Check for updates

A complementary approach for detecting biological signals through a semi-automated feature selection tool

Gabriel Santos Arini[1,2,3], Luiz Gabriel Souza Mencucini[1,2], Rafael de Felício[4], Luís Guilherme Pereira Feitosa[2], Paula Rezende-Teixeira[5], Henrique Marcel Yudi de Oliveira Tsuji[1,2], Alan Cesar Pilon[6], Danielle Rocha Pinho[2], Letícia Veras Costa Lotufo[5], Norberto Peporine Lopes[2], Daniela Barretto Barbosa Trivella[4] and Ricardo Roberto da Silva[1,2,3]*

[1]Department of Biomolecular Sciences, Computational Chemical Biology Laboratory, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil, [2]NPPNS, Department of Biomolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil, [3]Department of Cellular and Molecular Biology of Ribeirão Preto, Cellular and Molecular Biology Program, School of Medicine, University of São Paulo, Ribeirão Preto, Brazil, [4]Brazilian Biosciences National Laboratory (LNBio), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Brazil, [5]Department of Pharmacology, Marine Pharmacology Laboratory, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil, [6]Department of Biochemistry and Organic Chemistry, Institute of Chemistry, Paulista State University, São Paulo, Brazil

Introduction: Untargeted metabolomics is often used in studies that aim to trace the metabolic profile in a broad context, with the data-dependent acquisition (DDA) mode being the most commonly used method. However, this approach has the limitation that not all detected ions are fragmented in the data acquisition process, in addition to the lack of specificity regarding the process of fragmentation of biological signals. The present work aims to extend the detection of biological signals and contribute to overcoming the fragmentation limits of the DDA mode with a dynamic procedure that combines experimental and in silico approaches.

Methods: Metabolomic analysis was performed on three different species of actinomycetes using liquid chromatography coupled with mass spectrometry. The data obtained were preprocessed by the MZmine software and processed by the custom package RegFilter.

Results and Discussion: RegFilter allowed the coverage of the entire chromatographic run and the selection of precursor ions for fragmentation that were previously missed in DDA mode. Most of the ions selected by the tool could be annotated through three levels of annotation, presenting biologically relevant candidates. In addition, the tool offers the possibility of creating local spectral libraries curated according to the user's interests. Thus, the adoption of a dynamic analysis flow using RegFilter allowed for detection optimization and curation of potential biological signals, previously absent in

the DDA mode, being a good complementary approach to the current mode of data acquisition. In addition, this workflow enables the creation and search of in-house tailored custom libraries.

1 Introduction

Metabolomics offers the possibility of molecular phenotyping and has been applied to diverse fields of knowledge, from biotechnology to environmental chemistry (Kuehnbaum and Britz-McKibbin, 2013). Unlike nucleic acids and even proteins, the metabolites produced in a given biological context help the understanding of rapid spatio-temporal changes triggered by the environment. To harness the metabolic complexity, metabolomics uses several analytical techniques, focusing on mass spectrometry (MS) (Wehrens and Saleck, 2019). Targeted metabolomics focuses on specific metabolites, and its main advantage, besides specificity, is its reproducible quantitative aspect. In contrast, untargeted metabolomics is characterized by the enormous size and complexity of the data sets generated. One of the major advantages of this approach is the possibility of profiling a large number of new molecules, which in turn would have the capacity to reveal unexplored metabolic pathways (Prosser et al., 2014; Banh et al., 2021), establish evolutionary parallels (Nunes et al., 2017; Roach et al., 2021), offer possible insights about what is happening in a given cellular scenario (Zaramela et al., 2021), and present new natural products of biotechnological and pharmacological interest in a wide and diverse range of organisms (Wilke et al., 2021; Kato et al., 2024). Currently, more than 60% of pharmaceutical assets are related directly or indirectly to natural products, i.e., chemical substances produced by living organisms (Newman and Cragg, 2020; Kim et al., 2021).

Often, the data generated in untargeted metabolomics experiments are not explored in their entirety. The recent advances in data sharing, and availability of public spectral data libraries, in platforms such as Metlin (Xue et al., 2020), Metabolights (Yurekten et al., 2024), MoNA (https://mona.fiehnlab.ucdavis.edu/), Metabolomics Workbench (Sud et al., 2016) and GNPS (Wang et al., 2016; Wang et al., 2020) as well as analytical and integrative tools such as Mzmine (Schmid et al., 2023), NAP (da Silva et al., 2018), MolNetEnhancer (Ernst et al., 2019), ChemWalker (Borelli et al., 2023), ModiFinder (Shahneh et al., 2024), and NP³ MS Workflow (Bazzano et al., 2024) greatly expanded the potential of metabolomics application. However, the highly redundant and noisy data of mass spectra from complex mixtures is a great challenge. A promising strategy would be to target selected signals, deepening studies within a given context of interest, rationally moving from an untargeted to a targeted approach. This strategy is justified by the fact that many of the detected ions are not fragmented in the mostly used fragmentation mode for untargeted metabolomics: the data-dependent acquisition (DDA) mode (Broeckling et al., 2018; Davies et al., 2021; Davies et al., 2021; Stincone et al., 2023).

Reasons for the preference for DDA in untargeted metabolomics experiments include the availability of optimized tools and workflows for acquiring and processing these data (Stincone et al., 2023; Guo and Huan, 2020), as well as the ready availability of the acquired spectra and their quality (McBride et al., 2023). However, one of the main limitations of using this data acquisition mode in MS^2 is the non-fragmentation of a number of features (Broeckling et al., 2018). Despite the progress that has been made in terms of increased fragmentation coverage, it is reasonable to assume that much relevant biological information is lost in this process.Moving from an untargeted approach to rationally selecting relevant features in a given context is a logical next step in metabolomics research. The construction of feature selection lists based on objective statistical criteria has been presented as a promising resource in this context (Zhang et al., 2023). However, the proper automation of the process, as well as the ability to store the fragmentation spectra resulting from this selection process, is still a task to be overcome. In this scenario, the present work aims to introduce an hybrid and dynamic approach which combines experimental and computational procedures, providing a feature selection tool that optimizes the detection of biological signals, and contributes to overcoming the fragmentation limits of the DDA mode and the ability to create in-house spectral libraries.

2 Materials and methods

2.1 Collection, extract preparation, and metabolomics analysis

The actinomycetes BRA006, BRA010, and BRA177, from MicroMarin collection (https://www.labbmar.ufc.br/micromarinbr) were cultured in A1 medium [Starch (10 g/L); Yeast extract (4 g/L); Peptone (2 g/L); Sea Water 75%] in a volume of 100 mL in 250 mL Erlenmeyer flasks. The liquid cultures were extracted with ethyl acetate, and the organic phase was dried under pressure and kept at 4°C. For the LC-MS/MS analyses, organic extracts were diluted in methanol and injected at 1.0 mg/mL, 0.5 mg/mL, 0.25 mg/mL, 0.125 mg/mL, and 0.0625 mg/mL concentrations in 1:1 H_2O:metanol (Bauermeister et al., 2016). The LC-MS/MS analysis itself was conducted in the Acquity UPLC H-Class (Waters, Milford, MA - US) hyphenated with Impact II mass spectrometer (Bruker Daltonics, Billerica - US). The mobile phase (flow 0.3 mL.min-1) consisted of water (A) and methanol (B) acidified with formic acid at 0.1% in the following gradient: 0.0–15.0 min (5%–20% B, curve 6); 15.0–30.0 min (20%–95% B, curve 6); 30.0–33.0 min - (100 B, curve 1); 33.0–40.0 min (5% B, curve 1). C18 - Luna (Phenomonex®–100 mm × 2.1 mm x 2.6 μm)

and the temperature adjusted to 35°C. The parameters adjusted for the spectrometer were: end plate offset of 500V; capillary voltage of 4.5kV; nitrogen (N_2) was used as gas; drying gas flow at 5.0L.min-1; drying gas temperature at 180°C; 4 bar nebulizer gas pressure; positive ESI mode. Spectra (*m/z* 30–2000) were recorded at a rate of 8 Hz. Accurate masses were obtained using a solution of sodium formic acid [$HCOO-Na^+$] at 10 mM as an internal standard.

