RESEARCH

rMSIfragment: improving MALDI-MSI lipidomics through automated in-source fragment annotation

Gerard Baquer[1*], Lluc Sementé[1], Pere Ràfols[1,2,3*], Lucía Martín-Saiz[4], Christoph Bookmeyer[1,5], José A. Fernández[4], Xavier Correig[1,2,3] and María García-Altares[1,2]

Abstract

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) spatially resolves the chemical composition of tissues. Lipids are of particular interest, as they influence important biological processes in health and disease. However, the identification of lipids in MALDI-MSI remains a challenge due to the lack of chromatographic separation or untargeted tandem mass spectrometry. Recent studies have proposed the use of MALDI in-source fragmentation to infer structural information and aid identification. Here we present rMSIfragment, an open-source R package that exploits known adducts and fragmentation pathways to confidently annotate lipids in MALDI-MSI. The annotations are ranked using a novel score that demonstrates an area under the curve of 0.7 in ROC analyses using HPLC–MS and Target-Decoy validations. rMSIfragment applies to multiple MALDI-MSI sample types and experimental setups. Finally, we demonstrate that overlooking in-source fragments increases the number of incorrect annotations. Annotation workflows should consider in-source fragmentation tools such as rMSIfragment to increase annotation confidence and reduce the number of false positives.

Keywords In-source fragmentation, In-source decay, Lipids, Annotation, Mass spectrometry imaging, MALDI, Computation, Cheminformatics, Bioinformatics

Introduction

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) is an analytical technique used in biochemical and clinical studies to reveal the chemical composition and spatial information of organic tissues [7, 25, 28, 38, 53]. It provides valuable information in many applications, including the understanding and diagnosis of complex diseases such as cancer [10, 13, 15, 29, 32, 35, 39], diabetes [6, 22, 33, 57], Alzheimer's [27, 30, 31] and infectious diseases [36, 54]. In particular, the study of lipids is pivotal, as they play important roles in different pathways in health and disease [34].

Despite the surge of MALDI-MSI's popularity, associating each mass-to-charge (*m/z*) signal with univocal molecular identifications remains challenging: (1) samples include thousands of molecules; (2) each molecule is responsible for several MS signals (e.g. isotopes, adducts and, in-source fragments); and (3) isomers and isobars cannot be resolved using only MS1 [8].

*Correspondence:
Gerard Baquer
gerard.baquer@alumni.urv.cat
Pere Ràfols
pere.rafols@urv.cat
[1] Department of Electronic Engineering, University Rovira I Virgili, Tarragona, Spain
[2] Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain
[3] Institut D'Investigacio Sanitaria Pere Virgili, Tarragona, Spain
[4] Department of Physical Chemistry, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Leioa, Spain
[5] Institute of Hygiene, University of Münster, Münster, Germany

Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 2 of 14

Traditional mass spectrometry techniques rely on chromatographic separation (LC–MS, GC–MS) for sample simplification [63]. However, MALDI-MSI does not include such separation steps. Alternatively, tandem mass spectrometry can augment the depth of the chemical analysis by providing fragmentation information on single molecules. Many MALDI-MSI instruments are equipped with tandem-MS capabilities (Bruker's ultrafleXtreme, Thermo Scientific's MALDI LTQ Orbitrap XL, or Waters' MALDI SYNAPT G2-Si) but untargeted imaging MS/MS is not routinely feasible due to (1) prohibitive running times, (2) limited parental ion selectivity and intensity, and (3) increased data size and complexity. For all these reasons, untargeted fragmentation of all ions in a sample is only possible using highly specialized instrumental setups [18, 24].

Traditionally, In-Source Decay (ISD) or In-Source Fragmentation (ISF) (i.e. the natural and unavoidable generation of fragments inside the MALDI ion source) has been considered an undesired artifact and thus minimized [26]. ISD depends mainly on the chemical structure of the analyte and ionization conditions such as ionization temperature or voltage [26]. It can be problematic in the study of lipids, as several fragmentation pathways lead to isobaric lipid species. As an example, phosphatidic acid (PA) fragments can be produced in-source from their phosphatidylserine (PS) counterparts, and phosphatidylcholine (PC) in-source fragments are isobaric to endogenous phosphatidylethanolamine (PE) species [26]. These known lipid fragmentation pathways result in falsely low concentrations of lipids suffering from ISD and falsely high concentrations of lipids overlapping with isobaric in-source fragments. Additionally, if not properly annotated and removed, in-source fragments can yield an increased number of incorrect annotations using common MALDI-MSI annotation tools such as LipoStar, METASPACE, and rMSIannotation [42, 49, 55].

Nevertheless, recent studies have advocated for the use of well-characterized MALDI-ISD as a fast way of obtaining complementary fragmentation information to assist identification in the analysis of large molecules [5, 12, 37] and even lipids [58]. Some of these studies use automated protein ISD annotation tools like ProteinProspector (Baker, P.R. and Clauser, K.R. http://prospector.ucsf.edu) or DataAnalysis (Bruker Daltonics) but heavily rely on manual annotation. The use of ISD to strengthen identification has also been applied in MALDI-MSI in the field of top-down proteomics. Debois et al. demonstrated the use of ISD for in-situ de novo sequencing of several proteins on a porcine eye lens and a mouse brain slice [14, 64]. Similarly, Ait-Belkacem et al. [2] used ISD to identify several tumorigenic proteins in glioblastoma mouse brain tissue with MALDI-MSI. Franceschi et al.

proposed a semi-automated workflow for in-source fragmentation annotation based on initial manual annotation of parental metabolites followed by an Intensity Correlation Analysis to find ions with a high spatial correlation and thus assumed to be in-source fragments. They used this approach to image flavonols and dihydrochalcones in golden apple samples [19].

Recently, Garate et al. [20], reported the MALDI-MSI in-source fragmentation pathways and adduct formation of the 17 main lipid classes (Additional file 1: Figure S1 and Table S1-S2). Here we propose rMSIfragment, a software solution that exploits these known in-source fragmentation pathways to increase confidence in lipid annotations. Our novel ranking score combines the times a given lipid has been found in the dataset (adducts and in-source fragments) and their spatial correlation to filter out unlikely lipids. After validation with HPLC and 2 different Target-Decoy approaches, rMSIfragment demonstrates an Area Under the Curve (AUC) of over 0.7 on multiple sample types and experimental conditions. We also find that ISD-agnostic annotation tools like METASPACE can falsely annotate in-source fragments as endogenous lipids.

Algorithm description

Input and output format

As input, the user should provide the ppm tolerance for exact mass searches against the database and an MSI dataset in the rMSIproc [45] peak matrix format ([# pixels] x [# *m/z*] intensity matrix). Refer to the public repository (https://github.com/prafols/rMSIproc) for instructions on how to convert profile and centroid mode.imZML files to the peak matrix format. The theoretical fragmentation pathways and adducts for each lipid class [20] (Additional file 1: Figure S1 and Table S1-S2) are already included in rMSIfragment.

The algorithm produces a table of annotations (lipid annotation, number of carbons, number of double bonds, and molecular formula) for each of the monoisotopic masses found in the data set. Each annotation is associated with a likelihood score to assist manual curation. The resulting table can be exported as a.csv file (Additional file 1: Table S3). Additionally, the annotation results can be used to generate a molecular network graph.

