raMSI for Ground-Truth Machine Learning of Mass Spectrometry Imaging Data

Jialing Zhang\textsuperscript{a1}, Hiu Lok Ngan\textsuperscript{a}, Liu Yang\textsuperscript{a}, Yike Guo\textsuperscript{b}, Xian Yang\textsuperscript{b}, and Zongwei Cai\textsuperscript{a1}

\textsuperscript{a} State Key Laboratory of Environmental and Biological Analysis, Department of Chemistry, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, PR China

\textsuperscript{b} Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, PR China

Corresponding authors: Dr. Jialing Zhang and Dr. Zongwei Cai

Email: jialingzhang@hkbu.edu.hk and zwcai@hkbu.edu.hk

Abstract

The application of machine learning (ML) in mass spectrometry imaging (MSI) becomes one of the most eye-catching fields due to the unparallel sensitivity and specificity of MS for molecular detection and the unmatched efficiency and accuracy of ML for pattern recognition. To get ML started, binning is the most common method used to preprocess the MSI data for the acquisition of millions of \textit{m/z} bins. However, after deep mining of this approach, we find it suffers strikingly serious ambiguity problem, which introduces a fundamental question: was the machine "learning" the intricate MS data properly? In this report, we provide a resolution adaptive method, raMSI, which can attain ground truth molecular features for large datasets and is compatible with different data formats from mainstream mass analyzers. Build on raMSI, a ML ecosystem is designed including data collection, data preparation, database construction, explorative data analysis, modeling and biological insights acquisition. We envision this platform serves for the purpose of motivating cross-disciplinary research involving chemistry, statistics, and biology.
Owing to the capability of spatially mapping thousands of biological molecules without labelling, mass spectrometry imaging (MSI) is gaining prevalence in precision medicine, single cell analysis and other fields.\(^1\) With the advent of high resolving power/resolution mass analyzers such as Orbitrap, Fourier-transform ion cyclotron resonance (FTICR), and quadrupole time of flight (QTOF), the measurement of the mass to charge \((m/z)\) ratio with mass error in the range of parts per million (ppm) becomes routine nowadays, inducing explosive increase of the volume of the MSI datasets.\(^6\) Herein, it becomes obvious that efficient and perceptive data analysis approaches are in high demand. ML breaks the boundary and forges multidisciplinary collaborations crossing physical and life science and has been widely used in chemistry for chemical design, cancer classification, biomarker discovery, and biological pathway recognition.\(^7\)\(^-\)\(^10\) To apply ML in MSI data, the very first step is the determination of the training variables which correspond to the \(m/z\) ratios measured by MS. The unambiguity of the variables \((m/z)\) undoubtedly holds a cardinal role in ensuring the validity of the ML outcomes in the succeeding steps. Unfortunately, this initiative procedure is often omitted in this field, leading to the generation of misrepresented molecular features and leaving the subsequent ML predictive models questionable.

For MSI data, to obtain training features/variables, the most common practice is to generate a series of \(m/z\) bins by constant binning.\(^11\)\(^-\)\(^12\) The width of each bin often relies on the mass resolution \((R = m/Δm)\) of the mass analyzer being used.\(^13\) An ideal \(m/z\) variable used for ML should only represent one molecule and cannot represent any other molecule. Strikingly, constant binning seriously suffers from peak aggregation and/or peak clipping (Supplementary Fig. 1-3), accounting for 93%, 55%, and 95% of the selected \(m/z\) bins when using the bin width at 0.1, 0.01, and 0.001 respectively (Supplementary Fig. 4). Full visualization of the three binning datasets is available at [https://bit.ly/3x3KV7g](https://bit.ly/3x3KV7g). The considerably distorted \(m/z\) bins are misleading information, and the error will be inherited and accumulated for the following ML calculations. Therefore, deep mining of the MSI data is fundamentally important to accurately select the \(m/z\) variables and warrant faultless ML statistical models.

Here we introduce an approach, resolution adaptive MSI (raMSI), which enables comprehensive analysis of intricate MSI data (Fig. 1) and is integrated into a lab-designed ML pipeline (Fig. 2). The core of raMSI is to adapt the varied mass resolution (Fig. 1a), consider the systematic mass shift in mass spectrometry, and provide accurate determination of the molecular features for the subsequent ML. Using the Orbitrap data as an example, we first show the dependency of mass resolution on \(m/z\) ratio and signal-to-noise (S/N) level (Fig. 1b). To afford unambiguous \(m/z\) features, we then derive an algorithm for calculating the resolution of a specific \(m/z\), identify the indistinguishable mass intervals, and eventually use massFusion function to attain the final \(m/z\) list (Fig. 1c). With raMSI, out of 2051 selected \(m/z\) features, 1938 (94.5%) hold ground truth.