2.2 Computational methods

2.2.1 Raw data conversion and preprocessing through MZmine

Raw MS files were converted to. mzXML format using MSConvert software (Chambers et al., 2012). Such converted files were pre-processed by MZmine software version 2.53. Data preprocessing in full scan and MS/MS modes was performed separately. The following parameters were used for the full scan mode: (i) Mass detection (MS level 1): mass detector - Centroid; noise level - 1E3; (ii) ADAP: min group size - 4; Group intensity threshould - 1E3; Min highest intensity - 3E3; *m/z* tolerance - 0.001*m/z* or 15ppm; (iii) Chromatogram deconvolution: algorithm - Baseline cut-off; Min peak height - 3E3; Peak duration range: 0–15; Baseline level - 1E3; *m/z* center calculation: Median; *m/z* range for MS² scan pairing (Da): 0.01; RT range for MS² scan pairing (min): 0.2; (iv) Isotopic peaks grouper: *m/z* tolerance: 0.001 *m/z* or 15ppm; Retention time tolerance: 0.2; Maximum charge: 3; Representative isotopes: Most intense; (v) Join aligner: *m/z* tolerance: 0.001*m/z* or 15ppm; Weight for *m/z*: 75; Retention time tolerance: 0.2; Weight for RT: 25; (v) Peak finder: Intensity tolerance: 10; *m/z* tolerance: 0.001*m/z* or 15ppm; Retentation time tolerance: 0.2; (vi) Export to CSV file: Export common elements: Export row ID, Export row *m/z*, Export row retention time; Export data file elements: Peak area; Peak *m/z* min; Peak *m/z* max; Filter rows: All.

The following parameters have been adopted for the MS/MS data acquired in DDA mode: (i) Mass detection (MS level 1): mass detector - Centroid; noise level - 1E3; (ii) Mass detection (MS level 2): mass detector - Centroid; noise level - 1E2; (iii) ADAP: min group size - 4; Group intensity threshould - 1E3; Min highest intensity - 3E3; *m/z* tolerance - 0.001*m/z* or 15ppm; (iv) Chromatogram deconvolution: algorithm - Baseline cut-off; Min peak height - 3E3; Peak duration range: 0–15; Baseline level - 1E3; *m/z* center calculation: Median; *m/z* range for MS² scan pairing (Da): 0.01; RT range for MS² scan pairing (min): 0.2; (v) Isotopic peaks grouper: *m/z* tolerance: 0.001*m/z* or 15ppm; Retention time tolerance: 0.2; Maximum charge: 3; Representative isotopes: Most intense; (vi) Join aligner: *m/z* tolerance: 0.001*m/z* or 15ppm; Weight for *m/z*: 75; Retention time tolerance: 0.2; Weight for RT: 25; (vii) Peak finder: Intensity tolerance: 10; *m/z* tolerance: 0.001*m/z* or 15ppm; Retentation time tolerance: 0.2; (viii) Export to CSV file: Export common elements: Export row ID, Export row *m/z*, Export row retention time; Export data file elements: Peak area; Peak *m/z* min; Peak *m/z* max; Filter rows: All; (ix) Export/Submit to GNPS-FBMN: Filter rows: Only with MS².

The following parameters have been adopted for the MS/MS mode of data obtained with the scheduled precursor list (SPL): (i) Mass detection (MS level 1): mass detector - Centroid; noise level -

1E2; (ii) Mass detection (MS level 2): mass detector - Centroid; noise level - 1E1; (iii) ADAP: min group size - 4; Group intensity threshould - 1E2; Min highest intensity - 3E2; *m/z* tolerance - 0.001*m/z* or 15ppm; (iv) Chromatogram deconvolution: algorithm - Baseline cut-off; Min peak height - 3E2; Peak duration range: 0–15; Baseline level - 1E2; *m/z* center calculation: Median; *m/z* range for MS² scan pairing (Da): 0.01; RT range for MS² scan pairing (min): 0.2; (v) Isotopic peaks grouper: *m/z* tolerance: 0.001*m/z* or 15ppm; Retentation time tolerance: 0.2; Maximum charge: 3; Representative isotopes: Most intense; (vi) Join aligner: *m/z* tolerance: 0.001*m/z* or 15ppm; Weight for *m/z*: 75; Retention time tolerance: 0.2; Weight for RT: 25; (vii) Peak finder: Intensity tolerance: 10; *m/z* tolerance: 0.001*m/z* or 15ppm; Retentation time tolerance: 0.2; (viii) Export to CSV file: Export common elements: Export row ID, Export row *m/z*, Export row retention time; Export data file elements: Peak area; Peak *m/z* min; Peak *m/z* max; Filter rows: All; (ix) Export/Submit to GNPS-FBMN: Filter rows: Only with MS².
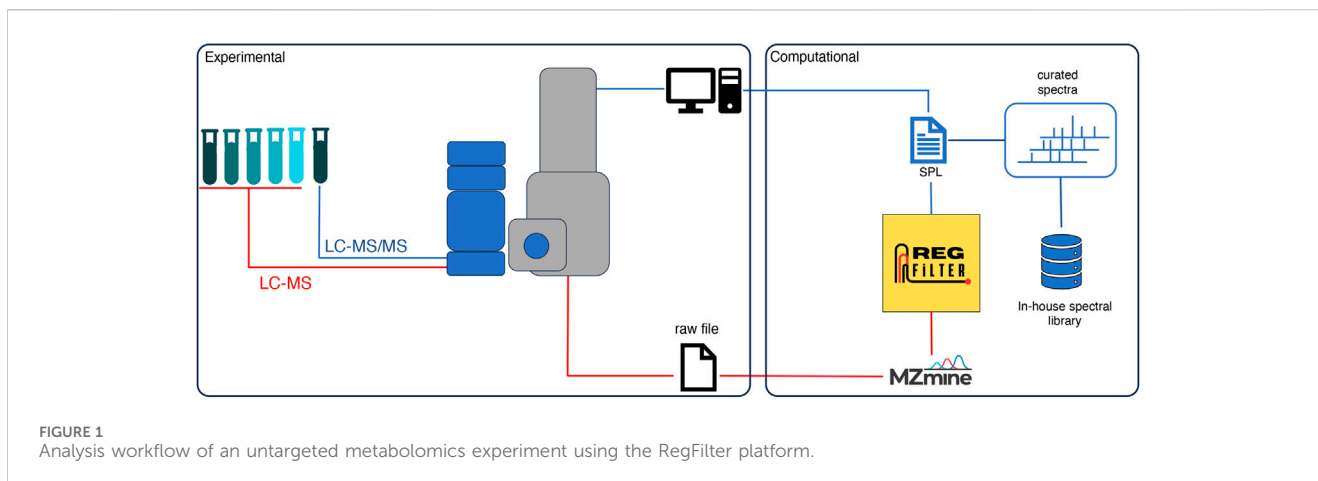
In all cases, a batch mode file was made available in the Zenodo repository in the Data Availability session.

