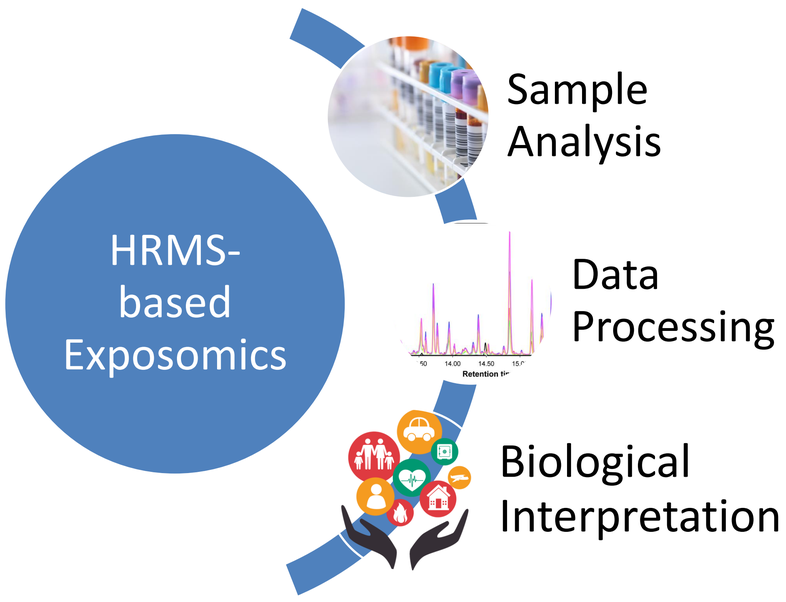
MUNI | RECETOX

Mass Spectrometry of Chemical Exposure

Helge Hecht[1], Jiří Novotný[2,5], Karolína Trachtová[1,2,3], Martin Čech[1,4], Maksym Skoryk[2,6], Ondřej Melichar[1,2], Aleš Křenek[2], Jana Klánová [1], Elliott James Price [1]

1 RECETOX Centre, Masaryk University, Brno, Czech Republic; 2 Institute of Computer Science, Masaryk University, Brno, Czech Republic; 3 Central European Institute of Technology, Masaryk University, Brno, Czech Republic; 4 Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czech Republic; 5 Faculty of Informatics, Masaryk University, Brno, Czech Republic; 6 Faculty of Chemistry, Brno University of Technology, Brno, Czech Republic



High-resolution mass spectrometry (HRMS) is increasingly being applied for the detection of chemicals in human biospecimen, including the application of non-target profiling methods to assess human population exposure to toxic chemicals and effects on health. These HRMS-based chemical profiling methods generate large data with the accompanying need for data processing pipelines.

The Galaxy platform was developed to provide an open source, community-driven, web-based platform for accessible, reproducible, and transparent computational research and training [1]. Originally developed for genomics data, Galaxy is now widely used in all areas of life sciences and provides an ideal platform to implement software and workflows for the processing of HRMS chemical exposure data (Figure 1).

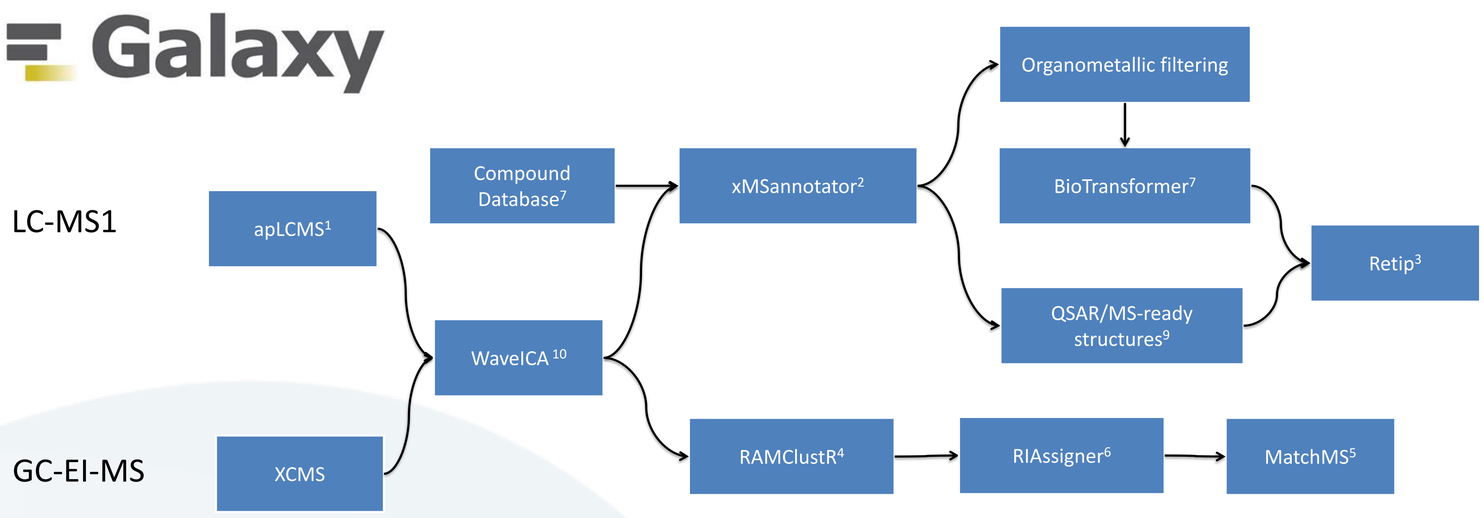
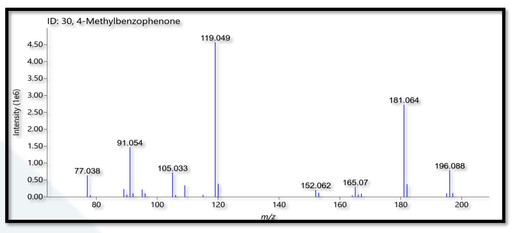


Figure 1. Overview of Galaxy MS data processing workflow development. 1. Hybrid peak detection for profile MS data; 2. LC-MS1 pathway annotation; 3. LC-MS1 retention filtering; 4. Ion-wise batch-wide deconvolution; 5. Spectral matching; 6. Retention index computation; 7. Custom interoperable database; 8. Transformation predictions; 9. QSAR-Ready & MS-Ready SMILES; 10. Batch correction

Application of machine learning techniques for HRMS data processing

Machine learning (ML) has advanced nearly all fields of science and is increasingly applied for MS data processing. The variable nature of the continuous data proves problematic when encoding high-resolution data, hindering HRMS focused developments. Finding fixed length encodings of high-resolution data without significant information loss is an open research objective.



Paired Mass Difference-based Binary Encoding

- The mass differences between individual peaks correspond to neutral losses in multi-step fragmentations.
- Paired mass differences between peaks of a single spectrum can be used to characterize and compare spectra [2] (Figure 2).
- The spectrum can be encoded according to presence of frequently observed losses as a binary vector and used for training of autoencoder networks (Figure 3).
- This method resembles the utilization of neutral losses (differences between peak and precursor m/z) but is also applicable for GC-EI-MS data.
- Neutral losses have been determined as an important source of information for a certain compound and incorporated into public databases [8].

