



Universiteit
Leiden
The Netherlands

The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation

Giera, M.; Aisporna, A.; Uritboonthai, W.; Siuzdak, G.

Citation

Giera, M., Aisporna, A., Uritboonthai, W., & Siuzdak, G. (2024). The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation. *Nature Metabolism*. doi:10.1038/s42255-024-01076-x

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](#)

Downloaded from: <https://hdl.handle.net/1887/4089830>

Note: To cite this publication please use the final published version (if applicable).

The hidden impact of in-source fragmentation in metabolic and chemical mass spectrometry data interpretation



Mass spectrometry, specifically liquid chromatography–tandem mass spectrometry (LC–MS/MS), stands at the forefront of metabolomic, pharmaceutical and clinical analyses, offering high sensitivity and specificity in detecting metabolites and other small molecules within biological samples. Yet, this powerful technique has unveiled a puzzling abundance of unidentified spectral features, coined the ‘dark metabolome’^{1,2}, which starkly contrasts with the known metabolic diversity. While humans are known to host approximately 20,000 protein encoding genes^{3,4}, with only a subset of these expressing enzymes, the presumed chemical diversity indicated by LC–MS/MS suggests hundreds of thousands to millions of metabolites remain to be characterized^{2,5}. This discrepancy raises a crucial point: might there be a technological explanation inflating the perceived complexity instead of a biological one?

The central dogma of biology describes the relation and information flow between the different omic layers from genome to transcriptome to proteome to metabolome, with the genome being the master code imprinted in any living system at its conception⁶. The code linking the first three layers is rather straightforward, with three nucleotides encoding one amino acid. However, such an encryption with the metabolome is yet to be revealed. This fact, combined with vast arrays of unannotated LC–MS/MS data, are two of the main reasons that the metabolome (being the entirety of small molecules within a biological system) cannot yet be defined. Current estimates suggest that less than 2% of observed LC–MS/MS spectra can be annotated, pointing to a potentially broad spectrum of unknown compounds².

The vast number of modern metabolomics studies (>90%) are conducted by LC–MS/MS analysis applying electrospray ionisation (ESI) as a so-called soft ionisation process. Following ionisation in the ESI source, charged molecules are sent into a collision cell and fragmented for structural characterization (Fig. 1). This process is ideally repeated for

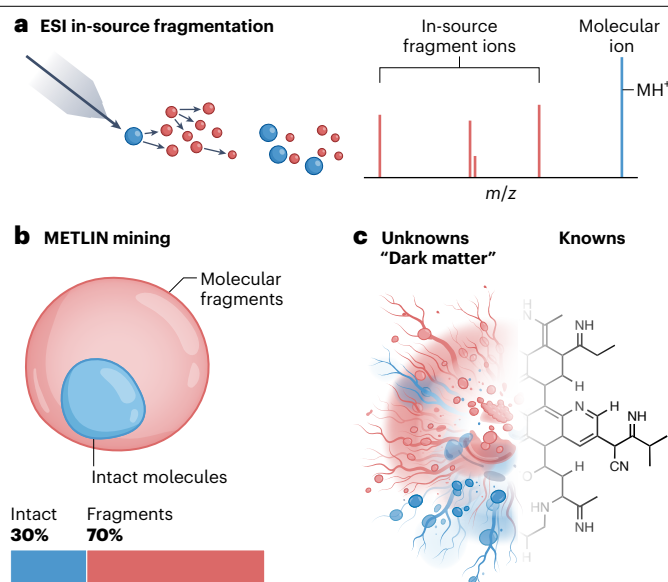


Fig. 1 | ESI ISF and its broad impact on mass spectrometry data analysis. **a**, ISF is a common phenomenon that occurs in the ESI source and has been investigated here across the METLIN MS/MS database of 931,000 molecular standards. **b**, The analysis reveals that widespread ISF (exceeding 70%) is responsible for most of the ESI peaks (also known as features), instead of molecular ions. **c**, This result indicates that the metabolome is in fact simpler than previously thought and that the ‘dark metabolome’ is largely made up of fragment ions generated during ESI.

all (charged) analytes of a sample. However, our laboratories (and others (ref. 7)) have observed and applied a phenomenon called in-source fragmentation (ISF)^{8,9}. ISF relates to the fragmentation of analytes during the initial ionisation process within the ESI source and hence before the collision cell. ISF basically generates a forest from a tree, in other words a single analyte can be presented as a molecular ion and one or many fragments. The employed mass analyser will blindly isolate and fragment (again) whatever is being sent into the collision cell. Given this fact, it appeared plausible to us that ISF might partially be responsible for the dark metabolome.

