IECM 2025 Conference
The 4th International Electronic Conference on Metabolomics
13–15 October 2025 | Online
MDPI

Mass Spectral Matching for Compound Identification in Metabolomics: An Open-Source Python/Shiny Package

Hunter Dlugas[1,2], Ikuko Kato[1,3,5], Jing Li[1,4], Xiang Zhang[6], Seongho Kim[1,2]

[1]Department of Oncology, School of Medicine, Wayne State University
[2]Biostatistics and Bioinformatics Core, Karmanos Cancer Institute, Wayne State University
[3]Department of Pathology, School of Medicine, Wayne State University
[4]Pharmacology and Metabolomics Core, Karmanos Cancer Institute, Wayne State University
[5]Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center
[6]Department of Chemistry, University of Louisville

INTRODUCTION

Accurate compound identification is essential for mass spectrometry-based metabolomics. The most common strategies are chemical structure library searching and spectral library searching, both of which depend on measuring the similarity between an unknown spectrum and reference spectra, either acquired experimentally or generated *in silico*. The quality of the similarity metric plays a central role in reliable identifications. To address this need, we developed **PyCompound**, a Python/Shiny package that provides an interactive graphical interface for mass spectral matching. **PyCompound** accommodates both nominal-resolution data (e.g., GC-MS) and high-resolution data (e.g., LC-MS/MS). It offers flexible preprocessing pipelines, including filtering, weight factor and low entropy transformations, centroiding, noise reduction, and customizable spectral matching. In addition to conventional cosine and binary similarity measures, **PyCompound** implements entropy-based metrics such as Shannon, Tsallis, and Rényi correlations, which can be combined into custom mixture measures with user defined weights. The package supports both untargeted and targeted workflows, includes a command line mode for batch analyses, and allows fine tuning of parameters when reference libraries with compound annotations are used.

METHODS

PyCompound offers a plethora of spectrum preprocessing transformations and similarity measures.

- Spectrum preprocessing transformations:
  - Filtering based on m/z and/or intensity
  - Weight factor transformation
  - Low-entropy transformation
  - Centroiding
  - Matching ion fragments of two spectra
- Similarity measures:
  - Cosine (aka dot product)
  - Three entropy-based similarity measures: Shannon [1], Rényi, and Tsallis [2]
  - 14 binary similarity measures: Jaccard, Dice, 3W-Jaccard, Sokal-Sneath, Cosine, Mountford, McConnaughey, Driver-Kroeber, Simpson, Braun-Banquet, Fager-McGowan, Kulczynski, Intersection, Hamming, and Hellinger [3]

PyCompound has three main utilities:

- Perform spectral library matching to identify unknown compounds
- Given a library of known compounds, perform (multithreaded) grid search to determine the hyperparameters which maximize identification accuracy.
- Generate a plot of two spectra before and after preprocessing transformations

Specifically, **PyCompound**'s novelty lies in (i) the novel Rényi Entropy Similarity Measure and (ii) its ability to allow a user to easily specify the order of spectrum preprocessing transformations without necessarily having to write code.

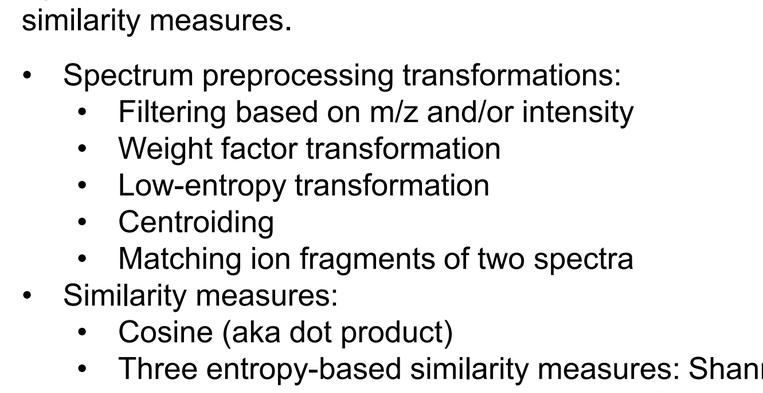

Figure 1. Examples of various spectrum preprocessing workflows capable with PyCompound. The spectrum preprocessing transformations available in the blue boxes can be performed in any user-specified order. NRMS: nominal-resolution mass spectrometry. HRMS: high-resolution mass spectrometry.