2.2.2 Ion filtering and schedule precursor list (SPL) construction

RegFilter is an open-source software written in Python and available from the laboratory's GitHub repository (https://github.com/computational-chemical-biology/regression_filter). To use it, the user must follow the installation steps described in the repository. Once installed, RegFilter uses as input material a table in.csv format containing the spectral data previously acquired by the user in full scan mode and duly pre-processed by MZmine (Schmid et al., 2023) software as described above. The package's instructions guide user on how to use the package and export a report showing the number of total features, the number of filtered features, and a new.csv table with the filtered features with their respective retention times, that can be formatted as a SPL for different mass spectrometers.

2.2.3 GNPS2 molecular networking and *in silico* annotation

From the raw data of the spectral analysis, we preprocessed these data through MZmine (Schmid et al., 2023), generating an area below the chromatographic curve quantification table (.csv) and a list of fragmentation spectra (.mgf) files which were imported into the GNPS (Wang et al., 2016) and GNPS2 platforms (gnps2.org/), where the molecular network and spectral pairing annotation was performed by the Feature Based Molecular Network (FBMN) workflow (Nothias et al., 2020). For the nodes that could not be annotated by spectral library matching, the MS/MS mass spectra were annotated in two layers of annotation. First, we propagated annotations based on the topology of the molecular network using the ChemWalker tool (Borelli et al., 2023) through its new graphical interface available on GNPS2 (https://gnps2.org/workflowinput?workflowname=chemwalker_nextflow_workflow). The criteria used to propagate the annotation were: (i) Adduct: [M + H]⁺; (ii) tolerance: 15 ppm. A second layer of *in silico* annotation was then performed using the SIRIUS tool (Dührkop et al., 2019) version 5.8. 5 for the nodes that could not be annotated in the previous steps. The recommended parameters for the Q-TOF instrument were adopted,

FIGURE 1
Analysis workflow of an untargeted metabolomics experiment using the RegFilter platform.

using the formulas available in all databases and taking into account the adducts $[M + H]^+$, $[M + K]^+$, and $[M + Na]^+$. In both cases of *in silico* annotation, we considered the best-ranked candidate as the annotation for the corresponding spectrum.

3 Results and discussion

3.1 Regression filter: optimizing the detection of biological signals through a semi-automated feature selection tool

Some of the bottlenecks that deserve to be highlighted in the processing and analysis of metabolomics data from untargeted experiments are (i) the selection of features directly related to the sample, i.e., biological signals, and (ii) the fragmentation of features that are lost in DDA. To overcome these limitations, we have created a dynamic analysis flow that combines experimental and computational approaches, as illustrated in Figure 1. In this analysis flow, a sample of interest is diluted serially (until 8-fold dilution). An LC-MS analysis is performed on the undiluted sample and each dilution in full scan mode. The raw data are converted to the.mzXML format and then pre-processed by the MZmine software (Schmid et al., 2023) with the parameters that best fit the user's necessities. The .csv file obtained in this step is subjected to feature filtering using an algorithm called Regression Filter (RegFilter). This algorithm is implemented in Python and describes the conditions to filter features. All the ions presented in the entire dilution series whose area under their curve decreases linearly, respecting a coefficient of determination (R^2) greater than or equal to 0.9 (default value that can be adjusted by the user) and a negative angular coefficient for the fitted regression model, are stored and reported to the user. The output is then exported as a new .csv file containing all the selected features with their respective retention times, which is then inserted into the mass spectrometer software as a scheduled precursor list (SPL), where the ions it contains are fragmented only in the undiluted sample.
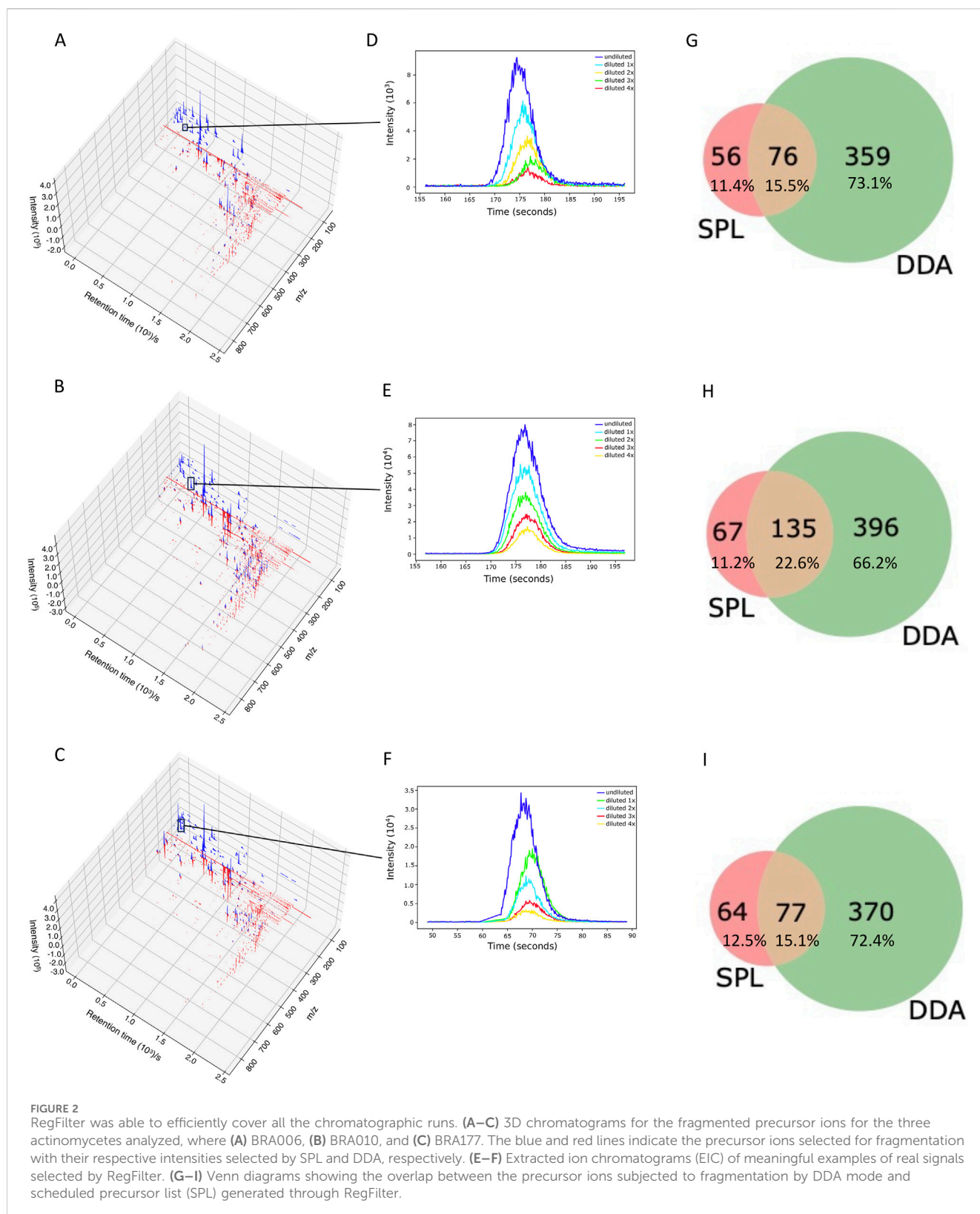
In addition to the feature filtering algorithm itself, the package provides several functions for data analysis and visualization as well as the possibility of storing spectra of interest for *in-house* library building. To test the tool's efficiency in extracting features of

potential biological relevance, as well as its potential to select features previously lost in the DDA mode, we performed a series of experiments in which the cellular extract of three marine actinomycetes (BRA006, BRA010, and BRA177) belonging to a Brazilian collection of bacterial isolates was subjected to the workflow, and the results obtained were compared with a dataset of fragmentation spectra obtained in DDA, the most common acquisition mode for untargeted experiments in complex biological samples. The choice to work on this biological matrix is due to the potential for the production of bioactive compounds that this group of bacteria represents, opening the possibility for the discovery, cataloging, and curation of new chemical entities (van Bergeijk et al., 2020; Chen et al., 2021).

3.2 RegFilter efficiency on chromatographic run coverage

We analyzed whether RegFilter was able to cover the entire chromatographic run. As can be seen in Figure 2 and Supplementary Figure S1 with the chromatograms 3D plot where we compared SPL and DDA chromatograms from those ions that led to fragmentation, RegFilter covered the entire analysis region from the three bacterial extracts. It is worth noting that ions with a low mass/charge ratio (m/z) were more frequently captured by the list compared to DDA, where the prevalence of these precursors destined for fragmentation throughout the chromatographic run is clear. Additionally, a less noisy baseline is observed in these chromatographic runs in all extracts, indicating that the noise filtering process was efficient.

Once verified that RegFilter could cover the entire analysis region, we proceeded with a first comparative analysis between the full scan and data-dependent acquisition (DDA) modes. The results are shown in Table 1. We can see that the total number of features detected in full scan mode is higher than the total number of features detected in DDA mode. This is an expected event because the instrument spends a considerable amount of time generating the spectra at the MS^2 level, which reduces the intensity of the features at the MS^1 level and, consequently, their detection, meaning that some of the ions previously detected in full scan mode are no longer detected (Guo and Huan, 2020).

FIGURE 2
RegFilter was able to efficiently cover all the chromatographic runs. **(A–C)** 3D chromatograms for the fragmented precursor ions for the three actinomycetes analyzed, where **(A)** BRA006, **(B)** BRA010, and **(C)** BRA177. The blue and red lines indicate the precursor ions selected for fragmentation with their respective intensities selected by SPL and DDA, respectively. **(E–F)** Extracted ion chromatograms (EIC) of meaningful examples of real signals selected by RegFilter. **(G–I)** Venn diagrams showing the overlap between the precursor ions subjected to fragmentation by DDA mode and scheduled precursor list (SPL) generated through RegFilter.

Applying RegFilter to the features detected in full scan mode, considering a regression fitted with at least three points (the experiments were conducted with 5 experimental points), it can be seen that 135 features for strain BRA006, 219 features for strain BRA010, and 150 features for strain BRA177 met the filtering requirements. Based on this result, we wondered how many and which features would be common to those detected by the DDA mode. As shown in the Venn diagrams (Figures 2G–I) and Table 1, we observed that 76 features, 135 features, and 77 features for strains BRA006, BRA010, and BRA177, respectively, were present both in

TABLE 1 Comparison between data obtained in untargeted metabolomics experiments for each of the actinomycetes BRA006, BRA010, and BRA177.

	BRA006		BRA010		BRA177	
Total features - full scan mode	1,384		2068		1,584	
Total features – DDA mode	1,068		1,222		1,147	
	SPL	DDA	SPL	DDA	SPL	DDA
Fragmented features	135	486	219	640	150	525
Common fragmented features	76		135		77	

TABLE 2 Comparison between parameters of the fragmentation spectra between SPL and DDA modes (p < 0.05, Student's t-test).

	BRA006		BRA010		BRA177	
	SPL	DDA	SPL	DDA	SPL	DDA
Number of peaks	5	22	4	55	5	20
Mean intensity	1	26	6	53	3	40
Maximum intensity	4	28	4	64	6	34
Minimum intensity	3	4	25	5	5	2
MS^2 events	54	1	114	0	64	8
Percentage of noisy spectra[a]	4,4	4,9	5,9	3,8	4,7	4,4

[a]Manually inspected.

the scheduled precursors list (SPL) selected by RegFilter and in the fragmented precursor ions in DDA mode.

It is worth noticing that there is a difference between the number of fragmented features shown in Table 1 and Figures 2G–I, both in the list and the DDA. Even with the pre-processing approach performed through MZmine, there may be redundancy in the detected features, which is common in data obtained in metabolomics experiments (Zhang et al., 2023). Therefore, the observed difference is due to the reduced redundancy verified by creating a unique identifier that combines the m/z values and retention times for each feature. Such diminished redundancy was achieved through a mass and retention time matching function contained in the RegFilter package. From these common precursor ions, we evaluate their fragmentation profile, taking as a basis for comparison (i) the number of peaks generated by fragmentation, (ii) the average, (iii) the maximum and (iv) minimum intensities of these peak heights, and of (v) the number of events in MS^2, which is how many times a given precursor ion was directed to fragmentation. Table 2 shows the number of precursor ions, with a higher average for those ions presented in three replicates ("upregulated"), showing statistically significant differences in the parameters described above (p < 0.05, Student's t-test).

Of those ions that showed a statistically significant difference in the mean, except the number of events in MS^2 for all strains and the minimum intensity for strains BRA010 and BRA177, the data-dependent acquisition (DDA) mode showed a better performance in obtaining fragmentation spectra compared to fragmentation

spectra generated by list targeting. The quality of fragmentation spectra can be measured by evaluating the number of fragments (Ausloos et al., 1999; Bern et al., 2004; Zhang et al., 2023) and the signal intensity of the detected features (Bern et al., 2004; Broeckling et al., 2018). Even with such quantitative prerogatives, the proper assessment of the quality of a given spectrum depends critically on the proper judgment and expertise of a spectrometrist, as well as the pairing of the spectrum of a given compound of interest with a reference spectrum (Ausloos et al., 1999).

Intuitively, since the number of events in MS^2 for SPL was higher for all strains compared to the number of events in MS^2 for DDA, we expected that the quality of the fragmentation spectra would, consequently, be higher for the fragmented ions in SPL than in DDA, which was not verified. However, it is important to highlight that not all ions common between SPL and DDA in the three bacterial strains showed a statistically significant difference in the parameters evaluated. In fact, for most of these common ions, except the events in MS^2, the spectral comparison based on the evaluated parameters did not show a statistically significant difference. Since the entire basis of the spectral comparison is based on the spectra of interest against reference spectra, we decided to go one step further in comparing these spectra. Therefore, we took the common ions fragmented by both the list and DDA and performed their spectral pairing separately in the GNPS library (Wang et al., 2016). For the spectra that were annotated and had the same annotation for the same precursor ion present in both the SPL and DDA, we evaluated the number of common peaks relative to the reference spectra. The results of this comparison are shown in Supplementary Figure S2.

Comparing the average number of common peaks of the same annotated ions between SPL and DDA, there is no significant statistical difference between them. Therefore, the quality of the fragmentation spectra generated by SPL and DDA has the same annotation potential. Taking into account the context of the analysis, the collision energy used was the same for the ions contained in both the SPL and the DDA, which helps to explain this absence of difference observed. Next, we decided to test the hypothesis that the use of fragmentation lists would lead to a selection of less noisy features. For this purpose, we used the construction of extracted ion chromatograms (EIC) for each of the precursor ions included in both the list generated by RegFilter and the precursor ions fragmented by DDA for each of the three model strains studied. The distinction between signal and noise is given by the shape of the curve of the extracted ion chromatogram. A true signal has a well-defined sinusoidal shape, whereas the noise is characterized by its random distribution profile (Jankevics et al., 2012).