Database search and likelihood score

As reported by Garate et al. [20] ISD can produce lipid fragments that overlap with endogenous lipids (Fig. 1A). rMSIfragment estimates the likelihood of each lipid by computing two metrics: lipid occurrences (LO) and spatial correlation (C). For each lipid in the library, rMSIfragment performs an exact mass search considering the adducts and in-source fragmentation pathways of the
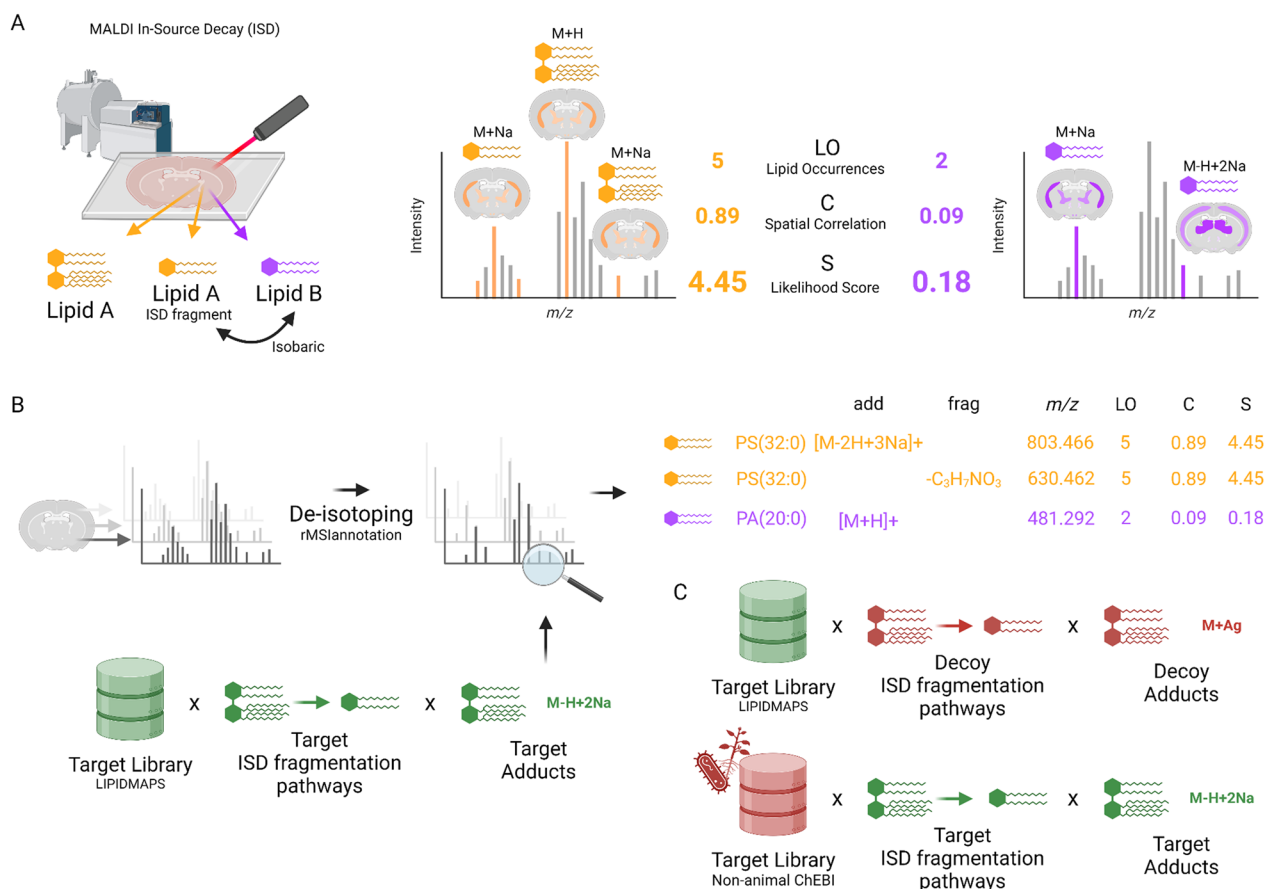
Fig. 1 Main algorithmic foundations for the annotation of in-source fragments. **A** ISD can generate in-source fragments that overlap with endogenous lipids. The use of Lipid Occurrences (LO) and Spatial Correlation (**C**) allows rMSIfragment to rank the likelihood of isobaric lipid annotations. **B** General rMSIfragment flux diagram. **C** Two alternative decoy libraries based on highly unlikely adducts and fragmentation pathways (top) and non-animal/xenobiotic compounds (bottom)

corresponding lipid class given by Garate et al. [20]. LO is the number of m/z features found. C is the mean spatial correlation (Pearson's r) of all possible distinct pairs of m/z features found. The final ranking score is computed using the following equation:

$$S = LO \cdot (1 + C)$$

Figure 1B summarizes the complete workflow. Initially, rMSIannotation [49] is used to perform deisotoping (removal of m/z features considered to be isotopes). Later, all remaining m/z features are matched against LIPIDMAPS considering all theoretically feasible adducts and fragments. Each lipid class has a different list of theoretical adducts and in-source fragmentation pathways. The results of the annotation are stored in an R "data.frame" that can be exported to.csv for manual inspection.

During validation, rMSIfragment uses two alternative decoy libraries (Fig. 1C) to estimate the False Discovery Rate (FDR) and performance (ROC AUC) of the annotation. The first decoy library is formed by generating highly unlikely adducts and fragmentation pathways. The second decoy library replaces LIPIDMAPS with a list of compounds found in non-animal specimens and thus highly unlikely to be found in animals. Both approaches will be described more in detail in Sect. "rMSIfragment shows high performance in a target-decoy validation".

Although by default, LIPIDMAPS [40] is used to perform database searches, the user can adjust the software to use any publicly available database (e.g. HMDB, MoNA, METLIN or, NIST) or an in-house made compound database.

Results

rMSIfragment matches HPLC–MS annotations in human nevi samples.

As initial validation, we challenged rMSIfragment with the annotation of human nevi samples G1-G15. In a previous study [20], extracts from the same tissue were analyzed by HPLC–MS, producing a list of identified lipids

Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 4 of 14

in positive-ion and negative-ion polarities (Garate et al. Additional file 1: Table S2 and S3). We use these lists to estimate the performance of our automatic annotation tool.

Figure 2 shows the estimated performance of rMSIfragment when comparing its automatic annotations (based only on MSI data) to the list of HPLC–MS-identified lipids. To control the number of rMSIfragment annotations we propose retaining the top N annotations with the highest S score for each *m/z* feature.
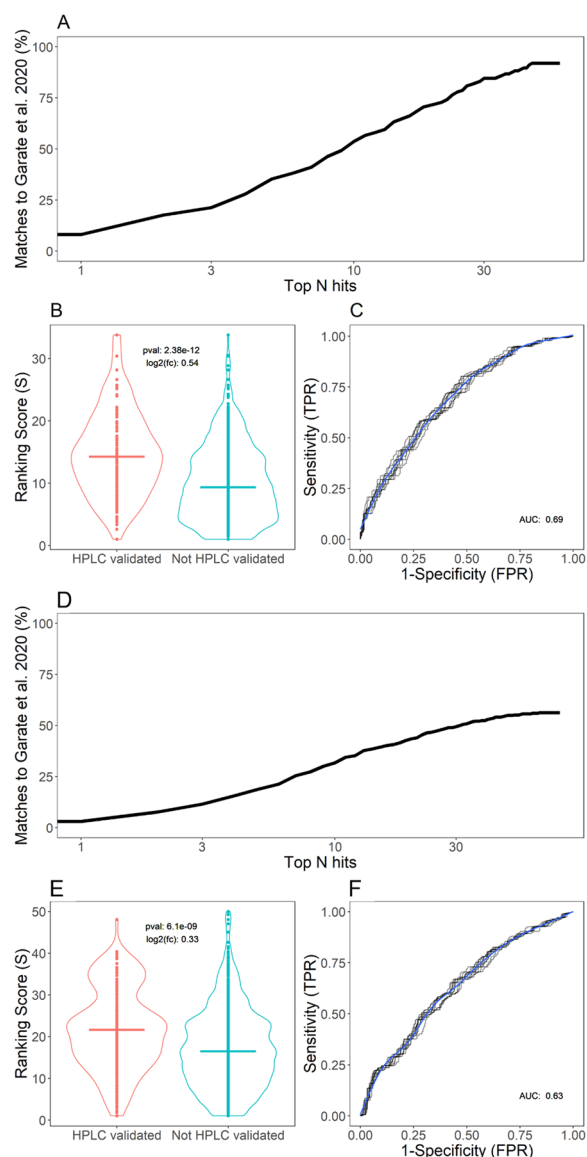

Fig. 2 Validation of rMSIfragment against HPLC validated annotations reported by Garate et al. **A** Percentage of HPLC matches, **B** S score distribution, and **C** ROC curve for the negative polarity datasets (G9-G15). **D** Percentage of HPLC matches, **E** S score distribution, and **F** ROC curve for the positive polarity datasets (G1-G8)

In the samples acquired in negative-ion polarity (G9-G15) (Fig. 2A) rMSIfragment retrieves 91.81% of the HPLC-validated annotations reported in the original publication. To retrieve at least 50% of the HPLC annotations we need to retain the top 9 hits with the highest S score in each *m/z* feature. When only keeping the top 5 annotations rMSIfragment returns 35.92% of the HPLC-validated annotations. Figure 2B shows that the HPLC-validated annotations obtain a significantly higher S score than the annotations not validated by HPLC (p-value < 0.01, t-test). Additionally, Fig. 2C shows an ROC area under the curve (AUC) of 0.7 when using the HPLC annotations as validation.