Notably, raMSI is suitable to process data from either one of the mass analyzers including Orbitrap, FTICR, and QTOF and is also broadly compatible for the common MSI data types including raw, imzML and comma-separated values (csv). Considering that the resolution information is available from neither imzML nor csv data, we introduce a method for the estimation of the resolution of a specific \(m/z\) based on the readily accessible instrumental calibration data. Upon following the five steps in raMSI, the true \(m/z′_\text{lower}\) and \(m/z′_\text{upper}\) for a molecule can be obtained, and any \(m/z\) ratio falling in such interval is represented by the mean value of the two as the ‘Effective \(m/z\)’. Detailed explanation of raMSI is presented in the Method.

Based on raMSI, we design a ML ecosystem, which involves data collection, data preparation, database construction, explorative data analysis (EDA), modeling, and biological insights acquisition (Fig. 2). In this system, the first three steps before raMSI are crucial to prepare the required data frame for the following ML analysis. The later three falls in the regime of interdisciplinary field, including chemistry, statistics, and biology.

In the first data collection section (Fig. 2a), both raw data from MSI experiments and public MSI datasets from repositories such as METASPACE and ProteomeXchange can be used. For data preparation (Fig. 2b), we provide specific guidelines and test all the three data formats, including raw, imzML, and csv. Hadoop and Spark are utilized to provide structured data storage, data query, and data transformation, which possess great scalability, flexibility, and more importantly the speed needed to handle large size...
MSI files (Fig. 2c). After the construction of database, the removal of random noises from the mass spectrometry instrument is conducted by filtering the m/z ratios with low S/N or frequencies. Once the noises are removed, the adaptive mass intervals are obtained, and the data frame based on the effective m/z is built.

Based on the generated data frame, the later three components in the ML pipeline can be sequentially used for gaining deeper insights of the MSI data. First, EDA is performed (Fig. 2d), including but not limited to description, visualization, correlation analysis, and transformation. The exploration empowers users to attain guidance such as removal of redundant features or outliers and the optimal selection of ML algorithms for modeling. Note that more functions are emerging, and users can implement their preferred EDA methods.

Owing to the wide compatibility of Python libraries, a variety of algorithms such as logistic regression, boosting, random forest, support vector machine (SVM), naïve bayes, and neural network, can be executed for training and testing (Fig. 2e). When selecting the ML algorithm, it is of significance to have the generated model with good interpretability. We test on the public ovary and breast cancer datasets obtained from METASPACE and present the best prediction matrix of the test set from each dataset for demonstration (Supplementary Fig. 11, and full presentation of the ML results can be found at https://bit.ly/3x3KV7g). The receiver operating characteristic (ROC) curve and selected important molecular features are also displayed.

For biological insights acquisition (Fig. 2f), we examine each of the selected m/z value from the ML model by checking the MSI images and matching online databases. For example, the top three selected molecular features by random forest in ovary cancer include m/z 174.0416, m/z 133.0148, and m/z 157.1240, which are tentatively assigned as N-acetyl aspartic acid (NAA), malic acid, and nonanoic acid respectively. The validation of the possible biomarkers is out of the scope of this work, but we envision the precisely selected molecular features by this system will boost closer cooperation between chemists and biologists. The outcome from this component will initiate new hypotheses or assumptions and stimulate next round of study using either MS or other approaches.

In summary, we present raMSI, an open-source method to statistically select the molecular features from high-resolution MSI data. And we develop a ML pipeline with the components including EDA, modeling, and model deployment for gaining biological insights. The significance of raMSI is its guarantee of the ground truth variables by minimizing peak aggregation and clipping encountered in constant binning. Importantly, raMSI can work with the three most common data types: raw file from the instrument, imzML file commonly shared on different online databases, and csv file from end users’ customized ROI selection, ensuring its adaptability to the diversified demands. The method promotes concerted effort from mass spectrometrist, computer scientist, statistician, and biologist to collaborate. Therefore, we anticipate wide applications of this approach in interdisciplinary research.
References

**Fig. 1** | Schematic illustration of raMSI for ground truth m/z acquisition. 

**a,** Resolution dependency on m/z ratio. 

**b,** The generation of resolution adaptive mass intervals through curve fitting and followed by massFusion. The curve fitting should be based on the signals from the molecular ions. 

**c,** The count distributions of selected biological related ions, represented by the ‘Effective m/z’ shown on the top. 

**d,** The 2051 adaptive mass intervals from the entire dataset with 67,496 MSI pixels.
Fig. 2 | The machine learning ecosystem based on raMSI.  

a, Data collection. The system can work on both laboratory data using Orbitrap, FTICR or QTOF, and public datasets.  
b, Data preparation. raMSI is compatible with raw, imzML and csv.  
c, Database construction. Hadoop and Spark are used to efficiently construct the database for subsequent ML.  
d, Exploratory Data Analysis. The data frame generated by raMSI is subjected to exploration of the main characteristics of the MSI dataset.  
e, Machine learning modeling. Different ML algorithms are employed to build predictive models.  
f, Biological insights. The ML models are compared, the selected important biological features are studied, and the feedback are provided to initiate further studies.
Methods

