Mass recalibration for desorption electrospray ionization mass spectrometry imaging using endogenous reference ions.

Paolo Inglese[1,2], Helen Xuexia Huang[2], Vincen Wu[2], Matthew R Lewis[1,2], Zoltan Takats[2]

[1]National Phenome Centre, Imperial College London, Hammersmith Campus, IRDB Building, London W12 0NN, United Kingdom

[2]Section of Bioanalytical Chemistry, Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

1 Abstract

Background

Mass spectrometry imaging (MSI) data often consist of tens of thousands of mass spectra collected from a sample surface. During the time necessary to perform a single acquisition, it is likely that uncontrollable factors alter the validity of the initial mass calibration of the instrument, resulting in mass errors of magnitude significantly larger than their nominal values. This phenomenon has a two-fold detrimental effect: a) it reduces the ability to interpret the results based on the observed signals, b) it can affect the quality of the observed signal spatial distributions.

Results

We present a post-acquisition computational method capable of reducing the observed mass drift up to 60 ppm in biological samples, exploiting the presence of

typical molecules with a known mass-to-charge ratio. The procedure, tested on time-of-flight (TOF) and Orbitrap mass spectrometry analyzers interfaced to a desorption electrospray ionization (DESI) source, improves the molecular annotation quality and the spatial distributions of the detected ions.

Conclusion

The presented method represents a robust and accurate tool for performing post-acquisition mass recalibration of DESI-MSI datasets and can help to increase the reliability of the molecular assignment and the data quality.

2 Introduction

In the last twenty years, mass spectrometry imaging (MSI) has attracted increasing attention as a technology capable of capturing molecular spatial patterns from sample surfaces. Among the available ionization sources, such as matrix-assisted laser desorption (MALDI) (Hillenkamp et al., 1991) or secondary ion mass spectrometry (SIMS) (Herzog & Viehböck, 1949), desorption electrospray ionization (DESI) (Takáts et al., 2004) has gained popularity and been widely applied thanks to its relatively simpler preparation process. A map of the detected ions' spatial distribution is usually represented as an image and analyzed for identifying informative spatial patterns. MSI has been successfully deployed in various application areas, from cancer research (Dória et al., 2016; Guenther et al., 2015; Mascini et al., 2018; Schwamborn et al., 2007) to pharmacology.(Drexler et al., 2011; Karlsson & Hanrieder, 2017; LaBonia et al., 2016; Nilsson et al., 2015; Prideaux & Stoeckli, 2012) However, as an emerging technology, MSI still faces implementation challenges for controlling the quality of data produced. (Ellis et

al., 2014; Hamm et al., 2012; Swales et al., 2018) The accuracy of mass-to-charge ratio (*m/z)* measurements under sustained use are chief among these concerns. MSI acquisitions often consist of tens of thousands of spectra (pixels), corresponding to several hours of operation, during which mass measurements can be subject to substantial drift. For example, a small image of 10 mm x 10 mm, acquired at a running speed of 1 scan/s, with a spatial resolution of 100 µm, requires about 2.8 hours. Shifting mass measurements are detrimental to the analyzed spatial patterns and can result in erroneous chemical annotations. Additionally, as MSI relies on only a single dimension of separation for detected ions, errors of a few parts per million (ppm) can result in hundreds of candidate molecular identities, interpreting in terms of biochemical hypotheses a practical challenge.

For these reasons, it is crucial to develop quality control approaches for MSI that mitigate this issue and yield accurate *m/z* measurements with high precision during extended instrument use. Solutions successfully applied in more routine hyphenated techniques (e.g., liquid chromatography-mass spectrometry LC-MS) include periodic recalibration or more frequent calibration based on the measurement of one or more standard reference materials during the acquisition. Such an approach is infeasible for the current MSI technology since the acquisition is performed continuously while the probe moves across the sample surface. However, this approach may be emulated by adding the standard material to the sample. (Ràfols et al., 2018) Although this approach can be used to facilitate post-acquisition *m/z* correction of mass accuracy, concerns may arise about the unknown effect on the ionization efficiency caused by these exogenous molecules'

presence. In the specific case of DESI-MSI, usually, only one reference molecule (lock mass) is added to the solvent, which can be insufficient to capture the non-linear drift across the measured m/z range.

Another class of approaches relies on using endogenous or background peaks as candidate reference ions to estimate the mass shift occurring in each spectrum and map the measured m/z onto their corresponding recalibrated values.(Li & Li, 2019) This class of methods requires prior knowledge about the molecular content of the sample.

Boskamp et al. showed that endogenous signal (chemical noise) could improve the mass accuracy in MALDI-MSI of peptide datasets. (Boskamp et al., 2020) Recently, La Rocca et al. have presented an algorithm to recalibrate MSI datasets using a linear fit on the observed mass errors from endogenous biological peaks. (La Rocca et al., 2021) In their method, they treated each spectrum as independent.

Here, we present a different approach based on a fixed set of reference ions across the entire MSI dataset, in which the individual mass shifts are modeled as a time series. The advantages of global references are that they 1) define a fixed boundary for the m/z extrapolation, 2) improve evaluation of the quality of the matching procedure by visualizing their spatial distribution, and 3) increase the robustness of the matching process as a time-dependent model.

The approach leverages the ubiquity of common structural components of tissues as dependable sources of reference values, using them as multiple points to estimate error and correct all observed masses in each spectrum. The approach is demonstrated on biological tissue datasets acquired using a DESI-MSI source

interfaced with time-of-flight (TOF) or Orbitrap ion analyzers, in both positive (ES+) and negative (ES-) ion mode.

For simplicity, we will refer to m/z values as "masses" throughout the manuscript.

3 DESI-MSI datasets

3.1 In-house DESI-MSI datasets

DESI-MSI data of six tissue sections from mouse (mus musculus) brain and pig (sus domesticus) liver were acquired using a Waters Xevo G2-XS QToF mass spectrometer.

Three brains of a C57BL/6 mouse model were purchased from Charles River Laboratories, while pig liver samples were obtained from a local supermarket.

We used the sample preparation procedure and DESI parameters reported in Tillner et al. (Tillner et al., 2017) The mouse brain samples were scanned at a rate of 75 µm/s horizontally, while the pig liver sample was scanned at a rate of 100 µm/s.

The RAW spectra from the six acquired TOF DESI-MSI were first converted into imzML using the MassLynx SDK (v4.7.0) and Python's pyimzML package[1]. Then, they were filtered from the baseline noise using a modified version of the *kneedle* algorithm (Satopää et al., 2011) (Section S1 of Supplementary Information) and smoothed using a Savitzky-Golay kernel convolution. Centroided peaks were detected and used as input for the recalibration procedure (Section S2 of Supplementary Information). Throughout the manuscript, we will refer to spectrum and list of centroid peaks as synonyms.

[1]https://github.com/alexandrovteam/pyimzML

3.2 Public DESI-MSI datasets

Twenty-four DESI-MSI datasets were downloaded from the public service METASPACE[2]. We considered various tissue types analyzed with either Orbitrap or TOF ion analyzers. Details about the datasets can be found in Table S1 of Supplementary Information.

The datasets consisted of centroided RAW peaks. No peak detection or denoising was applied.

4 Methods

4.1 Selection of sample pixels

The first step of the presented workflow consists of determining the pixels associated with the biological sample. Since our method uses endogenous molecules that are expected to be detected by DESI-MS in tissue, it is crucial to remove all pixels that do not correspond to tissue-related signals.

The procedure aims at discriminating the sample-related from off-sample pixels, using a supervised classifier trained on a set of user-defined labeled pixels.

We first define the set of features from the RAW peaks, applying a uniform binning with a bin size of 1 m/z. Subsequently, through a graphical user interface (GUI), the user manually labels a set of pixels as either 'sample' or 'background' (Figure S1 of Supplementary Information). Then, a linear support vector machines (SVM) model is fitted on this set of features and labels and used to predict all pixels' labels, providing a binary map of the region-of-interest (ROI). The user can

[2]https://metaspace2020.eu/

manually refine the segmentation through the GUI. Finally, connected regions (8-neighborhood) smaller than a given size are assigned to the background class. In all the experiments, we set the size threshold equal to 50 pixels.

The ROI mask is saved in a comma-separated values (CSV) file to be used in the workflow's following steps.

4.2 General recalibration workflow

In this section, a general description of the recalibration method is given, with details reported in the following sections.

Let us consider a DESI-MSI dataset representing a collection of N sequentially acquired spectra (where $N = N^{(\mathrm{ROI})} + N^{(\mathrm{off})}$) is the sum of the number of sample ROI and off-sample pixels).

The calibration procedure aims to estimate the nominal (calibrated) mass from the observed mass of the detected ions.

One common strategy is based on reference points, called *lock masses*. (Michelsen & Karlsson, 1999) This method assumes a statistical model g between the nominal $\boldsymbol{\mu} = (\mu_k)$ and detected $\boldsymbol{\mu}^{\#} = \left(\mu_k^{\#}\right)$ masses:

$$\boldsymbol{\mu} = g_i\left(\boldsymbol{\mu}_i^{\#}, \boldsymbol{\theta}_i\right) \tag{1}$$

where the index i indicates that the observed masses $\boldsymbol{\mu}^{\#}$, the model g and its parameters $\boldsymbol{\theta}$ are spectrum (or pixel)-specific.

Given a set of reference masses and their observed values, the parameters are fitted from data, and the model is used to predict the calibrated values of all detected masses.

When using endogenous signals as a reference, lock masses can be spectrum-specific ($\boldsymbol{\mu} = \boldsymbol{\mu}_i$) (La Rocca et al., 2021) or be defined globally. Our method uses the latter approach.

The recalibration model can only be fitted if all $\boldsymbol{\mu}^{\#}$ are available. However, due to the molecular heterogeneity of biological samples, this property is not guaranteed. To overcome this challenge, we model the observed $\boldsymbol{\mu}^{\#}$ as a smooth function of the acquisition time (or, equivalently, the pixel order), which is equivalent to assuming that $\boldsymbol{\mu}^{\#}$ depend on the actual conditions of the instrument, and that these smoothly vary in a controlled environment. In practice, we model each mass in $\boldsymbol{\mu}^{\#}$ as a time series, where the pixel order is a proxy for the acquisition time.

Given a reference mass $M \in \boldsymbol{\mu}$, let $\mathbf{P}_M = (p_1, p_2, \dots) = (p_i)$ be the subset of ROI pixels where it is detected with mass $\mathbf{M}^{\#} = (M_1^{\#}, M_2^{\#}, \dots) = \left(M_i^{\#}\right)$, where $M_i^{\#} = M_{p_i}^{\#}$. We model the time series trend using a generalized additive model (GAM) (Hastie & Tibshirani, 1986):

$$\begin{aligned} M_i^{\#} | \eta_i, \sigma^2 &\sim \text{Normal}(\eta_i, \sigma^2) \\ \eta_i &= \beta_0 + \sum_{j=1}^{J} s_k(p_i) \end{aligned} \tag{2}$$

where s_j are penalized cubic spline functions and J is equal to 20.

The procedure, repeated for all $M \in \boldsymbol{\mu}$, separately, provides the models for all reference masses. As for any regression models, if $\mathbf{P}_M$ corresponds to 'a large portion' of the ROI, the fitted model can accurately predict the values of M, $\widehat{\mathbf{M}}^{\#} = \widehat{\boldsymbol{\eta}}$, that would have been observed (up to a random error term) in all ROI pixels, subject to the instrumental condition at the acquisition time.

In the following section, we describe the procedure to detect and select the reference ions signals.

4.3 Detection and selection of reference ions

Since the molecular composition of the sample is largely unknown, it is practically impossible to define a general set of lock masses that works for all datasets. Before describing the procedure aimed at identifying them, we need to introduce properties that define 'good' endogenous reference masses.

The candidate reference mass is optimal if: a) it is detected in a 'large' portion of the sample, b) its peak is 'isolated' in the m/z space. Property (a) allows us to use the model described in Eq. 2, while property (b) reduces the uncertainty of the matching procedure.