In positive-ion polarity (G1-G8) (Fig. 2D) rMSIfragment retrieves 56.18% of all the HPLC-validated annotations. Despite the lower performance when compared to the negative-ion polarity samples, rMSIfragment is still capable of retrieving 18.56% of the annotations when focusing on the top 5 annotations for each MS feature. HPLC-validated annotations still report a significantly higher S score (p-value < 0.01, t-test) (Fig. 2E) but present a slightly worse performance at 0.63 AUC.

An alternative representation of the same results is shown in Additional file 1: Figure S2. Instead of selecting the top N annotations per *m/z* feature, we set a global likelihood score (S) threshold. A complete list of rMSIfragment annotations is provided in Additional file 2: Tables S5 and Additional file 3: Table S6.

The current validation assumes that all HPLC-validated lipids should be found in the samples and the rest should not. HPLC-MSI and MALDI-MSI are fundamentally different analytical techniques [6] and some lipids found with one technique could not be found with the other. We consider that HPLC annotations can accurately estimate the number of true positives (present and matched by rMSIfragment) and false negatives (present but not matched by rMSIfragment), but they may fail to estimate true negatives (not present and not matched by rMSIfragment) and false positives (not present but matched by rMSIfragment). Figure 3 highlights some of these differences in the annotation of three different phosphatidylcholines. PC 34:1 was found by both HPLC–MS and MALDI-MSI, PC 33:1 showed a similarly high S score but was not found by HPLC–MS, and PC 38:3 was found by HPLC–MS but shows a low S score.

Performances of 0.7 AUC might be considered modest in other fields. For instance, scores above 0.85 AUC are common in chest X-ray classification [50, 60]. However, performance depends on task complexity [51]. As rMSIfragment is the first tool addressing in-source fragmentation in MALDI-MSI, a direct benchmark against comparable tools is not possible. When compared to top MSI annotation tools like MSM [42] and
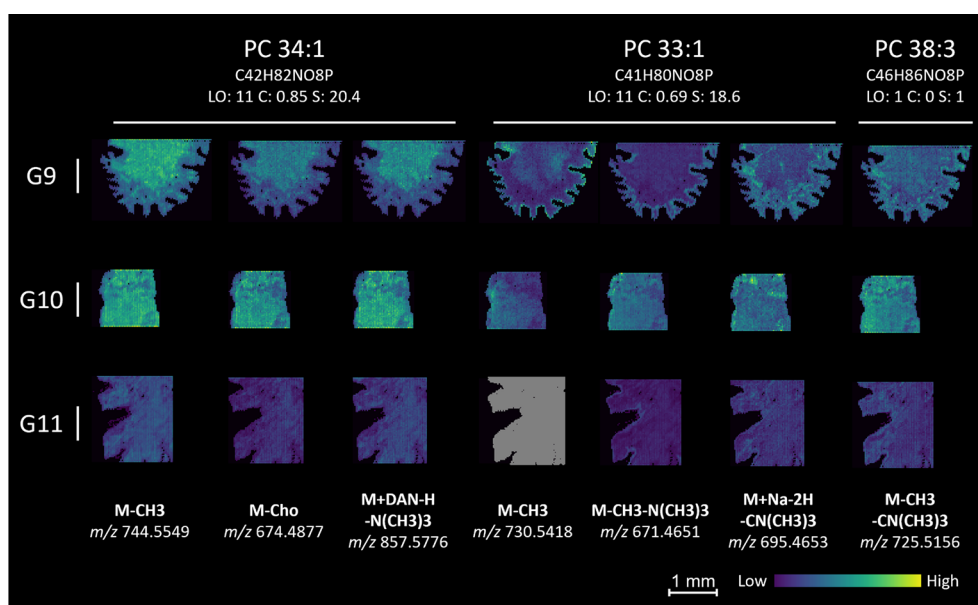
Fig. 3 Example annotations by rMSIfragment. PC 34:1 reports a high S score and was also found by HPLC–MS in the lipid extracts. PC 33:1 shows a similarly high S score (with worse spatial correlation) but was not found by HPLC–MS. PC 38:3 was found by HPLC–MS but shows a low S score

METASPACE-ML [56], our 0.7 AUC performance ranks in the top quartile [56]. While not directly comparable due to their different focus, these performances offer a reliable estimate of the state of MALDI-MSI molecular annotation. Overall, rMSIfragment demonstrates confident annotation of lipid in-source fragments, attaining performances that approach the theoretical upper limit of current MSI annotation tools.

rMSIfragment shows high performance in a target-decoy validation

In an attempt to gauge and overcome the limitations of the HPLC validation, we propose a second validation based on a target-decoy search strategy, a commonly used approach in MSI [21, 42]. This strategy runs rMSI-fragment on the same MSI data using our target library (LIPIDMAPS) and a decoy library containing compounds that should not be found in the sample. The decoy library matches the size and distribution of masses of the target library, to ensure that randomly generated masses are equally likely to hit either of the two databases. The rate of matches in the decoy can then be used to estimate measures such as true positives, true negatives, false positives, false negatives, and False Discovery Rate (FDR) [17].

Figure 4 shows the results of the target-decoy validation on the human nevi datasets (G1-G15) using a decoy library composed of highly unlikely adducts and fragmentation pathways, an approach adapted from pySM [42]. The classification performance obtained a value of 0.72 AUC for the negative-ion polarity datasets (Fig. 4A) and 0.6 AUC for the positive-ion polarity datasets (Fig. 4C). These results are consistent with the performances obtained in HPLC validation. When retaining the top 10 matches for each MS feature the FDR is estimated to be 14.93% in negative-ion polarity (Fig. 4B) and 34.24% in positive-ion polarity (Fig. 4D). Similarly, when only retaining the top 5 matches per MS feature the FDR is 4.5% and 17.95% respectively.

As extra validation, we created a second decoy library, a subset of ChEBI [23] containing only metabolites found in non-animal specimens (plants, algae, fungi, and bacteria) and xenobiotics. The results are summarized in Additional file 1: Figure S3. The performance on the negative-ion polarity samples is consistent with the one estimated in previous approaches (0.73 AUC). In the positive-ion polarity samples, the performance is estimated to be higher than in previous validations (0.75 AUC). In both cases, the FDR is estimated to be under 10% when retaining the top 10 matches.

