Published in final edited form as:
Nat Methods. 2020 October ; 17(10): 953–954. doi:10.1038/s41592-020-0942-5.

METLIN MS^2 Molecular Standards Database: A Comprehensive Chemical and Biological Resource

Jingchuan Xue[†,#], **Carlos Guijas**[†,#], **H. Paul Benton**[†], **Benedikt Warth**[£], **Gary Siuzdak**[†,‡,*]

[†]Scripps Center for Metabolomics and Mass Spectrometry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037

[£]Department of Food Chemistry and Toxicology, Faculty of Chemistry, University of Vienna, Währinger Str. 38, 1090, Vienna, Austria

[‡]Department of Molecular and Computational Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037

Editor's note:

This article has been peer reviewed.

To the Editor - Tandem mass spectrometry (MS^2) data provides high confidence molecular identification of known molecules and preliminary characterization of novel, unknown molecules (unknowns). However, in order for databases to be an effective resource, broad chemical space coverage is necessary. Consequently, we have created METLIN (http://metlin.scripps.edu) a highly annotated and structurally diverse database of over 850,000 molecular standards. METLIN's tandem mass spectral library covers almost 1% of PubChem's 93 million compounds, essentially a number that can be characterized as the currently known chemical space.

The utility of MS^2 data acquisition is especially advantageous when coupled with liquid chromatography mass spectrometry (LC-MS) analysis of complex samples[1, 2]. Such datasets typically have tens of thousands of features[3] and while accurate mass measurements (MS^1) of precursor molecular ions can provide putative identifications, these MS^1 measurements alone are not sufficient to structurally characterize the number of compounds having identical or similar molecular weights. Therefore, characterizing every feature in an LC-MS dataset is challenging and currently not possible. However, implementation of MS^2 provides structural information that greatly increases the confidence of molecular identifications[2]. The recent expansion of the METLIN MS^2 database of molecular standards offers an opportunity to quantify this improved confidence (Figure 1). METLIN now hosts over 850,000 molecular standards with MS^2 data generated in both positive and negative

[*]Corresponding author: Gary Siuzdak, PhD, Scripps Center for Metabolomics, The Scripps Research Institute, La Jolla, CA 92037, Tel.: 858 784 9113, siuzdak@scripps.edu.
[#]authors contributed equally

AUTHOR CONTRIBUTIONS
J.X., C.J., H.P.B., B.W. and G.S. contributed to data collection, analysis, and manuscript writing. J.X. and C.J. share first authorship.

COMPETING FINACIAL INTERESTS
The authors declare no competing financial interests.

ionization modes at multiple collision energies, collectively containing over 3,000,000 curated high-resolution tandem mass spectra. Thus, the size of METLIN makes molecular annotation and identification more feasible. In comparison, the NIST MS^2 database, the next largest molecular standards database, contains 15,000 standards (Figure 1a).

The combination of METLIN's molecular standards and systematically acquired experimental data allows for the examination of the impact that MS^2 data has on the identification of known molecules. For example, when METLIN is searched against precursor *m/z* values at varying part per million (ppm) errors, the number of hits typically ranges from tens to hundreds of compounds. However, with the addition of MS^2 data the false positive rate can be minimized to only a few compounds. Beyond providing molecular identification through its multiple search capabilities (e.g. MS^2, batch, name, and elemental composition), METLIN's expansion will facilitate similarity searching. The similarity searching algorithm was originally developed to aid in the identification of unknowns and the discovery of novel molecules (unknowns)[6] and operates by using fragment ion data to help align an unknown molecule to compounds with similar fragmentation data within a database to help further identify and characterize them[6]. METLIN's fragment ion similarity searching (FISS) and neutral loss similarity search (NLSS) is applied in identifying endogenous metabolites, drugs, drug metabolites, as well as biotransformation of xenobiotics[7]. METLIN facilitates both endogenous and exogenous compound identification. For example, it contains MS^2 data for over 60,000 pyrimidine analogues and over 6000 purine analogues (Figure 1b) among others.

The expanded METLIN database will enable new types of analyses. First, we expect that MS^2 data of this size can significantly reduce the magnitude of false positives that molecular identification based solely on molecular ion values can generate. Second, while very high accuracy MS^2 data is useful it does not significantly enhance identification confidence. Therefore, low-resolution instrumentation can be more broadly utilized for relatively sophisticated experiments by chemists and biologists that do not have access to high-end equipment[8]. And finally, METLIN can be applied for identification of unknown compounds via fragment ion and neutral loss similarity searching (FISS and NLSS). For example, synthetic chemists can apply METLIN toward the structure elucidation of unexpected products, while biochemists can use it in identifying the plethora of bacterial and human metabolites in microbiome and exposome-related studies, and it has unexplored potential in the chemical, toxicological, and pharmaceutical sciences[9, 10].

Acknowledgements

This research was partially funded by National Institutes of Health grants R35 GM130385 (G.S.), P30 MH062261 (G.S.), P01 DA026146 (G.S.), and U01 CA235493 (G.S.) and by Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory for the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under contract number DE-AC02-05CH11231 (G.S.).

Data availability

The data in METLIN database is available at http://metlin.scripps.edu. The data in other databases mentioned in this study were obtained from their websites [accessed on Februrary

2020] or published papers: MONA (https://mona.fiehnlab.ucdavis.edu/); mzCloud (https://www.mzcloud.org/), GNPS (https://gnps.ucsd.edu/; ref4), HMDB (https://hmdb.ca/; ref5), and NIST 17 (https://chemdata.nist.gov/).

References

1. Guijas C et al. Anal. Chem 90, 3156–3164 (2018). [PubMed: 29381867]
2. Tautenhahn R et al. Nat. Biotechnol 30, 826–828 (2012).
3. Kafader JO et al. Nat. Methods 17, 391–94 (2020). [PubMed: 32123391]
4. Wang M et al. Nat. Biotechnol 38, 23–26 (2020). [PubMed: 31894142]
5. Wishart DS et al. Nucleic Acids Res 46, D608–D617 (2018). [PubMed: 29140435]
6. Benton HP, Wong DM, Trauger SA & Siuzdak G Anal. Chem 80, 6382–6389 (2008). [PubMed: 18627180]
7. Flasch M et al. ACS Chem. Bio 15, 970–981 (2020). [PubMed: 32167285]
8. Xue J et al. Anal. Chem 92, 6051–6059 (2020). [PubMed: 32242660]
9. Quinn RA et al. Nature 579, 123–129 (2020). [PubMed: 32103176]
10. Cllayton TA et al. Proc. Natl. Acad. Sci. USA 106, 14728–14733 (2009). [PubMed: 19667173]

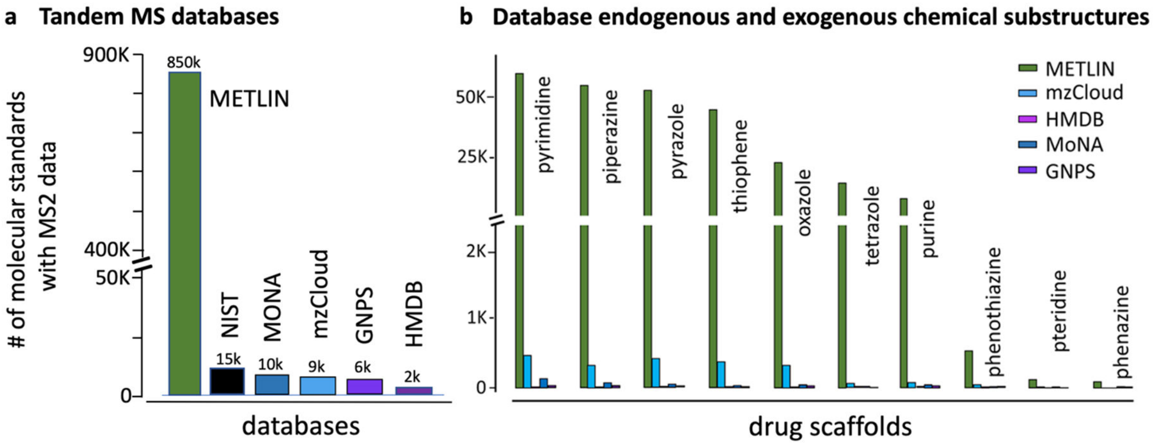

Figure 1.

The METLIN MS^2 database. (a) Comparison of databases containing MS^2 data from molecular standard,[4, 5]. (b) MS^2 database comparison of commonly observed endogenous substructures and drug scaffolds.