Developing software tools for multidimensional mass spectrometry measurements

Samiha Chabane and Aivett Bilbao

Project report for the science undergraduate laboratory internship (SULI) program

ABSTRACT: With advancements in Mass Spectrometry (MS) technology, high-throughput data generation has become feasible, offering unprecedented opportunities for comprehensive molecular characterization. The integration of artificial intelligence and machine learning methods (AI/ML) has the potential to revolutionize data analysis in the field of MS-based omics. The goal of this project is to continue the work implementing and evaluating false discovery rate methods in Python for MS omics data and contributing to the development of AI-based algorithms for molecular annotation. Methods will be integrated in AI-based workflows for processing and annotating molecular signatures of microbes in complex environmental samples, such as soil and plants, to support the study of the role of molecular processes in transformative science. Clustering methods are applied to the data to extract patterns. The generation of heatmaps for outlier detection methods results are also part of the analysis. Venn diagrams are generated to visualize common outliers detected by different algorithms. Finally, we built a graphical user interface (GUI) to offer an interactive data visualization to the user, which generates a histogram of confidence for targets/decoys matching outputs and display bar plots to visualize abundances and mass-to-charge errors. It includes an option for the user to choose a specific threshold for the false discovery rate estimation. The interface contains tow buttons to select csv result files and functional buttons to establish bar plots adequate

for abundance and mass to charge errors. We dive into the fundamental concepts of clustering methods, normalization techniques, and outlier detection algorithms. We employed K-means and hierarchical clustering techniques to partition a dataset of peptides into groups. This means that data points within the same cluster are more similar to each other compared to those in different clusters. However, before applying clustering algorithms, we preprocessed the data normalization techniques to ensure that features are on a similar scale using Min-Max scaling and Z-score normalization techniques. For the outlier detection methods, we adapted outlier Isolation Forest, Local Outlier Factor (LOF), and One-class SVM technics.

INTRODUCTION:

Identification of chemical analytes and species is a common application of mass spectrometry (MS). During the process, the sample undergo ionization to generate molecular ions, and acceleration through an electric field. A magnetic field deflects the species which are accelerated by an electric field and get detected by a sensor to measure their abundance. This produces a mass spectrum, which is a plot of ion abundance versus mass-to-charge ratio. The species can subsequently be determined using the charge to mass ratio. The mass spectrum obtained provides information about the mass of the ions present in the sample. By analyzing the peaks in the spectrum, scientists can determine the molecular weight of compounds present in the sample. Additionally, the fragmentation pattern of ions can provide information about the structure of molecules.

MS is helpful for finding new compounds and recognizing known molecules since it enables the chemical identification of an unknown species. It plays a crucial role in pharmaceutical analysis for drug discovery, development, and quality control. It is used for characterizing drug molecules, studying metabolism, and detecting impurities.

While mass spectrometry is highly sensitive for many compounds, it may not be equally sensitive for all types of molecules. Some compounds may not ionize efficiently or produce detectable signals at low concentrations. Classical machine learning approaches are of limited capabilities to analyze original mass spectrometry data at full spectral dimensions. Mainly, because those approaches suffer a common issue known as curse-of-dimensionality that deteriorates the clustering/classification accuracy on high-dimensional data [1].

PeakDecoder is an algorithm that automatically calculates error rates for metabolite identification, independently of spectral annotations or libraries [2]. This proposed method introduces an alternative approach to generating decoys from raw data-independent acquisition (DIA) spectra, integrating principles from DIA and spectral library searching into a machine learning (ML) framework that merges unbiased false discovery (UFD) and target-decoy competition strategies. In order to showcase the effectiveness of our metabolomics workflow and highlight its practicality, we implement this method in the analysis of microbial samples sourced from diverse strains. Our method relies on an object detection technique that assigns a unique color (RGB) to each ion precursor and fragment. By utilizing YOLOv8, an image detection tool, we trained the tool to distinct colors for overlapping regions of these ions (red, green, blue). The shaped white regions are considered for further processing and analysis. We apply some data visualization to explore our datasets efficiently by providing graphical representations that can reveal patterns and outliers that might not be immediately apparent in raw data.