The two decoy libraries provided comparable estimations of performance and FDR. Nevertheless, the definition of a decoy library based on non-animal/xenobiotic compounds (Additional file 1: Figure S3) is sample-dependent. The use of highly unlikely adducts and fragments, on the other hand, has already been discussed [42] and can be assumed to be much more sample-independent. Additionally, its performance estimates are more conservative and closely match the results obtained in the HPLC validation. For all these reasons,
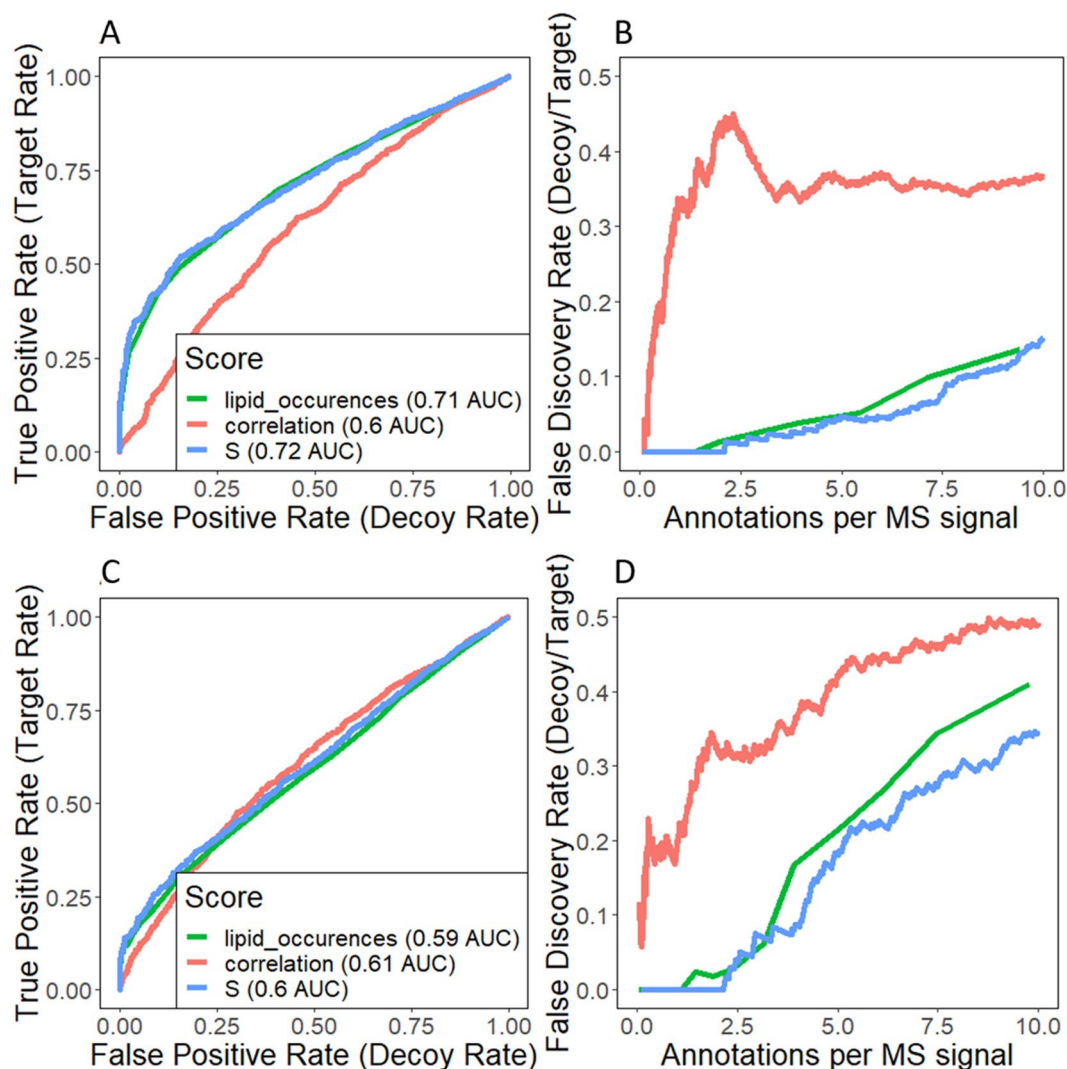
Fig. 4 Performance estimation of the ranking scores proposed using a target-decoy validation approach. The decoy database is composed of adducts and fragmentation pathways highly unlikely to be found in the lipid classes considered. Lipid occurrences (LO): number of times a given lipid is found (including parental adducts and in-source fragments); Spatial correlation (C): The weighted mean Pearson's correlation of all *m/z* features annotated as the same lipid. Final ranking score (S): LO * C. **A** ROC curve and **B** FDR estimation for the negative polarity datasets (G9-G15) **C** ROC curve and **D** FDR estimation for the positive polarity datasets (G1-G8)

further validations use a decoy library composed of highly unlikely adducts and fragments.

These results further confirm that rMSIfragment can confidently annotate lipids and their fragments.

rMSIfragment can be used under multiple experimental conditions

To determine its applicability to other experimental conditions we challenged rMSIfragment with the annotation of 12 different publicly available datasets from METAS-PACE (Datasets M1-M12) [4]. Since the real lipid composition of these datasets is unknown, we use the target

and decoy approach based on highly unlikely adducts and fragments to estimate the performance of rMSIfragment in each dataset. Figure 5 summarizes the results. The performances ranged between 0.65 AUC and 0.84 AUC ($\mu = 0.74$ AUC, $\sigma = 7.5\%$) (Fig. 5A). These performances are comparable to the performance on human nevi samples validated with HPLC.

Additionally, when grouping the datasets according to different experimental parameters (ionization polarity, matrix, tissue, analyzer, specimen, or mass range) the differences in performance were not found to be significant (p-value > 0.1, f-test). This demonstrates that
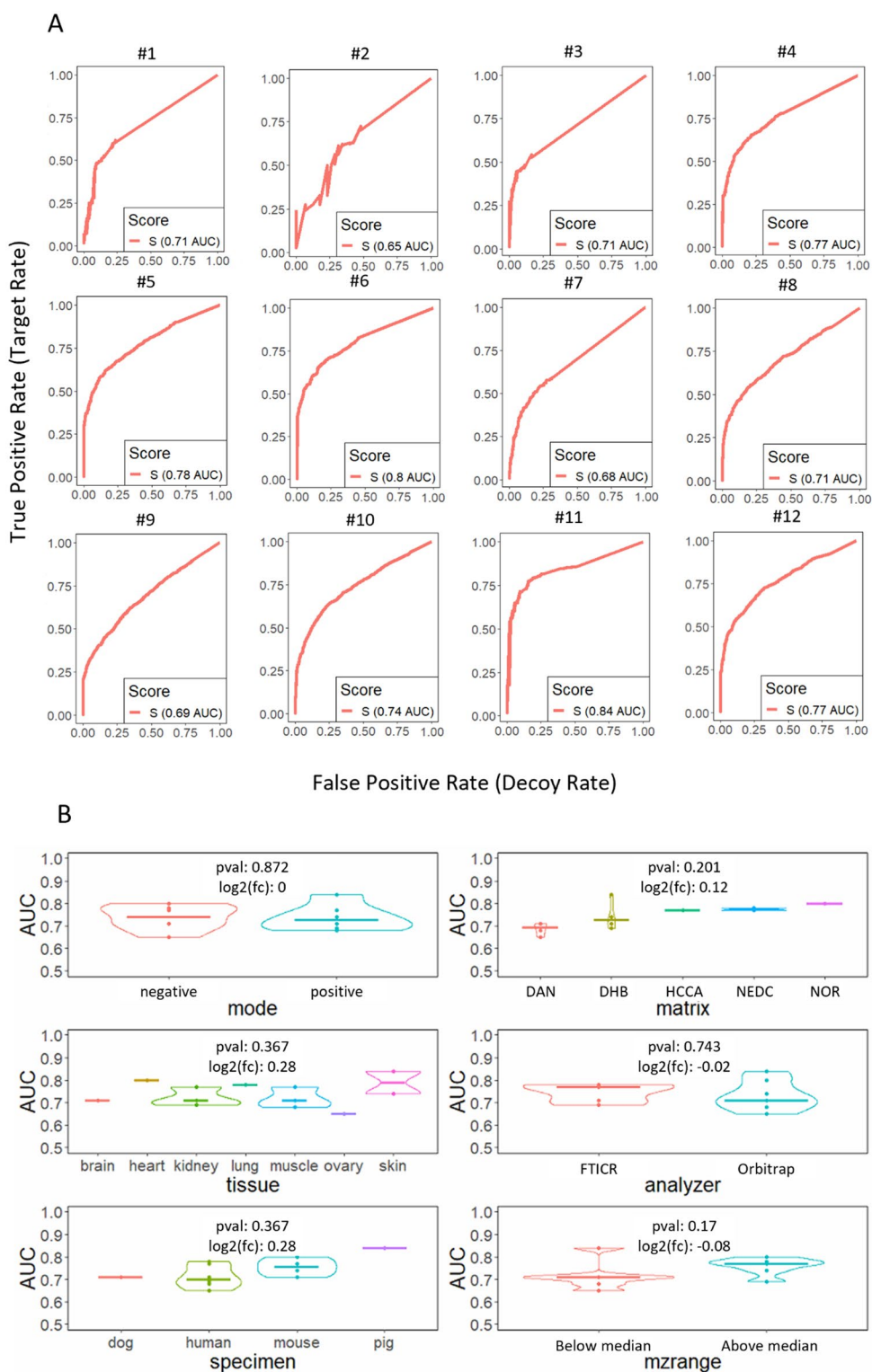
Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 7 of 14

Fig. 5 Target and decoy validation on 12 datasets publicly available on METASPACE [4]. **A** ROC curves and Area Under the Curve (AUC) for each dataset M1-M12. **B** Statistical significance tests of performance (AUC) differences across different biological (tissue and specimen), sample preparation (matrix), and instrumental (ion polarity, analyzer, *m/z* range) parameters

rMSIfragment applies to a wide range of experimental conditions.

Annotation software must take into account in-source fragmentation

Major automatic annotation tools such as pySM [42], LipostarMSI [55] or rMSIannotation [49] completely overlook in-source fragmentation and almost exclusively focus on protonated and alkali ions. Ignoring in-source fragmentation during annotation could potentially lead to false annotations [8]. This is a particular concern in lipidomics where several lipids fragment in-source and become isobaric to other lipids [20].

To assess the impact of in-source fragmentation on lipid annotation we annotate datasets M1-M12 with METASPACE [4], only considering traditional MALDI adducts ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$ in positive-ion polarity and $[M-H]^-$, $[M+Cl]^-$ in negative polarity). We then annotate the same datasets using rMSIfragment, considering all adducts and fragmentation pathways specified in Additional file 1: Fig. S1 and Table S1. The same version of LIPIDMAPS is used with both tools.

