), they can be used to set mass range and $m/z$ values binning, approximate noise parameters for the simulated spectra. Thus, the proposed simulation configuration extraction from the experimental datasets includes the following steps: instrument information (i.e. resolution), mass range and statistics extraction. The scheme for such module is illustrated in Figure 4.
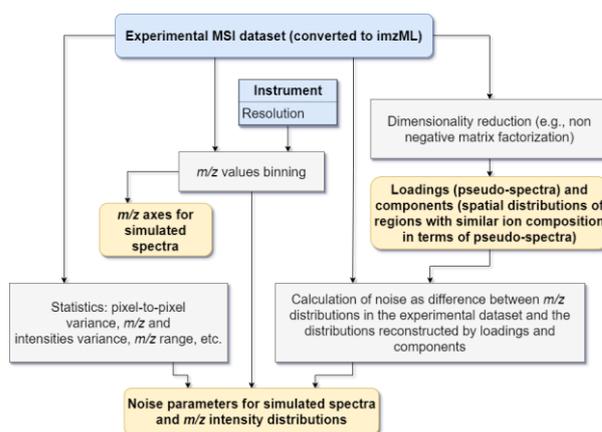


Figure 4: The module for the extraction of simulation parameters from an experimental MSI dataset.

The input morphological components (distinct tissue regions) are provided by the user as separate
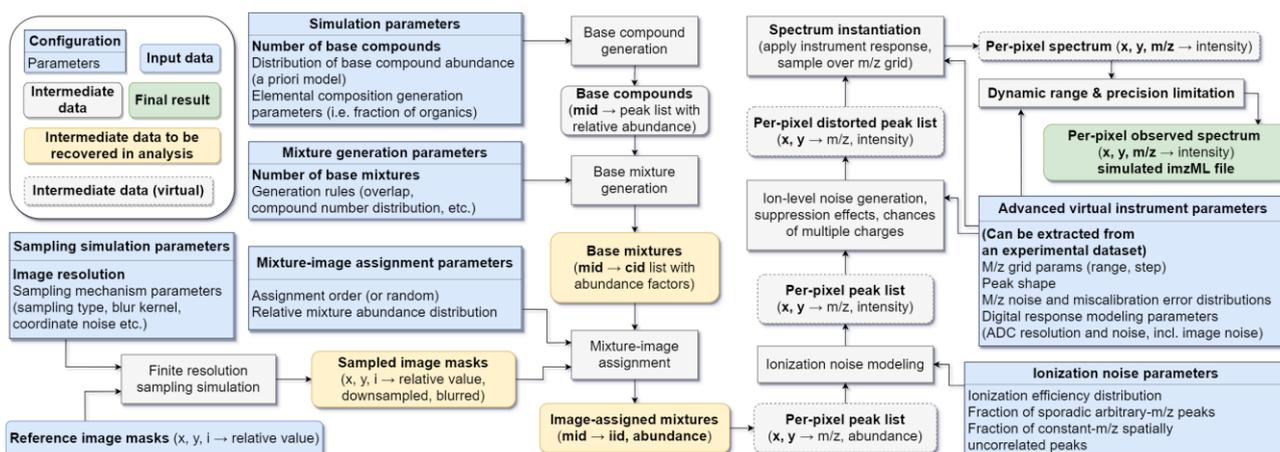
Figure 5: The MSI data simulation pipeline.

grayscale images (could be drawn in with a graphical editor or generated computationally). If the experimental dataset is not provided, the simulation can use preset parameters which correspond to different types of instruments. Parameters can be also set by a user in the configuration file, e.g. analog-to-digital converter resolution, mass range, resolution, etc.

The general steps of the proposed simulation pipeline (parameters can be extracted from the experimental data, see Figure 4) are illustrated in Figure 5.

Let us consider an example of the simulation using the proposed pipeline. We used an experimental MSI dataset of the macaque cerebellum section (22430 single ion distributions 100×80 pixels) as input for the module for extraction of simulation parameters: the m/z values binning was 0.01; the number of ground truth distinct spectral compositions (each corresponds to the simulated tissue) acquired via NMF was set to 6 (Figure 6 A); noise was extracted as the difference between the experimental spectra and the spectra reconstructed by NMF components and loadings. We drew 6 grayscale reference image masks (99×72 pixels each, Figure 6 B) which correspond to 6 overlapping morphological regions with simulated spectral compositions. With added noise, the resulting simulated spectra closely resemble the experimental ones (Figure 6 D). The ground truth data, the exact spectral composition in each spatial location of the simulated imzML, as well as the spectral composition for each reference mask are saved separately.

