Combined LC-MS/MS feature grouping, statistical prioritization, and interactive networking in msFeaST

Kevin Mildau[1,2,3,*], Christoph Büschl[4], Jürgen Zanghellini[2], andJustin J.J. van der Hooft[1,5,*].

April 16, 2024

1 Bioinformatics Group, Wageninen University, Droevendaalsesteeg, 6708PB, Gelderland, The Netherlands
2 Department of Analytical Chemistry, University of Vienna, Währinger Straße,1090, Vienna, Austria
3 Doctoral School in Chemistry (DOSCHEM), University of Vienna, Währinger Straße,1090, Vienna, Austria
4 Institute of Bioanalytics and Agro-Metabolomics, University of Natural Resources and Life Sciences, Konrad-Lorenz-Straße, 3430, Lower Austria, Austria
5 Department of Biochemistry, Auckland Park Kingsway Campus, 2006, University of Johannesburg, Gauteng Province, South Africa
∗ Corresponding authors: kevin.mildau@wur.nl, justin.vanderhooft@wur.nl

Abstract

Computational metabolomics workflows have revolutionized the untargeted metabolomics field. However, the organization and prioritization of metabolite features remains a laborious process. Organizing metabolomics data is often done through mass fragmentation-based spectral similarity grouping, resulting in feature sets that also represent an intuitive and scientifically meaningful first stage of analysis in untargeted metabolomics. Exploiting such feature sets, feature-set testing has emerged as an approach that is widely used in genomics and targeted metabolomics pathway enrichment analyses. It allows for formally combining groupings with statistical testing into more meaningful pathway enrichment conclusions. Here, we present msFeaST (mass spectral Feature Set Testing), a feature-set testing and visualization workflow for LC-MS/MS untargeted metabolomics data. Feature-set testing involves statistically assessing differential abundance patterns for groups of features across experimental conditions. We developed msFeaST to make use of spectral similarity-based feature groupings generated using k-medoids clustering, where the resulting clusters serve as a proxy for grouping structurally similar features with potential biosynthesis pathway relationships. Spectral clustering done in this way allows for statistical testing on the scale feature-sets using the globaltest package, which provides high power to detect small concordant effects via joint modeling and reduced multiplicity adjustment penalties. Hence, msFeaST provides interactive integration of the semi-quantitative experimental information with mass-spectral structural similarity information, enhancing the prioritization of features and feature sets during exploratory data analysis. The msFeaST workflow is available through `https://github.com/kevinmildau/msFeaST` and built to work on MacOS and Linux systems.

1 Introduction

Untargeted metabolomics deals with the comprehensive characterization of the composition of small chemicals, or metabolites, in biological samples. Typically, high-resolution LC-MS/MS (Liquid Chromatography Tandem Mass Spectrometry) workflows are used to provide comprehensive snapshots of the metabolome (Wolfender et al., 2018). However, despite recent advances in computational metabolomics, the reliable annotation of ms/ms spectral data remains a challenge (de Jonge et al., 2022). Hence, further complementary analyses, laborious manual annotations, and detailed validations of putative structure hypotheses remain a necessity (Wolfender et al., 2018; Beniddir et al., 2021). The exploratory analysis of typical untargeted metabolomics data is a daunting task due to large data heterogeneity and largely unknown chemical spaces. Here, molecular networking has established itself as a vital tool enabling researchers to organize and prioritize features for in-depth evaluations (Watrous et al., 2012; Nothias et al., 2020). Molecular Networking comprises a combination of data subdivision and exploratory data visualization. Large heterogeneous datasets are subdivided into smaller, more manageable feature groups based on spectral similarity scoring (Nothias et al., 2020; Mildau et al., 2024). Those feature groups are further presented to the user as subnetworks (molecular families) for exploration of relationships across features within them (Nothias et al., 2020). While advantageous as a first analysis step, relationships between subnetworks are lost completely (Olivon et al., 2018; Mildau et al., 2024), parameter setting and optimization are time-consuming and opaque, and further customization of resulting molecular networks in Cytoscape is usually necessary to highlight statistical features for sub-network for feature prioritization (Pakkir Shah et al., 2023).