Ground truth molecular feature selection by raMSI

To make raMSI adaptable to the most common mass analyzers (Orbitrap, FTICR, and QTOF), we develop the five steps shown below, following which users can obtain the raMSI m/z(s) (Supplementary Fig. 5). For Orbitrap/FTICR, the resolutions across the whole mass range are not constant, leading to varied $\Delta m$ (the smallest width for two close m/z ratio being discriminated from each other), which is the key parameter needed for generating the adaptive intervals. However, the corresponding mass resolutions are missing in the imzML or csv file, so regression is performed to compute the $\Delta m$. Note that if raw data is used, we can directly apply MSFilereader to get the resolution information and use it for the $\Delta m$ calculation. For QTOF type analyzer, the corresponding resolution of a specific m/z is relatively constant throughout the mass range, so a direct calculation of $\Delta m$ for a specific m/z can be achieved by skipping step 3.

Step 1. Visualization of the relationship between the m/z and mass resolution

The first step is to obtain a mass spectrum using standard calibration solutions provided by the vendors of the mass spectrometer. Users need to make sure when they calibrate the instrument, the mass resolution are consistent with the MSI experiments. For Orbitrap, there are two sub-curves shown after plotting the m/z against the resolution (Supplementary Fig. 6a-c), which is irrelevant to the tissue type. The top curve originates from the ions with higher intensities, and the bottom curve originates from the ions with lower intensities (Supplementary Fig. 6d).

Step 2. Removal of instrumental random noises

We find the bottom curve consists mainly low-abundance instrument noises with low S/N, which should be removed to enable accurate calculation of the mass resolutions of the molecular ions. After showing the S/N ratios from 1 to 6 (Supplementary Fig. 7), we adopt a more conservative value of S/N=5 for complete separation of the two sub-curves and filtration of random instrumental noises. The S/N threshold at 5 are tested on different Orbitrap analyzers and showed its efficiency in noise removal (Supplementary Fig. 8). Concerning the double curve phenomenon, our explanation is when the S/N is close to the limit of detection (S/N=3), insufficient number of ions captured by the mass analyzer will lead to compromised mass resolution.

Step 3. Curve fitting between the m/z and mass resolution

Based on the physical principle of FTICR, the resolution of the FTICR and the m/z ratio follows the Eq. 1:

$$ R(FT) = \frac{-eB}{\Delta \omega} (m/z)^{-1} \approx C(FT) \times (m/z)^{-1} \quad Eq.(1) $$

In which $e$ represents the electron charge, $B$ represents the magnetic field, $\Delta \omega$ represents the derivative of the frequency of axial oscillations (in rad/s), and $C(FT)$ represents a constant. As is shown, $R(FT)$ is the inversely proportional to the m/z (Supplementary Fig. 9).

Based on the physical principle of Orbitrap, the resolution of the Orbitrap and the m/z ratio follows the Eq. 2:

$$ R(Orbi) = \frac{\sqrt{ek}}{-2\Delta \omega} (m/z)^{-0.5} \approx C(Orbi) \times (m/z)^{-0.5} \quad Eq.(2) $$

In which $e$ represents the electron charge, $k$ represents the field curvature, $\Delta \omega$ represents the derivative of the frequency of axial oscillations (in rad/s), and $C(Orbi)$ represents a constant. $R(Orbi)$ is the inversely proportional to the square root of the m/z.
Computationally, $C(Orbi)$ and $C(FT)$ for Orbitrap and FTICR can be calculated by using scipy.optimize.curve_fit in Python. For Orbitrap at different resolutions, we use calibration solution provided by the vendor to obtain a spectrum and calculate the $C(Orbi)$ for resolution at 15 000, 30 000, 60 000, and 120 000 respectively (Supplementary Fig. 10). Within raMSI, we provide the default $C(Orbi)$ of each resolution for users.

Step 4. Acquisition of lower and upper $m/z$ limits

Based on the acquired resolution $R$ from the above steps for Orbitrap/FTICR or the constant resolution $R$ for QTOF, the $\Delta m$ can be computed. We define the lower and upper $m/z$ limit and form an interval for each $m/z$ (Eq. 3 and 4), any $m/z$ value falls in such interval cannot be resolved from each other (Eq. 5).

$$\Delta m = m/z_{\text{upper}} - m/z_{\text{lower}} \quad \text{Eq. (3)}$$

$$m/z = (m/z_{\text{upper}} + m/z_{\text{lower}})/2 \quad \text{Eq. (4)}$$

$$m/z_{\text{lower}} = (m/z) - (m/z)/2R, \quad m/z_{\text{upper}} = (m/z) + (m/z)/2R \quad \text{Eq. (5)}$$

Step 5. Formation of the adaptive intervals by massFusion using entire MSI dataset

As we described in step 1, random instrumental noises exist in all the mass spectra of an MSI image. For thousands or tens of thousands of pixels (mass spectra) in MSI, these noises will have low counts due to their randomness. With raw data, the removal of these random noises can be straightforwardly realized by filtering the S/N lower than 10 to make sure the $m/z$ ratios used for further ML are from molecular ions. For imzML or csv data, the noise information is not available, and we suggest using 1% as the cutoff to filter the $m/z$ values which show less than 1% of the total pixels, or at least 5 counts for small dataset, whichever is larger. Note that both the S/N for raw data and the cutoff value for imzML/csv data can be adjusted to control the number of $m/z$ values being subjected to the subsequent ML.