I. METHODS AND RESULTS:

The work during this internship centered around data analysis using different statistical methods. The goal is to rate the false positive outcomes from data identifications of proteomics and metabolomics in biological samples. Our work involved programming code that would help doing the required operations and show meaningful results related to actual data experiments. We used some biological material and Python programming language to create a decoy database to use it for the false discovery rate (FDR) analysis. Using Python, we created a friendly useful graphical user interface (GUI) (see Figure 1) to contain boutons to generate histogram of confidence for target/decoys thresholds the user can tape in (see Figure 2-a). Another selection button is set to select a file from which the user will specify one peptide he is interested in. Two other boutons are set to generate abundance and m/z error bar plots for the peptide that the user selected (see Figure 2-b and Figure 2-c).

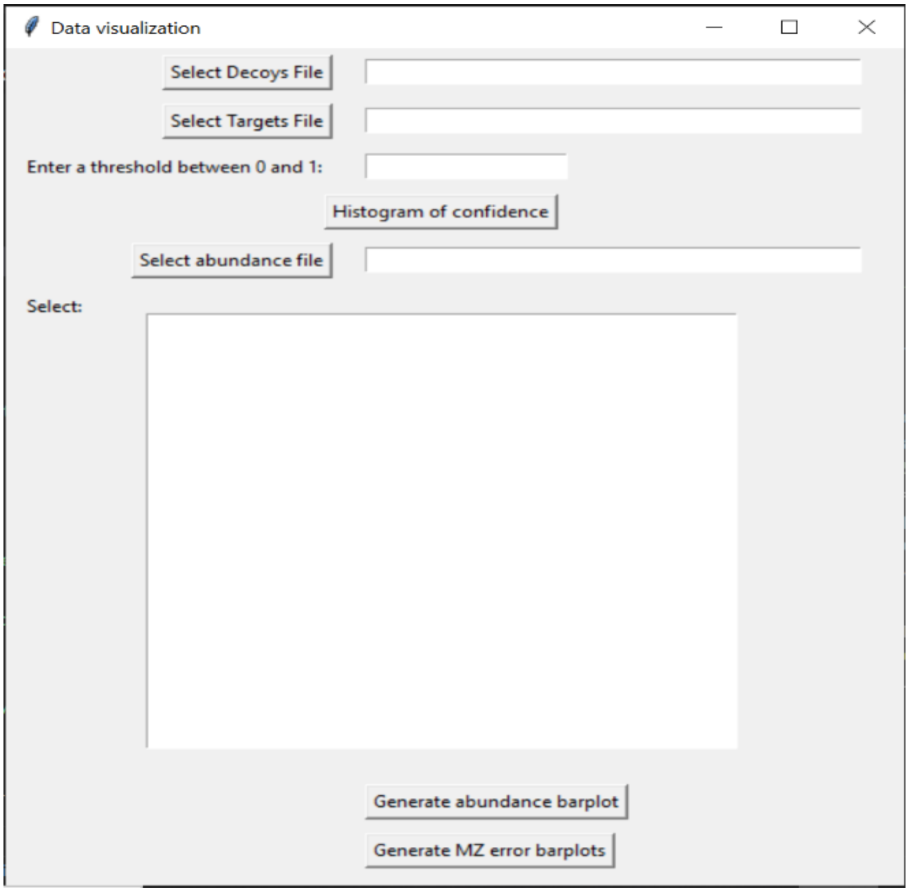

Figure 1: Data visualization graphical user interface.

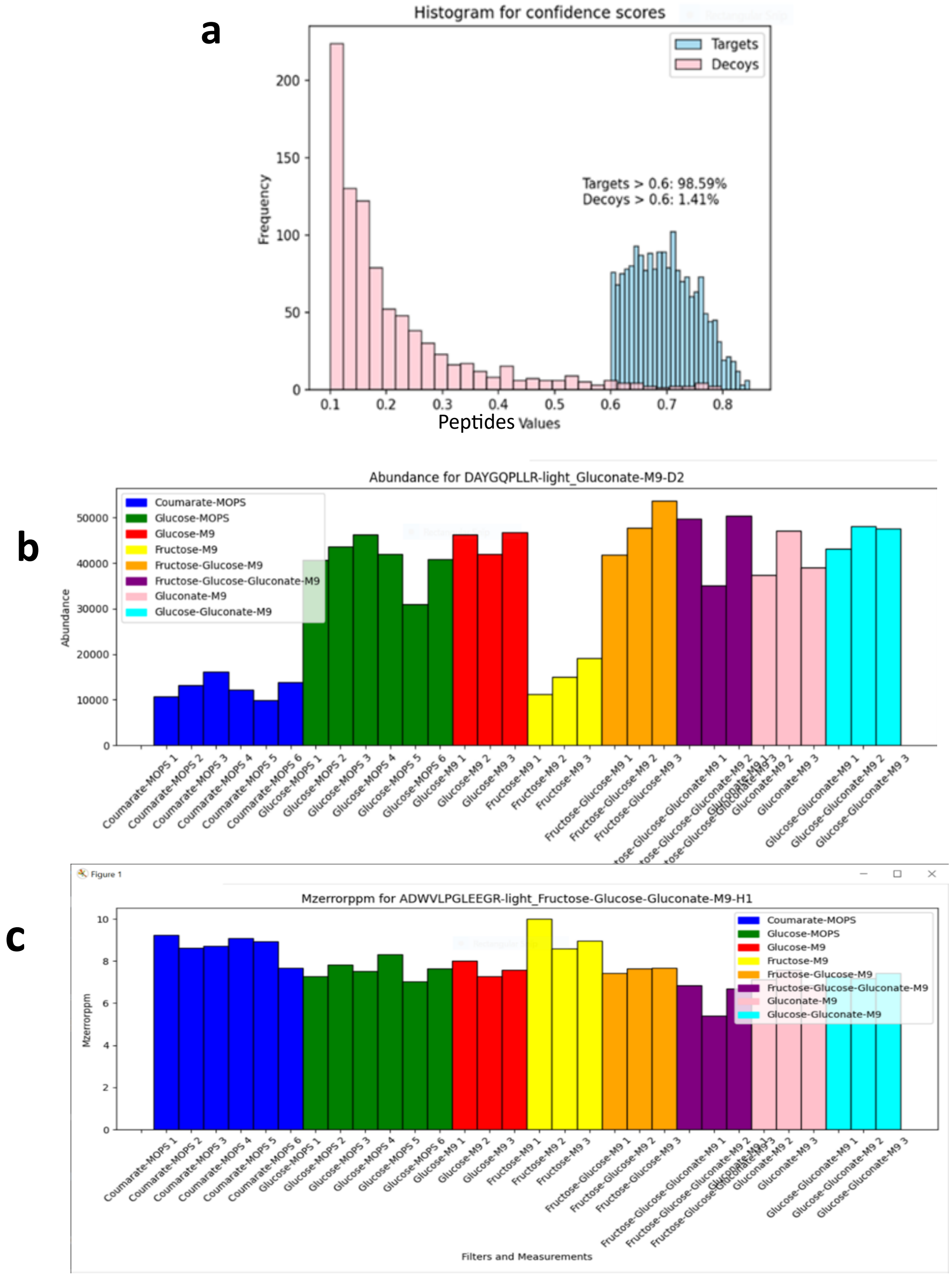

Figure 2: Histograms generated by the Data visualization GUI. a: Histogram of confidence for peptides. b: Histogram of abundance for a peptide. c: Histogram of Mzerrors for a peptide.

1. *Normalization and Clustering methods:*
   Clustering methods are commonly used in peptides and metabolomics classifications to group similar peptides or metabolites together based on their properties, such as mass-to-charge ratio, retention time, abundance, or structural features. We applied hierarchical clustering and k-means clustering to find subsets that are representative for the data identifying inherent patterns and group the data into clusters without prior knowledge of class labels. This reveals the natural groupings for retention time (Rt) against mass to charge ratio (Mz) within the data. We normalized the data before clustering, which rescales the data so that all variables have a similar scale, preventing variables with larger ranges from dominating the clustering process (see figure 3).

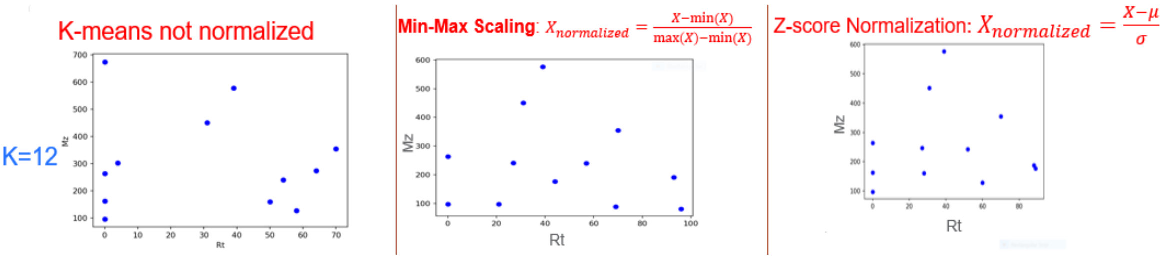

Figure 3: peptides clustering with and without normalization with k=12.

2. *Outlier detection methods:*
   Outlier detection methods are crucial in the analysis of peptides and metabolomics data to identify data points that deviate significantly from the rest of the dataset. These outliers can represent errors in data collection, sample contamination, or biologically interesting observations. We applied Isolation Forest, Local Outlier Factor and One-class SVM methods on our runs to classify the m/z errors on peptides. The outputs are fed into venn diagram to explicitly show how these results are close or different from each other.

II. FUTUR WORK:

During my science undergraduate laboratory internship (SULI) at PNNL, I have learned substantial advancements in programming and data analysis, yet there remains possibility for enhancements and further implementations within our project. While our model exhibits notable proficiency in training for proteins and metabolites, there's a necessity to extend its scope to include other biological data types. To enhance the reliability and comprehensibility of our findings, integration of false discovery rate, histogram of occurrences, clustering methodologies with machine learning models is imperative. Additionally, optimizing time efficiency for processing high-dimensional data is among our project's primary objectives. Our ambition is to develop a comprehensive graphical user interface tool in Python, facilitating data visualization using explicitly comparative histograms of different components results and FDR result interpretation. This work is dedicated for assisting biologists and researchers in their collaborative activities.

III. IMPACT ON LABORATORY AND NATIONAL MISSIONS:

This project contributed to augmented PNNL’s capabilities to analyze metabolomics and proteomics samples in large-scale studies. This capability will be directly beneficial to DOE and PNNL efforts to characterize and analyze compounds in microbial and plant communities.

IV. CONCLUSION

Analyzing small molecules through mass spectrometry is vital for comprehending biochemical processes across various domains such as the environment, oceans, and individual organisms. Transitioning from manual selection and statistical calculations to automated methods holds immense promise for researchers, facilitating more efficient and accurate analyses. The integration of artificial intelligence tools for collecting biological samples and deriving meaningful statistical insights represents a significant benefit for laboratory work. Through this second research internship, I acquired proficiency in coding and deepened my understanding of fundamental biological principles. I dived more in machine learning and data visualization, the tools for interpretation and explanation for molecules processing results. This internship not only provided insights into the working of national laboratories but also equipped me with valuable professional skills essential for my future career vocations.

Upon its completion, this new project will stand as a groundbreaking attempt, packed with innovative features designed to enhance the efficacy of biologists' endeavors. It serves as a significant contribution to the Pacific Northwest National Laboratory by rationalizing the time and cost associated with proteomics and metabolomics analysis. By facilitating preprocessing, peak picking, feature extraction, and noise reduction, it elevates the quality and accuracy of data. The fusion of mass spectrometry with machine learning presents compelling opportunities to glean invaluable insights from complex datasets, potentially driving advancements in environmental analysis and various other scientific domains pertinent to the laboratory's mission.

V. ACKNOWLEDGEMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers, and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

VI. APPENDIXES

Participants

	Aivett Bilbao	Mentor Computational Scientist

	Ashfiqur Rahman	Post master’s Research Associate
	Samiha Chabane	SULI Intern
	Andrea Harrisson	SULI Intern

Scientific Facilities

This work was performed at the Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830.

VII. REFERENCES

1. Abdelmoula, W. M. (2021, September 20). Peak learning of mass spectrometry imaging data using artificial neural networks. https://doi.org/10.1038/s41467-021-25744-8.

2. Bilbao, Aivett. PeakDecoder Enables Machine Learning-based Metabolite Annotation and Accurate Profiling in Multidimensional Mass Spectrometry Measurements. 28 Apr. 2023.