MatchMass: A Web-Based Tool for Efficient Mass Spectrometry Data Analysis

Authors:
Lukas Ustrnul [a,*], Tatsiana Jarg [a], Martin Jantson [b], Irina Osadchuk [a], Lauri Anton [b], Riina Aav [a,*]

Affiliations:
[a] Department of Chemistry and Biotechnology, School of Science, Tallinn University of Technology, Akadeemia tee 15, 12618, Tallinn, Estonia
[b] IT College , School of Information Technologies, Tallinn University of Technology, Raja 4C, 12616, Tallinn, Estonia
[*] corresponding authors

Highlights:

- MatchMass is a free web application for identifying molecules in complex mixtures studied with mass spectrometry (MS).
- MatchMass allows upload of theoretical masses and multiple experimental data, followed by comparison, matching and aggregating of results.
- The tool offers features for setting matching criteria (ions of interest, mass accuracy, and abundance threshold), visualizing results, and downloading reports.

Abstract
MatchMass is a user-friendly web-based application designed to streamline the analysis of mass spectrometry (MS) data for researchers studying complex mixtures of molecules. Users upload experimental data (*m/z* and abundance) and a list of theoretical monoisotopic masses. MatchMass then identifies matching molecules within the sample based on user-defined parameters (mass accuracy, abundance threshold, and ions of interest). The tool generates a visual comparison of experimental and theoretical *m/z* values, facilitates result exploration, and allows download of comprehensive reports. MatchMass eliminates the need for manual data processing, improving efficiency and accuracy in MS data analysis, particularly for researchers working with complicated mixtures.

Keywords: MatchMass, Web application, Mass spectrometry, User-friendly, Complex mixtures analysis, High-throughput analysis

1. Introduction

Mass spectrometry (MS) is a powerful analytical technique widely used across various scientific disciplines for the identification and quantification of molecules. It offers high sensitivity and resolution, allowing researchers to analyze a broad range of samples, from simple single-component systems to complex mixtures containing numerous molecules, which is crucial for materials science, and life sciences fields such as proteomics, metabolomics, and mechanistic research in chemistry [1]. However, analyzing data from complex mixtures measured by MS is a significant challenge. Therefore, many different tools were developed in past decades to streamline MS data analysis, often employing various databases and more recently machine learning [2–5]. The wide range of possible sample matrices as well as large number of MS techniques — including various ionization options and tandem MS — led to large extent of specialization of these tools [6] and at the same time made their use more difficult for non-frequent users.

When we studied oligomerization reactions of functionalized urea-based monomers, there was no free-to-use user-friendly tool, which could simplify our data analysis in desired extent, for example MS tool for analyzing polymerization reactions [7] provided efficient analysis for each dataset but could not aggregate results from multiple parallel measurements. Hence, we found need for a simpler tool which would utilize users provided list of masses of interest. We introduce MatchMass [8,9], a user-friendly web application designed to streamline the analysis of MS data (Figure 1) for moderately complex mixtures. MatchMass eliminates the need for manual data processing and offers a suite of

features to facilitate efficient and accurate molecule identification and aggregation of results from multiple measurements.

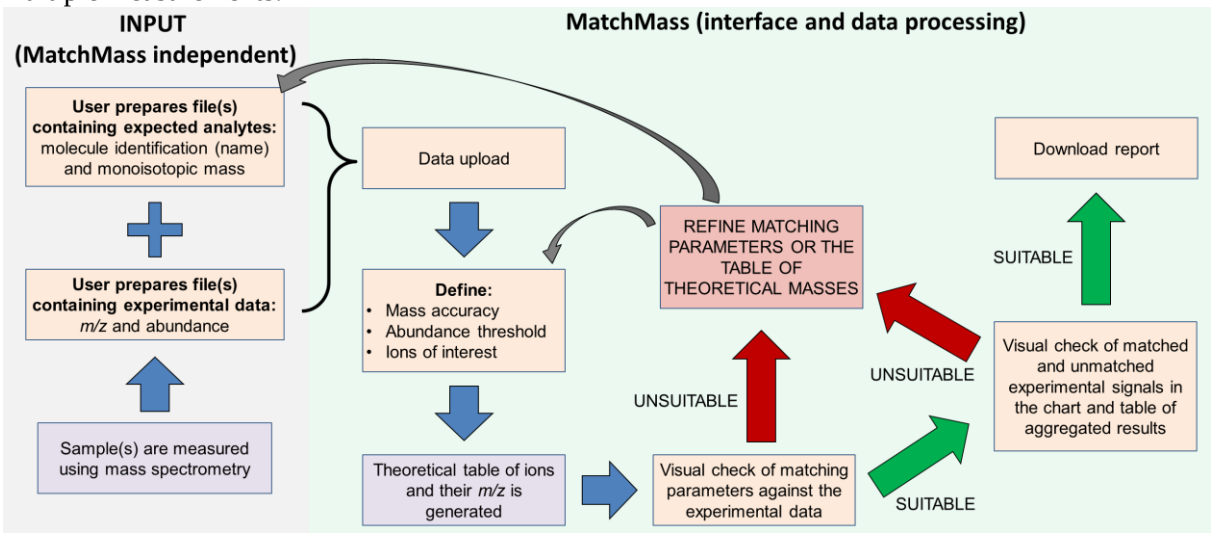

Figure 1: General workflow for MS data analysis in MatchMass web application.

2. Results and Discussion

MatchMass is a web application written in Python [10] that performs calculations, data matching, grouping, and visualization for MS data. MatchMass uses the Pandas [11] and NumPy [12] libraries to process the experimental MS data and the table of theoretical monoisotopic masses, perform their matching and aggregate results from multiple files containing experimental data.

MatchMass utilizes the Plotly [13] library for generating plots and is bundled in the Streamlit [14] library to present them in a form of web application. MatchMass can be used and accessed at https://matchmass.taltech.ee[1] after logging in with ORCID profile [15]. It can process experimental data from any type of MS if the file containing theoretical masses involves all molecules and fragments of interest. However, MatchMass is best suited for evaluating of data from MS methods employing soft ionization techniques.

We tested its capabilities by evaluating MS data from reaction mechanism study which employed two different urea-based monomers in oligomerization and macrocyclization reaction under solid-state synthesis (ball-milling) conditions. Specifically, the influence of ball-milling duration and aging on the content of intermediates and products was studied and the results will be discussed in other article focused on synthesis of new derivatized macrocycles. MatchMass allowed us to efficiently evaluate all experimental data simultaneously and observe the changes in reaction mixture composition.

2.1. MatchMass functionality and data processing

The core functionality of MatchMass relies on a matching algorithm that compares the experimental mass-to-charge (*m/z*) values from the uploaded data file with the theoretical masses provided by the user and expanded with ions of interest. This algorithm takes into account the user-defined mass accuracy parameter. For each experimental *m/z* value, the algorithm searches within a defined window around the theoretical mass to identify potential matches. The width of this window is determined by the mass accuracy value. A higher mass accuracy (its lower value) setting translates to a narrower matching window, resulting in more precise but potentially fewer matches. Conversely, a lower mass accuracy allows for a wider matching window, potentially capturing more matches but also increasing the likelihood of including false positives. In addition, each matched signal which falls within matching window of more than one ion is assigned with warning note. However, the algorithm always provides match to the nearest *m/z* in the theoretical table.

The matching process considers the abundance threshold defined by the user. Only experimental signals exceeding this threshold are included in the matching process. This helps to eliminate

[1] As of April 2023, MatchMass will not be fully functional in the taltech domain until the ORCID login implementation is complete. MatchMass can be temporarily tested at Streamlit community cloud on the link: https://matchmass.streamlit.app/

background noise and signals from low-abundance species that may interfere with the identification of target molecules.
The selection of ions (i.e., positive or negative), which should be added to the theoretical table of masses is as simple as possible via ticking checkboxes for the most common ions. Adduct ions are calculated and added for each molecule in the user provided table. The available ions were chosen based on comprehensive quantitative study [16] and our own practical needs. The calculation of *m/z* for adduct ions considers the electron mass [17] and leverages values from the Fiehn Lab's Excel-based mass spectrometry adduct calculator [18].
MatchMass generates two visual representations of the data. First, allowing users to compare the experimental and theoretical *m/z* values (Figure 2a). This visualization helps users assess the quality of the match and identify potential areas of interest before the matching procedure. Second, it employs color-coding to clearly differentiate between experimental signals that found a match in the theoretical table and those that didn't found match (Figure 2b).