After the removal of random noises, the $m/z_{\text{lower}}$ and $m/z_{\text{upper}}$ can be determined for each of the molecular ions. Then using the customized massFusion function, the mass intervals which have overlaps with others will be unified, resulting in the final true $m/z'_{\text{lower}}$ and $m/z'_{\text{upper}}$ for each $m/z$ value. We use the mean value of the $(m/z'_{\text{lower}} + m/z'_{\text{upper}})/2$ as the ‘Effective $m/z$’ to represent the mass interval.

**Application of raMSI in the Machine Learning Ecosystem**

To put raMSI in use, we design a ML ecosystem including data collection, data preparation, database construction, exploratory data analysis, ML modeling, biological insights acquisition and its feedback for the next cycle.

a. Data Collection

In this step, users can perform data collection either with their own imaging setup or from public database. For the lab collected data, raw files from the instrument can be used directly. Note that with raw data, maximum reservation of the mass spectrometry information, including $m/z$ ratios, intensity, resolution and noise, is possible. For public database, imzML and csv files are commonly accessible, in which resolution and noise are not available.

b. Data Preparation

For users who wish to use the calibration solution and calculate the constant of the $C(Orbi)$ or $C(FT)$ on their own, we provide the “MassResolutionConstantCalculation.ipynb” so that a more precise calculation can be achieved. The obtained constant can then be inputted into the “raMSI.ipynb”.

c. Database Construction
In “raMSI.ipynb”, we provide different options for building the database with different data formats including raw, imzML or csv and with different mass analyzers including Orbitrap, QTOF, and FTICR. We create two tables, one for storing the sample type and the other for storing the information in each MSI pixel (spectrum). The common information being stored from a pixel for all the three data types include ‘source’ which is the data type being chosen, ‘filename’ which provides the identity for the sample, ‘x’ and ‘y’ which are the locations in the image, ‘mass’ which is the m/z ratio, and ‘intensity’ which is the corresponding intensity. Particularly, for raw data, ‘resolution’ which is the corresponding mass resolution of the m/z ratio, and ‘noise’ which is the corresponding noise level for the m/z ratio are stored in the database. Based on the adaptive mass intervals obtained in the raMSI, the data frame is formed by using the ‘Effective_m/z’ as the columns names and pixel information from the samples with the corresponding sample type as the last column. The data frame named “raMSI_df” is stored in the database for the following EDA analysis.

d. Exploratory Data Analysis

Currently, we add commonly used EDA functions. For visualization, various of plotting methods including scatterplot, histogram, boxplot, pairplot, and heatmap are implemented, to enable basic understanding of the data in the perspective of data deviation and skewness. We suggest users to perform interactive analysis using libraries like Plotly. For correlation analysis, it can be used to study the strength of the linear relationship between two m/z ratios, which is useful for the removal of isotopic peaks (redundant features). Scaling and logarithmic transformation can be used for skewness correction and more efficient subsequent ML analysis. More functions are being implemented and users can perform their own EDA to advance their perceptions of the data.

e. Machine Learning Modeling

The EDA enables the summarization of the main characteristics of a specific MSI data set. The data set now can be divided into the training set and test set for building ML models. We provide the common algorithms including parametric algorithms such as logistic regression, naive bayes, and simple neural networks, and nonparametric algorithms such as random forest and SVM. The choice of certain types of ML algorithms should be based on the previous EDA analysis. Note that due to limited number of samples in the public datasets, without independent test set, overfitting/underfitting may happen. More thorough studies can be carried out together with statisticians regarding the fine tuning of the parameters of different algorithms. We anticipate this section encourages collaborations between chemists, computer scientists, and statisticians.

f. Biological Insight Acquisition

The final section in this ecosystem is to obtain biological insights from the well-performed prediction models. Important m/z features selected by the ML algorithm can be visualized. The MSI images from these m/z ratios should be displayed to double confirm the validity of the significance. The molecular patterns revealed in these models such as up/down regulation of a specific group of molecules can then be employed for further biological experiments. We expect the biologists and chemists to work together to have better understanding of the biological insights, which eventually can feedback to the next round of study design.

