Databases are an effective and efficient method for storage and access of mass-spectrometry data

William Kumler

Anitra E. Ingalls

<https://researcher-resources.acs.org/publish/author_guidelines?coden=jprobs>

# Abstract

(200 words max)

# Introduction

Mass spectrometry (MS) still lacks a performant data access format. The mzML file type1, a result of over a decade of interlaboratory collaboration and workshopping, struggles to provide rapid computational access to the m/z and intensity tuples. This is the crucial component in nearly all mass spectrometry analysis, but mzML’s text-based XML format requires time-consuming decompression performed one scan at a time. This is largely due to its preservation of the scan as the unit of transaction while the field moves increasingly away from single-scan analysis2.

As a result, alternative file formats are proposed practically every year. These include direct improvements to the mzML format with indexing3 and better internal encoding of the data4, HDF5-based alternatives4–8, relational databases9–13, or fully custom alternatives14,15. Fundamentally, these alternatives implement tradeoffs between user sanity in exchange for access speed and/or size on disk with clever compression algorithms and modern data structures that move away from the human-readable format of the mzML. These optimized formats are inherently more difficult to understand and usually lack comprehensive documentation or examples (particularly across programming languages) making it difficult for new users to enjoy their benefits or extend their functionality. This steep learning curve, coupled with a lack of support in conversion tools such as Proteowizard’s msconvert16, has prevented widespread adoption of these new formats despite their clear computational advantages. Such formats are also fragile in the sense that without community support, their continued development depends entirely on the original developers and easily become deprecated (i.e. YAFMS, Shaduf, mz5). A simple, speedy, and small MS data format remains very much in demand.

Relational databases are not new for MS workflows (see references above) and compete predominantly with HDF5-based methods. Both of these systems are widely used for big data and can be applied to MS data in a plethora of ways, leading to the proliferation of implementations we see today. Both backends provide excellent universality, larger-than-memory support, and rapid access to data, but HDF5-based systems excel at self-description and hierarchical structures8 while the relational database model is optimized for multi-table queries using a consistent syntax**ref?** original database paper?. Relational databases are increasingly seen in MS workflows for both raw and processed data, with SQLite backends now supported in the popular peakpicking software xcms17 via the Spectra package18 (though in-memory and HDF5 options are also supported) and on MetabolomicsWorkbench19 while the development of MassQL20 demonstrates the increasing comfort that MS analysts have with the adoption of SQL.

Relational databases also have several distinct advantages over hierarchical or text-based systems, particularly in performing searches for subsets of data via indices. Importantly, this indexing differs from the byte-offset indexes that already exist in the indexed mzML and HDF5 formats because the search for a particular subset cannot be done efficiently with a byte-offset index when the m/z data is encoded, though access to a particular scan can be incredibly rapid. Additionally, data from multiple samples can be stored together in a single table, differing from formats like mzDB, mzTree, and mzMD and allowing queries of all dataset samples to be performed without looping through each file in turn, thereby avoiding the associated computational overhead and query complexity.

SQL databases also allow mass spectrometrists to access the continual improvements and long-term stability produced by the industries who specialize in these. While HDF5 is a common scientific data format, databases are constantly under development by industry titans deeply invested in their maintenance and optimization. Online analytical processing (OLAP) methods are particularly well suited for MS data given their optimization for read speed, making modern systems such as DuckDB21 or Apache Arrow’s columnar formats highly appealing while preserving the familiar serverless approach.

Our previous work showed how the ragged arrays of MS data can be converted into a tidy database table22 and here we logically extend that method into proper database storage and access. Here, we test the hypothesis that a “vanilla” implementation of a relational database which exposes the raw m/z and intensity tuples is an intuitive and performant way of storing MS data for exploratory analysis, visualization, and quality control. We compare the time and space required to extract a representative data subset under three conditions (single scan, ion chromatogram, and all scans within a retention time range) and perform these tests on multiple databases as well as mzML and other MS data formats across multiple MS experiments with varying magnitudes for comparison.

# Experimental section

This “vanilla” database deprioritizes the metadata associated with a given file and focuses on performance and simplicity. Thus, this method does not replace the existing vendor-specific or mzML files but instead constructs a parallel data structure and represents our intuition that metadata is rarely the main target of MS analysis and that labs typically preserve the original files anyway.