As the reference ions consist of molecules that are expected to be found in tissue samples and be detected by DESI-MS, we perform a database search to determine the list of candidates. The database $\mathbf{\Omega}$ includes phospholipids, fatty acids, mono/di/triglycerides, cholesterols, and ceramides from Lipidmaps (Sud et al., 2007) and Human Metabolite DataBase (Wishart et al., 2007) (HMDB) databases. Deprotonated and chlorine adducts ($[M-H]^-$, $[M+Cl]^-$) are considered for negative polarity mode, while protonated, sodium and potassium adducts ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$) are considered for positive polarity mode.

Given a database mass $M \in \mathbf{\Omega}$, a mass m_i among pixel p_i detected masses is considered a candidate match if $m_i \in [M - \lambda, M + \lambda]$, where $\lambda = W \times M \times 10^{-6}$ is the mass distance from M corresponding to the relative error W in ppm units. All masses satisfying the condition are considered possible matches, meaning that

more matches per pixel can be found. For each candidate match, we retrieve its peak index π_i, observed mass $M_i^{\#} = m_i$, and intensity $\iota_i^{\#}$.

To satisfy property (a), for all $M \in \mathbf{\Omega}$, we remove the candidate matched peaks that are detected in less than 75% of the sample ROI pixels.

Subsequently, we further filter the list of candidate reference ions using a method based on a kernel density estimator (KDE) following a similar procedure to that described in Smirnov et al. (Smirnov et al., 2019) If candidate reference peaks for M are represented as points with coordinates $\mathbf{x} = (x, y) = \left(p_i, M_i^{\#}\right)$, we assume that highly dense connected regions represent the shift trends of M.

Given a candidate reference mass $M \in \boldsymbol{\mu} \subseteq \mathbf{\Omega}$ and its observed values $\mathbf{M}^{\#} = \left(M_i^{\#}\right)$, we estimate the density f_h of the (x, y)-points using a 2D Fast Fourier Transform (FFT) KDE (Silverman, 1982), with a triangular kernel K, defined as

$$K(\mathbf{x}) = 1 - |\mathbf{x}|, \qquad |\mathbf{x}| \leq 1 \tag{3}$$

The kernel is fitted on the coordinates of the points projected onto the [0,1] interval.

The triangular kernel is chosen because of its computational efficiency. The kernel bandwidth is set to $h = 2.575 \times \sigma \times N^{-1/5}$, where σ represents the standard deviation of the whole set of points and N represents the number of points $\mathbf{x}$. (Silverman, 2018) The kernel is fitted on a regular grid of size $G \times G$, with $G = 200$. Given the estimated 2D density $\hat{f}_h = \hat{f}_h(x, y)$, we identify the curve passing through its local maxima as follows. For each $x' \in \{1, \dots, G\}$, we determine the local density maximum $y' = \underset{x=x'}{\operatorname{argmax}}\, \hat{f}_h(x, y)$. If $\hat{f}_h(x', y) = 0$ for all y, no maximum is considered for that value of x'. Subsequently, a cubic spline $S(x)$ is fitted on the vector of local maxima $(\hat{x}, \hat{y})$ after transforming them back to the original (pixel,

mass)-space. Using the predicted spline values $S(p_i)$ for all p_i, we calculate the absolute residuals r_i and the dispersion d_i:

$$r_i = \left|M_i^{\#} - S(p_i)\right| \tag{4}$$

$$d_i = \frac{2 \times \mathrm{mad}(\mathbf{r})}{S(p_i)} \times 10^6 \tag{5}$$

Points with $r_i \geq 2 \times \mathrm{mad}(\mathbf{r})$, $\mathbf{r} = (r_i)$, where "*mad*" is the median absolute deviation multiplied by the inverse of cumulative Normal distribution $1/\Phi^{-1}(3/4) \approx 1.4826$, are considered outliers and removed from the list of matched peaks (Figure 1A). Finally, the candidate reference M is kept if: the number of distinct inlier pixels is greater or equal than 75% of the ROI size and $\max_i(d_i) \leq 10$ ppm. In this way, we select isolated signals (property (b)).

The procedure is repeated for all candidate reference masses.

The final set of reference masses is denoted as $\boldsymbol{\mu}^*$. The time series trends (Eq. 2) for each $M^* \in \boldsymbol{\mu}^*$ are fitted using the corresponding inlier points, and the models are used to predict the observed values $\widehat{\mathbf{M}}^{\#} = \left(\widehat{M}_i^{\#}\right)$ in all ROI pixels (Figure 1B). The intensities $\iota_i^{\#}$ of the matched peaks can be plotted to reveal their spatial distribution (Figure 1C). Visual inspection provides additional validation of the consistency of the matched peaks. When multiple matches are available per pixel, the intensities of the peaks with the smallest r_i are plotted.

4.4 Pixel-wise mass recalibration

Although the selected reference ions have passed the filtering procedure, mismatched peaks may still be present if: a) same peaks fall within two search

windows so that they are assigned to two reference ions, b) shifted peaks fall within the search window by chance (especially if the mass shift is large).

To reduce the chance of using mismatched reference ions masses, we apply a pixel-specific mass filter, based on the assumption that most of the true matches share a similar relative mass error. The details of the filtering procedure are reported in Section S4 of Supplementary Information.

Once the set of reference masses $\boldsymbol{\mu}_p^* = \left(M_{1,p}^*, M_{2,p}^* \dots\right) = \left(M_{k,p}^*\right)$ is determined for pixel p together with the values predicted by GAMs, we fit a model g (Eq. 1) specific for each type of ion analyzer (Suppl. Information S3). For Orbitrap analyzers, we fit a linear regression model (Wenger et al., 2010)

$$\begin{aligned} M_{k,p}^* &\sim \text{Normal}\left(\eta_{k,p}, \sigma_p^2\right) \\ \eta_{k,p} &= \beta_{0,p} + \beta_{1,p}\widehat{M}_{k,p}^{\#} \end{aligned} \tag{7}$$

Instead, for the TOF analyzer, we use a polynomial regression model (Christian et al., 2000; Vera et al., 1996)

$$\begin{aligned} \sqrt{M_{k,p}^*} &\sim \text{Normal}\left(\eta_{k,p}, \sigma_p^2\right) \\ \eta_{k,p} &= \beta_{0,p} + \sum_{d=1}^{D} \beta_{d,p}\sqrt{\left(\widehat{M}_{k,p}^{\#}\right)^d} \end{aligned} \tag{8}$$

where the optimal polynomial degree D, $D \leq 5$, corresponds to the smallest Bayesian Information Criterion (BIC). (Schwarz, 1978)

The fitted model is then used to predict the calibrated mass for all the detected ions in the specific pixel (Figure 1D).