The main advantages of the proposed pipeline:
1) the ability to produce large artificial datasets in reasonable time; 2) the distinct tissue regions with significantly different spectral composition to be generated (and mixed if overlapped) are provided by the user as simple grayscale images; 3) flexibility of parameters: can be set or extracted from the experimental dataset.

The algorithm is implemented in Python 3.7 and uses the following libraries: imageio, numpy, pandas, pyimzml, scikit-image, sklearn, scipy.
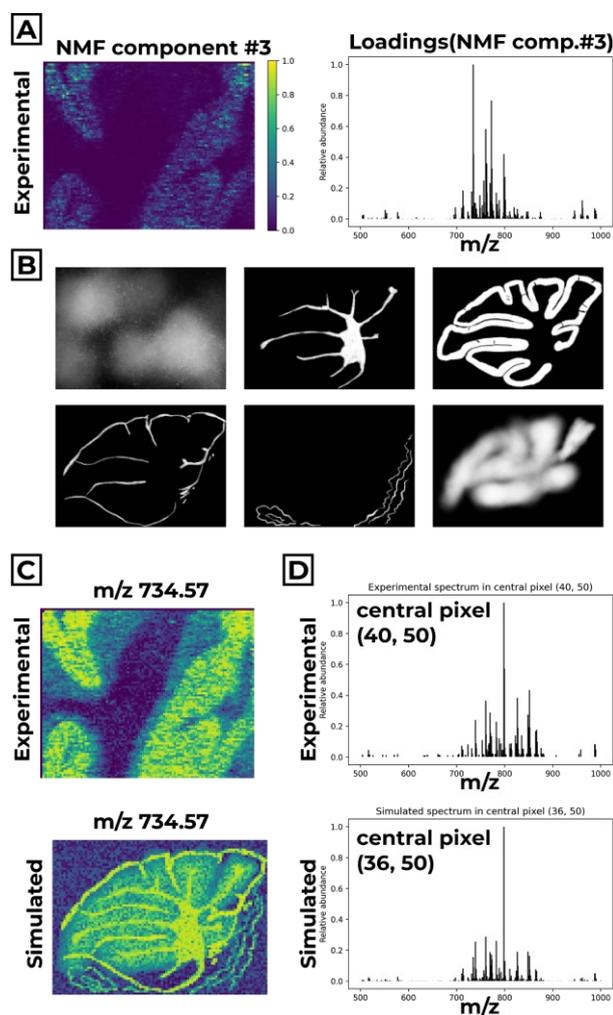


Figure 6: (A) Extraction of simulation parameters: illustration of one of 6 NMF components with corresponding loadings. (B) Input reference image masks (morphological components/tissues to simulate). (C) Comparison of experimental and simulated single ion images. (D) Comparison of experimental spectrum and simulated spectrum in the central pixel of the data cubes.

## CONCLUSION

We have found that there is no convenient and universal simulation tool available for large MSI datasets. We proposed a versatile pipeline, which includes the extraction of simulation parameters from the experimental dataset to tailor the simulation to a specific MSI setup. The initial implementation of the proposed pipeline was tested on a high-mass-resolution Orbitrap-based imaging setup as a parameter source with a set of hand-drawn masks as reference images. The implemented algorithm allowed for the generation of large datasets suitable for quantitative testing of algorithms for enhancement, decomposition, and visualization of MSI data.

## REFERENCES

Alexandrov, T.; K. Ovchinnikova; A. Palmer; V. Kovalev; A. Tarasov; L. Stuart; ... and S. Shahidi-Latham. 2019. METASPACE: A community-populated knowledge base of spatial metabolomes in health and disease. *BioRxiv*, 539478.

Awan, M.G. and F. Saeed. 2018. MaSS-Simulator: A highly configurable MS/MS simulator for generating test datasets for big data algorithms. *bioRxiv*, 302489.

Baker, T.C.; J. Han; and C.H. Borchers. 2017. Recent advancements in matrix-assisted laser desorption/ionization mass spectrometry imaging. *Current opinion in biotechnology*, 43, 62-69.

Bemis, K.D. and A. Harry. 2017. Cardinal: Analytic tools for mass spectrometry imaging.

Bemis, K.D.; A. Harry; L.S. Eberlin; C. Ferreira; S.M. van de Ven; P. Mallick; M. Stolowitz; and O. Vitek. 2015. Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics*, 31(14), 2418-2420.