While statistical data can often be manually integrated into Cytoscape networks, there is currently a lack of workflows in untargeted metabolomics that provide both spectral clustering and statistical data integration on the spectral group level aimed at feature prioritization (McLuskey et al., 2021). As a notable exception, the PALS (Pathway Activity Level Scoring) workflow makes use of molecular families, i.e. feature groupings representing "pathways" via linking together chemical analogues, for latent variable summation-based statistical comparisons across treatment groups (McLuskey et al., 2021; Tomfohr et al., 2005). However, PALS stops short of visually integrating such analysis results back into the network representations it is based on. The field is thus still lacking an integrated approach providing spectral data clustering, statistical analysis at the group level, and integrated visualization of the results of both of these features for streamlined exploratory data analysis.

To fill this gap, we developed msFeaST (**M**ass **S**pectral **Fea**ture **S**et **T**esting), a comprehensive workflow for integrated analysis and visual exploration. msFeaST draws inspiration from gene-set testing and pathway enrichment analyses, which shift the focus from the often large and cluttered feature space to aggregated groups such as gene ontologies or pathways. This approach provides the basis for meaningful scientific conclusions in group-based comparisons of experimental conditions that can then be further explored in detail (Maleki et al., 2020; Rosato et al., 2018; Chong et al., 2020). To achieve a similar approach in untargeted metabolomics, where genes are replaced by metabolite features with unknown pathway membership, we make use of k-medoids clustering on the pairwise similarity matrix to provide more homologous subdivisions of the data with higher implied structural relationship as in molecular networking (Mildau et al., 2024; Schubert and Rousseeuw, 2021). Feature-set testing is then performed for each cluster using globaltest, providing high power to detect small concordant treatment-specific effects across cluster members via a single statistical test (Goeman et al., 2004). msFeaST draws visualization inspiration from both specXplore and MetGem making use of t-SNE embeddings to represent the entire spectral dataset based on the spectral similarity matrix in one overview (Olivon et al., 2018; Mildau et al., 2024; Maaten, 2008). Furthermore, it draws inspiration from EdgeMaps and specXplore emphasizing topological views of neighborhoods via top-k neighbor edge drawings and group-based color overlays (Mildau et al., 2024; Dork et al., 2011).

2 Methods & Implementation

The msFeaST workflow is a combination of data clustering, statistical testing at the cluster level, and interactive visualization using overlays of two-dimensional embeddings with an ego-network, i.e., a node-centric neighborhood graph. An overview of the Python and R-based processing pipeline can be seen in Fig. 1.

2.1 Data Processing for msFeaST

The msFeaST pipeline requires spectral data, feature quantification tables, and statistical group indicator data for samples to work. When working with feature based molecular networking data, preprocessing functionalities are available to load the data accordingly (Nothias et al., 2020). Once the data are loaded, the msFeaST pipeline method is initialized and handles all intermediate data structures. The workflow guides the user through the data processing steps in the form of spectral similarity computations using the desired similarity score, k-medoid clustering based on the resulting distance matrix, and t-SNE embedding using the same structure. We refer to supplementary sections 1.1 to 1.4 for further details. Here, we will focus specifically on the statistical testing and interactive visualization components.

2.2 Statistical Testing via globaltest

Within this step of the pipeline the clustered data are passed to a R script for group-wise testing. The msFeaST workflow makes use of the well established globaltest method, a feature-set testing approach based on a so-called self-contained null hypothesis (Goeman et al., 2004; Goeman and Oosting, 2023; Goeman and Bühlmann, 2007; Maleki et al., 2020).

The test was originally developed for genetic analyses, but has since found use in the field of targeted metabolomics as well (Rosato et al., 2018; Chong et al., 2020). The test is based on a generalized linear model which aims to assess whether any of the feature-specific effects in a pre-specified group of features have predictive utility in differentiating treatment groups, a task that is closely related to testing whether treatment group-specific effects are not null (for further details see supporting information section 4). This test does not require concordance of effects but is considered powerful at detecting even small concordant effects (Goeman et al., 2004). Within msFeaST, each feature-set and contrast combination, that is control vs [treatment group 1, treatment group 2, etc.], is tested using this model, and p-values for the group as a whole, as well as feature-specific p-value contributions, are extracted. Multiple testing correction is done using the Bonferroni method at the family size given by the number of groups times the number of contrasts. This correction is usually substantially smaller than the correction needed if feature-specific testing was performed, resulting in less stringent cutoff adjustments. However, special attention is needed when using globaltest with small sample sizes as results may be unreliable when using very small numbers of statistical units per group (Maleki et al., 2019).

2.3 Visual Integration via Interactive Dashboard

In this final step of the processing workflow results from the workflow are integrated into a .json formatted file that can be read by the msFeaST visualization dashboard. The visualization dashboard is run as a local browser-based javascript tool and makes use of the plotly and visjs libraries (Inc., 2015; visjs community, 2024). Upon opening the website in a modern browser (e.g., Firefox, Chrome, Safari, Edge, etc.), the user may load the generated json file.

Loading the data initiates the interactive visualization tab of the dashboard (Fig. 1). The leftmost panel contains the t-SNE embedding of the spectral similarity matrix, which serves as both a global data view and an exploratory analysis medium. Node sizes encode feature-specific statistical mappings in the form of either feature-specific p-value contributions from globaltest, or log-2 fold changes (absolute) for the contrast in question. Dropdown menus allow toggling between measures displayed and contrasts. Users may hover over nodes to receive node information (e.g., node identifier, group identifier), and click to receive node details including feature-specific and group-specific statistics results, as well as local topology via edge overlays giving a glimpse into the top-K neighbors of the feature. Edge overlays make use of a discrete similarity mapping allowing quick assessment of similarity, while quantitative similarity score labels provide more precise information. Clicking successively on different nodes adds additional edge overlays until a click on the empty canvas is used to reset edge overlays. In this way, focused local node neighborhoods can be explored easily without causing computational bottlenecks or visual overload. In addition to the t-SNE embedding, the tool allows the user to make use of iterations of force-directed layout stabilization to mitigate overlapping nodes caused by the t-SNE embedding. To assist group-based prioritization, the network view is supported by a connected heatmap representation of group-level statistics which provides a quick reference to groups with statistically significant deviations for the different selected contrasts. Clicking on a group entry in the heatmap highlights the respective group in the t-SNE embedding.

3 Illustrative Example

To showcase msFeaST we make use of the high-resolution LC-MS/MS data from the study of Kathib et al. (Khatib et al., 2024). Specifically, we make use of the contrast of the *Pleurotus eryngii* fruiting body samples grown on 0% olive mill solid waste substrate against those grown on 80% olive mill solid waste substrate. While each treatment group contained only 3 samples each, substantial differences in the metabolome can be observed with msFeaST, where many metabolites show elevated log-2 fold changes, and a lower number showing corresponding statistical enrichment patterns. We note that msFeaST allows generating a snapshot overview of features with differential abundance, both using set-wise p-values and feature-specific descriptive fold changes and p-value contributions within the globaltest set statistic. Feature subsets of potential interest are easily spotted and their local topology can be explored (see supplementary figure 1). For example, the differential feature in Figure 1 B is visualized with its top 30 neighbor nodes, highlighting relatively strong connectivity of this feature beyond its small feature cluster highlighted in color. Edges discretely encode similarity while providing quantitative labels with exact similarity values, allowing inspection not only of top-k neighbors but also the degree of their potential relationship.

The processed data files are available on github with the respective processing notebooks and can be used to inspect the results within the visualization dashboard. No installation is required for this, only the pre-processed .json file and the html dashboard bundle that can be used using web-browsers are needed (see github readme quickstart). In addition to files for the illustrative example shown here, additional runs using the modified cosine score, as well as a modified cosine score comparison using a comparison between the different mushroom types are included. The latter show large differential intensity patterns resulting in wide scale feature highlighting.

4 Conclusion

The msFeaST workflow separates visual design considerations from data subdivision and clustering criteria. Clustering is dealt with separately, using its own optimality criteria. It thus allows for a more principled approach to subdividing spectral data that is not impacted by network visualization settings aimed at limiting visual overload (such as edge thresholds, top-K edge limits on each node, maximum cluster sizes). This separation is made possible by making use of its network visualization approach that remains manageable via “interactive details on demand”. Here, interactive local topological neighborhood explorations guided by statistical information integration provide a robust means of exploring and visualizing data that can adapt to different similarity scores without extensive visual tuning. The msFeaST workflow thus seamlessly combines formal spectral similarity-based feature clustering, feature cluster prioritization using statistical contrast information, and local topological neighborhood exploration in a molecular networking like context. We believe that msFeaST can assist researchers in better understanding their untargeted metabolomics data and identifying relevant chemistry by leveraging both qualitative spectral and quantitative relative intensity information.

5 Competing interests

J.J.J.v.d.H. is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy, and consults for Corteva Agriscience, Indianapolis, IN, USA. All other authors declare to have no competing interests.

6 Author contributions statement

K.M. and J.J.J.v.d.H conceived the tool. K.M. implemented code for the tool, wrote the initial manuscript, and implemented the illustrative example. J.J.J.v.d.H, C.B., and J.Z are part of the supervision team of K.M and reviewed and edited the manuscript.

7 Acknowledgments

The authors would like to thank Soliman Khatib for providing the mushroom data used in the illustrative example. KM would like to thank Henry Ehlers and Mara Meisenburg for their valuable feedback on the project.

References

M. A. Beniddir, K. B. Kang, G. Genta-Jouve, F. Huber, S. Rogers, and J. J. J. van der Hooft. Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. *Natural Product Reports*, 38(11):1967–1993, 2021. ISSN 1460-4752. doi: 10.1039/d1np00023c. URL http://dx.doi.org/10.1039/D1NP00023C.

J. Chong, P. Liu, G. Zhou, and J. Xia. Using microbiomeanalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature Protocols*, 15(3):799–821, Jan. 2020. ISSN 1750-2799. doi: 10.1038/s41596-019-0264-1. URL http://dx.doi.org/10.1038/s41596-019-0264-1.

N. F. de Jonge, K. Mildau, D. Meijer, J. J. R. Louwen, C. Bueschl, F. Huber, and J. J. J. van der Hooft. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics*, 18(12), Dec. 2022. ISSN 1573-3890. doi: 10.1007/s11306-022-01963-y. URL http://dx.doi.org/10.1007/s11306-022-01963-y.

M. Dork, S. Carpendale, and C. Williamson. Visualizing explicit and implicit relations of complex information spaces. *Information Visualization*, 11(1):5–21, Nov. 2011. ISSN 1473-8724. doi: 10.1177/1473871611425872. URL http://dx.doi.org/10.1177/1473871611425872.

J. Goeman and J. Oosting. Globaltest: testing association of a group of genes with a clinical variable, 2023. URL https://bioconductor.org/packages/globaltest.

J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 02 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm051. URL https://doi.org/10.1093/bioinformatics/btm051.

J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 01 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg382. URL https://doi.org/10.1093/bioinformatics/btg382.

P. T. Inc. Collaborative data science, 2015. URL https://plot.ly.

S. Khatib, I. Pereman, E. Kostanda, M. M. Zdouc, N. Ezov, R. Schweitzer, and J. J. J. van der Hooft. Olive mill solid waste induces beneficial mushroom-specialized metabolite diversity: a computational metabolomics study. *bioRxiv*, 2024. doi: 10.1101/2024.02.09.579616. URL https://www.biorxiv.org/content/early/2024/02/09/2024.02.09.579616.

L. Maaten. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579, 2008.

F. Maleki, K. Ovens, I. McQuillan, and A. J. Kusalik. Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Human Genomics*, 13(1):42, Oct 2019. ISSN 1479-7364. doi: 10.1186/s40246-019-0226-2. URL https://doi.org/10.1186/s40246-019-0226-2.

F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik. Gene set analysis: Challenges, opportunities, and future research. *Frontiers in Genetics*, 11, June 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00654. URL http://dx.doi.org/10.3389/fgene.2020.00654.