Figure 6A summarizes the overall comparison of annotations between the 2 tools. On average, 48.6% of the annotations returned by METASPACE are also found with rMSIfragment. Interestingly, crossing the annotations also allows us to determine that, on average, 54.21% of METASPACE annotations are overlapped with at least one in-source fragment found by rMSIfragment. Additional file 1: Figure S4 shows the same results color-coded based on the different sample and experimental parameters.

To exemplify this overlap we highlight the annotation of *m/z* 887.57 from a human lung cancer biopsy prepared with NEDC and analyzed in negative polarity with an FTICR (Dataset M5) (Fig. 6B). Both tools reliably annotate this *m/z* feature as PI 38:3 (M-H)$^-$. In the same dataset, rMSIfragment also finds 3 adducts ($[M-OH]^-$, $[M-CH_3]^-$, $[M+Na-2H]^-$) and 3 in-source fragments ($[M-CH_3-NH_2]^-$, $[M-H-C_4H_{10}O_5]^-$, $[M-2H_2O-NH_2]^-$) with high spatial correlation to the parental ion. Two of these in-source fragments ($[M-H-C_4H_{10}O_5]^-$, $[M-2H_2O-NH_2]^-$) are overlapped with 2 METASPACE annotations ($[PA 40:5 M-H]^-$, $[PS 40:5 M-H]^-$).

These results are not a comparison between tools but rather a quantification of the negative impact that in-source fragmentation has on automatic annotation tools. Annotation tools in MSI need to take into account in-source fragmentation. Due to the fundamental limitations of MS1, software tools are unable to resolve an endogenous PA from an isobaric PA originating from the in-source fragmentation of its PS counterpart. However,

rMSIfragment mitigates the issue by (1) making the user aware of the potential overlap and (2) giving a higher score to the lipid found forming other adducts and in-source fragments.

rMSIfragment provides a molecular network to visually interpret the results

Figure 7 shows an example exploration of the annotation results using the rMSIfragment GUI. The top 3 annotations for *m/z* 744.55 are shown as individual molecular networks including adducts (purple) and in-source fragments (yellow) (Fig. 7A). The number of lipid occurrences (LO) and spatial correlation (C) are shown as the main metrics to filter out unlikely lipids. When selecting the desired molecular network the user can view the spatial distribution of all adducts and in-source fragments.

Discussion and conclusions

We have demonstrated the performance of rMSIfragment on 15 human nevi datasets with two orthogonal approaches: (1) matching its annotations to HPLC and (2) using a target-decoy approach. Both approaches yield similar performance estimates (0.7 AUC and 0.6 AUC for the samples acquired in negative and positive polarity). rMSIfragment is the first tool addressing in-source fragmentation and cannot be directly benchmarked. However, our results rank in the upper quartile of performances reported by leading tools in metabolite annotation [42, 56], a closely related task.

We also annotated 12 publicly available datasets covering a wide combination of samples and experimental setups. The performances obtained are comparable and often better than the ones obtained on the human nevi datasets. Additionally, rMSIfragment shows a high lipidome coverage overlap comparable to available annotation tools like METASPACE [4]. These findings suggest that rMSIfragment can consistently annotate lipids and their in-source fragments under different experimental conditions.

One key highlight of our study is the importance of considering in-source fragmentation pathways when performing molecular annotation. We have found that overlooking ISD pathways can lead to up to 75% of the reported lipid annotations to be overlapped with at least one in-source fragment. rMSIfragment mitigates this issue through two mechanisms: (1) unlikely lipids with low occurrences (number of adducts and in-source fragments) and poor spatial correlation are filtered out, and (2) the user is aware of the overlap, allowing them to be cautious with their interpretation of the automated annotations.

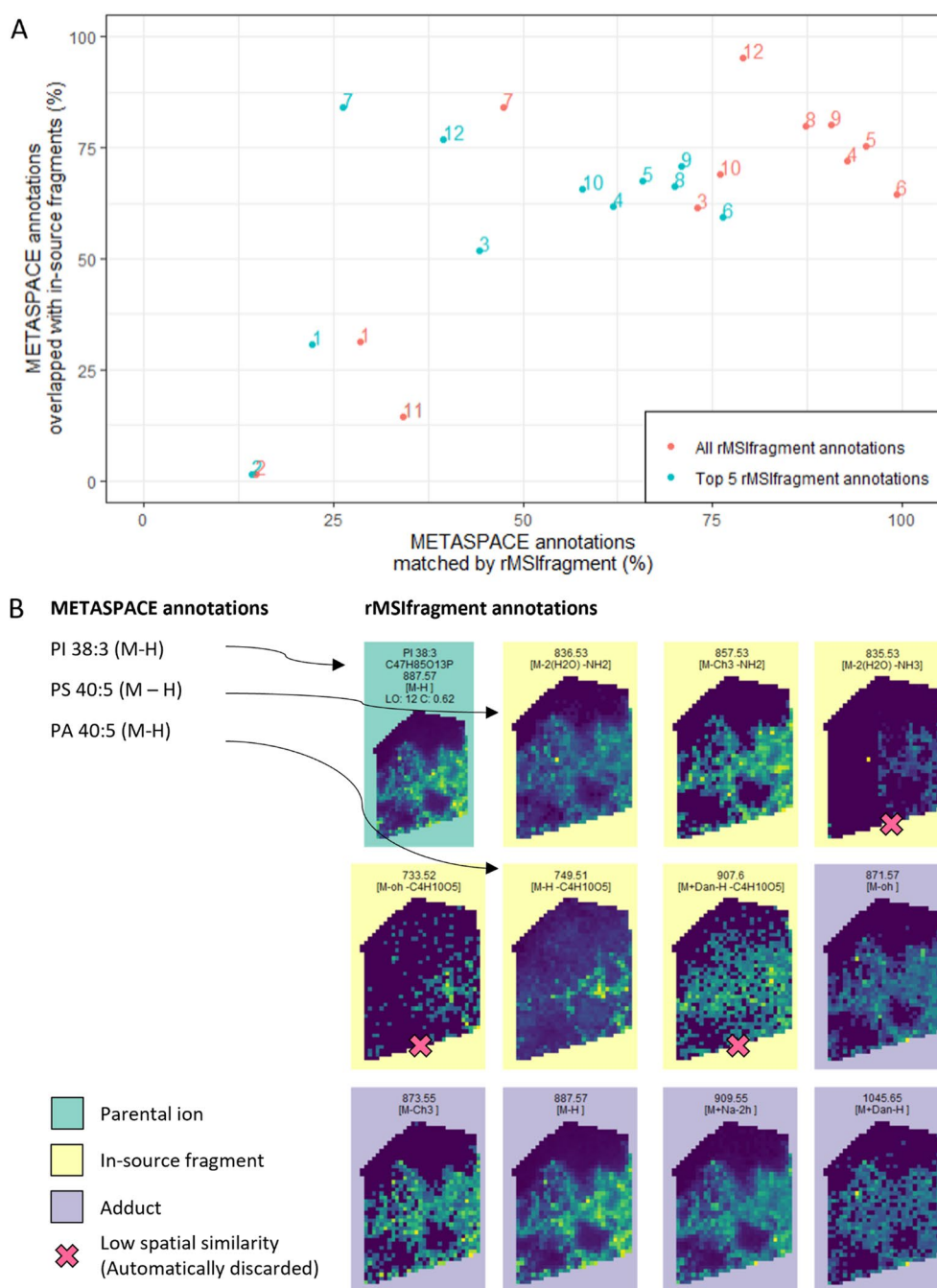We propose three avenues to increase the performance of the future generation of MALDI-MSI in-source

Fig. 6 Comparison of annotation results between rMSIfragment and METASPACE **A** Bulk comparison using 12 datasets publicly available in METASPACE. The horizontal axis shows the percentage of METASPACE annotations that are matched by rMSIfragment. The vertical axis indicates the percentage of METASPACE annotations that are overlapped with at least one in-source fragment annotated by rMSIfragment. The standard FDR threshold of 0.2 was used for METASPACE annotations. rMSIfragment annotations without any threshold (red) and retaining the top 5 annotations per MS feature (blue). **B** Example comparison for a human lung biopsy (Dataset M5) where *m/z* 887.57 is annotated by both tools as PI 38:3 (M + H) (C47H85O13P). In the same dataset, rMSIfragment also finds 3 adducts and 3 in-source fragments with high spatial correlation to the parental ion. Two of these in-source fragments are overlapped with 2 METASPACE annotations ([PA 40:5 M-H]⁻, [PS 40:5 M-H]⁻)

fragmentation annotation tools: (1) leveraging known ion suppression effects between different lipid classes, (2) compiling MALDI-ISD or MALDI-MS/MS libraries and, (3) deploying Machine Learning (ML) and Deep Learning (DL) models.