Using this profile as an evaluation reference, we manually inspected and analyzed each of the chromatograms, and quantified the percentage of noise for all three strains, both from the precursors included in the fragmentation list (SPL) and DDA mode. These results are presented in Table 2 (percentage of noisy spectra). It can be seen that there is no real difference in the amount of noise between the ions generated from the list (SPL) and those from the DDA. Detecting noisy features is an inherent challenge in MS-based metabolomics experiments (Beisken et al., 2015). One of the pitfalls of analyses performed with this analytical technique, especially when considering time-of-flight analyzers, is that signals of reproducible intensities tend to be generated in noisy regions
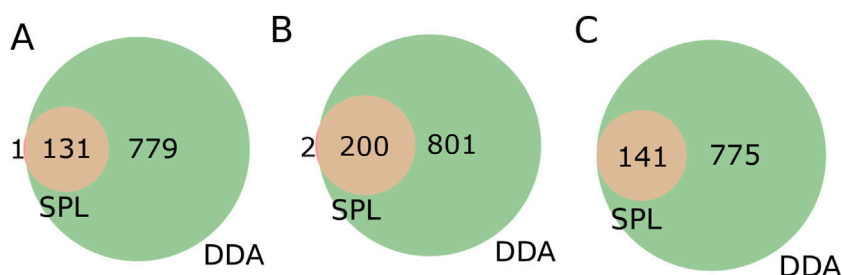
FIGURE 3
Venn diagrams of features present in the fragmentation list (SPL) and detected among the total features contained in DDA mode. **(A)** BRA006, **(B)** BRA010 and **(C)** BRA177.

(Sands et al., 2021). In this sense, the practice of serial dilution is widely used as a resource to identify such noisy features, allowing the selection of quality features and acting as more reliable and interpretable data (Jankevics et al., 2012; Sands et al., 2021). The reproducibility of such noisy signals described by Sands et al. (2021) was confirmed in our results, as even with a five-point regression model, these noisy signals persisted throughout the dilution curve. In other words, the filtering process based on a linear regression model alone does not guarantee the exclusion of noisy features, requiring an additional selection step. In this sense, RegFilter proved to be an excellent resource in the process of presenting, detecting, and selecting quality features, as it allows the user to visualize and discriminate what these features are.

It is important to note that the reduction of noise and better-quality fragmentation spectra has been previously reported in the literature (Jankevics et al., 2012; Broeckling et al., 2018). The fact that these observations were not replicated in the present study is partly due to the low complexity of the bacteria extract used. More complex matrices, such as plants or environmental samples, may present higher complexity and benefit from noise removal and increased spectral quality in addition to the complementary coverage demonstrated here.
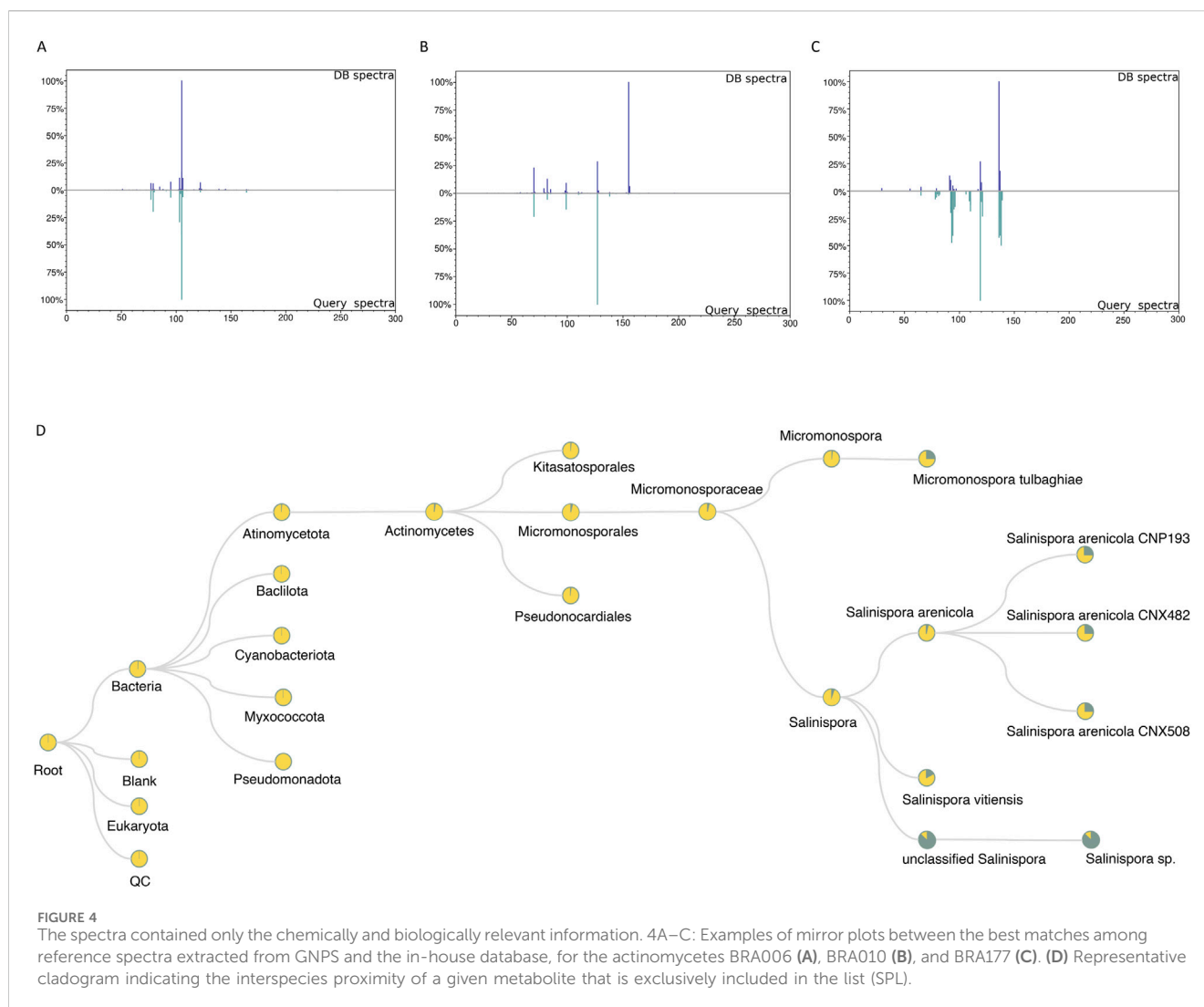
3.3 Flying under the radar: detecting once-missed signals in data-dependent acquisition mode and exploring the potential of in-house MS libraries

Turning our attention to the features contained only in SPL, we decided to look for annotations for precursor ions, and their respective fragmentation spectra, contained only in the list. To investigate the consistency of the features, we further inspected the presence of the precursor ions (m/z) and the corresponding retention times for each of the precursor ions present in the list in the DDA MS^1 data, i.e., the total features. The overlap between the precursor ions in the list and the total features detected in DDA was 99.2%, 99.0%, and 100.0% for strains BRA006, BRA010, and BRA177, respectively (Figure 3). In other words, almost all of the precursor ions selected in the fragmentation list were present in the MS^1 of the DDA method, but not all were fragmented. We decided to go one step further in analyzing this result. Therefore, we manually examined each of these ions that were exclusively

fragmented by the list and were present in MS^1 of the DDA, but not fragmented in the latter mode. We found that in 33.9% of cases for strain BRA006, 22.4% of cases for strain BRA010% and 29.7% of cases for strain BRA177 the base peak (most intense ion on the MS^1 spectrum) was the ion fragmented by the list, and not fragmented by DDA. For the other ions fragmented by the list, the peaks selected were not the base peak, having intensities from 1.1 to 86.2 times lower than the highest intensity ion for BRA006, from 1.2 to 46.5 for BRA010, and from 1.2 to 57.1 for BRA177, which explains their lack of fragmentation by the data-dependent acquisition mode. One of the limitations of the data-dependent fragmentation method is precisely its stochastic nature, where ions with similar intensities may or may not be fragmented (Davies et al., 2021). This finding was only possible by manual verification and analysis of these data. Even in the case of the smallest fraction of fragmented ions included in the list, it is still noteworthy because even those ions that would theoretically be good candidates for fragmentation by the DDA mode were not fragmented and, most importantly, RegFilter was able to select these ions.

Now, to look for annotations in the fragmentation spectra of the ions selected from the list, we took three approaches: first, we performed their spectral pairing using the GNPS2 library (Wang et al., 2016), as it comes from one of the largest public libraries of spectral data (Bittremieux et al., 2022). For those features that could not be annotated using this approach, we decided to go one step further and use ChemWalker to propagate the annotations through the topology of the molecular network (Borelli et al., 2023). Finally, we used the SIRIUS tool (Dührkop et al., 2019) on the fragmentation spectra of each of the candidates that could not be annotated through GNPS spectral pairing as well as ChemWalker annotation propagation. It should be noted that none of the annotated spectra was considered noise, but rather a true signal. As a result, 43 annotations were recovered for BRA006, 46 annotations for BRA010, and 47 annotations for BRA177. Proportionally, they represent 76.8%, 68.7%, and 73.4% of the features uniquely contained in the SPL for strains BRA006, BRA010, and BRA177, respectively. All the annotations are presented in the Supplementary Table S1.

To gain further information on the biological interpretation of the identified molecules, we cross-referenced the annotation data of each of the molecules contained in the three bacteria with the KEGG database (Kanehisa and Goto, 2000) to assess whether the set of molecules detected had already been described and cataloged in the

FIGURE 4
The spectra contained only the chemically and biologically relevant information. 4A−C: Examples of mirror plots between the best matches among reference spectra extracted from GNPS and the in-house database, for the actinomycetes BRA006 **(A)**, BRA010 **(B)**, and BRA177 **(C)**. **(D)** Representative cladogram indicating the interspecies proximity of a given metabolite that is exclusively included in the list (SPL).

encyclopedia along with its metabolic pathway presented. The compounds annotated with their respective pathway are presented in Supplementary Table S2. From the total set, 39 annotated compounds were identified using KEGG with their respective metabolic pathway associated. Of which, nine for BRA006, 10 for BRA010, and 20 for BRA177 were described along known pathways. The relationship between the compounds described and their respective pathways is shown in Supplementary Table S2. We can see that there are compounds common to the three strains, such as styrene, and others, such as 2-methylbenzaldehyde, that are common to BRA006 and BRA010. In turn, it is interesting to note that there are common metabolic pathways between the three strains but with different intermediates, such as purine metabolic pathways. Particularly in the case of the BRA177 compound, it is interesting to note that of the 20 annotated features, three had the same annotation for three different features, indicating the possible presence of isomers and/or adducts. However, the most important aspect of this result is that list-directed feature fragmentation makes it possible to annotate biological signals previously lost in DDA.

As mentioned in the tool description, with the RegFilter package, it is possible to build in-house libraries to store spectra of interest using a html-based graphical interface included in the

package. The sequence of steps to create such a library is shown in Supplementary Figure S1. We created a small library by focusing our attention on the ions targeted for fragmentation by the list, and with this repository in hand, we followed two different approaches, as shown in Figure 1. First, using the MASST tool (Wang et al., 2020), we searched the individual spectra against public spectrometric data, which greatly expands the potential of public spectral libraries. Of those that matched the platform, the spectrum with the highest score was downloaded, and, through a local search function, that is part of the package, we performed the reverse path. In other words, we show how one could take a spectrum obtained from a public database, perform spectral matching, calculate the similarity score, and evaluate the number of peaks shared between this reference spectrum and all those contained in one's local library. Figures 4A–C shows the results of examples to showcase the potential of local spectral collections, where the cosine score values and number of paired peaks for actinomycetes BRA006 (6A), BRA010 (6B), and BRA177 (6C) were 0.9565 and 25, 0.3228 and 10, 0.6048 and 13, respectively. In addition, we analyzed the spectra contained in the in-house libraries for the three strains with microbeMASST (Zuffa et al., 2024) to evaluate the potential of these spectra to provide evidence for the taxonomy of these actinomycetes through this

selected list of spectra. It was possible to find similarities between them and other spectra described for other microorganisms (Figure 4D). These results illustrate the potential and advantages of building custom libraries to obtain chemical information. It also demonstrates the potential of the Regfilter tool for obtaining spectra of relevant biological importance.

Among the current challenges in the design of experiments and tools that allow both the acquisition and the efficient use of metabolomics data based on mass spectrometry, the selection of features relevant to a context of interest, allowing a targeted experiment from an untargeted, is one of them. In this sense, the results obtained by the hybrid approach proposed in this work, with emphasis on the RegFilter package, have shown that this tool is not only capable of selecting biologically relevant features but also allows the fragmentation of features previously lost by the DDA mode as well as to visualize the profile of the selected features, presenting itself as a complementary resource to data acquisition in DDA. Proposals for the use of serial dilution in mass spectrometry-based metabolomics experiments have been suggested in the literature with different goals (Jankevics et al., 2012; Duan et al., 2016; Sands et al., 2021; Zhang et al., 2023). The focus of Duan and co-workers (2016) was to overcome the problem of noise inherent in experiments using mass spectrometry as an analytical resource. To overcome this problem, the authors used serially diluted quality control samples to detect signals that show a linear pattern of quantification in the spectrometry and are, therefore, not due to random noise. The authors reported that the method was able to reduce signals detected from a mixture of 20 known compounds, with 1,342 mass peaks detected. After filtering, only 102 mass peaks (7.6%) were recovered and assigned to the 20 original compounds. When considering our results, RegFilter was able to select and recover 9.3% of BRA006, 9.9% of BRA010, and 9.0% of BRA177 biological signals from the total number of features observed in MS^1 full scan mode. Despite the difference in recovering signals from standard samples, these numbers align with the finding that mass spectra are highly redundant and noisy. Still, in terms of recovering potentially relevant signals, this is a considerable recovery since they are true signals and can be considered of higher quality and reliability (Sands et al., 2021). It should be noted that the process was carried out in a semi-automated way, and these are relatively low complexity samples if compared with plant or environmental extracts, where more signal and more noise are expected. The results highlight the potential for applying linear regression methods for more complex samples where the proportion of signal/noise is smaller.

RegFilter was able to select ions of low relative abundance, overcoming one of the limitations already described for this acquisition mode itself (Zhang et al., 2023). Another important difference in using this analytical tool is that RegFilter proved to be effective in what could be called the enrichment of data with potential biological value since the features detected by it proved to be stable throughout the dilution steps and, therefore, exclusively related to the biological sample. Several works point to the need to increase the coverage of ions to be fragmented in untargeted metabolomics experiments. Multiple injections of the same sample (Broeckling et al., 2018), selecting the totality of ions contained in specific regions of the chromatogram (Davies et al., 2021), and automating the exclusion of ions to avoid redundancy (McBride et al., 2023) are some of the many strategies to increase the number of ions to be fragmented in a given context of interest. However, in addition to increasing the fragmentation coverage, as important as the number of fragmented ions is the choice of which of these ions are targeted for fragmentation. In this sense, the debate on the fragmentation of specific ions mediated by inclusion lists has been a recent topic (Koelmel et al., 2017; Broeckling et al., 2018; Zhang et al., 2023). In the case of economically expensive experiments (Aksenov et al., 2017), a more targeted selection would allow the optimization of time and resources in discovering molecules of biological interest.