Figure 2: MatchMass offers two charts to aid analysis visually. The first chart (**a**) lets you verify how your chosen matching settings, like mass accuracy and adduct ions, align with the experimental data. The second chart (**b**), visible only after a successful match, shows matched and unmatched signals.

Following the matching process, MatchMass provides users with visualization mentioned above, the table of result aggregated for all provided experimental files, and downloadable report. The aggregated table provides information such as the theoretical *m/z* of the matched molecule adducts, the corresponding average experimental *m/z* value with standard deviation, the abundance of the detected ions in each experiment, and total abundance of all ions within each molecule and experiment.
The downloadable report (.xlsx format) provides comprehensive results and matching details on separate sheets. The first sheet lists uploaded experimental files with their applied mass accuracy and abundance thresholds. The second sheet includes the expanded theoretical table with adduct ions and their *m/z* values. It also includes a warning about potential mismatches, generated based on the largest mass accuracy value. The third sheet mirrors the aggregated results shown in the MatchMass interface after analysis. Finally, individual matching results for each experimental file are provided in a dedicated sheet.
The use of MatchMass offers automatic matching process which minimizes the risk of human error and improves the overall accuracy of results. The user-friendly interface and clear data visualization features make MatchMass accessible to researchers with varying levels of experience in MS data

analysis. Finally, the web-based nature of the tool eliminates the need for software installation and allows users to access it from any device with an internet connection.

2.2. MatchMass workflow

MatchMass offers a user-friendly interface that guides researchers through the data analysis process (Figure 1, detailed instructions in SI). Users begin by preparing two key pieces of information: **I**) Experimental data file containing two columns. The first column represents the *m/z* values of the detected ions, and the second column represents the abundance (intensity of the signal) for each *m/z* value. **II**) Theoretical masses file containing a list of names (or other identification) for the molecules of interest and corresponding theoretical monoisotopic masses. In addition, MatchMass offers dummy data for testing its functionality without need for your own data.

After uploading data, MatchMass allows users to define key parameters that influence the matching process: mass accuracy, abundance threshold, and adduct ions of interest (Figure 3). These parameters are reflected in a chart that overlays experimental signals with colored bands representing theoretical *m/z* values (Figure 2a). The width of these bands corresponds to twice the set mass accuracy.

Users can refine the matching parameters if needed. Otherwise, they can proceed by initiating the match and reviewing the results. These include a chart with color-coded signals for successful matches (Figure 2b) and a table summarizing the findings. Finally, users can either repeat the match with different settings or download the comprehensive report.

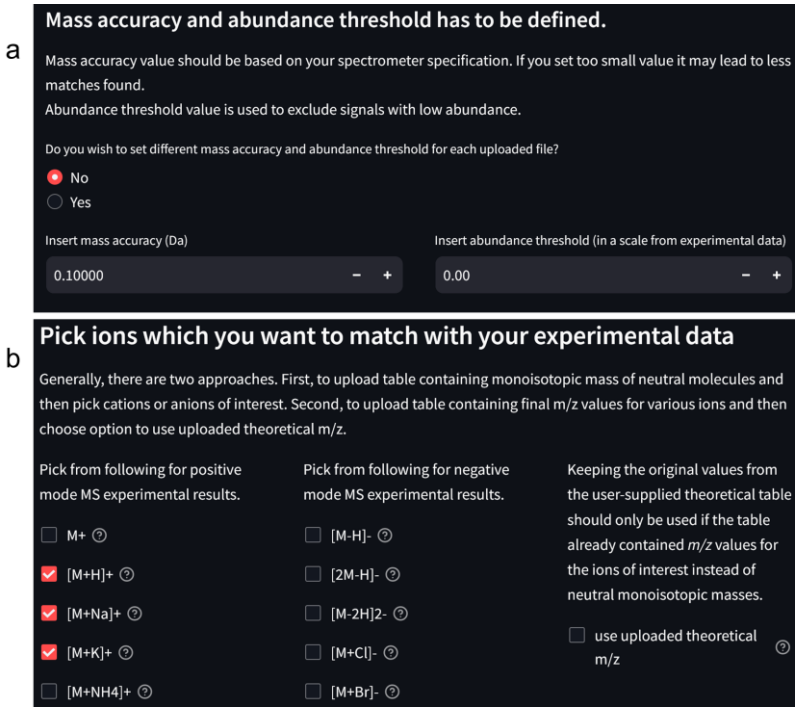

Figure 3: The MatchMass matching process hinges on three user-defined parameters. Setting the mass accuracy (**a**) according to your spectrometer's specifications ensures precise matching of theoretical and experimental *m/z* values. The abundance threshold (**a**) acts as a filter, removing background noise and low-abundance signals in the experimental data. Finally, selecting relevant adduct ions from the checkboxes (**b**) expands the theoretical *m/z* table.

