Reproducible and scalable annotation of fragmentation spectra using the matchms Galaxy tools ecosystem

Helge Hecht[1], Maksym Skoryk[1,2], Matej Troják[1], Martin Čech[1], Zargham Ahmad[1], Jana Klánová[1] & Elliott James Price[1]

helge.hecht@recetox.muni.cz

1 RECETOX, Faculty of Science, Masaryk University, Kotlářská 2, Brno, Czech Republic
2 Institute of Computer Science, Masaryk University, Šumavská 15, Brno, Czech Republic

Introduction

Annotation of spectra with a chemical identity is considered as one of the major bottlenecks for exploratory mass spectrometry analysis of small molecules. Compound annotations can be assigned based on scores from matching experimentally acquired spectra to references in a spectral library. The matchms (1) python library has fostered the development of a larger ecosystem of connected tools focusing on matching of fragmentation spectra for compound identification. This includes machine learning based scores such as Spec2Vec (2) and MS2DeepScore (3) as well as other means of compound identification, such as molecular and spectral networking or analog search via MS2Query (4). Machine learning based scores have been shown to outperform the traditional cosine similarity and are already implemented in other cloud platforms such as GNPS (2, 5). Molecular networking based on spectral- and metadata similarity allows propagating annotations to un-identified compounds or putatively assigning an unknown spectrum to a compound class. We are taking the functionalities of matchms - spectral matching, molecular networking, and spectral library processing - to the cloud with the matchms Galaxy tool suite for reproducible and scalable compound identification.

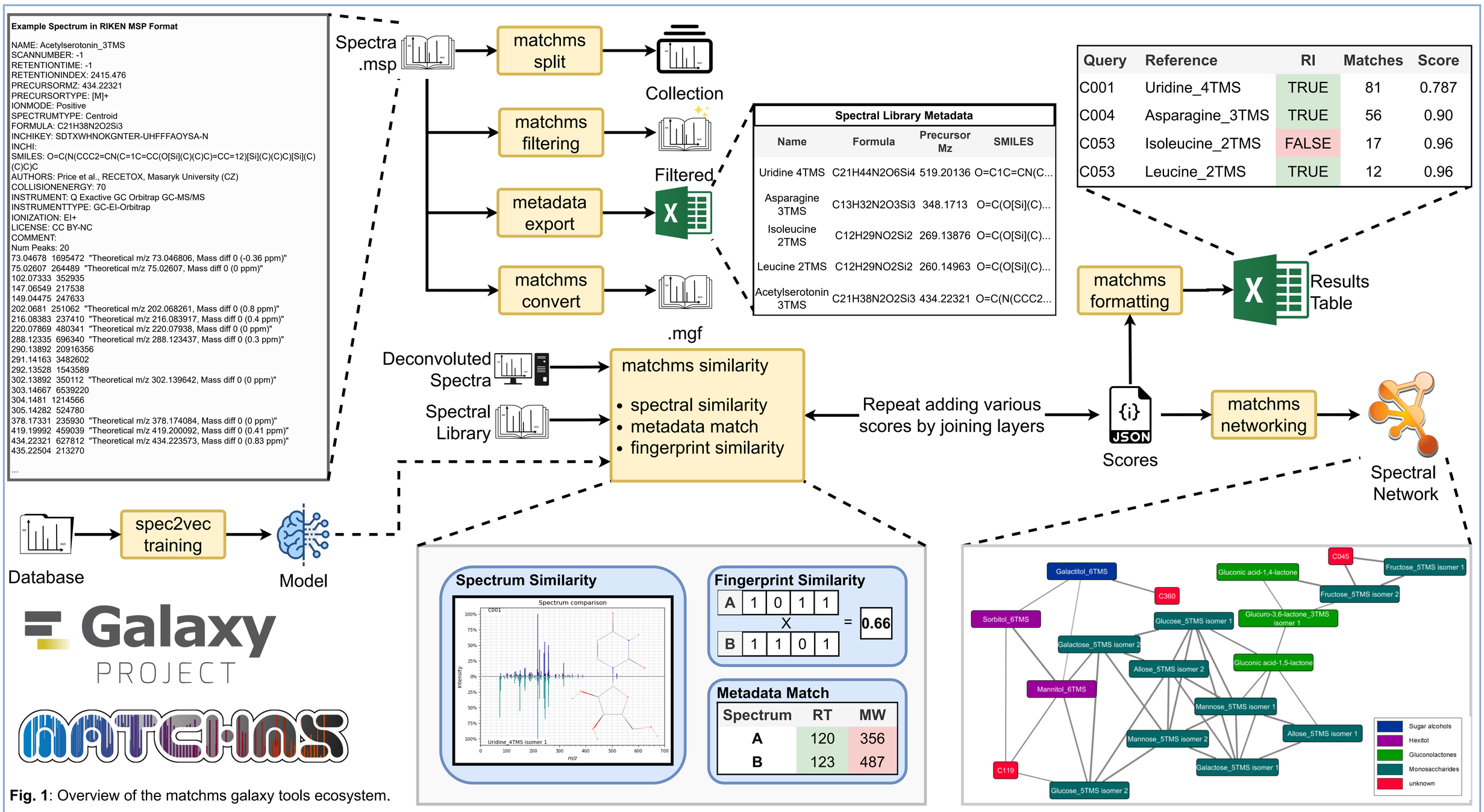

Fig. 1: Overview of the matchms galaxy tools ecosystem.