Recently, Zhang et al. (2023) presented improvements and advantages of using a list of pre-selected precursors to gain biological information in untargeted metabolomics experiments. To this end, the list of precursors is constructed by analyzing the samples of interest in full scan mode, and based on these data and which ions would be statistically different in MS^1 full scan mode between contrasting experimental groups of samples, the list is constructed, and the ions contained therein are fragmented. The approach presented by the authors has important parallels with our approach in RegFilter, where both first use MS^1 full scan data to construct the inclusion lists. However, the inclusion criteria are different and this is where RegFilter stands out for the novelty of its approach. Firstly, the selection of features is semi-automatic, allowing full automation, as the software can be applied to a serial dilution, and the list generated can be communicated to the instrument to fragment samples that are obtained in sequence (Broeckling et al., 2018), which itself represents an optimization in the time for data acquisition and analysis. The feature selection criteria promoted by RegFilter are essentially based on two aspects: provided that the features contained in a dilution series are consistent with the construction of a curve with a coefficient of determination (R^2) greater than or equal to 0.9 and with a negative slope coefficient, these features are included in the fragmentation list. Zhang et al. (2023), on the other hand, used the construction of lists with control samples to extrapolate the data to real samples. In this sense, RegFilter can build lists from the samples themselves simply by diluting them.

These different approaches can be complementary, as the filter points out noisy features that may not be removed by a differentially regulated approach. The main drawbacks of serial dilutions are additional data collection and features that have low intensity and may not present in at least three less diluted samples to fit a curve. Of all the challenges in mass spectrometry-based metabolomics, one of the greatest is undoubtedly the design and execution of experiments that allow for more effective targeting when collecting and navigating the sea of data generated in this field. Even with increasing access to cutting-edge techniques, improved data analysis capabilities, and advances in technology to build thinner and more accurate spectrometers, a lot of information is lost in untargeted analysis. Or worse, a lot of data is discarded, and data replication using the untargeted approach on the same biological matrix is performed repeatedly. In this respect, RegFilter is a promising tool, not only to guide the acquisition of data relevant to the sample of interest but especially to allow the acquisition and selection of data to be curated and stored, allowing the construction of spectral libraries of specific interest.

4 Conclusion

In conclusion, the adoption of an analysis flow that dynamically integrates experimental and computational approaches, with the latter introducing the RegFilter feature selection tool, has made it possible to optimize the detection of biological signals belonging to the sample of interest that were previously lost in the traditional Data-Dependent Fragmentation (DDA) mode. Furthermore, RegFilter package opens up the possibility of creating curated spectral libraries with relevant biological information. Thus, the use of RegFilter proves to be an effective and complementary tool to the DDA mode of data acquisition, both in directing the features to be fragmented from a non-directed approach, curating and in storing these spectra.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://massive.ucsd.edu, MSV000095042, MSV000095044, and MSV000095129. RegFilter package can be found at https://github.com/computational-chemical-biology/regression_filter. The parameters used for preprocessing in MZmine are available at Zenodo https://doi.org/10.5281/zenodo.10822821 as .xml forms. The links for the jobs made on GNPS2 for spectral pairing through FBMN are available on: https://gnps2.org/status?task=834d956aa8b54757915e1a93af2866b4; https://gnps2.org/status?task=1abad20ab0b44181a0421a4371b5ca00; https://gnps2.org/status?task=f1e2242d9aad4fca9d5bec3746fae479. The links for the jobs made on ChemWalker are available on: https://gnps2.org/status?task=b7da5a5d55334f3090d07712971df236; https://gnps2.org/status?task=96c2d2be678a45718050fe4aca5a5bfc; https://gnps2.org/status?task=d6fd31a18cc942739096b7cf7f338cde. The links for the jobs made on GNPS (MASST) are available at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=51abc1f1bfbe4a4fa3da3ace237f90d0; https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e39178dcf8f142649a00f425f9a52b47; https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8d5285bb93e44c149ba09c9a86930f53.

Author contributions

GA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. LM: Data curation, Methodology, Software, Validation, Writing–review and editing. RdF: Data curation, Investigation, Validation, Writing–review and editing. LF: Data curation, Validation, Writing–review and editing. PR-T: Data curation, Writing–review and editing. HT: Data curation, Software, Writing–review and editing. AP: Data curation, Writing–review and editing. DP: Data curation, Writing–review and editing. LC-L: Funding acquisition, Project administration, Resources, Writing–review and editing. NL: Funding acquisition, Project administration, Supervision, Writing–review and editing. DB: Funding acquisition, Project administration, Resources, Writing–review and editing. RR: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2024.1477492/full#supplementary-material

References

Aksenov, A., da Silva, R., Knight, R., Lopes, N. P., and Dorrestein, P. (2017). Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* 1, 0054. doi:10.1038/s41570-017-0054

Ausloos, P., Clifton, C. L., Lias, S. G., Mikaya, A. I., Stein, S. E., Tchekhovskoi, D. V., et al. (1999). The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* 10 (4), 287–299. doi:10.1016/S1044-0305(98)00159-7

Banh, R. S., Kim, E. S., Spillier, Q., Biancurd, D. E., Yamamoto, K., Sohn, A. S. W., et al. (2021). The polar oxy-metabolome reveals the 4-hydroxymandelate CoQ10 synthesis pathway. *Nature* 597 (7876), 420–425. doi:10.1038/s41586-021-03865-w

Bauermeister, A., Zucchi, T. D., and Moraes, L. A. (2016). Mass spectrometric approaches for the identification of anthracycline analogs produced by actinobacteria. *J. Mass Spectrom.* 51 (6), 437–445. doi:10.1002/jms.3772

Bazzano, C. F., de Felicio, R., Alves, L. F. G., Costa, J. H., Ortega, R., Vieira, B. D., et al. (2024). NP3 MS workflow: an open-source software System to empower natural product-based drug discovery using untargeted metabolomics. *Anal. Chem.* 96 (19), 7460–7469. doi:10.1021/acs.analchem.3c05829

Beisken, S., Eiden, M., and Salek, R. M. (2015). Getting the right answers: understanding metabolomics challenges. *Expert Rev. Mol. Diagn* 15 (1), 97–109. doi:10.1586/14737159.2015.974562

Bern, M., Goldberg, D., McDonald, W. H., and Yates, J. R. 3rd. (2004). Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* 4, i49–i54. doi:10.1093/bioinformatics/bth947

Bittremieux, W., Wang, M., and Dorrestein, P. C. (2022). The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* 18 (12), 94. doi:10.1007/s11306-022-01947-y

Borelli, T. C., Arini, G. S., Feitosa, L. G. P., Dorrestein, P. C., Lopes, N. P., and da Silva, R. R. (2023). Improving annotation propagation on molecular networks through random walks: introducing ChemWalker. *Bioinformatics* 39 (3), btad078. doi:10.1093/bioinformatics/btad078

Broeckling, C. D., Hoyes, E., Richardson, K., Brown, J. M., and Prenni, J. E. (2018). Comprehensive tandem-mass-spectrometry coverage of complex samples enabled by data-set-dependent acquisition. *Anal. Chem.* 90 (13), 8020–8027. doi:10.1021/acs.analchem.8b00929

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 30 (10), 918–920. doi:10.1038/nbt.2377

Chen, J., Xu, L., Zhou, Y., and Han, B. (2021). Natural products from actinomycetes associated with marine organisms. *Mar. Drugs* 19 (11), 629. doi:10.3390/md19110629

da Silva, R. R., Wang, M., Nothias, L. F., van der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Fox, E., et al. (2018). Propagating annotations of molecular networks using *in silico* fragmentation. *PLoS Comput. Biol.* 14 (4), e1006089. doi:10.1371/journal.pcbi.1006089

