








MSMetaEnhancer: A Python package for mass spectra metadata annotation

Matej Troják ¹, Helge Hecht ^{1¶}, Martin Čech ¹, and Elliott James Price ¹

¹ RECETOX, Faculty of Science, Masaryk University, Kotlářská 2, Brno, Czech Republic ¶
Corresponding author

DOI: [10.21105/joss.04494](https://doi.org/10.21105/joss.04494)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Charlotte Soneson](#)  

Reviewers:

- [@marshallmcdonnell](#)
- [@chryswoods](#)

Submitted: 30 May 2022

Published: 24 October 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

MSMetaEnhancer is a Python software package for the metadata enrichment of records in mass spectral library files commonly used as reference for chemical identification via mass spectrometry. Each record contains spectral information, i.e., peak mass to charge ratio (m/z) and intensities, alongside chemical & structural metadata, e.g., identifiers. The package uses `matchms` ([Huber et al., 2020](#)) for data IO and supports the open, text-based `.msp` format. It annotates given mass spectra records in the library file by adding missing metadata such as SMILES, InChI, and CAS numbers to the individual entries. The package retrieves the respective information by querying several external databases using existing metadata (e.g., SMILES or CAS number), converting different representations or database identifiers. Multiple databases and services are included, currently supporting the chemical identifier resolver (CIR), chemical translation service (CTS) ([Wohlgemuth et al., 2010](#)), ChemIDplus ([Tomasulo, 2002](#)), the Integrated Database for Small Molecules (IDSM) ([Galgonek & Vondrášek, 2021](#)), PubChem ([Kim et al., 2021](#)), and BridgeDb ([van Iersel et al., 2010](#)). Additionally, instead of querying external databases, computing the identifiers is also supported (e.g. molecular weight from SMILES).

Statement of need

Mass spectra stored in a library need to be enriched with metadata (e.g. chemical formula, SMILES code, InChI, the origin of the spectrum, etc.) to (1) combine spectral and structural information, (2) make the identification process more robust and reproducible, and (3) leverage the interoperability capabilities of chemical databases ([Wallace et al., 2017](#)). While this metadata is mostly accessible from public chemical databases, they are not always present in mass spectral library records. Therefore, the data needs to be post-processed via enhancement with metadata. Such a process usually cannot be fully automated, and assistance from the user is required to specify particular annotation steps and sources ([Ausloos et al., 1999](#)). Moreso, manual curation and addition of metadata while creating a compound library is labour intensive and error-prone ([Price et al., 2021](#); [Stravs et al., 2013](#)).

State of the field

There are several packages within the Python and R ecosystems which support querying external databases. For example, there are R packages that provide an interface to PubChem ([Cao et al., 2008](#); [Guha, 2016](#)), and a package with interface to wikidata ([Keys & Shafee, 2021](#)). Then, there are packages unifying several sources – `webchem` ([Szöcs et al., 2020](#)) allows to automatically query chemical data from several web sources (similar to MSMetaEnhancer) and

to interconvert between identifiers. The MetaFetcher (Yones et al., 2021) package focuses on database-specific identifiers and links metabolite data from several small-compound databases (e.g., PubChem, the Human Metabolome Database (HMDB) (Wishart et al., 2022)), trying to resolve inconsistencies. Similarly, RaMP cross-references multiple database specific identifiers via their internal RaMP_ID to integrate various pathway and compound databases (Zhang et al., 2018). BridgeDb is an ELIXIR project providing mapping functionality of different identifiers present in HMDB (e.g., PubChemCID, ChEBI and InChIKey), gene information and several pathway databases in an organism centric manner, exposing a Java and REST API (van Iersel et al., 2010; Willighagen et al., 2022).

On the Python side, there are packages providing direct API access for PubChem (Swain, 2017), ChemSpider (Swain, 2018), or CIR (Swain, 2016). PubChem's public API limits programmatic access to less than ~5 requests per second, limiting the ability of advanced users to effectively mine the database.

However, to the best of our knowledge, there is no Python package connecting these sources into a single tool, allowing straightforward metadata annotation of large mass spectral libraries with various identifiers and cross-references to different databases in a user-friendly way.

The software package

MSMetaEnhancer is a tool to enhance the metadata content of records in mass spectral library files. It takes as input a single .msp file with multiple mass spectra records and a list of annotation steps. These steps consist of specifying what service should be used to obtain a particular metadata attribute based on another already existing attribute. To improve the performance of the tool, we use services with high-throughput APIs when available (e.g. IDSM (Galgonek & Vondrášek, 2021), which can be used to access the PubChem database). The supported metadata attributes include InChI, InChIKey, SMILES, IUPAC chemical name, chemical formula, CAS number, and others. The particular available conversions can be found in the documentation via <https://msmetaenhancer.readthedocs.io/> and are open to extension. Finally, the obtained metadata are validated to ensure their correct form (currently, matchms validators are employed for this task). For an overview of the annotation process, see Figure 1.