Software specifications

We exclusively code in Jupyter Notebook (Version 6.4.5) using Python to process the MSI data. For database construction, libraries including pandas, numpy, ImzMLParser, MSFileReader, pyspark, pathlib, and re are used. For EDA, libraries including matplotlib, seaborn, plotly, and sklearn are used. For ML modeling, libraries including sklearn, xgboost, lightgbm, and tensorflow are used. For Windows system, Hadoop (Version 3.2.2) is needed to pre-installed. To process the raw data from the Orbitrap, MSFileReader 3.0 SP2 needs to be installed. Microsoft C++ is also needed and can be installed using vs_BuildTools.
Brain, liver and kidney tissues were harvested from rats (purchased from The Chinese University of Hong Kong), and subsequently stored at ~80°C. All housing and experimental procedures were performed according to the guidelines for the use of experimental animals of Hong Kong Baptist University (HKBU) and approved by the HKBU Committee on the Use of Human & Animal Subjects in Teaching and Research. The kidney tissue was homogenized and remained frozen in the freezer at -80°C before use. All the tissue blocks were sliced into sections with a thickness of 10 µm using a cryostat (CryoStar NX70, Thermo Scientific, U.S.). The kidney homogenates experiments were conducted for day 1, 2, 3, 4, 13, 33 for the collection of 67,496 pixels of MSI data.

Air-flow-assisted desorption electrospray ionization (AFA-DESI) imaging platform (Beijing Victor, Beijing, China) was coupled with both a Hybrid Q Exactive Orbitrap (Thermo Fisher, San Jose, CA, USA) and Exploris 120 Orbitrap (Thermo Fisher, Bremen, Germany) mass spectrometer. The experiments were performed in negative ion mode. The spatial resolution of biological tissue imaging was set at 200 µm. The parameters of AFA-DESI stage include vacuum pressure at -50 psi, spray voltage at -4K Volts and gas pressure at 180 psi.

For the identification of the detected m/z values, multiple approaches including high mass accuracy, tandem mass spectrometry and database matching (METASPACE, HMDB, and LIPID MAPS) were employed to increase the confidence of annotation.

Data availability

Laboratory kidney homogenate tissue dataset, public breast and ovary cancer datasets from METASAPCE can be found at https://bit.ly/3x3KV7g. The comparison of selected molecular features by raMSI and binning approach, and machine learning results are available at https://bit.ly/3x3KV7g. The open-source codes including raMSI, EDA, ML Modeling are available at https://bit.ly/3R9W9iL.

Acknowledgments

We thank the collaboration research fund from the research grants council of Hong Kong SAR (C2011-21GF) and the Tier2 fund (RC-OFSGT_20_21_SCI_007) from Hong Kong Baptist University. We thank Beijing Viktor Technology Co., LTD. for the instrument support. Special thanks to Mrs. Junru Xia for her valuable discussions.

Author Contributions: J.Z. and Z.C. designed and conceptualized the research. J.Z., H.N. and L.Y. performed programming, analyzed data and drafted the manuscript. H.N. and L.Y. performed the MSI experiments. X.Y., Y.G. and Z.C. edited and revised this paper.
Supplementary Data

Supplementary Fig. 1 | Presentation of peak aggregation and peak clipping by using a constant bin of 0.01. Visualization of all the selected 1365 m/z bins are available at https://bit.ly/3x3KV7g. **a**, Scatter plot of the calculated actual bin sizes by applying confidence interval of 99.7%. The effective bin size using the binning method was obtained by calculating the difference between ultimate right and ultimate left peak(s) within each bin. **b**, Histogram of the count distributions of m/z peaks within each bin. Note: A total of 67,496 mass spectra were obtained by analyzing kidney homogenates sections using AFA-DESI MSI. The constant bin of 0.01 has two critical defects: 1) for two or more different molecules with very close m/z values, if the bin size is not small enough, they will be binned together (peak aggregation); 2) for a single molecule, if the bin size is not big enough, the mass shifts of the m/z which is a persistent issue in MSI will lead to the split of this m/z into two bins (peak clipping).
Supplementary Fig. 2 | Presentation of peak aggregation and peak clipping by using a constant bin of 0.1. Visualization of all the selected 812 m/z bins are available at https://bit.ly/3x3KV7g. a, Scatter plot of the calculated actual bin sizes by applying confidence interval of 99.7%. b, Histogram of the count distributions of m/z peaks within each bin.
Supplementary Fig. 3 | Presentation of peak clipping by using a constant bin of 0.001. Visualization of all the selected 1859 m/z bins are available at https://bit.ly/3x3KV7g. a, Scatter plot of the calculated actual bin sizes by applying confidence interval of 99.7%. b, Histogram of the count distributions of m/z peaks within each bin.
Supplementary Fig. 4| Summary of peak aggregation and peak clipping using constant bins at 0.1, 0.01 and 0.001 respectively, and the comparison with the effective m/z ratios acquired by raMSI.
Supplementary Fig. 5| Decision tree depending on the data type and mass analyzer.

Is Database source 'RAW'?

Yes

- S/N can be calculated, and the random noises can directly be filtered by removing mz with low S/N.

- Select:
  ```
  round(mass, 4) as mass, avg(noise) as noise, avg(intensity) as intensity, avg(resolution) as resolution
  from (self.schema) (self.table_name)
  where intensity/noise > 10
  group by mass
  ```

- Save the results in a new table.