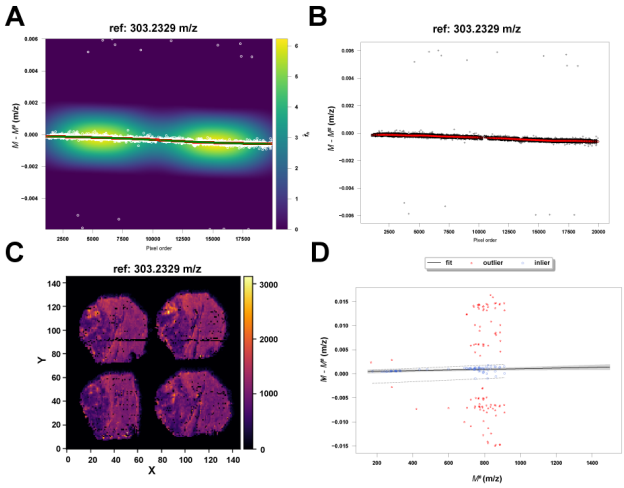

Figure 1 - Example of the recalibration process for the Orbitrap dataset liver ES-. (A) FFT KDE is fitted from the matched masses for a candidate reference (303.2329 m/z). The white points represent the difference in m/z between the observed and the theoretical values of the reference. The curve passing through the density maxima (points in red) is used to fit a regression spline (blue points). (B) Points with large residuals from the spline are labeled as outliers (in grey) and excluded from the matched set. The remaining inliers are then used to fit a GAM (red line). (C) The intensity corresponding to the inlier points can be plotted to reveal their spatial distribution. This allows validating the consistency of the selected matches visually. (D) Finally, a regression model is fitted in each pixel using the only reference masses with similar errors in ppm (blue circles) within the residual intervals defined by the filter (grey dashed lines). The degree of the polynomial corresponds to the smallest BIC value. The predicted values (black line, the grey bands represent 95% confidence intervals) are used to correct all observed masses in the pixel.

5 Results and discussion

As a first experiment, we tested the validity of the pixel-wise recalibration model and the reference mass filtering described in Section S4 of Supplementary Information. We simulated a series of pixel-wise reference masses and their observed values following the method described in Section S5 of Supplementary

Information. In all simulations, we observed a significant reduction of the median relative error, in both Orbitrap and TOF models. No significant difference in terms of performances was observed between the two models. Also, the simulations confirmed a reduction of the heteroskedasticity of the selected masses (Figure S5 of Supplementary Information).

We then tested the DESI-MSI datasets after removing the pixels outside of the ROI as described in Section 4.1.

All recalibration parameters were kept fixed within the Orbitrap or TOF datasets, as described in the Methods section. For Orbitrap datasets, we used a search window $W = 20$ ppm, while for TOF datasets we used $W = 100$ ppm.

FFT KDEs for matched points were fitted using the "KDEpy" package for Python[3] (v. 1.1.0). The cubic splines are fitted using the function *UnivariateSpline* available in "SciPy" package for Python (v 1.7.1). (Virtanen et al., 2020) The smoothing parameter s of the function is determined by 5-fold cross-validation, among 20 values from 0.0001 to 0.1 evenly spaced in the $\log_{10}$ scale, corresponding to the smallest mean squared error.

The KDE-based model of the reference mass shift showed its robustness in the case of close peaks, correctly discarding wrongly matched peaks (Figure 2).

The GAMs were fitted using the pyGAM package (v. 0.8.0) for Python. (Servén et al., 2018) The spline penalization parameter was chosen among eleven values varying between 0.001 and 1000 evenly spaced in the $\log_{10}$ scale, corresponding to the smallest generalized cross-validation value (GCV). (Wood, 2017)

[3]https://github.com/tommyod/KDEpy

310 In all tested DESI-MSI datasets, except for Orbitrap pancreas ES-, the selected
311 reference masses for the pixel-wise recalibration well covered their acquisition
312 m/z range (Figure S2-S3 of Supplementary Information). In general, the reference
313 masses were more concentrated in the m/z ranges below 400 and above 600,
314 corresponding to small molecules and phospholipids, respectively.
315 In five TOF datasets, the optimal pixel-wise recalibration models were polynomial
316 with a degree greater or equal to two. In particular, the distribution of the
317 regression models coefficients revealed that the TOF mass shifts were greater
318 than Orbitrap (Orbitrap error within 5 ppm, while TOF error up to 65 ppm)
319 (Figure 3). This observation may indicate a higher tendency of one class of
320 analyzers to be subject to the changes of the environmental conditions, although
321 these may depend on the unknown actual instrumental conditions.
322 To evaluate the recalibration efficacy, we looked at two main effects: 1) the number
323 of putative molecular annotations, 2) mass error of test ions.
324 The number of putative annotations was generated using the METASPACE
325 website. METASPACE assigns molecule identities to MSI datasets based on a
326 metabolite-signal match score (MSM) calculated from spectral and spatial
327 measures and an FDR-estimation using a decoy strategy. Each assignment is
328 characterized by an FDR equal to 5%, 10%, 20%, and 50%. The annotation is
329 performed by database search. Because of the biological nature of the analyzed
330 samples, we selected the following four databases among those available: a)
331 "ChEBI 2018-01", b) "CoreMetabolome v3", c) "HMDB v4", d) "LipidMaps 2017-12-
332 12".