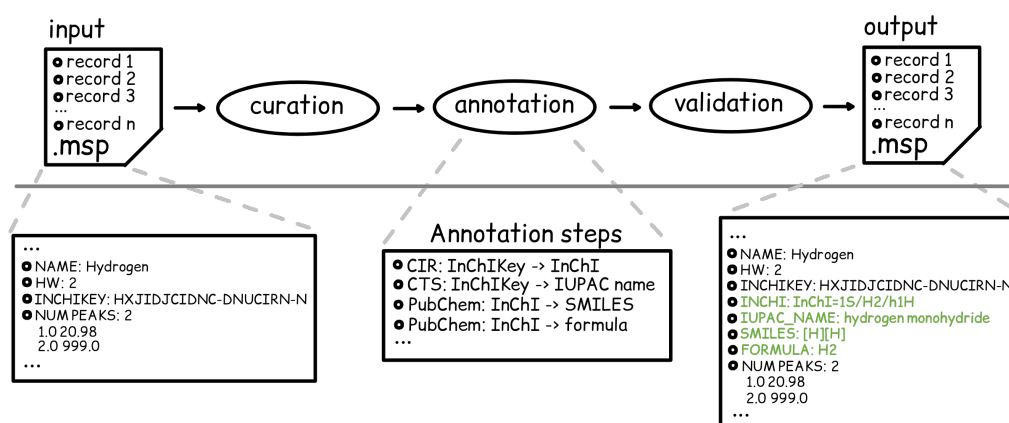


Figure 1: Schematic overview of MSMetaEnhancer annotation workflow.

The tool processes the spectral library by iteratively executing all steps for each entry until no new metadata is found. This happens for each spectral record in the provided file. Since it takes some non-trivial time for the services to respond to a query, this task is suitable for the asynchronous approach, making the tool computationally efficient. Additionally, results containing all metadata related to a compound are cached, making access to all available

metadata for a compound result in only a single query. For services with limited access rate (i.e., PubChem), we implemented a throttling mechanism – maximizing performance while mitigating restrictions from the requested webservice. Besides querying external services, we also support converters to compute identifiers based on existing ones. For demonstration, we employed computation of molecular weight from SMILES using RDKit ([Landrum et al., 2022](#)). Any issues regarding the annotation process (such as the absence of target data or unavailability of a service) are recorded in a detailed log file, which can be specified as an optional output of the tool.

To improve the usability of the tool, a Galaxy ([Afgan et al., 2018](#)) wrapper was created to provide a user-friendly interface and a simple way of reproducible data processing and analysis. The tool is hosted on the Galaxy instance available at <https://umsa.cerit-sc.cz/>, among others ([Spectrometric Data Processing and Analysis & Institute of Computer Science, 2022](#)). Moreover, the tool is available from bioconda ([Grüning et al., 2018](#)) as a standalone package.

Example workflow

Performing annotation of a .msp file is straightforward and requires to specify services to be used and a list of annotation steps.

```
import asyncio
from MSMetaEnhancer import Application

app = Application()
# import your .msp file
app.load_spectra('input_spectra_file.msp', file_format='msp')
# specify services
services = ['CIR', 'CTS', 'IDSM']

# specify annotation steps
jobs = [('inchikey', 'inchi', 'CIR'),
        ('inchikey', 'iupac_name', 'CTS'),
        ('inchi', 'canonical_smiles', 'IDSM'),
        ('inchi', 'formula', 'IDSM')]

# run asynchronous annotation of spectra data
asyncio.run(app.annotate_spectra(services, jobs))
# export .msp file
app.save_spectra('annotated_spectra_file.msp', file_format='msp')
```

Author's Contributions

MT wrote the manuscript and developed the software. HH contributed to the manuscript and via code reviews and implementation guidance. MČ contributed via code reviews and implementation guidance. EJP provided conceptual oversight and contributed to the manuscript.

Acknowledgements

Authors thank to Research Infrastructure RECETOX RI (No LM2018121) financed by the Ministry of Education, Youth and Sports, and OP RDE project CETOCOEN EXCELLENCE (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). This project was supported from the European Union's Horizon 2020 research and innovation

programme under grant agreement No 857560. 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.