2.3. Limitations of MatchMass

MatchMass has limitations in two main areas: technological aspects and data analysis capabilities.

On the technological side, file size is limited to 200MB due to processing demands for larger files. Additionally, while users can define individual mass accuracy and abundance thresholds for each experimental file, the theoretical table is generated using the largest individual value. This leads to a single, generalized warning for all files regarding potential mismatches arising from overlapping theoretical *m/z* values within the set mass accuracy range.

For data analysis, MatchMass may not be ideal for highly complex mixtures with overlapping signals within mass accuracy ranges. While searching for isotopic patterns could improve identification in such cases, this feature is currently not available. Additionally, MatchMass doesn't connect to external databases, requiring users to create their own theoretical mass tables using other tools and calculators. Furthermore, the selection of available adduct ions is limited to most common ones. However, researchers can address this by providing a custom table containing their desired *m/z* values for specific adduct ions. They can then proceed with matching using these values without adding further ions within MatchMass. Finally, the matching process excludes unmatched experimental signals from the final report. However, users who might find these signals relevant can inspect them in the chart visualizing matching results.

3. Conclusion

MatchMass offers user-friendly solution for researchers working with moderately complex MS data, streamlines analysis through automated data processing, customizable matching criteria, and clear result visualization. Users provide a table of theoretical masses, which MatchMass expands with user-selected ions and compares to experimental data within set parameters. It can handle and aggregate experimental data from multiple files. Complete results can be downloaded in a format of comprehensive report. MatchMass has shallow learning curve and thus allows user to easily analyze MS data and interpret results.

Hence, the main benefits of MatchMass are its ease of use, fast generation of *m/z* for ions of interest, aggregation of results from multiple files, comprehensive reporting, and clear uncomplicated web-based user-friendly interface. By introducing these pipelines to the community, we wish to enable the easy analysis of the results to specialist who desires simpler and more universal tool for casual use than the most of currently available specialized software and online applications.

In MatchMass, we use common file formats (.csv, .xlsx, .xls) as an input. This allows easy data preparation and independence on file formats provided by spectrometer producers. In addition, MatchMass is written in Python and is open source; therefore, a further development to include additional calculations or features is simple. Moreover, MatchMass is not necessarily limited to MS data. In principle, any data stored in supported format can be loaded and matched in MatchMass-based pipelines.

4. Methods

MatchMass software development. MatchMass is written in Python [10] using the following libraries, such as NumPy [12] for working with arrays, Pandas [11] to serve the data as a table or data frame, and Matplotlib [19] and Plotly [13] to visualize the results and to provide users with necessary interactivity. These scripts are utilized by a user-friendly interface built with Streamlit [14] environment. MatchMass is hosted and deployed in a local server (in virtual machine running in Ubuntu Server) in Tallinn University of Technology (Taltech) accessible at https://matchmass.taltech.ee domain.[2] Data for analysis can be uploaded only by logged-in users (login utilizes ORCID account). MatchMass prioritizes temporary data storage. User-uploaded files are processed only within the current session and automatically deleted from the server after the session ends or the web page reloads. The most recent version of MatchMass's source code is available in main author's GitHub repository https://github.com/lukasustrnul/MatchMass.

Writing of the article. The initial draft of this paper was written with the assistance of Gemini, a large language model from Google AI, to streamline the writing process.

CRediT authorship contribution statement

Lukas Ustrnul: Conceptualization, Methodology, Software, Validation, Formal Analysis, Data Curation, Writing – Original Draft, Visualization, Project Administration. **Martin Jantson:** Software. **Tatsiana Jarg:** Conceptualization, Validation, Formal Analysis, Investigation, Writing - Review &

[2] As of April 2023, MatchMass will not be fully functional in the taltech domain until the ORCID login implementation is complete. MatchMass can be temporarily tested at Streamlit community cloud on the link: https://matchmass.streamlit.app/

Editing. **Irina Osadchuk:** Data Curation. **Lauri Anton:** Project Administration. **Riina Aav:** Fund Acquisition, Writing - Review & Editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by Estonian Research Council grants PRG399, MOBJD592, and PRG2169. We also thank Dr. Mario Öeren (Tallinn University of Technology) for kind advice and suggestions on code improvements.

References

[1] D.A. Boiko, K.S. Kozlov, J.V. Burykina, V.V. Ilyushenkova, V.P. Ananikov, Fully Automated Unconstrained Analysis of High-Resolution Mass Spectrometry Data with Machine Learning, J. Am. Chem. Soc. 144 (2022) 14590–14606. https://doi.org/10.1021/jacs.2c03631.

[2] F. Zhang, W. Ge, G. Ruan, X. Cai, T. Guo, Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020, PROTEOMICS 20 (2020) 1900276. https://doi.org/10.1002/pmic.201900276.

[3] B.B. Misra, New software tools, databases, and resources in metabolomics: updates from 2020, Metabolomics 17 (2021) 49. https://doi.org/10.1007/s11306-021-01796-1.

[4] A. Halder, A. Verma, D. Biswas, S. Srivastava, Recent advances in mass-spectrometry based proteomics software, tools and databases, Drug Discov. Today Technol. 39 (2021) 69–79. https://doi.org/10.1016/j.ddtec.2021.06.007.

[5] Q. Yang, H. Ji, Z. Xu, Y. Li, P. Wang, J. Sun, X. Fan, H. Zhang, H. Lu, Z. Zhang, Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale in-silico library, Nat. Commun. 14 (2023) 3722. https://doi.org/10.1038/s41467-023-39279-7.

[6] Online Mass Spectrometry Tools: The ISIC- EPFL mstoolbox, (2019). https://ms.epfl.ch/ (accessed April 22, 2024).

[7] J.S. Desport, G. Frache, L. Patiny, MSPolyCalc: A web-based App for polymer mass spectrometry data interpretation. The case study of a pharmaceutical excipient, Rapid Commun. Mass Spectrom. 34 (2020) e8652. https://doi.org/10.1002/rcm.8652.

[8] L. Ustrnul, MatchMass, (2024). https://github.com/lukasustrnul/MatchMass (accessed April 22, 2024).

[9] L. Ustrnul, M. Jantson, MatchMass, (n.d.). https://matchmass.taltech.ee/ (accessed April 22, 2024).

[10] Welcome to Python.org, Python.Org (2024). https://www.python.org/ (accessed April 23, 2024).

[11] J. Reback, W. McKinney, jbrockmendel, J.V. den Bossche, T. Augspurger, P. Cloud, gfyoung, S. Hawkins, Sinhrks, M. Roeschke, A. Klein, T. Petersen, J. Tratner, C. She, W. Ayd, S. Naveh, M. Garcia, patrick, J. Schendel, A. Hayden, D. Saxton, V. Jancauskas, R. Shadrach, M. Gorelli, A. McMaster, P. Battiston, S. Seabold, K. Dong, chris-b1, h-vetinari, pandas-dev/pandas: Pandas 1.2.3, (2021). https://doi.org/10.5281/zenodo.4572994.

[12] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362. https://doi.org/10.1038/s41586-020-2649-2.

[13] P.T. Inc, Collaborative data science, (2015). https://plot.ly.

[14] Streamlit • A faster way to build and share data apps, (2021). https://streamlit.io/ (accessed April 23, 2024).

[15] ORCID, ORCID (n.d.). https://info.orcid.org/ (accessed April 23, 2024).

[16] M.R. Blumer, C.H. Chang, E. Brayfindley, J.R. Nunez, S.M. Colby, R.S. Renslow, T.O. Metz, Mass Spectrometry Adduct Calculator, J. Chem. Inf. Model. 61 (2021) 5721–5725. https://doi.org/10.1021/acs.jcim.1c00579.
[17] I. Ferrer, E.M. Thurman, Importance of the electron mass in the calculations of exact mass by time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 21 (2007) 2538–2539. https://doi.org/10.1002/rcm.3102.
[18] O. Fiehn, Mass Spectrometry Adduct Calculator, (n.d.). https://fiehnlab.ucdavis.edu/staff/kind/metabolomics/ms-adduct-calculator (accessed April 23, 2024).
[19] Matplotlib documentation — Matplotlib 3.8.4 documentation, (n.d.). https://matplotlib.org/stable/index.html (accessed April 23, 2024).