No

- Save the mz, intensity, and mz_count

- Take the round(mass, 4) as mass, avg(intensity) as intensity, count(round(mass, 4)) as mz_count.

- Group by mass:

- Save the results in a new table.

Because noise information is not available in either of the data format, noise can only be removed by filtering mz with lower counts.

Resolutions are in the table, so no need to be calculated.

TOF

- Resolution can be calculated directly by the experiment setting.

- Resolution can be calculated by acquiring a calibrant spectrum with the same MSI setting*:

  \[ R(\text{trim}) = C(\text{trim}) - \left( \frac{m}{c} \right)^{0.5} \]

  \[ C(\text{trim}) \text{ can be obtained} \]

  Users can use the \( C(\text{trim}) \) provided by us, with options at 15,000, 30,000, 60,000, and 120,000.

FT-ICR

- Resolution can be calculated by acquiring a calibrant spectrum with the same MSI setting*:

  \[ R(\text{FF}) = C(\text{FF}) - \left( \frac{m}{c} \right)^{0.5} \]

  \[ C(\text{FF}) \text{ can be obtained} \]

  Users can use the \( C(\text{FF}) \) provided by us, with options at 40,000.

- Mass resolutions can be used for calculating the Δm.

- Mass resolving power is now attained which can be used for calculating the Δm.

Calculation of the lower and upper mass limit which can't be differentiated under current mass resolution:

\[ m/z_{\text{lower}} = (m/z) - (m/z) \cdot 2R \]

\[ m/z_{\text{upper}} = (m/z) + (m/z) \cdot 2R \]

For the entire dataset, massFusion can be called to merge the mz intervals with overlaps.

Acquisition of Effective m/z to represent the mass interval:

\[ \text{Effective}(m/z) = \frac{m/z_{\text{lower}} + \text{true } m/z_{\text{upper}}}{2} \]
Supplementary Fig. 6 | Investigation of the mass resolution relevant parameters. a, Mass spectra from mouse brain and liver tissue sections. b, The consistent trend between mass resolution and m/z ratio among three tissue types. c, The scatter plot of m/z and resolution of the entire mass spectrum from brain. d, Separation of the two sub-curves by two intensity levels. The intensity levels were separated into two categories: the low level is colored in blue with the intensity smaller than 1000, the high level is colored in red with the intensity larger than 1000.
Supplementary Fig. 7| The dependency of mass resolution with m/z and signal to noise ratio. Considering different mass spectrometry vendors may setup their own rules of the intensity level, which hinders the wide application of an absolute intensity threshold, we include the noise information from the raw data, calculate the signal to noise (S/N) ratios for each detected m/z, and plotted against the mass resolution. Scatter plot of m/z and resolution of the whole mass spectrum from the kidney using different S/N ratios ranging from 1 to 6 as the thresholds.
Supplementary Fig. 8 | S/N cutoff at 5 for Orbitrap Exploris 120, Orbitrap Fusion Tribrid, LTQ-Orbitrap Elite and Q Exactive Orbitrap. In each figure, the scattering markers with small bubble sizes represent the signal from the random instrumental noises (at the bottom) and the markers with large bubble sizes are those from the molecular ions.
Supplementary Fig. 9| Curve fitting of mass resolution with m/z ratios on FTICR. The FT-ICR data were acquired with a 9.4 T superconducting magnet, resulting in a resolution of roughly 420 000 at m/z 400. The signal-to-noise ratio and dynamic range were improved by summing 128-time domain transients for each spectrum.

\[
R(FT) = 1.64E8 \times \left(\frac{m}{z}\right)^{-1}
\]
Supplementary Fig. 10 | Curve fitting of mass resolution at 15 000, 30 000, 60 000, and 120 000 with m/z ratios on Orbitrap Exploris 120 using calibration solution. Note that the S/N of 5 is applied to remove the random noises from the instrument and we only fit on the molecular ions.
Supplementary Fig. 11: Case studies using raMSI ML ecosystem for public datasets on METASPACE. Full presentations of the ML results for both datasets are available at https://bit.ly/3x3KV7g.

**a** Ovary cancer prediction by Random Forests using DESI-Orbitrap Data

<table>
<thead>
<tr>
<th>True Pred</th>
<th>Normal</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3309</td>
<td>12</td>
</tr>
<tr>
<td>Cancer</td>
<td>5</td>
<td>18816</td>
</tr>
</tbody>
</table>

AUC=1.0

**b** Breast cancer prediction by SVM using DESI-Orbitrap Data

<table>
<thead>
<tr>
<th>True Pred</th>
<th>Normal</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>628</td>
<td>2</td>
</tr>
<tr>
<td>Cancer</td>
<td>1</td>
<td>3665</td>
</tr>
</tbody>
</table>

AUC=1.0

**Dataset Information on METASPACE:**
- Sample information
- Organism: Homo sapiens (human)
- Organism part: Ovary
- Sample growth conditions: Surgical specimen
- Sample stabilisation: Fresh frozen
- MS analysis
- Polarity: Negative
- Ionisation source: DESI
- Analyzer: Orbitrap
- Resolving power: 60000
**Supplementary Table 1.** Selected biological ions detected in kidney homogenate using AFA-DESI MSI.