333 All METASPACE default parameters were used. Specifically, the interval for the
334 image generation was set to 3 ppm, and the decoy set size was set to 20. Therefore,
335 we compared the number of unique annotations corresponding to the stringent
336 criterion of FDR = 5%. In addition, we considered distinct annotations
337 corresponding to different molecular formulas plus their adduct among the four
338 used databases. This choice was justified by the fact that the actual molecules may
339 be present in only specific databases. The annotations were performed using the
340 ROI filtered raw and recalibrated peaks without applying any peak binning or
341 spatial filter.
342 We checked whether the largest number of unique annotations was found in the
343 original or recalibrated dataset. Also, we calculated the relative difference
344 between the number of annotations in the two datasets:

$$v_r = \frac{n_r - n_0}{n_0} \tag{9}$$

345 In twenty-two cases (≈ 73%), the recalibrated dataset received a larger number of
346 assignments compared to the raw dataset (**Error! Reference source not**
347 **found.**).
348 The recalibration increased the number of annotations by up to 246. In contrast,
349 in TOF liver ES- 3 we observed the largest relative decrease equal to 23% of the
350 original annotations, corresponding to 18 fewer annotations, while the largest
351 absolute decrease was observed in Orbitrap necrosis ES+ with 40 fewer
352 annotations, equivalent to about 7% of the original annotations (Figure 4).
353 We tested the mass accuracy of a set of test masses to evaluate the effect of the
354 recalibration. We followed the same idea described in Boskamp et al. (Boskamp et
355 al., 2020) Since the actual molecular content of a sample is unknown, we compared

356 the mass accuracy of a list of ions expected to be detected by DESI-MSI in
357 biological tissue samples.
358 The test ions list was generated from the available annotated molecules in the
359 public datasets of METASPACE. The list consisted of the monoisotopic form plus
360 the first three identified isotopes detected in more than 10% of the datasets from
361 the same tissue type and ion mode of the DESI-MSI dataset. We used a Python
362 script from LaRocca et al.[4] (version available in June 2021) to generate the list of
363 candidate test masses. The mass values common to the set used for fitting the
364 pixel-wise recalibration models were excluded. In all datasets, the test masses
365 covered the acquisition m/z range, confirming that they represented a good set of
366 probes for evaluating the recalibration effects (Figure S2-S3 of Supplementary
367 Information).
368 Using the procedure described in Section S6 of Supplementary Information, we
369 selected the test masses within 1 ppm from the most abundant relative errors and
370 calculated the median of the pixel-wise median difference between the absolute
371 mass errors $\tilde{\Delta}$. We tested if we could reject the null hypothesis $H_0: \tilde{\Delta} = 0$, using
372 bootstrapping (number of repetitions equal to 10^4). We observed in nineteen
373 datasets a significant (Benjamini-Hochberg corrected p-value < 0.05) decrease of
374 the median relative error after the recalibration, with values varying between 0.02
375 and 63.6 ppm. In all these datasets, the final median relative error was below 2.7
376 ppm for Orbitrap, and below 4.9 ppm for TOF. In one case (TOF liver ES- 3), the
377 median relative error increased by about 1.3 ppm, with a final median relative

[4]https://github.com/LaRoccaRaphael/MSI_recalibration/blob/master/recalibration/internal_calibrants_generation.ipynb

error equal to 4 ppm. In ten datasets, $\tilde{\Delta}$ was not significantly different from zero (Figure 5, Table S2-S3 of Supplementary Information).

Consistently, the datasets with the largest increase in the number of annotations corresponded to those with the largest decrease in the mass error.

Furthermore, the recalibration improved the ion image quality. After removing the mass shifts, all pixels correctly displayed the expected molecular spatial distributions (Figure 6).

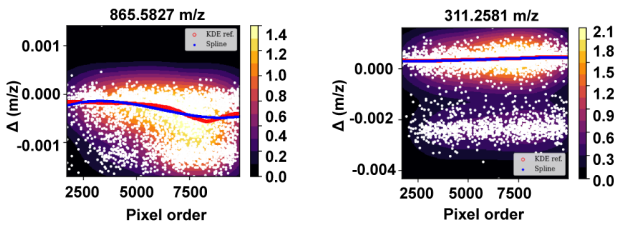

Figure 2 – Left: Two peaks are too close to be resolved by the KDE. This reference is correctly discarded because of the large dispersion around the regression spline. Right: in this case, the KDE can resolve the two close peaks. Here, the method correctly fits the spline on the points belonging to one of the peaks. After filtering the outliers, the dispersion is below the threshold.

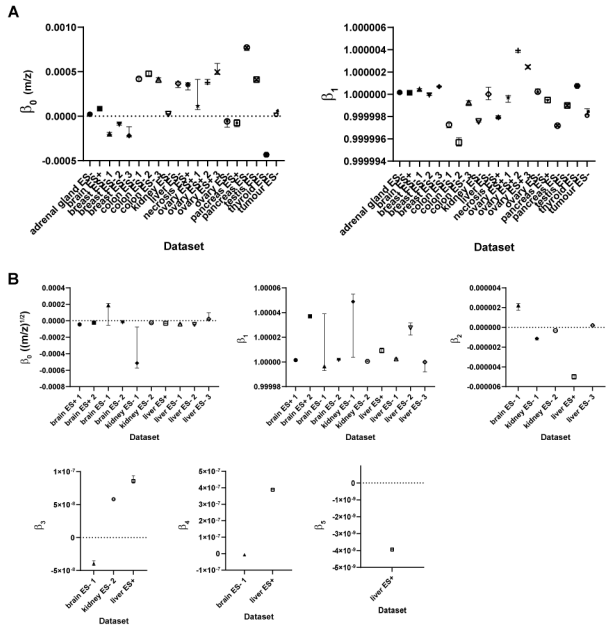

Figure 3 – Distribution of the pixel-wise recalibration regression model coefficients. Each dataset is represented by the median and 0.95 quantile interval of the model coefficient values. Intercept and slope of Orbitrap linear models are shown in (A). In five TOF datasets (B), the models were polynomial with a degree greater or equal to two. In all cases, the linear coefficient β_1 was significantly larger than the intercept β_0, with $\beta_1 M \gg \beta_0$.

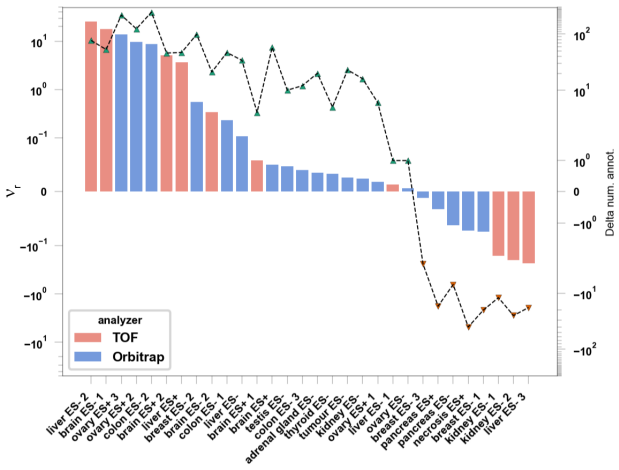

Figure 4 - Results of the METASPACE annotations with an FDR=0.05. The bars represent the relative difference of the number of annotated molecules between the original and recalibrated datasets. The bars corresponding to TOF datasets are colored red, while those corresponding to Orbitrap datasets are colored in blue. The dashed line and triangular markers represent the actual difference between the number of annotated molecules in the original and recalibrated datasets (Eq. 9). The triangles are colored in green if the recalibrated

data generated more annotations than the original data, in red otherwise. In eight cases (five Orbitrap and three TOF), the recalibrated data generated fewer annotations.

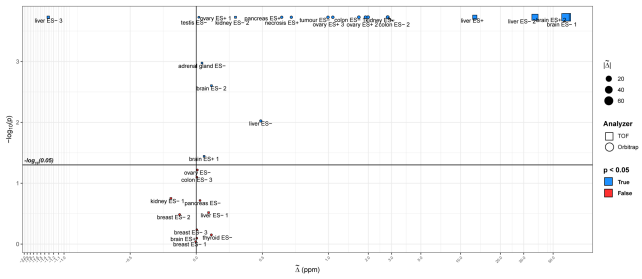

Figure 5 – Scatter plot representing the bootstrapping results for the median difference between the mass errors in the original and recalibrated datasets. Each dataset is characterized by its $\tilde{\Delta}$ value (calculated following the method described in Section S6 of Supplementary Information) and the significance value of the bootstrapping test. Most datasets show a significant reduction of the test masses absolute error, with values up to about 60 ppm.

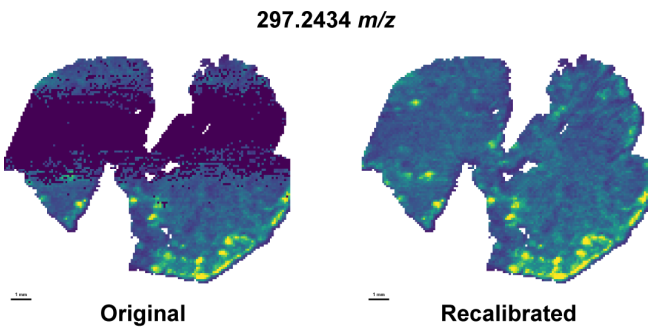

Figure 6 – Example of the effect of recalibration on the quality of an assigned metabolites' spatial distribution by METASPACE for the ORBITRAP colon ES- 2. The original dataset shows a band of missing intensities caused by the masses' nonlinear distribution across the pixels, which is corrected in the recalibrated data.

Finally, we tested the effect of modeling the reference mass shifts as time series. We repeated the pixel-wise recalibration, considering the pixels' spectra independent. We searched the candidate reference using the same method

described in Section 4.3, with a tolerance $W = 20$ ppm for Orbitrap, and $W = 100$ ppm for TOF datasets. The masses that were matched in less than 10 pixels were removed. Thus, in each pixel we applied the pixel-wise recalibration procedure described in Section 4.4. Subsequently, we tested the mass accuracies for the METASPACE masses, using the procedure described in Section S6 of Supplementary Information. Finally, we tested the null hypothesis H_0: $\tilde{\Delta} = 0$, using bootstrapping (number of repetitions equal to 10^4). The results showed that in fifteen datasets there was a significantly (Benjamini-Hochberg corrected p-value < 0.05) smaller relative error when the references were modeled as a time series. In eleven cases, the relative error was smaller in the model with independent pixel-wise regressions. In nine datasets, we did not observe a significant difference (Figure S6). However, the values of $\tilde{\Delta}$ varied between 0.004 and 0.2 ppm in the cases favorable to the independent models, while they varied between 0.03 and 35.2 ppm for the results favorable to the time series models. These results suggested that modeling the reference masses as a time series was more beneficial than considering independent models for each pixel spectrum.

6 Conclusion

Despite its enormous potential to capture the spatial characteristics of the metabolic and proteomic mechanisms in a wide range of biological samples, MSI remains a relatively young technology. Advancements in fundamental data quality control are necessary to transform it into a more reliable approach.

Accuracy of measured masses is particularly challenging in imaging approaches since usually a run consists of tens of thousands of pixels and may require several

455 hours to complete. Thus, the changes of the instrumental condition during this
456 time may invalidate its initial configuration, resulting in mass drifts that correlate
457 with time or pixel order.
458 Here, we have presented a computational workflow that exploits the presence of
459 typical biological molecules in most of the sample spectra and uses them as a set
460 of reference ions for applying a spectra-wise lock mass correction.
461 Using a global set of reference ions for the entire MSI dataset allowed us to test
462 the assignment's validity by visualizing the corresponding spatial distributions.
463 Furthermore, we modeled the mass drifts as a smooth function of the acquisition
464 time, as expected in acquisitions occurring in a controlled environment. Thus,
465 these models represent an additional level of evidence about the assignments'
466 correctness.
467 FFT KDE-based reference match filtering, together with GAMs, proved robust
468 against outliers and false-positive reference ions, for instance, in the presence of
469 close peaks with a complementary spatial distribution to that of the actual
470 reference ions peaks.
471 We showed that the presented approach could improve the mass accuracy,
472 removing the nonlinear fluctuations of the measured masses. Additionally, our
473 approach improved the molecular assignment to the detected peaks, using state-
474 of-the-art molecular annotation methods, such as METASPACE, and their quality
475 in terms of assigned spatial distributions.
476 Modeling the drifts of the reference masses as a time series resulted in better
477 performances compared to considering the pixels' spectra independent.

478 In the dataset with already accurate masses, the recalibration can introduce errors
479 due to its statistical nature. For this reason, it is crucial to evaluate the test
480 masses accuracies before and after performing the recalibration, using the
481 methods described here.
482 We employed pixel-wise recalibration models specific to the physics of the MS
483 analyzer. This is an essential aspect of the procedure since these characteristics
484 influence the statistical properties of the signal generated from the detected ions.
485 It is also important to underline that we designed the mass shift models to be
486 simple, considering only the time-dependency of the acquisition. Unfortunately,
487 this means that several complex aspects of the phenomena driving the mass drifts
488 are not captured. This is a necessary trade-off between generality and reduction
489 of the risk of overfitting. More realistic models may be designed if additional
490 variables are tracked, such as temperature, humidity, just to name a few.
491 We showed that computational models could effectively reduce the mass drifts that
492 affect MSI datasets. However, it is crucial to stress that they can represent a
493 temporary solution to this problem until technological solutions capable of
494 performing an accurate online calibration will not be available.
495 The presented work has limitations. Although we used an extensive list of publicly
496 available molecular masses as references, some peculiar molecules for unknown
497 samples may be missing. For this reason, building a database of accurately
498 identified molecules in previous studies is crucial.
499 Another limitation of the approach is that it can only be applied to the sample-
500 related ROI pixels. Thus, when selecting these pixels, the spatial information
501 necessary to identify possible non-sample-related signals is lost. However, this

difficulty can be easily overcome by applying spatial-aware filters, such as SPUTNIK (Inglese et al., 2019), before the recalibration. In this way, most of the signal correlated with regions outside of the ROI is removed before performing the pixels selection. This aspect is crucial for the correct biological interpretation of the observed molecular spatial patterns.

In the future, we will work to extend the method by including automatic reference ion selection based on the sample properties and will study the possible integration with the usage of external lock mass reference ions.

7 Acknowledgments

We thank Clara L Feider, Meghan P. Friis, Kyana Garza, Emrys Jones, James McKenzie, Nicole Strittmatter, Jialing Zhang, for publicly sharing their datasets on METASPACE.

PI is supported by the NIHR Imperial Biomedical Research Centre (BRC). HH is supported by the UK Dementia Research Institute which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

8 Code and data availability

The Python scripts are freely distributed under the MIT license and can be downloaded from https://github.com/paoloinglese/DESI_MSI_recal, together with the reference masses databases.

9 References

Boskamp, T., Lachmund, D., Casadonte, R., Hauberg-Lotte, L., Kobarg, J. H., Kriegsmann, J., & Maass, P. (2020). Using the Chemical Noise Background in MALDI Mass Spectrometry Imaging for Mass Alignment and Calibration. *Analytical Chemistry*, *92*(1), 1301–1308. https://doi.org/10.1021/acs.analchem.9b04473

Christian, N. P., Arnold, R. J., & Really, J. P. (2000). Improved calibration of time-of-flight mass spectra by simplex optimization of electrostatic ion calculations. *Analytical Chemistry*, *72*(14), 3327–3337. https://doi.org/10.1021/ac991500h

Dória, M. L., McKenzie, J. S., Mroz, A., Phelps, D. L., Speller, A., Rosini, F., Strittmatter, N., Golf, O., Veselkov, K., Brown, R., Ghaem-Maghami, S., & Takats, Z. (2016). Epithelial ovarian carcinoma diagnosis by desorption electrospray ionization mass spectrometry imaging. *Scientific Reports*, *6*(November), 1–11. https://doi.org/10.1038/srep39219

Drexler, D. M., Reily, M. D., & Shipkova, P. A. (2011). Advances in mass spectrometry applied to pharmaceutical metabolomics. In *Analytical and Bioanalytical Chemistry* (Vol. 399, Issue 8, pp. 2645–2653). https://doi.org/10.1007/s00216-010-4370-8

Ellis, S. R., Bruinen, A. L., & Heeren, R. M. A. (2014). A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry. *Analytical and Bioanalytical Chemistry*, *406*(5), 1275–1289. https://doi.org/10.1007/s00216-013-7478-9

Guenther, S., Muirhead, L. J., Speller, A. V. M., Golf, O., Strittmatter, N., Ramakrishnan, R., Goldin, R. D., Jones, E., Veselkov, K., Nicholson, J., Darzi, A., & Takats, Z. (2015). Spatially resolved metabolic phenotyping of breast cancer by desorption electrospray ionization mass spectrometry. *Cancer Research*, *75*(9), 1828–1837. https://doi.org/10.1158/0008-5472.CAN-14-2258

Hamm, G., Bonnel, D., Legouffe, R., Pamelard, F., Delbos, J. M., Bouzom, F., & Stauber, J. (2012). Quantitative mass spectrometry imaging of propranolol and olanzapine using tissue extinction calculation as normalization factor. *Journal of Proteomics*, *75*(16), 4952–4961. https://doi.org/10.1016/j.jprot.2012.07.035

Herzog, R. F. K., & Viehböck, F. P. (1949). Ion source for mass spectrography. In *Physical Review* (Vol. 76, Issue 6, pp. 855–856). American Physical Society. https://doi.org/10.1103/PhysRev.76.855

Hillenkamp, F., Karas, M., Beavis, R. C., & Chait, B. T. (1991). Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Analytical Chemistry*, *63*(24), 1193A-1203A. https://doi.org/10.1021/ac00024a716

Inglese, P., Correia, G., Takats, Z., Nicholson, J. K., & Glen, R. C. (2019). SPUTNIK: An R package for filtering of spatially related peaks in mass spectrometry imaging data. *Bioinformatics*, *35*(1), 178–180. https://doi.org/10.1093/bioinformatics/bty622

Karlsson, O., & Hanrieder, J. (2017). Imaging mass spectrometry in drug development and toxicology. In *Archives of Toxicology* (Vol. 91, Issue 6, pp. 2283–2294). https://doi.org/10.1007/s00204-016-1905-6

La Rocca, R., Kune, C., Tiquet, M., Stuart, L., Eppe, G., Alexandrov, T., De Pauw, E., & Quinton, L. (2021). Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Analytical Chemistry*, *93*(8), 4066–4074. https://doi.org/10.1021/acs.analchem.0c05071

LaBonia, G. J., Lockwood, S. Y., Heller, A. A., Spence, D. M., & Hummon, A. B. (2016). Drug penetration and metabolism in 3D cell cultures treated in a 3D printed fluidic device: assessment of irinotecan via MALDI imaging mass spectrometry. *Proteomics*, *16*(11–12), 1814–1821. https://doi.org/10.1002/pmic.201500524

Li, Y., & Li, L. (2019). Mass Accuracy Check Using Common Background Peaks for Improving Metabolome Data Quality in Chemical Isotope Labeling LC-MS. *Journal of the American Society for Mass Spectrometry*, *30*(9), 1733–1741. https://doi.org/10.1007/s13361-019-02248-w

Mascini, N. E., Teunissen, J., Noorlag, R., Willems, S. M., & Heeren, R. M. A. (2018). Tumor classification with MALDI-MSI data of tissue microarrays: A case study. *Methods*, *151*, 21–27. https://doi.org/10.1016/j.ymeth.2018.04.004

Michelsen, P., & Karlsson, A. Å. (1999). Accurate mass determination of a metabolite of a potential diagnostic imaging drug candidate by high performance liquid chromatography with time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, *13*(21), 2146–2150. https://doi.org/10.1002/(SICI)1097-0231(19991115)13:21<2146::AID-RCM766>3.0.CO;2-3

Nilsson, A., Goodwin, R. J. A., Shariatgorji, M., Vallianatou, T., Webborn, P. J. H., & Andrén, P. E. (2015). Mass spectrometry imaging in drug development. *Analytical Chemistry*, *87*(3), 1437–1455. https://doi.org/10.1021/ac504734s

Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., Kovalev, V., Fuchser, J., Nikolenko, S., Pineau, C., Becker, M., & Alexandrov, T. (2016). FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, *14*(1), 57–60. https://doi.org/10.1038/nmeth.4072

Prideaux, B., & Stoeckli, M. (2012). Mass spectrometry imaging for drug distribution studies. In *Journal of Proteomics* (Vol. 75, Issue 16, pp. 4999–5013). https://doi.org/10.1016/j.jprot.2012.07.028

Ràfols, P., Castillo, E. del, Yanes, O., Brezmes, J., & Correig, X. (2018). Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered

Ag nanolayer. *Analytica Chimica Acta*, *1022*, 61–69. https://doi.org/10.1016/j.aca.2018.03.031

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “kneedle” in a haystack: Detecting knee points in system behavior. *Proceedings - International Conference on Distributed Computing Systems*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Schwamborn, K., Krieg, R. C., Reska, M., Jakse, G., Knuechel, R., & Wellmann, A. (2007). Identifying prostate carcinoma by MALDI-imaging. *International Journal of Molecular Medicine*, *20*(2), 155–159. https://doi.org/10.3892/ijmm.20.2.155

Silverman, B. W. (1982). Algorithm AS 176: Kernel Density Estimation Using the Fast Fourier Transform. *Applied Statistics*, *31*(1), 93. https://doi.org/10.2307/2347084

Silverman, B. W. (2018). Density estimation: For statistics and data analysis. In *Density Estimation: For Statistics and Data Analysis*. CRC Press. https://doi.org/10.1201/9781315140919

Smirnov, K. S., Forcisi, S., Moritz, F., Lucio, M., & Schmitt-Kopplin, P. (2019). Mass Difference Maps and Their Application for the Recalibration of Mass Spectrometric Data in Nontargeted Metabolomics. *Analytical Chemistry*, *91*(5), 3350–3358. https://doi.org/10.1021/acs.analchem.8b04555

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., Merrill, A. H., Murphy, R. C., Raetz, C. R. H., Russell, D. W., & Subramaniam, S. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Research*, *35*(SUPPL. 1), D527–D532. https://doi.org/10.1093/nar/gkl838

Swales, J. G., Dexter, A., Hamm, G., Nilsson, A., Strittmatter, N., Michopoulos, F., Hardy, C., Morentin-Gutierrez, P., Mellor, M., Andren, P. E., Clench, M. R., Bunch, J.,

Critchlow, S. E., & Goodwin, R. J. A. (2018). Quantitation of Endogenous Metabolites in Mouse Tumors Using Mass-Spectrometry Imaging. *Analytical Chemistry*, *90*(10), 6051–6058. https://doi.org/10.1021/acs.analchem.7b05239

Takáts, Z., Wiseman, J. M., Gologan, B., & Cooks, R. G. (2004). Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science*, *306*(5695), 471–473. https://doi.org/10.1126/science.1104404

Vera, C. C., Zubarev, R., Ehring, H., Hakansson, P., & Sunqvist, B. U. R. (1996). A three-point calibration procedure for matrix-assisted laser desorption/ionization mass spectrometry utilizing multiply charged ions and their mean initial velocities. *Rapid Communications in Mass Spectrometry*, *10*(12), 1429–1432. https://doi.org/10.1002/(SICI)1097-0231(199609)10:12<1429::AID-RCM600>3.0.CO;2-W

Wenger, C. D., McAlister, G. C., Xia, Q., & Coon, J. J. (2010). Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass spectrometer. *Molecular and Cellular Proteomics*, *9*(5), 754–763. https://doi.org/10.1074/mcp.M900541-MCP200

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M. A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., … Querengesser, L. (2007). HMDB: The human metabolome database. *Nucleic Acids Research*, *35*(SUPPL. 1), D521–D526. https://doi.org/10.1093/nar/gkl923