Ion suppression effects strongly favor certain classes of lipids, difficulting the analysis of suppressed species
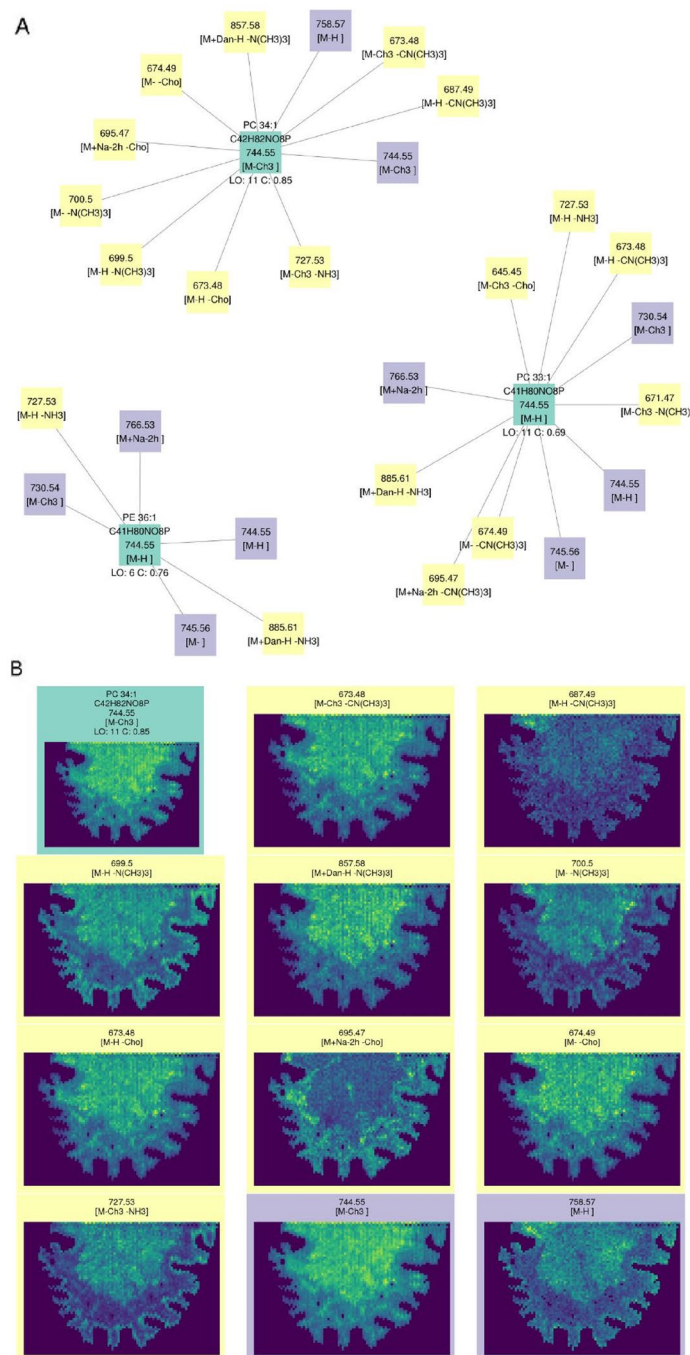
Fig. 7 Example application and interpretation of rMSIfragment results on Dataset G9. **A** Top 3 annotations for *m/z* 744.55. Each parental annotation (green) is the center of a network including all adducts (purple) and in-source fragments (yellow) annotated in the sample. **B** Spatial representation of the top annotation (PC 34:1 [M-CH3])

[11]. In positive polarity, PC species display stronger signals than other lipids (PE, PS, PG, or PI). In negative polarity, the effect is reversed and PC species show lower signals than other lipids. These interactions have been characterized in the past [11] and could be leveraged to define a new ranking score to filter out unlikely lipid annotations.

Previous LC–MS studies [61] induce increased ESI in-source fragmentation to yield fragmentation patterns similar to those present in MS/MS libraries like

Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 11 of 14

METLIN [52] to aid in molecular identification. In a preliminary exploration, we concluded that public MS/MS libraries were not directly applicable to MALDI-MSI for two reasons: (1) MALDI and ESI are not always directly comparable, and (2) the fragmentation of common MALDI adducts such as $[M+Na]+$ and $[M+K]+$ are underrepresented in MS/MS libraries (< 10%). The compilation of MALDI-ISD libraries could overcome these limitations. This community-wide effort would help better characterize MALDI-ISD in a wide range of biomolecules. Alternatively, a MALDI-MS/MS library, which may be a more urgent interest of the MALDI community, could already provide enough information. These two libraries would be invaluable tools to foster the development of the next generation of ISD annotation algorithms and models in MALDI-MSI.

Finally, in spite of their success in MALDI-MSI tasks like tumor classification [9], clustering [62], and image registration [44], ML and DL models have yet to be consistently deployed for molecular annotation [3, 8]. The performance of rMSIfragment could potentially be improved l,leveraging ML to suggest optimal functions of our predefined metrics (LO and C to compute the final score (S, following the strategy proposed by METAS-PACE-ML [56]. Taking a step further, new metrics could be defined to better capture spectral and spatial similarities of in-source fragments. However, DL has only attained modest improvements in MS image colocalization [41]. From our perspective, creating DL models for MALDI in-source fragmentation holds the greatest promise for enhancing annotation performance, building upon the achievements of in-silico MS/MS fragmentation tools like Metfrag [47], CFM-ID (F. [57, 59], and Sirius [16].

In conclusion, neglecting in-source fragmentation leads to an increased number of false lipid annotations. rMSIfragment mitigates this effect by prioritizing annotations of lipids found forming multiple adducts and in-source fragments.

Materials and methods

A total of 27 different datasets were used to validate this study. Human nevi samples acquired in positive (G1-G8) and negative (G9-G15) ion polarity from a previous study [20] were used to perform HPLC validation and determine the best target-decoy strategy for further validation. Publically available METASPACE [4] datasets M1-M12 were used to demonstrate the applicability of rMSIfragment to different sample types and experimental conditions. Additional file 1: Table S4 summarizes the main processing parameters for each of the 27 datasets.

Sample preparation
Human nevi tissue sections G1-G15 were already used in a previous study [20]. The Euskadi Ethics Committee approved the study protocol and it conformed to the Helsinki Declaration. The biopsies were embedded in OCT and sectioned at 16 μm thickness. The sections were covered with MALDI matrices 2-Mercaptobenzothiazole (MBT) and 1,5-Diaminonaphthalene (DAN) for positive and negative-ion modes respectively. Both matrices were sublimated onto the sample using an Ace Glass 8023 Glass Sublimator.

MALDI-MSI acquisition
Mass spectra were acquired using an LTQ-Orbitrap XL mass spectrometer (ThermoFisher, MA, USA), equipped with a custom MALDI source with an N2 laser [20]. Low laser energy was used to prevent excessive fragmentation. The spectrum at each pixel is the result of averaging two micro-scans of 10 shots. Positive-ion polarity experiments had an *m/z* range of $480-1100$ Da while negative-ion polarity experiments had a range of $550-1200$ Da. The mass resolving power was set to 30 000 at *m/z* 400. The step size was 25 μm.

MSI data processing
Datasets G1-G15, originally in.RAW format (Thermo Fischer), were exported to.mzML using ProteoWizard msConvert [1], and later converted to.imzML [48] using imzMLConverter [43]. The software rMSIproc [45] was used to process the data and generate a peak matrix in centroid mode. The default processing parameters were used. The Signal-to-Noise Ratio (SNR) threshold was set to 5 and the Savitzky– Golay smoothing had a kernel size of 7. Peaks appearing in less than 5% of the pixels were filtered out. Peaks within a window of 6 data points or scans were binned together as the same mass peak.

Datasets M1-M12, already in centroid mode.imZML, were imported using rMSIproc [45].

Deisotoping was performed using rMSIannotation [49]. No data normalization was performed. Data were visualized and explored using rMSI [46].

Statistical significance is tested using a t-test or an f-test for the comparison of 2 and 3 + groups respectively.