# Results

# Discussion

# Conclusion

# Acknowledgements

# Data availability

# References

# Supplement

(1) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. mzML—a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics* **2011**, *10* (1), R110.000133. <https://doi.org/10.1074/mcp.R110.000133>.

(2) Ting, Y. S.; Egertson, J. D.; Payne, S. H.; Kim, S.; MacLean, B.; Käll, L.; Aebersold, R.; Smith, R. D.; Noble, W. S.; MacCoss, M. J. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & Cellular Proteomics* **2015**, *14* (9), 2301–2307. <https://doi.org/10.1074/mcp.O114.047035>.

(3) Röst, H. L.; Schmitt, U.; Aebersold, R.; Malmström, L. Fast and Efficient XML Data Access for Next-Generation Mass Spectrometry. *PLOS ONE* **2015**, *10* (4), e0125108. <https://doi.org/10.1371/journal.pone.0125108>.

(4) Bhamber, R. S.; Jankevics, A.; Deutsch, E. W.; Jones, A. R.; Dowsey, A. W. mzMLb: A Future-Proof Raw Mass Spectrometry Data Format Based on Standards-Compliant mzML and Optimized for Speed and Storage Requirements. *Journal of Proteome Research* **2021**, *20* (1), 172–183. <https://doi.org/10.1021/acs.jproteome.0c00192>.

(5) Wilhelm, M.; Kirchner, M.; Steen, J. A. J.; Steen, H. Mz5: Space- and Time-efficient Storage of Mass Spectrometry Data Sets. *Molecular & Cellular Proteomics* **2012**, *11* (1), O111.011379. <https://doi.org/10.1074/mcp.O111.011379>.

(6) Bilbao, A.; Ross, D. H.; Lee, J.-Y.; Donor, M. T.; Williams, S. M.; Zhu, Y.; Ibrahim, Y. M.; Smith, R. D.; Zheng, X. MZA: A Data Conversion Tool to Facilitate Software Development and Artificial Intelligence Research in Multidimensional Mass Spectrometry. *Journal of Proteome Research* **2023**, *22* (2), 508–513. <https://doi.org/10.1021/acs.jproteome.2c00313>.

(7) Tully, B. Toffee – a Highly Efficient, Lossless File Format for DIA-MS. *Scientific Reports* **2020**, *10* (1), 8939. <https://doi.org/10.1038/s41598-020-65015-y>.

(8) Askenazi, M.; Ben Hamidane, H.; Graumann, J. The Arc of Mass Spectrometry Exchange Formats Is Long, but It Bends Toward HDF5. *Mass Spectrometry Reviews* **2017**, *36* (5), 668–673. <https://doi.org/10.1002/mas.21522>.

(9) Shah, A. R.; Davidson, J.; Monroe, M. E.; Mayampurath, A. M.; Danielson, W. F.; Shi, Y.; Robinson, A. C.; Clowers, B. H.; Belov, M. E.; Anderson, G. A.; Smith, R. D. An Efficient Data Format for Mass Spectrometry-Based Proteomics. *Journal of the American Society for Mass Spectrometry* **2010**, *21* (10), 1784–1788. <https://doi.org/10.1016/j.jasms.2010.06.014>.

(10) Bouyssié, D.; Dubois, M.; Nasso, S.; Gonzalez De Peredo, A.; Burlet-Schiltz, O.; Aebersold, R.; Monsarrat, B. mzDB: A File Format Using Multiple Indexing Strategies for the Efficient Analysis of Large LC-MS/MS and SWATH-MS Data Sets \*. *Molecular & Cellular Proteomics* **2015**, *14* (3), 771–781. <https://doi.org/10.1074/mcp.O114.039115>.

(11) Handy, K.; Rosen, J.; Gillan, A.; Smith, R. Fast, Axis-Agnostic, Dynamically Summarized Storage and Retrieval for Mass Spectrometry Data. *PLOS ONE* **2017**, *12* (11), e0188059. <https://doi.org/10.1371/journal.pone.0188059>.

(12) Yang, R.; Ma, J.; Zhang, S.; Zheng, Y.; Wang, L.; Zhu, D. mzMD: Visualization-Oriented MS Data Storage and Retrieval. *Bioinformatics* **2022**, *38* (8), 2333–2340. <https://doi.org/10.1093/bioinformatics/btac098>.