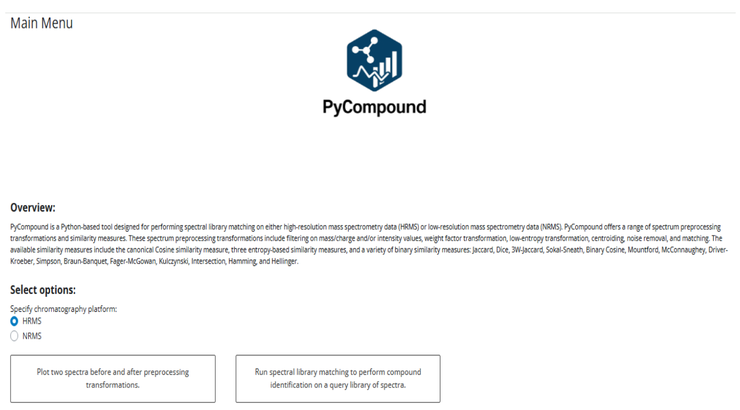

Figure 2. Shiny PyCompound interface.

RESULTS

PyCompound was demonstrated using two well established spectral libraries: the NIST GC-MS Webbook and a GNPS derived LC-MS/MS collection. These case studies show that the tool can be applied across both nominal and high-resolution data, and that it successfully integrates entropy-based similarity measures alongside traditional approaches.

Table 1. Example compound identification output from GC-MS data generated by PyCompound. The top three matches for each query compound are shown.

Query Spectrum ID	RANK.1.PRED	RANK.2.PRED	RANK.3.PRED	RANK.1.SIMILARITY	RANK.2.SIMILARITY	RANK.3.SIMILARITY
ID_1	616386	922689	1912283	0.752204811	0.742321126	0.680327002
ID_2	503286	877894	107255	0.961996437	0.806852602	0.791218167

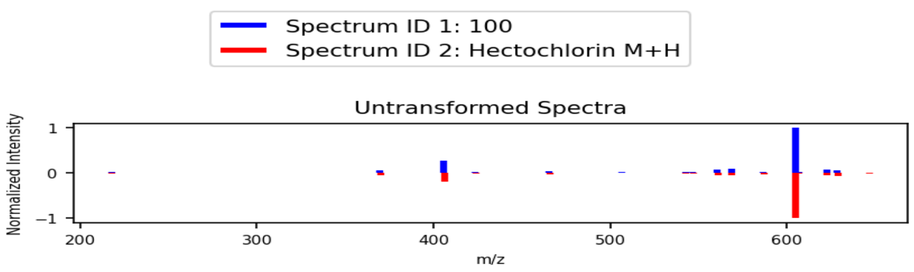

Figure 3. Example plot generated by PyCompound.

- New software tool **PyCompound** was developed to assist wet-lab researchers in performing compound identification on mass spectrometry data (https://github.com/hdlugas/pycompound).
  - Shiny web application (https://fy7392.shinyapps.io/pycompound/)
  - Python package 'pycompound' (https://pypi.org/project/pycompound/).
  - Command line version capable of finding optimal hyperparameters given known compound identities.
- **PyCompound** was applied to two publicly available real-world datasets (https://zenodo.org/records/12786324) [4]:
  - WebNIST GC-MS
  - GNPS LC-MS/MS

CONCLUSION

PyCompound provides a user-friendly framework for mass spectral matching. By combining established and novel similarity metrics with customizable preprocessing options, it offers a practical resource for enhancing compound identification in both research and applied metabolomics workflows.

REFERENCES

1. Li, Y., Kind, T., Folz, J. et al. (2021) Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods, 18 1524–1531. https://doi.org/10.1038/s41592-021-01331-z.

2. Dlugas H, Zhang X, Kim S. Comparative analysis of continuous similarity measures for compound identification in mass spectrometry-based metabolomics. Chemometr Intell Lab Syst. 2025 Aug 15;263:105417. doi: 10.1016/j.chemolab.2025.105417. Epub 2025 May 3. PMID: 40453508; PMCID: PMC12121958.

3. Kim S, Kato I, Zhang X. Comparative Analysis of Binary Similarity Measures for Compound Identification in Mass Spectrometry-Based Metabolomics. Metabolites. 2022 Jul 26;12(8):694. doi: 10.3390/metabo12080694. PMID: 35893261; PMCID: PMC9394311.

4. Dlugas, H., Zhang, X., & Kim, S. (2024). Liquid Chromatography - Tandem Mass Spectrometry (LC-MS/MS) and Gas Chromatography - Mass Spectrometry (GC-MS) Reference Libraries from Global Natural Products Social Molecular Networking (GNPS) and National Institute of Standards and Technology (NIST) WebBook Processed for Spectral Library Matching (V1.0) [Data set]. Zenodo.