K. McLuskey, J. Wandy, I. Vincent, J. J. J. van der Hooft, S. Rogers, K. Burgess, and R. Daly. Ranking metabolite sets by their activity levels. *Metabolites*, 11(2):103, Feb. 2021. ISSN 2218-1989. doi: 10.3390/metabo11020103. URL http://dx.doi.org/10.3390/metabo11020103.

K. Mildau, H. Ehlers, I. Oesterle, M. Pristner, B. Warth, M. Doppler, C. Bueschl, J. Zanghellini, and J. J. J. van der Hooft. Tailored mass spectral data exploration using the specxplore interactive dashboard. *Analytical Chemistry*, Apr 2024. ISSN 0003-2700. doi: 10.1021/acs.analchem.3c04444. URL https://doi.org/10.1021/acs.analchem.3c04444.

L.-F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, A. M. Caraballo-Rodriguez, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, C. Martin H., L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers,

B. Shrestha, A. Tripathi, J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang, and P. C. Dorrestein. Feature-based molecular networking in the gnps analysis environment. *Nature Methods*, 17(9):905–908, Aug. 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0933-6. URL http://dx.doi.org/10.1038/s41592-020-0933-6.

F. Olivon, N. Elie, G. Grelier, F. Roussi, M. Litaudon, and D. Touboul. Metgem software for the generation of molecular networks based on the t-sne algorithm. *Analytical Chemistry*, 90(23): 13900–13908, Oct. 2018. ISSN 1520-6882. doi: 10.1021/acs.analchem.8b03099. URL http://dx.doi.org/10.1021/acs.analchem.8b03099.

A. K. Pakkir Shah, A. Walter, F. Ottosson, F. Russo, M. Navarro-Díaz, J. Boldt, J.-C. Kalinski, E. E. Kontou, J. Elofson, A. Polyzois, C. González-Marín, S. Farrell, M. R. Aggerbeck, T. Pruksatrakul, N. Chan, Y. Wang, M. Pöchhacker, C. Brungs, B. Cámara, A. M. Caraballo-Rodríguez, A. Cumsille, F. de Oliveira, K. Dührkop, Y. El Abiead, C. Geibel, L. G. Graves, M. Hansen, S. Heuckeroth, S. Knoblauch, A. Kostenko, M. C. Kuijpers, K. Mildau, S. Papadopoulos Lambidis, P. W. Portal Gomes, T. Schramm, K. Steuer-Lodd, P. Stincone, S. Tayyab, G. A. Vitale, B. C. Wagner, S. Xing, M. T. Yazzie, S. Zuffa, M. de Kruijff, C. Beemelmanns, H. Link, C. Mayer, J. J. van der Hooft, T. Damiani, T. Pluskal, P. C. Dorrestein, J. Stanstrup, R. Schmid, M. Wang, A. T. Aron, M. Ernst, and D. Petras. The hitchhiker's guide to statistical analysis of feature-based molecular networks from non-targeted metabolomics data, Nov. 2023. URL http://dx.doi.org/10.26434/chemrxiv-2023-wwbt0.

A. Rosato, L. Tenori, M. Cascante, P. R. De Atauri Carulla, V. A. P. Martins dos Santos, and E. Saccenti. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, 14(4), Feb. 2018. ISSN 1573-3890. doi: 10.1007/s11306-018-1335-y. URL http://dx.doi.org/10.1007/s11306-018-1335-y.

E. Schubert and P. J. Rousseeuw. Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804, Nov. 2021. ISSN 0306-4379. doi: 10.1016/j.is.2021.101804. URL http://dx.doi.org/10.1016/j.is.2021.101804.

J. Tomfohr, J. Lu, and T. B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1), Sept. 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-225. URL http://dx.doi.org/10.1186/1471-2105-6-225.

visjs community. A dynamic, browser based visualization library., 2024. URL https://visjs.org/.

J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26), May 2012. ISSN 1091-6490. doi: 10.1073/pnas.1203689109. URL http://dx.doi.org/10.1073/pnas.1203689109.