<table>
<thead>
<tr>
<th>Measured m/z</th>
<th>Resolution</th>
<th>Theoretical m/z</th>
<th>Absolute Error (Da)</th>
<th>Tentative Attribution</th>
<th>Proposed Formula</th>
<th>Tissue Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>133.0132</td>
<td>99102</td>
<td>133.0142</td>
<td>-0.0010</td>
<td>Malic acid</td>
<td>C_4H_6O_5</td>
<td>Kidney</td>
</tr>
<tr>
<td>195.0505</td>
<td>79602</td>
<td>195.0510</td>
<td>-0.0005</td>
<td>Gliconic acid</td>
<td>C_6H_11O_7</td>
<td>Kidney</td>
</tr>
<tr>
<td>215.0324</td>
<td>77406</td>
<td>215.0326</td>
<td>-0.0002</td>
<td>Glucose</td>
<td>C_6H_12O_6Cl</td>
<td>Kidney</td>
</tr>
<tr>
<td>255.2329</td>
<td>71307</td>
<td>255.2330</td>
<td>-0.0001</td>
<td>FA (16:0)</td>
<td>C_18H_33O_2</td>
<td>Kidney</td>
</tr>
<tr>
<td>281.2486</td>
<td>67807</td>
<td>281.2486</td>
<td>0</td>
<td>FA (18:1)</td>
<td>C_18H_33O_2</td>
<td>Kidney</td>
</tr>
<tr>
<td>327.2329</td>
<td>63007</td>
<td>327.2330</td>
<td>-0.0001</td>
<td>FA (22:6)</td>
<td>C_22H_31O_2</td>
<td>Kidney</td>
</tr>
<tr>
<td>537.4887</td>
<td>48307</td>
<td>537.4888</td>
<td>-0.0001</td>
<td>FAHFA (34:0)</td>
<td>C_34H_65O_4</td>
<td>Kidney</td>
</tr>
<tr>
<td>563.5043</td>
<td>46807</td>
<td>563.5045</td>
<td>-0.0002</td>
<td>FAHFA (36:1)</td>
<td>C_36H_67O_4</td>
<td>Kidney</td>
</tr>
<tr>
<td>627.4761</td>
<td>44502</td>
<td>627.4761</td>
<td>0</td>
<td>DG (34:2)</td>
<td>C_37H_68O_5Cl</td>
<td>Kidney</td>
</tr>
<tr>
<td>747.5178</td>
<td>41407</td>
<td>747.5182</td>
<td>-0.0004</td>
<td>PG (34:1)</td>
<td>C_40H_70O_10P</td>
<td>Kidney</td>
</tr>
<tr>
<td>766.5378</td>
<td>40507</td>
<td>766.5392</td>
<td>-0.0014</td>
<td>PE (38:4)</td>
<td>C_43H_77NO_4P</td>
<td>Kidney</td>
</tr>
<tr>
<td>810.5286</td>
<td>39507</td>
<td>810.5291</td>
<td>-0.0005</td>
<td>PS (38:4)</td>
<td>C_45H_77NO_10P</td>
<td>Kidney</td>
</tr>
<tr>
<td>857.5179</td>
<td>37707</td>
<td>857.5186</td>
<td>-0.0007</td>
<td>PS (36:4)</td>
<td>C_45H_77O_13P</td>
<td>Kidney</td>
</tr>
<tr>
<td>885.5491</td>
<td>38207</td>
<td>885.5499</td>
<td>-0.0008</td>
<td>PI (38:4)</td>
<td>C_47H_82O_13P</td>
<td>Kidney</td>
</tr>
<tr>
<td>1034.2474</td>
<td>32706</td>
<td>1034.2517</td>
<td>-0.0043</td>
<td>CL + DG (108:10)</td>
<td>C_126H_215O_22P_2</td>
<td>Kidney</td>
</tr>
<tr>
<td>1447.9631</td>
<td>28306</td>
<td>1447.9650</td>
<td>-0.0019</td>
<td>CL (72:8)</td>
<td>C_61H_141O_17P_2</td>
<td>Kidney</td>
</tr>
<tr>
<td>1541.0455</td>
<td>87802</td>
<td>146.0459</td>
<td>-0.0014</td>
<td>Glutamic acid</td>
<td>C_5H_6NO_4</td>
<td>Brain</td>
</tr>
<tr>
<td>174.0396</td>
<td>80406</td>
<td>174.0408</td>
<td>-0.0012</td>
<td>NAA</td>
<td>C_6H_8NO_5</td>
<td>Brain</td>
</tr>
<tr>
<td>255.2321</td>
<td>71106</td>
<td>255.2330</td>
<td>-0.0009</td>
<td>FA (16:0)</td>
<td>C_18H_31O_2</td>
<td>Brain</td>
</tr>
<tr>
<td>283.2634</td>
<td>67902</td>