References

- Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 update. *Nucleic Acids Research*, 46(W1), W537–W544. <https://doi.org/10.1093/nar/gky379>
- Ausloos, P., Clifton, C. L., Lias, S. G., Mikaya, A. I., Stein, S. E., Tchekhovskoi, D. V., Sparkman, O. D., Zaikin, V., & Zhu, D. (1999). The critical evaluation of a comprehensive mass spectral library. *Journal of the American Society for Mass Spectrometry*, 10(4), 287–299. [https://doi.org/10.1016/S1044-0305\(98\)00159-7](https://doi.org/10.1016/S1044-0305(98)00159-7)
- Cao, Y., Charisi, A., Cheng, L.-C., Jiang, T., & Girke, T. (2008). ChemmineR: A compound mining framework for R. *Bioinformatics*, 24(15), 1733–1734. <https://doi.org/10.1093/bioinformatics/btn307>
- Galgonek, J., & Vondrášek, J. (2021). IDSM ChemWebRDF: SPARQLing small-molecule datasets. *Journal of Cheminformatics*, 13(1), 1–19. <https://doi.org/10.1186/s13321-021-00515-1>
- Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7), 475–476. <https://doi.org/10.1038/s41592-018-0046-7>
- Guha, R. (2016). Rpubchem: Interface to the PubChem collection. *R Package Version 1.5.10*. <https://CRAN.R-project.org/package=rpubchem>
- Huber, F., Verhoeven, S., Meijer, C., Spreeuw, H., Castilla, E., Geng, C., van der Hooft, J., Rogers, S., Belloum, A., Diblen, F., & Spaaks, J. (2020). matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5(52), 2411. <https://doi.org/10.21105/joss.02411>
- Keys, O., & Shafee, T. (2021). WikidataR: API client library for wikidata. *R Package Version 2.0.0*. <https://CRAN.R-project.org/package=rpubchem>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., & others. (2021). PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, 49(D1), D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>
- Landrum, G., Tosco, P., Kelley, B., Ric, sriniker, gedec, Vianello, R., NadineSchneider, Kawashima, E., Cosgrove, D., Dalke, A., N, D., Jones, G., Cole, B., Swain, M., Turk, S., AlexanderSavelyev, Vaucher, A., Wójcikowski, M., ... DoliathGavid. (2022). *Rdkit/rdkit: 2022_03_5 (Q1 2022) release*. <https://doi.org/10.5281/zenodo.6961488>
- Price, E. J., Palát, J., Coufalíková, K., Kukučka, P., Codling, G., Vitale, C. M., Koudelka, Š., & Klánová, J. (2021). Open, High-Resolution EI+ Spectral Library of Anthropogenic Compounds. *Frontiers in Public Health*, 9(March). <https://doi.org/10.3389/fpubh.2021.622558>
- Spectrometric Data Processing and Analysis, & Institute of Computer Science. (2022). *RECETOX/galaxytools: Release v0.2.0 (Version v0.2.0) [Computer software]*. Zenodo.

- <https://doi.org/10.5281/zenodo.6035335>
- Stravs, M. A., Schymanski, E. L., Singer, H. P., & Hollender, J. (2013). Automatic recalibration and processing of tandem mass spectra using formula annotation. *Journal of Mass Spectrometry*, 48(1), 89–99. <https://doi.org/10.1002/jms.3131>
- Swain, M. (2016). CIRpy. *Python Package Version 1.0.2*. <https://github.com/mcs07/CIRpy>
- Swain, M. (2017). PubChemPy. *Python Package Version 1.0.4*. <https://github.com/mcs07/PubChemPy>
- Swain, M. (2018). ChemSpiPy. *Python Package Version 2.0.0*. <https://github.com/mcs07/ChemSpiPy>
- Szöcs, E., Stirling, T., Scott, E. R., Scharmüller, A., & Schäfer, R. B. (2020). Webchem: An R package to retrieve chemical information from the web. *Journal of Statistical Software*, 93(1), 1–17. <https://doi.org/10.18637/jss.v093.i13>
- Tomasulo, P. (2002). ChemIDplus—super source for chemical and drug information. *Medical Reference Services Quarterly*, 21(1), 53–59. https://doi.org/10.1300/j115v21n01_04
- van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., & Evelo, C. T. (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1), 5. <https://doi.org/10.1186/1471-2105-11-5>
- Wallace, W. E., Ji, W., Tchekhovskoi, D. V., Phinney, K. W., & Stein, S. E. (2017). Mass Spectral Library Quality Assurance by Inter-Library Comparison. *Journal of the American Society for Mass Spectrometry*, 28(4), 733–738. <https://doi.org/10.1007/s13361-016-1589-4>
- Willighagen, E., Kutmon, M., Martens, M., & Slenter, D. (2022). BridgeDb and Wikidata: a powerful combination generating interoperable open research (BridgeDb). *Research Ideas and Outcomes*, 8. <https://doi.org/10.3897/rio.8.e83031>
- Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., ... Gautam, V. (2022). HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1), D622–D631. <https://doi.org/10.1093/nar/gkab1062>
- Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T., & Fiehn, O. (2010). The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, 26(20), 2647–2648. <https://doi.org/10.1093/bioinformatics/btq476>
- Yones, S., Csombordi, R., Komorowski, J., & Diamanti, K. (2021). MetaFetchR: An R package for complete mapping of small compound data. *bioRxiv*. <https://doi.org/10.1101/2021.02.28.433248>
- Zhang, B., Hu, S., Baskin, E., Patt, A., Siddiqui, J., & Mathé, E. (2018). RaMP: A Comprehensive Relational Database of Metabolomics Pathways for Pathway Enrichment Analysis of Genes and Metabolites. *Metabolites*, 8(1), 16. <https://doi.org/10.3390/metabo8010016>