Supplementary Information

Additional file 1: Figure S1. Lipid fragmentation pathways. Reproduced with the permission of Garate et al. 2020. **Figure S2.** Automatic annotation with rMSIfragment validated with HPLC in human nevi samples (Garate et al. 2020). Percentage of HPLC validated matches (Garate et al. 2020) against increasing ranking score (S) threshold (blue). **(A)** Samples G9-G15 (negative-ion polarity). **(B)** Samples G1-G8 (positive-ion polarity). **Figure S3.** Performance estimation of the Ranking scores proposed using

Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 12 of 14

a Target Decoy Validation approach. The Decoy database is composed of metabolites and lipids unlikely to be found in non-animal specimens (plants, algae, fungi, and bacteria) and xenobiotics. (A) ROC and (B) FDR estimation on samples G9-G15 (negative-ion polarity). (C) ROC and (D) FDR estimation on samples G1-G8 (positive-ion polarity). **Figure S4.** METASPACE annotations overlapped with in-source fragments vs METAS-PACE annotations matched by rMSIfragment color-coded based on: **(A)** Ion polarity **(B)** MALDI matrix **(C)** Tissue type **(D)** Analyzer **(E)** Mean *m/z*. **Table S1.** Lipid adduct formation in positive-ion polarity. Reproduced with the permission of Garate et al. 2020. **Table S2.** Lipid adduct formation in negative-ion polarity. Reproduced with the permission of Garate et al. 2020. **Table S3.** Example rMSIfragment output. **Table S4.** List of the 27 MALDI MSI datasets used for validation. Sample type, sample preparation, and MALDI-MSI acquisition parameters.

Additional file 2: Table S5: Complete list of annotations using rMSIfragment (Samples G1-G8).

Additional file 3: Table S6: Complete list of annotations using rMSIfragment (Samples G9-G15).

Acknowledgements

Author contributions

GB, PR, XC, and MG developed the concept for the study. LM, CB, and JAF designed and performed mass spectrometry imaging experiments. GB developed the computational workflow in collaboration with LS, PR, XC, and MG. GB and MG wrote the original manuscript with substantial edits and contributions from all authors. JAF and XC provided supervision, project administration, and funding.

Funding

Availability of data and materials

The platform-independent R package rMSIfragment presented in this publication is freely available under the terms of the GNU General Public License v3.0 at https://github.com/gbaquer/rMSIfragment. The datasets supporting the conclusions of this article are available in the Mendeley Data repository: https://doi.org/10.17632/53grw3ys6y.1 (Datasets G1—G15). Datasets M1-M12 are available at https://metaspace2020.eu/ (References provided in Additional file 1: Table S1). rMSIfragment is available under the GPLv3 at www.github.com/gbaquer/rMSIfragment.

Declarations

Ethics approval and consent to participate

The Euskadi Ethics Committee approved the study protocol and it conformed to the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to declare.

References

1. Adusumilli R, Mallick P (2017) Data conversion with ProteoWizard msConvert. Methods Mol Biol 1550:339–368
2. Ait-Belkacem, Rima, Caroline Berenguer, Claude Villard, L 'houcine Ouafik, Dominique Figarella-Branger, Olivier Chinot, and Daniel Lafitte. MALDI imaging and in-source decay for top-down characterization of glioblastoma. Proteomics. 2014; 14 (10): 1290–1301.
3. Alexandrov T (2020) Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. Ann Rev Biomed Data Sci 3(July):61–87
4. Alexandrov T, Ovchinnikova K, Palmer A, Kovalev V, Tarasov A, Stuart L, Nigmetzianov R, Fay D, Contributors KM (2019) METASPACE: a community-populated knowledge base of spatial metabolomes in health and disease. bioRxiv. https://doi.org/10.1101/539478
5. Antone AJ, Liang Q, Sherwood JA, Weiss JC, Wilson JM, Deb S, Cassady CJ, Bao Y (2019) Surface effects of iron oxide nanoparticles on the MALDI in-source decay analysis of glycans and peptides. ACS Appl Nano Mater 2(6):3999–4008
6. Awad H, Khamis MM, El-Aneed A (2015) Mass spectrometry, review of the basics: ionization. Appl Spectrosc Rev 50(2):158–175
7. Baijnath S (2022) Mass spectrometry imaging: the future is now. Bioanalysis 14(7):383–386
8. Baquer, Gerard, Lluc Sementé, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in mass spectrometry imaging. mass spectrometry reviews. 2022; July, e21794.
9. Behrmann J, Etmann C, Boskamp T, Casadonte R, Kriegsmann J, Maaß P (2018) Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics 34(7):1215–1223
10. Berghmans E, Boonen K, Maes E, Mertens I, Pauwels P, Baggerman G (2020) Implementation of MALDI mass spectrometry imaging in cancer proteomics research: applications and challenges. J Personal Med. https://doi.org/10.3390/jpm10020054
11. Boskamp MS, Soltwisch J (2020) Charge distribution between different classes of glycerophospolipids in MALDI-MS imaging. Anal Chem 92(7):5222–5230
12. van der Burgt YE, Kilgour DP, Tsybin YO, Srzentic K, Fornelli L, Beck A, Wuhrer M, Nicolardi S (2019) Structural analysis of monoclonal antibodies by ultrahigh resolution MALDI in-source decay FT-ICR mass spectrometry. Anal Chem 91(3):2079–85
13. Coy S, Wang S, Stopka SA, Lin J-R, Yapp C, Ritch CC, Salhi L et al (2022) Single cell spatial analysis reveals the topology of immunomodulatory purinergic signaling in glioblastoma. Nat Commun 13(1):4814
14. Debois D, Bertrand V, Quinton L, De Pauw-Gillet M-C, De Pauw E (2010) MALDI-in source decay applied to mass spectrometry imaging: a new tool for protein identification. Anal Chem 82(10):4036–4045
15. Denti V, Andersen MK, Smith A, Bofin Anna M, Nordborg A, Magni F, Moestue SA, Giampà M (2021) Reproducible lipid alterations in patient-derived breast cancer xenograft FFPE tissue identified with MALDI MSI for pre-clinical and clinical application. Metabolites. https://doi.org/10.3390/metabo11090577
16. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Dorrestein PC, Rousu J, Böcker S (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16(4):299–302
17. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4(3):207–214
18. Ellis SR, Paine MRL, Eijkel GB, Pauling JK, Husen P, Jervelund MW, Hermansson M, Ejsing CS, Heeren RMA (2018) Automated, parallel mass

spectrometry imaging and structural identification of lipids. Nat Methods 15(7):515–518

19. Franceschi P, Dong Y, Strupat K, Vrhovsek U, Mattivi F (2012) Combining intensity correlation analysis and MALDI imaging to study the distribution of flavonols and dihydrochalcones in golden delicious apples. J Exp Bot 63(3):1123–1133

20. Garate J, Lage S, Martín-Saiz L, Perez-Valle A, Begoña Ochoa M, Boyano D, Fernández R, Fernández JA (2020) Influence of lipid fragmentation in the data analysis of imaging mass spectrometry experiments. J Am Soc Mass Spectrom 31(3):517–526

21. Guo G, Papanicolaou M, Demarais NJ, Wang Z, Schey KL, Timpson P, Cox TR, Grey AC (2021) Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP. Nat Commun 12(1):3241

22. Harkin C, Smith KW, Logan MacKay C, Moore T, Brockbank S, Ruddock M, Cobice DF (2022) Spatial localization of β-unsaturated aldehyde markers in murine diabetic kidney tissue by mass spectrometry imaging. Anal Bioanal Chem 414(22):6657–6670

23. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44(D1):D1214–D1219

24. Heijs B, Potthoff A, Soltwisch J, Dreisewerd K (2020) MALDI-2 for the enhanced analysis of N-linked glycans by mass spectrometry imaging. Anal Chem 92(20):13904–13911

25. He Q, Sun C, Liu J, Pan Y (2021) MALDI-MSI analysis of cancer drugs: significance, advances, and applications. Trends Anal Chem TRAC 136(March):116183

26. Hu C, Luo W, Jie Xu, Han X (2022) Recognition and avoidance of ion source-generated artifacts in lipidomics analysis. Mass Spectrom Rev 41(1):15–31

27. Ikegawa M, Nirasawa T, Kakuda N, Miyasaka T, Kuzuhara Y, Murayama S, Ihara Y (2019) Visualization of amyloid β deposits in the human brain with matrix-assisted laser desorption/ionization imaging mass spectrometry. J Vis Exp. https://doi.org/10.3791/57645

28. Israr MZ, Bernieh D, Salzano A, Cassambai S, Yazaki Y, Suzuki T (2020) Matrix-assisted laser desorption ionisation (MALDI) mass spectrometry (MS): basics and clinical applications. Clin Chem Lab Med 58(6):883–896

29. Janßen, Charlotte, Tobias Boskamp, Lena Hauberg-Lotte, Jens Behrmann, Sören-Oliver Deininger, Mark Kriegsmann, Katharina Kriegsmann, et al. Robust subtyping of non-small cell lung cancer whole sections through MALDI mass spectrometry imaging. PROTEOMICS Clin Appl. 2022; 2100068.

30. Kaya I, Jennische E, Lange S, Baykal AT, Malmberg P, Fletcher JS (2020) Brain region-specific amyloid plaque-associated myelin lipid loss, APOE deposition and disruption of the myelin sheath in familial Alzheimer's disease mice. J Neurochem 154(1):84–98

31. Kaya I, Zetterberg H, Blennow K, Hanrieder J (2018) Shedding light on the molecular pathology of amyloid plaques in transgenic Alzheimer's disease mice using multimodal MALDI imaging mass spectrometry. ACS Chem Neurosci 9(7):1802–1817

32. Ma, Xin, and Facundo M. Fernández. 2022. "Advances in Mass Spectrometry Imaging for Spatial Cancer Metabolomics." *Mass Spectrometry Reviews*, September, e21804.

33. Miyamoto S, Hsu C-C, Hamm G, Darshi M, Diamond-Stanic M, Declèves A-E, Slater L et al (2016) Mass spectrometry imaging reveals elevated glomerular ATP/AMP in diabetes/obesity and identifies sphingomyelin as a possible mediator. EBioMedicine 7(May):121–134

34. Muro E, Ekin Atilla-Gokcumen G, Eggert US (2014) Lipids in cell biology: how can we understand them better? Mol Biol Cell 25(12):1819–1823

35. Melo N, Martins L, Lesner NP, Sabatier M, Ubellacker JM, Tasdogan A (2022) Emerging metabolomic tools to study cancer metastasis. Trends in Cancer Res. https://doi.org/10.1016/j.trecan.2022.07.003

36. Nguyen TD, Lan Y, Kane SS, Haffner JJ, Liu R, McCall L-I, Yang Z (2022) Single-cell mass spectrometry enables insight into heterogeneity in infectious disease. Anal Chem 94(30):10567–10572

37. Nicolardi S, Danuser R, Dotz V, Domínguez-Vega E, Kaabi AA, Beurret M, Anish C, Wuhrer M (2022) Glycan and protein analysis of glycoengineered bacterial E. Coli vaccines by MALDI-in-source decay FT-ICR mass spectrometry. Anal Chem 94(12):4979–4987

38. Nishidate M, Hayashi M, Aikawa H, Tanaka K, Nakada N, Miura S-I, Ryu S et al (2019) Applications of MALDI mass spectrometry imaging for pharmacokinetic studies during drug development. Drug Metab Pharmacokinet 34(4):209–216

39. Notarangelo G, Spinelli JB, Perez EM, Baker GJ, Kurmi K, Elia I, Stopka SA et al (2022) Oncometabolite D-2HG alters T cell metabolism to impair CD8+ T cell function. Science 377(6614):1519–1529

40. O'Donnell VB, Dennis EA, Wakelam MJO, Subramaniam S (2019) LIPID MAPS: serving the next generation of lipid researchers with tools, resources, data, and training. Sci Signaling. https://doi.org/10.1126/scisignal.aaw2964

41. Ovchinnikova K, Stuart L, Rakhlin A, Nikolenko S, Alexandrov T (2020) ColocML: machine learning quantifies Co-localization between mass spectrometry images. Bioinformatics 36(10):3215–3224

42. Palmer A, Phapale P, Chernyavsky I, Lavigne R, Fay D, Tarasov A, Kovalev V et al (2017) FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods 14(1):57–60

43. Race AM, Styles IB, Bunch J (2012) Inclusive sharing of mass spectrometry imaging data requires a converter for all. J Proteomics 75(16):5111–5112

44. Race AM, Sutton D, Hamm G, Maglennon G, Morton JP, Strittmatter N, Campbell A et al (2021) Deep learning-based annotation transfer between molecular imaging modalities: an automated workflow for multimodal data integration. Anal Chem 93(6):3061–3071

45. Ràfols P, Heijs B, Del Castillo E, Yanes O, McDonnell LA, Brezmes J, Pérez-Taboada I, Vallejo M, García-Altares M, Correig X (2020) RMSIproc: an R package for mass spectrometry imaging data processing. Bioinformatics 36(11):3618–3619

46. Ràfols P, Torres S, Ramírez N, Del Castillo E, Yanes O, Brezmes J, Correig X (2017) RMSI: an R package for MS imaging data handling and visualization. Bioinformatics 33(15):2427–2428

47. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminf 8:3

48. Schramm T, Hester A, Klinkert I, Both JP, Heeren RMA, Brunelle A, Laprévote O et al (2012) ImzML—a common data format for the flexible exchange and processing of mass spectrometry imaging data. J Proteomics 75(16):5106–5110

49. Sementé L, Baquer G, García-Altares M, Correig-Blanchar X, Ràfols P (2021) rMSIannotation: a peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios. Anal Chim Acta 1171(August):338669

50. Seyyed-Kalantari, Laleh, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2020. "CheXclusion: Fairness gaps in deep chest x-ray classifiers." In *Biocomputing 2021*, 232–43. WORLD SCIENTIFIC.

51. Sidey-Gibbons JAM, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19(1):64

52. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27(6):747–751

53. Stoeckli M, Farmer TB, Caprioli RM (1999) Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight instrument. J Am Soc Mass Spectrom 10(1):67–71

54. Tans R, Dey S, Dey NS, Calder G, O'Toole P, Kaye PM, Heeren RMA (2021) Spatially resolved immunometabolism to understand infectious disease progression. Front Microbiol 12(August):709728

55. Tortorella S, Tiberi P, Bowman AP, Claes BSR, Ščupáková K, Heeren RMA, Ellis SR, Cruciani G (2020) LipostarMSI: comprehensive, vendor-neutral software for visualization, data analysis, and automated molecular identification in mass spectrometry imaging. J Am Soc Mass Spectrom 31(1):155–163

56. Wadie B, Stuart L, Rath CM, Alexandrov T (2023) METASPACE-ML: metabolite annotation for imaging mass spectrometry using machine learning. biorxiv. https://doi.org/10.1101/2023.05.29.542736

57. Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart DS (2021) CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. Anal Chem 93(34):11692–11700

58. Wang HY, Hsu FF (2022) Structural characterization of phospholipids and sphingolipids by in-source fragmentation MALDI/TOF mass spectrometry. Anal Bioanal Chem 414(6):2089–2102

59. Wang Z, Wenqing Fu, Huo M, He B, Yaqi Liu Lu, Tian WL et al (2021) Spatial-resolved metabolomics reveals tissue-specific metabolic

Baquer *et al. Journal of Cheminformatics*   (2023) 15:80

Page 14 of 14

reprogramming in diabetic nephropathy by using mass spectrometry imaging. Acta Pharm Sinica B 11(11):3665–3677

60. Wu E, Kevin Wu, Daneshjou R, Ouyang D, Ho DE, Zou J (2021) How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 27(4):582–584

61. Xue J, Domingo-Almenara X, Guijas C, Palermo A, Rinschen MM, John Isbell H, Benton P, Siuzdak G (2020) Enhanced in-source fragmentation annotation enables novel data independent acquisition and autonomous METLIN molecular identification. Anal Chem 92(8):6051–6059

62. Zhang W, Claesen M, Thomas Moerman M, Groseclose R, Waelkens E, De Moor B, Verbeeck N (2021) Spatially aware clustering of ion images in mass spectrometry imaging data using deep learning. Anal Bioanal Chem. https://doi.org/10.1007/s00216-021-03179-w

63. Zhou B, Xiao JF, Tuli L, Ressom HW (2012) LC-MS-based metabolomics. Mol BioSyst 8(2):470–481

64. Zimmerman TA, Debois D, Mazzucchelli G, Bertrand V, De Pauw-Gillet M-C, De Pauw E (2011) An analytical pipeline for MALDI in-source decay mass spectrometry imaging. Anal Chem 83(15):6090–6097

Publisher's Note