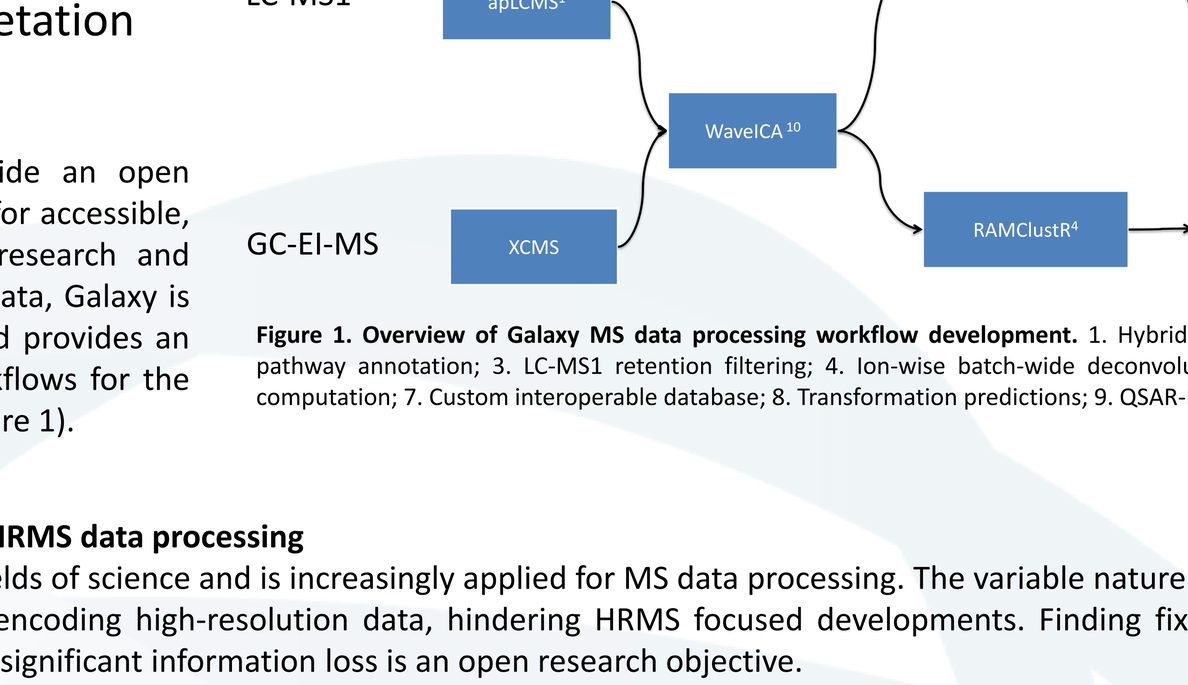


Figure 2. Example mass spectrum and paired mass difference encoding. The connections between m/z differences and the molecule substructures are shown. The m/z differences corresponding to losses of specific groups during fragmentation are depicted in red.

Figure 3. Representing mass spectra for neural network. Typically, neural netowrks handle only fixed size inputs, presenting need for uniform representation of mass spectra.

Application: ML-based Spectral Similarity Metrics

- Unidentified spectra can be matched into reference libraries of known compounds using spectral similarity metrics (Figure 4).
- ML based metrics capture data intrinsic properties and outperform traditional metrics like cosine similarity [3,4].
- Comparison of similarity based molecular networks and clustering in learned embedding.

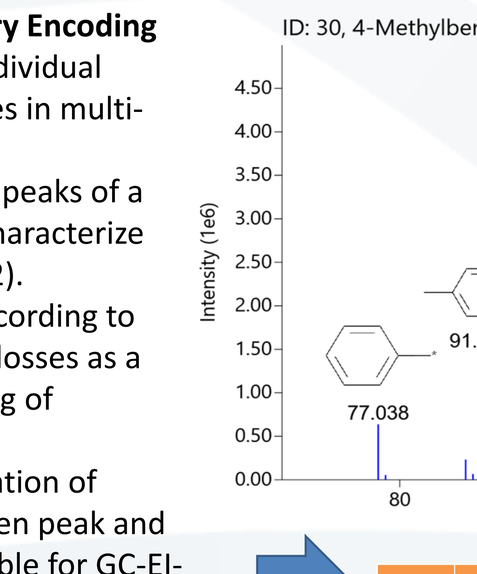


Figure 4. Comparison of mass spectra. Spectrum A plotted against spectrum B for visual comparison.

Reference Set Similarity-based Encoding

- A fixed set of reference spectra can be used to characterize spectra with varying peak-counts using a set number of features [7].
- Arrangement of these features in a spatial matrix (Figure 5) facilitates application of techniques originating from image processing, such as convolutional neural networks (CNNs).