J.-L. Wolfender, J.-M. Nuzillard, J. J. J. van der Hooft, J.-H. Renault, and S. Bertrand. Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography–high-resolution tandem mass spectrometry and nmr profiling, in silico databases, and chemometrics. *Analytical Chemistry*, 91(1):704–742, Nov. 2018. ISSN 1520-6882. doi: 10.1021/acs.analchem.8b05112. URL http://dx.doi.org/10.1021/acs.analchem.8b05112.

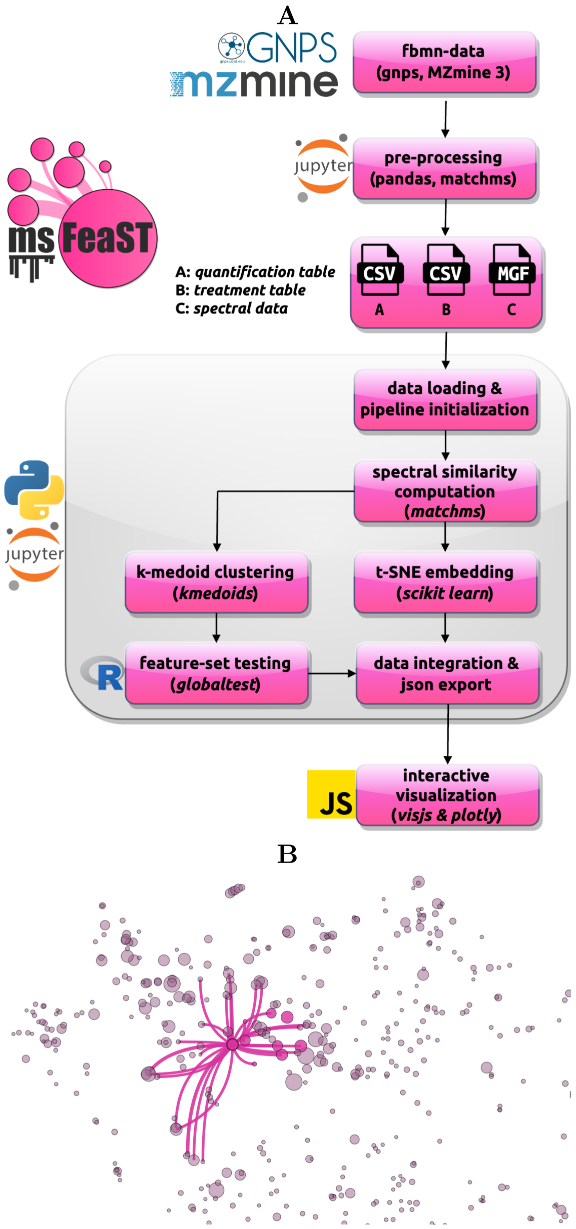
A
GNPS
mzmine
fbmn-data (gnps, MZmine 3)
jupyter
pre-processing (pandas, matchms)
ms FeaST
A: quantification table
B: treatment table
C: spectral data
CSV A
CSV B
MGF C
data loading & pipeline initialization
spectral similarity computation (matchms)
k-medoid clustering (kmedoids)
t-SNE embedding (scikit learn)
R
feature-set testing (globaltest)
data integration & json export
JS
interactive visualization (visjs & plotly)
B

Figure 1: msFeaST workflow overview and dashboard network visualization example. (i) Feature-based molecular networking data is translated into the expected msFeaST input data format. Once loaded, a msFeaST pipeline instance is created. Using msFeaST pipeline methods, spectral similarities are computed, clustering is performed (python package kmedoids), t-SNE embedding is performed (using scikit-learn), feature-set testing is performed (via an embedded R script, globaltest), and data is integrated into a json format compatible with the interactive javascript based visualization (visjs, plotly). The user only needs to run a sequence of commands while intermediate data structures are handled by the pipeline object. (ii) msFeaST network visualization example run on the illustrative example mushroom data using ms2deepscore as the scoring approach. Example of a feature within a feature set that shows differential abundance across fruiting bodies of *Pleurotus eryngii* cultivated using 0% and 80% olive mill solid waste mixed into their substrate. The selected node is shown with its top 30 neighbors, connecting to other clusters within the local t-SNE area.