Davies, V., Wandy, J., Weidt, S., van der Hooft, J. J. J., Miller, A., Daly, R., et al. (2021). Rapid development of improved data-dependent acquisition strategies. *Anal. Chem.* 93 (14), 5676–5683. doi:10.1021/acs.analchem.0c03895

Duan, L., Molnár, I., Snyder, J. H., Shen, G. A., and Qi, X. (2016). Discrimination and quantification of true biological signals in metabolomics analysis based on liquid chromatography-mass spectrometry. *Mol. Plant* 9 (8), 1217–1220. doi:10.1016/j.molp.2016.05.009

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., et al. (2019). SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* 16 (4), 299–302. doi:10.1038/s41592-019-0344-8

Ernst, M., Kang, K. B., Caraballo-Rodríguez, A. M., Nothias, L. F., Wandy, J., Chen, C., et al. (2019). MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* 9 (7), 144. doi:10.3390/metabo9070144

Guo, J., and Huan, T. (2020). Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography-mass spectrometry based untargeted metabolomics. *Anal. Chem.* 92 (12), 8072–8080. doi:10.1021/acs.analchem.9b05135

Jankevics, A., Merlo, M. E., de Vries, M., Vonk, R. J., Takano, E., and Breitling, R. (2012). Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics* 8 (Suppl. 1), 29–36. doi:10.1007/s11306-011-0341-0

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kato, N. N., Arini, G. S., Silva, R. R., Bichuette, M. E., Bitencourt, J. A. P., and Lopes, N. P. (2024). The world of cave microbiomes: biodiversity, ecological interactions, chemistry, and the multi-omics integration. *J. Braz. Chem. Soc.* 35, e20230148. doi:10.21577/0103-5053.20230148

Kim, L. J., Ohashi, M., Zhang, Z., Tan, D., Asay, M., Cascio, D., et al. (2021). Prospecting for natural products by genome mining and microcrystal electron diffraction. *Nat. Chem. Biol.* 17 (8), 872–877. doi:10.1038/s41589-021-00834-2

Koelmel, J. P., Kroeger, N. M., Gill, E. L., Ulmer, C. Z., Bowden, J. A., Patterson, R. E., et al. (2017). Expanding lipidome coverage using LC-MS/MS data-dependent acquisition with automated exclusion list generation. *J. Am. Soc. Mass Spectrom.* 28 (5), 908–917. doi:10.1007/s13361-017-1608-0

Kuehnbaum, N. L., and Britz-McKibbin, P. (2013). New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era. *Chem. Rev.* 113 (4), 2437–2468. doi:10.1021/cr300484s

McBride, R., Wandy, J., Weidt, S., Rogers, S., Davies, V., Daly, R., et al. (2023). TopNEXt: automatic DDA exclusion framework for multi-sample mass spectrometry experiments. *Bioinformatics* 39 (7), btad406. doi:10.1093/bioinformatics/btad406

Newman, D. J., and Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83 (3), 770–803. doi:10.1021/acs.jnatprod.9b01285

Nothias, L. F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., et al. (2020). Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* 17 (9), 905–908. doi:10.1038/s41592-020-0933-6

Nunes, T., Oldroyd, B., Elias, L., Mateus, S., Turatti, I. C., and Lopes, N. P. (2017). Evolution of queen cuticular hydrocarbons and worker reproduction in stingless bees. *Nat. Ecol. Evol.* 1 (0185), 0185. doi:10.1038/s41559-017-0185

Prosser, G. A., Larrouy-Maumus, G., and de Carvalho, L. P. (2014). Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO Rep.* 15 (6), 657–669. doi:10.15252/embr.201338283

Roach, T. N. F., Dilworth, J, H, C. M., Jones, A. D., Quinn, R. A., and Drury, C. (2021). Metabolomic signatures of coral bleaching history. *Nat. Ecol. Evol.* 5 (4), 495–503. doi:10.1038/s41559-020-01388-7

Sands, C. J., Gómez-Romero, M., Correia, G., Chekmeneva, E., Camuzeaux, S., Izzi-Engbeaya, C., et al. (2021). Representing the metabolome with high fidelity: range and response as quality control factors in LC-MS-based global profiling. *Anal. Chem.* 93 (4), 1924–1933. doi:10.1021/acs.analchem.0c03848

Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T. S., et al. (2023). Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* 41 (4), 447–449. doi:10.1038/s41587-023-01690-2

Shahneh, M. R. Z., Strobel, M., Vitale, G. A., Geibel, C., Abiead, Y. E., Garg, N., et al. (2024). ModiFinder: tandem mass spectral alignment enables structural modification site localization. *J. Am. Soc. Mass Spectrom.* doi:10.1021/jasms.4c00061

Stincone, P., Pakkir Shah, A. K., Schmid, R., Graves, L. G., Lambidis, S. P., Torres, R. R., et al. (2023). Evaluation of data-dependent MS/MS acquisition parameters for non-targeted metabolomics and molecular networking of environmental samples: focus on the Q exactive platform. *Anal. Chem.* 95 (34), 12673–12682. doi:10.1021/acs.analchem.3c01202

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44 (D1), D463–D470. doi:10.1093/nar/gkv1042

van Bergeijk, D. A., Terlouw, B. R., Medema, M. H., and van Wezel, G. P. (2020). Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* 18 (10), 546–558. doi:10.1038/s41579-020-0379-y

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* 34 (8), 828–837. doi:10.1038/nbt.3597

Wang, M., Jarmusch, A. K., Vargas, F., Aksenov, A. A., Gauglitz, J. M., Weldon, K., et al. (2020). Mass spectrometry searches using MASST. *Nat. Biotechnol.* 38 (1), 23–26. doi:10.1038/s41587-019-0375-9

Wehrens, R., and Saleck, R. (2019). *Metabolomics: practical guide to design and analysis.* Boca Raton, Florida: CRC Press.

Wilke, D. V., Jimenez, P. C., Branco, P. C., Rezende-Teixeira, P., Trindade-Silva, A. E., Bauermeister, A., et al. (2021). Anticancer potential of compounds from the Brazilian blue amazon. *Planta Med.* 87 (1-02), 49–70. doi:10.1055/a-1257-8402

Xue, J., Guijas, C., Benton, H. P., Warth, B., and Siuzdak, G. (2020). METLIN MS2 molecular standards database: a broad chemical and biological resource. *Nat. Methods* 17 (10), 953–954. doi:10.1038/s41592-020-0942-5

Yurekten, O., Payne, T., Tejera, N., Amaladoss, F. X., Martin, C., Williams, M., et al. (2024). MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* 52 (D1), D640–D646. doi:10.1093/nar/gkad1045

Zaramela, L. S., Moyne, O., Kumar, M., Zuniga, C., Tibocha-Bonilla, J. D., and Zengler, K. (2021). The sum is greater than the parts: exploiting microbial communities to achieve complex functions. *Curr. Opin. Biotechnol.* 67, 149–157. doi:10.1016/j.copbio.2021.01.013

Zhang, Y., Liao, J., Le, W., Wu, G., and Zhang, W. (2023). Improving the data quality of untargeted metabolomics through a targeted data-dependent acquisition based on an inclusion list of differential and preidentified ions. *Anal. Chem.* 95 (34), 12964–12973. doi:10.1021/acs.analchem.3c02888

Zuffa, S., Schmid, R., Bauermeister, A., P Gomes, P. W., Caraballo-Rodriguez, A. M., El Abiead, Y., et al. (2024). microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat. Microbiol.* 9 (2), 336–345. doi:10.1038/s41564-023-01575-9