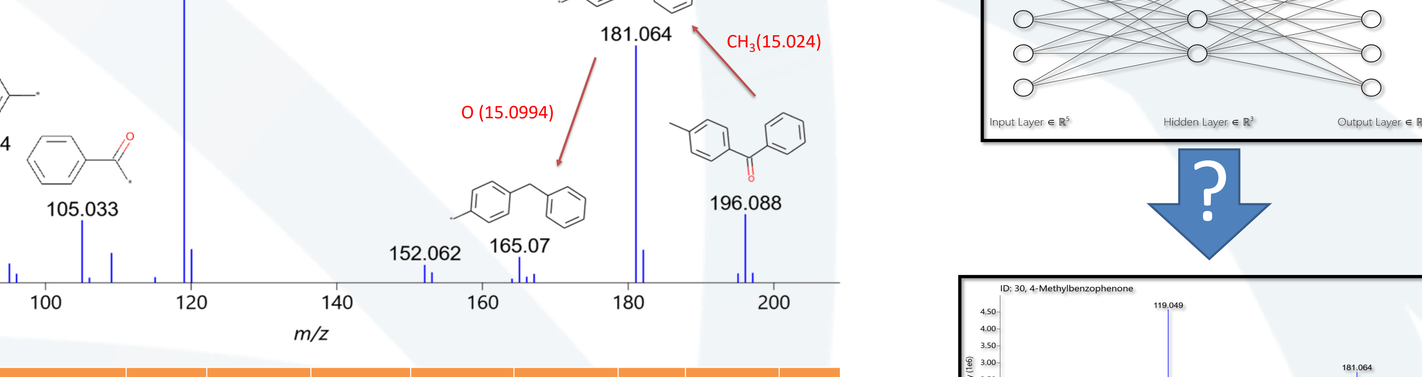


Figure 5. An example spectrum (left) and a plot of a similarity matrix (right). Similarity score of query spectra are as rows for a fixed set of reference spectra on the columns. Bright color indicates high spectral similarity, a dark color a low value. The values of a row can be used as fixed size representation of the respective query spectrum.

[1] Afgan, Enis, et al. “Galaxy: A Gateway to Tools in e-Science.” Guide to E-Science, edited by Xiaoyu Yang et al., Springer, 2011, pp. 145–77, doi:10.1007/978-0-85729-439-5_6.
[2] Yu, Miao, et al. “Structure/Reaction Directed Analysis for LC-MS Based Untargeted Analysis.” *Analytica Chimica Acta*, vol. 1050, Elsevier Ltd, Mar. 2019, pp. 16–24, doi:10.1016/j.aca.2018.10.062.
[3] Huber, Florian, et al. “Spec2Vec: Improved Mass Spectral Similarity Scoring through Learning of Structural Relationships.” BioRxiv, 2020, p. 2020.08.11.245928, doi:10.1101/2020.08.11.245928.
[4] Huber, Florian, et al. MS2DeepScore - a Novel Deep Learning Similarity Measure for Mass Fragmentation Spectrum Comparisons. 2021, doi:10.1101/2021.04.18.440324.
[5] Treen, C., et al. “SIMILE Enables Alignment of Fragmentation Mass Spectra with Statistical Significance.” BioRxiv, Cold Spring Harbor Laboratory, Feb. 2021, p. 2021.02.24.432767, doi:10.1101/2021.02.24.432767.
[6] Li, Mike, and X. Rosalind Wang. “Peak Alignment of Gas Chromatography–Mass Spectrometry Data with Deep Learning.” Journal of Chromatography A, vol. 1604, Elsevier B.V., Oct. 2019, p. 460476, doi:10.1016/j.chroma.2019.460476.
[7] May, Damon H., et al. “A Learned Embedding for Efficient Joint Analysis of Millions of Mass Spectra.” BioRxiv, 2018, pp. 1–17, doi:10.1101/483263.
[8] Aisporna, Aries, et al. METLIN Neutral Loss Database Enhances Similarity Analysis. 2021, pp. 8–12, doi:10.1101/2021.04.02.438066.