(13) Beagley, N.; Scherrer, C.; Shi, Y.; Clowers, B. H.; Danielson, W. F.; Shah, A. R. Increasing the Efficiency of Data Storage and Analysis Using Indexed Compression. In *2009 Fifth IEEE International Conference on e-Science*; IEEE: Oxford, United Kingdom, 2009; pp 66–71. <https://doi.org/10.1109/e-Science.2009.18>.

(14) Römpp, A.; Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Stöckli, M.; Spengler, B. imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. In *Data Mining in Proteomics*; Hamacher, M., Eisenacher, M., Stephan, C., Eds.; Humana Press: Totowa, NJ, 2011; Vol. 696, pp 205–224. <https://doi.org/10.1007/978-1-60761-987-1_12>.

(15) Lu, M.; An, S.; Wang, R.; Wang, J.; Yu, C. Aird: A Computation-Oriented Mass Spectrometry Data Format Enables a Higher Compression Ratio and Less Decoding Time. *BMC Bioinformatics* **2022**, *23* (1), 35. <https://doi.org/10.1186/s12859-021-04490-0>.

(16) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nature Biotechnology* **2012**, *30* (10), 918–920. <https://doi.org/10.1038/nbt.2377>.

(17) Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78* (3), 779–787. <https://doi.org/10.1021/ac051437y>.

(18) Rainer, J.; Vicini, A.; Salzer, L.; Stanstrup, J.; Badia, J. M.; Neumann, S.; Stravs, M. A.; Verri Hernandes, V.; Gatto, L.; Gibb, S.; Witting, M. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* **2022**, *12* (2), 173. <https://doi.org/10.3390/metabo12020173>.

(19) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools. *Nucleic Acids Research* **2016**, *44* (D1), D463–D470. <https://doi.org/10.1093/nar/gkv1042>.

(20) Jarmusch, A. K.; Aron, A. T.; Petras, D.; Phelan, V. V.; Bittremieux, W.; Acharya, D. D.; Ahmed, M. M. A.; Bauermeister, A.; Bertin, M. J.; Boudreau, P. D.; Borges, R. M.; Bowen, B. P.; Brown, C. J.; Chagas, F. O.; Clevenger, K. D.; Correia, M. S. P.; Crandall, W. J.; Crüsemann, M.; Damiani, T.; Fiehn, O.; Garg, N.; Gerwick, W. H.; Gilbert, J. R.; Globisch, D.; Gomes, P. W. P.; Heuckeroth, S.; James, C. A.; Jarmusch, S. A.; Kakhkhorov, S. A.; Kang, K. B.; Kersten, R. D.; Kim, H.; Kirk, R. D.; Kohlbacher, O.; Kontou, E. E.; Liu, K.; Lizama-Chamu, I.; Luu, G. T.; Knaan, T. L.; Marty, M. T.; McAvoy, A. C.; McCall, L.-I.; Mohamed, O. G.; Nahor, O.; Niedermeyer, T. H. J.; Northen, T. R.; Overdahl, K. E.; Pluskal, T.; Rainer, J.; Reher, R.; Rodriguez, E.; Sachsenberg, T. T.; Sanchez, L. M.; Schmid, R.; Stevens, C.; Tian, Z.; Tripathi, A.; Tsugawa, H.; Nishida, K.; Matsuzawa, Y.; Van Der Hooft, J. J. J.; Vicini, A.; Walter, A.; Weber, T.; Xiong, Q.; Xu, T.; Zhao, H. N.; Dorrestein, P. C.; Wang, M. A Universal Language for Finding Mass Spectrometry Data Patterns, 2022. <https://doi.org/10.1101/2022.08.06.503000>.

(21) Raasveldt, M.; Mühleisen, H. DuckDB: An Embeddable Analytical Database. In *Proceedings of the 2019 International Conference on Management of Data*; ACM: Amsterdam Netherlands, 2019; pp 1981–1984. <https://doi.org/10.1145/3299869.3320212>.

(22) Kumler, W.; E. Ingalls, A. Tidy Data Neatly Resolves Mass-Spectrometry’s Ragged Arrays. *The R Journal* **2022**, *14* (3), 193–202. <https://doi.org/10.32614/RJ-2022-050>.