METLIN MS^2 molecular standards database: a broad chemical and biological resource

To the Editor — Tandem mass spectrometry (MS^2) data provide high-confidence molecular identification of known molecules and preliminary characterization of novel, unknown molecules (unknowns). However, 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 numerically covers almost 1% of PubChem's 93 million compounds, essentially a number that can be characterized as the 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 single-stage MS measurements of precursor molecular ions can provide putative identifications, these 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 (Fig. 1). METLIN now hosts over 850,000 molecular standards with MS^2 data generated in both positive and negative ionization modes at multiple collision energies, collectively containing over 4,000,000 curated high-resolution tandem mass spectra. Thus, the size of METLIN makes molecular annotation and identification more feasible. In comparison, the National Institute of Standards and Technology (NIST) MS^2 database, the next largest molecular standards database, contains 15,000 standards (Fig. 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 parts 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 (for example, 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[4] 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 it[4]. METLIN's fragment ion similarity searching and neutral loss similarity searching are applied in identifying endogenous metabolites, drugs and drug metabolites, as well as biotransformation products of xenobiotics[5]. METLIN facilitates both endogenous and exogenous compound identification. For example, it contains MS^2 data for over 60,000 pyrimidine analogs and over 6,000 purine analogs (Fig. 1b), among others.

The expanded METLIN database will enable new types of analyses. First, we expect that a MS^2 database of this size can substantially 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 are useful, they do not substantially enhance identification confidence. Therefore, low-resolution instrumentation can be more broadly used for relatively sophisticated experiments by chemists and biologists who do not have access to high-end equipment[6]. And finally, METLIN can be applied for identification of unknown compounds via fragment ion and neutral loss similarity

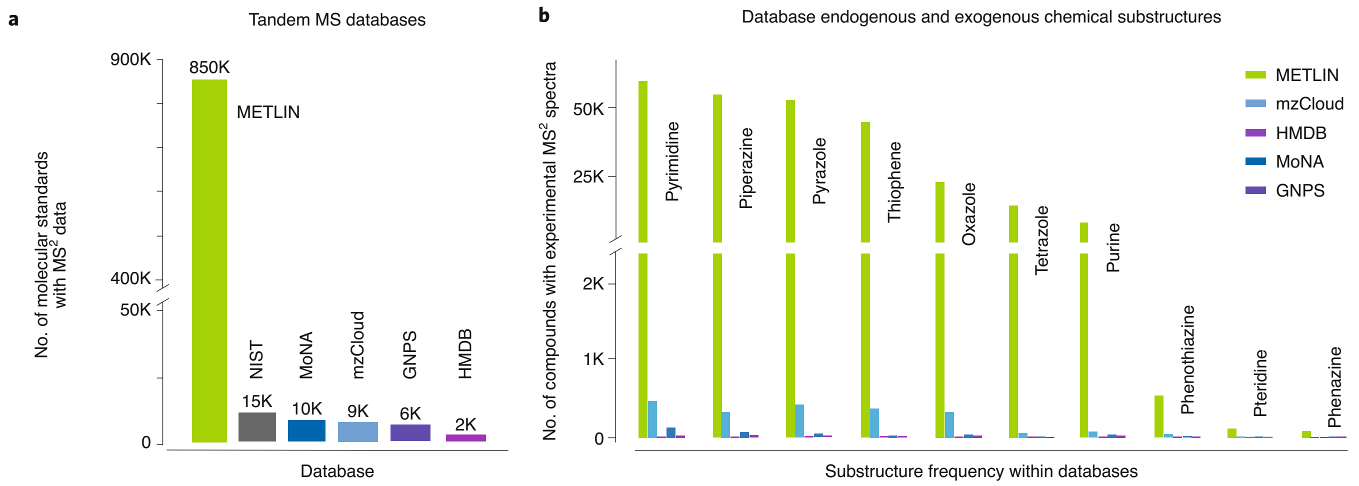

Fig. 1 | The METLIN MS^2 database. a, Comparison of databases containing MS^2 data from molecular standards[9,10]. **b**, MS^2 database comparison of commonly observed endogenous substructures and drug scaffolds.

searching. For example, synthetic chemists can apply METLIN to the structural elucidation of unexpected products, while biochemists can use it to identify the plethora of bacterial and human metabolites in microbiome and exposome studies, and it has unexplored potential in the chemical, toxicological and pharmaceutical sciences[7,8].

Data availability

The data in the METLIN database are available at http://metlin.scripps.edu. The data in other databases mentioned in this study were obtained from their websites (accessed in February 2020) or published papers: MoNA (https://mona.fiehnlab.ucdavis.edu/), mzCloud (https://www.mzcloud.org/), GNPS (https://gnps.ucsd.edu/)[9], HMDB (https://hmdb.ca/)[10] and NIST 17 (https://chemdata.nist.gov/). ❐

Editorial note: This article has been peer reviewed.

Jingchuan Xue[1,4], Carlos Guijas[1,4], H. Paul Benton[1], Benedikt Warth[2] and Gary Siuzdak[1,3] ✉

[1]Scripps Center for Metabolomics and Mass Spectrometry, Scripps Research Institute, La Jolla, CA, USA. [2]Department of Food Chemistry and Toxicology, Faculty of Chemistry, University of Vienna, Vienna, Austria. [3]Department of Molecular and Computational Biology, Scripps Research Institute, La Jolla, CA, USA. [4]These authors contributed equally: Jingchuan Xue, Carlos Guijas.
✉*e-mail: siuzdak@scripps.edu*

Published online: 24 August 2020
https://doi.org/10.1038/s41592-020-0942-5

References

1. Guijas, C. et al. *Anal. Chem.* **90**, 3156–3164 (2018).
2. Tautenhahn, R. et al. *Nat. Biotechnol.* **30**, 826–828 (2012).
3. Kafader, J. O. et al. *Nat. Methods* **17**, 391–394 (2020).
4. Benton, H. P., Wong, D. M., Trauger, S. A. & Siuzdak, G. *Anal. Chem.* **80**, 6382–6389 (2008).
5. Flasch, M. et al. *ACS Chem. Biol.* **15**, 970–981 (2020).
6. Xue, J. et al. *Anal. Chem.* **92**, 6051–6059 (2020).
7. Quinn, R. A. et al. *Nature* **579**, 123–129 (2020).
8. Clayton, T. A., Baker, D., Lindon, J. C., Everett, J. R. & Nicholson, J. K. *Proc. Natl Acad. Sci. USA* **106**, 14728–14733 (2009).
9. Wang, M. et al. *Nat. Biotechnol.* **38**, 23–26 (2020).
10. Wishart, D. S. et al. *Nucleic Acids Res.* **46**, D608–D617 (2018). D1.

Acknowledgements

This research was partially funded by US 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 US Department of Energy, Office of Science, Office of Biological and Environmental Research, under contract number DE-AC02-05CH11231 (G.S.).

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0942-5.