To test our hypothesis, we examined the METLIN MS/MS database. METLIN consists of 931,000 molecular standards representing over 350 chemical classes. For the purposes of this study, METLIN data was mined at 0eV,

an energy designed to simulate the absence of collision-induced dissociation. To retrieve ISF from the METLIN database, we queried all MS/MS data giving peak intensities $\geq 5\%$, with the charge setting being positive or negative or both, and the collision energy set to 0eV. The obtained hits were collected and summed up. A python script was used to perform this data analysis from all 931,000 molecules in METLIN. The analysis revealed that ISF could account for over 70% of the peaks observed in typical LC–MS/MS metabolomic datasets. Further to this, we reviewed data from fragmentation across different instrument platforms¹⁰ and found that, generally, ISF accounts for greater than 70% of the peaks observed, and in some instances this value was even greater (depending on the instrument).

This finding disrupts the prevailing assumption that the majority of peaks in mass spectra

correspond to unique metabolites. Instead, it suggests that the spectra may be significantly populated by fragment ions generated during the ionisation process, before any collision-induced dissociation. Such fragments, if not recognized as such, could be misclassified as distinct molecular entities, thus artificially expanding the metabolome's perceived complexity.

This insight challenges traditional MS/MS spectral interpretations that define any unidentified MS/MS spectrum as unknown (for example, the 'dark metabolome'). Moreover, the fact that several 'unknown' features will become intrinsically correlated potentially impacts statistical analysis. Our finding sheds an entirely different light on the potential size of the dark metabolome and questions if a technological rather than a biochemical phenomenon explains the overabundance of unknown metabolic space. Moreover, as LC-MS/MS data are also instrumental to metabolic networking and related computational approaches, our finding questions the outcomes of such methods on a broad biological scale. And while there remains a significant number of metabolites that are yet undiscovered, it is imperative to re-evaluate

the prevalent utilization and interpretation of LC-MS/MS in light of these insights. Nevertheless, our findings suggest a notable positive outcome: the metabolome appears to be distinctly more defined than earlier assumptions indicated. This clarity brings the possibility of fully understanding the metabolome within closer reach than we have previously envisioned.

Code availability

The python code to analyse the METLIN library is available at <https://metlin.scripps.edu/>.

Martin Giera ^{1,2} , **Aries Aisporna**³, **Winnie Uritboonthai**³ & **Gary Siuzdak** ^{3,4} 

¹Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands. ²The Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW), Leiden University Medical Center, Leiden, Netherlands. ³Scripps Center of Metabolomics and Mass Spectrometry, La Jolla, CA, USA. ⁴Department of Chemistry, Molecular and Computational Biology Scripps Research Institute, La Jolla, CA, USA.

 e-mail: M.A.Giera@lumc.nl; siuzdak@scripps.edu

Published online: 25 June 2024

References

1. Peisl, B. Y. L., Schymanski, E. L. & Wilmes, P. *Anal. Chim. Acta* **1037**, 13–27 (2018).
2. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
3. Amaral, P. et al. *Nature* **622**, 41–47 (2023).
4. Varabyou, A. et al. *Genome Biol.* **24**, 249 (2023).
5. Crow, J. M. *Nature* **599**, S14–S15 (2021).
6. Crick, F. *Nature* **227**, 561–563 (1970).
7. Xu, Y. F., Lu, W. & Rabinowitz, J. D. *Anal. Chem.* **87**, 2273–2281 (2015).
8. Chen, L. et al. *Rapid Commun. Mass Spectrom.* **37**, e9519 (2023).
9. Bernardo-Bermejo, S. et al. *Nat. Protocols* **18**, 1296–1315 (2023).
10. Hoang, C. et al. *Anal. Chem.* **96**, 5478–5488 (2024).

Acknowledgements

This research was partially funded by the Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW) supported by Novo Nordisk Foundation grant (NNF21CC0073729) (M.G.) and by the National Institutes of Health R35 GM130385 (G.S.).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Metabolism* thanks Caroline Bartman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.