<td>283.2643</td>
<td>-0.0009</td>
<td>FA (18:0)</td>
<td>C_18H_31O_2</td>
<td>Brain</td>
</tr>
<tr>
<td>303.2322</td>
<td>65306</td>
<td>303.2330</td>
<td>-0.0008</td>
<td>FA (20:4)</td>
<td>C_20H_31O_2</td>
<td>Brain</td>
</tr>
<tr>
<td>673.4792</td>
<td>42806</td>
<td>673.4814</td>
<td>-0.0022</td>
<td>PA (34:1)</td>
<td>C_37H_78O_2P</td>
<td>Brain</td>
</tr>
<tr>
<td>747.5173</td>
<td>41306</td>
<td>747.5182</td>
<td>-0.0009</td>
<td>PG (34:1)</td>
<td>C_40H_70O_10P</td>
<td>Brain</td>
</tr>
<tr>
<td>788.5422</td>
<td>40206</td>
<td>788.5447</td>
<td>-0.0025</td>
<td>PS (36:1)</td>
<td>C_42H_79NO_10P</td>
<td>Brain</td>
</tr>
<tr>
<td>834.5265</td>
<td>39506</td>
<td>834.5291</td>
<td>-0.0026</td>
<td>PS (40:6)</td>
<td>C_46H_77NO_10P</td>
<td>Brain</td>
</tr>
<tr>
<td>888.6207</td>
<td>38106</td>
<td>888.6240</td>
<td>-0.0033</td>
<td>Sulfatide (42:2)</td>
<td>C_48H_90O_11S</td>
<td>Brain</td>
</tr>
<tr>
<td>1477.9996</td>
<td>23100</td>
<td>1478.0119</td>
<td>-0.0123</td>
<td>CL (74:7)</td>
<td>C_63H_147O_17P_2</td>
<td>Brain</td>
</tr>
<tr>
<td>124.0063</td>
<td>10107</td>
<td>124.0074</td>
<td>-0.0011</td>
<td>Taurine</td>
<td>C_7H_8NO_5S</td>
<td>Liver</td>
</tr>
<tr>
<td>215.0324</td>
<td>78506</td>
<td>215.0326</td>
<td>-0.0002</td>
<td>Glucose</td>
<td>C_6H_12O_6Cl</td>
<td>Liver</td>
</tr>
<tr>
<td>255.2329</td>
<td>70707</td>
<td>255.2330</td>
<td>-0.0001</td>
<td>FA (16:0)</td>
<td>C_18H_31O_2</td>
<td>Liver</td>
</tr>
<tr>
<td>279.2330</td>
<td>68207</td>
<td>279.2330</td>
<td>0</td>
<td>FA (18:2)</td>
<td>C_18H_31O_2</td>
<td>Liver</td>
</tr>
<tr>
<td>327.2330</td>
<td>62207</td>
<td>327.2330</td>
<td>0</td>
<td>FA (22:6)</td>
<td>C_22H_31O_2</td>
<td>Liver</td>
</tr>
<tr>
<td>535.4734</td>
<td>47607</td>
<td>535.4732</td>
<td>+0.0002</td>
<td>FAHFA (34:1)</td>
<td>C_36H_68O_4</td>
<td>Liver</td>
</tr>
<tr>
<td>627.4763</td>
<td>44007</td>
<td>627.4761</td>
<td>+0.0002</td>
<td>DG (34:2)</td>
<td>C_37H_68O_4Cl</td>
<td>Liver</td>
</tr>
<tr>
<td>773.5334</td>
<td>38107</td>
<td>773.5340</td>
<td>-0.0006</td>
<td>PG (36:2)</td>
<td>C_42H_76O_10P</td>
<td>Liver</td>
</tr>
<tr>
<td>810.5292</td>
<td>39307</td>
<td>810.5291</td>
<td>+0.0001</td>
<td>PS (38:4)</td>
<td>C_42H_77NO_10P</td>
<td>Liver</td>
</tr>
<tr>
<td>885.5490</td>
<td>38307</td>
<td>885.5499</td>
<td>-0.0009</td>
<td>PI (38:4)</td>
<td>C_47H_82O_13P</td>
<td>Liver</td>
</tr>
</tbody>
</table>
| 1033.2414 | 32806 | 1033.2439 | -0.0025 | CL + DG  
(108:11) | C_{120}H_{211}O_{22}P_2 | Liver |
| 1447.9647 | 28707 | 1447.9650 | -0.0003 | CL (72:8) | C_{81}H_{141}O_{17}P_2 | Liver |

Note: NAA=N-acetyl aspartic acid; FA=fatty acid; CL=cardiolipin; DG=diacylglycerol; PG=phosphatidylglycerol; PA=phosphatidic Acid; FAHFA=fatty acid ester of a hydroxy fatty acid; PE=phosphatidylethanolamine; PS=phosphatidylserine; PI=phosphatidylinositol