Bemis, K.A. 2020. Cardinal 2: User guide for mass spectrometry imaging analysis. (http://bioconductor.org/packages/release/bioc/vignettes/C ardinal/inst/doc/Cardinal-2-guide.html#advanced-operations-on-msimagingexperiment)

Bielow, C.; S. Aiche; S. Andreotti; K. Reinert. 2011. MSSimulator: Simulation of mass spectrometry data. *Journal of proteome research*, 10(7), 2922–9.

Booij, T. 2021. Data-Driven Soft Discriminant Maps: Class-aware Linear Feature Extraction in Imaging Mass Spectrometry. (Master thesis, Delft University of Technology)

Buchberger, A.R.; K. DeLaney; J. Johnson; and L. Li. 2018. Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical chemistry*, 90(1), 240.

Coombes, K.R.; J.M. Koomen; K.A. Baggerly; J.S. Morris; and R. Kobayashi. 2005. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer informatics*, 1, 117693510500100103.

De Hoffmann, E. 2000. Mass spectrometry. *Kirk-Othmer Encyclopedia of Chemical Technology.*

Dexter, A. 2018. Developing computational methods for fundamentals and metrology of mass spectrometry imaging (Doctoral dissertation, University of Birmingham).

Du, P.; G. Stolovitzky; P. Horvatovich; R. Bischoff; J. Lim; and F. Suits. 2008. A noise model for mass spectrometry based proteomics. *Bioinformatics*, 24(8), 1070-1077.

Gatto, L.; K.D. Hansen; M.R. Hoopmann; H. Hermjakob; O. Kohlbacher; and A. Beyer. 2016. Testing and validation of computational methods for mass spectrometry. *Journal of proteome research*, 15(3), 809-814.

Guo, D.; K. Bemis; C. Rawlins; J. Agar; and O. Vitek. 2019. Unsupervised segmentation of mass spectrometric ion images characterizes morphology of tissues. *Bioinformatics*, 35(14), i208-i217.

Lieb, F.; T. Boskamp; and H.G. Stark. 2020. Peak detection for MALDI mass spectrometry imaging data using sparse frame multipliers. *Journal of Proteomics*, 225, 103852.

Makarov, A.; E. Denisov; A. Kholomeev; W. Balschun; O. Lange; K. Strupat; and S. Horning. 2006. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry*, 78(7), 2113-2120.

Morris, J.S.; K.R. Coombes; J. Koomen; K.A. Baggerly; and R. Kobayashi. 2005. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9), 1764-1775.

Noyce, A.B.; R. Smith; J. Dalgleish; R.M. Taylor; K.C. Erb; N. Okuda; and J.T. Prince. 2013. Mspire-Simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. *Journal of proteome research*, 12(12), 5742-5749.

Palmer, A.D. 2014. Information processing for mass spectrometry imaging (Doctoral dissertation, University of Birmingham).

Palmer, A.; P. Phapale; I. Chernyavsky; R. Lavigne; D. Fay; A. Tarasov; V. Kovalev; J. Fuchser; S. Nikolenko; C. Pineau; and M. Becker. 2017. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature methods*, 14(1), 57-60.

Römpp, A. and B. Spengler. 2013. Mass spectrometry imaging with high resolution in mass and space. *Histochemistry and cell biology*, 139(6), 759-783.

Sarycheva, A., Grigoryev, A., Sidorchuk, D., Vladimirov, G., Khaitovich, P., Efimova, O., ... and Kostyukevich, Y. 2020. Structure-Preserving and Perceptually Consistent Approach for Visualization of Mass Spectrometry Imaging Datasets. Analytical Chemistry.

Schramm, T., Z. Hester; I. Klinkert; J.P. Both; R.M. Heeren; A. Brunelle; O. Laprévote; N. Desbenoit; M.F. Robbe; M. Stoeckli; and B. Spengler. 2012. imzML—a common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of proteomics*, 75(16), 5106-5110.

Schulz-Trieglaff, O.; N. Pfeifer; C. Grpl; O. Kohlbacher; K. Reinert. 2008. LC-MSsim – a simulation software for liquid chromatography mass spectrometry data. *BMC Bioinformatics*, 9, 423

Verbeeck, N. (2014). Datamining of imaging mass spectrometry data for biomedical tissue exploration. (Doctoral dissertation, KU Leuven).

Verbeeck, N.; R.M. Caprioli; and R. Van de Plas. 2020. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass spectrometry reviews*, 39(3), 245-291.

Wijetunge, C.D.; I. Saeed; B.A. Boughton; U. Roessner; and S.K. Halgamuge. 2015. A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model. *BMC genomics*, 16(12), 1-12.