Users can train their own Spec2Vec model in Galaxy for subsequent use with the matchms similarity module. Besides similarity matching, metadata information, e.g. retention indices or precursor masses can be matched to serve as a filter, leveraging those efficient metadata scores as a mask to avoid time consuming spectral similarity computations between non-matching pairs. This leverages the matchms pipeline mechanism which allows layering and combining subsequent scores computed on the same query and reference datasets. Structural similarity between identified compounds can be computed using the molecular fingerprint similarity module. We implemented a matchms networking wrapper and connected the outputs with the Cytoscape Galaxy plugin for seamless visualization. The library handling tools enable conversion between different formats, exporting spectrum metadata, splitting a large library into smaller subsets and applying filters to a library (e.g., to remove low intensity peaks or compute molecular fingerprints).

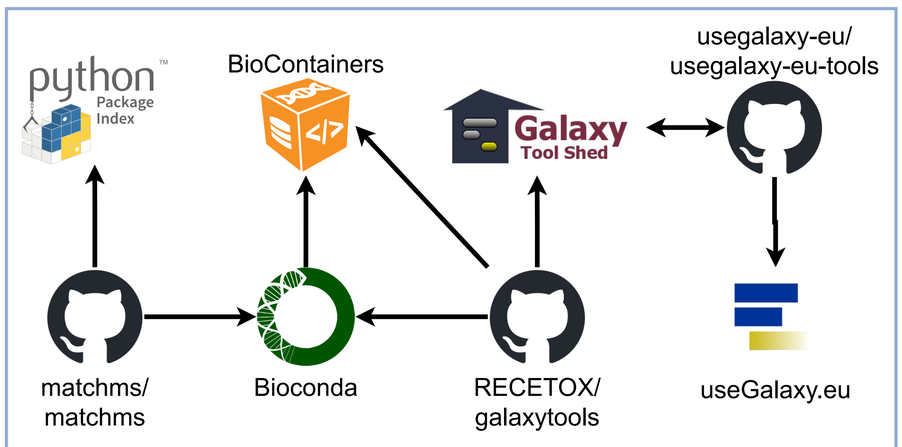

Fig. 2: Technology stack for tool development & deployment.

The matchms package is pushed automatically to PyPI and Bioconda when a new release is drafted on GitHub. The galaxy tool wrapper uses the bioconda package or biocontainer to provide access to the matchms API. The galaxytool is published on the toolshed via a GitHub action. The usegalaxy-eu-tools respository is then leveraged to automatically install the matchms tools on usegalaxy.eu.

Results

- Splitting large reference libraries enables parallel matching of query spectra against smaller subsets, leveraging Galaxy's built-in job scheduling to distribute the computational load across nodes.
- Package versions are consistent across tools through using the same biocontainer, improving interoperability, reproducibility and maintainability.
- Interoperability with other resources is ensured through use of standard file formats (e.g., msp and mgf) and file-based operations.

Discussion

- With its general spectrum handling capacities, matchms serves as a base package for many more demanding or specialized tasks. The ecosystem of tools built around it includes further tools such as RIAssigner (6) and MSMetaEnhancer (7) which are used to improve the compound identification process and curation of spectral libraries.
- Future developments include the addition of further spectral similarity scores, such as MS2DeepScore and spectral entropy (8), improved molecular networking capabilites and more versatile scores handling by implementing operations like *join* and *append*.

1. F. Huber et al., “matchms - processing and similarity evaluation of mass spectrometry data.,” J. Open Source Softw., vol. 5, no. 52, p. 2411, Aug. 2020, doi: 10.21105/joss.02411.
2. F. Huber et al., “Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships,” PLOS Comput. Biol., vol. 17, no. 2, p. e1008724, Feb. 2021, doi: 10.1371/journal.pcbi.1008724.
3. F. Huber, S. van der Burg, J. J. J. van der Hooft, and L. Ridder, “MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra,” J. Cheminform., vol. 13, no. 1, p. 84, Dec. 2021, doi: 10.1186/s13321-021-00558-4.
4. N. F. de Jonge et al., “MS2Query: reliable and scalable MS2 mass spectra-based analogue search,” Nat. Commun., vol. 14, no. 1, p. 1752, Mar. 2023, doi: 10.1038/s41467-023-37446-4.
5. M. Wang et al., “Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking,” Nat. Biotechnol., vol. 34, no. 8, pp. 828–837, Aug. 2016, doi: 10.1038/nbt.3597.
6. M. Troják, H. Hecht, M. Čech, and E. J. Price, “MSMetaEnhancer: A Python package for mass spectra metadata annotation,” J. Open Source Softw., vol. 7, no. 79, p. 4494, Nov. 2022, doi: 10.21105/joss.04494.
7. H. Hecht, M. Skoryk, M. Čech, and E. J. Price, “RIAssigner: A package for gas chromatographic retention index calculation,” J. Open Source Softw., vol. 7, no. 75, p. 4337, Jul. 2022, doi: 10.21105/joss.04337.
8. Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta, and O. Fiehn, “Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification,” Nat. Methods, Dec. 2021, doi: 10.1038/s41592-021-01331-z.

SCAN ME

The authors thank to Research Infrastructure RECETOX RI (No LM2018121) financed by the Ministry of Education, Youth and Sports, H2020 CETOCOEN Excellence 857560 and OP RDE project (No CZ.02.1.01/0.0/0.0/17_043/0009632) for supportive background. EJP was supported from OP RDE - Project ”MSCAfellow4@MUNI” (No. CZ.02.2.69/0.0/0.0/20_079/0017045). Supported by FR CESNET (679R1/2021). Computational resources provided by the e-INFRA CZ project (90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This publication reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains.