nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-023-01985-4

Fast mass spectrometry search and clustering of untargeted metabolomics data

Received: 23 March 2023

Accepted: 12 September 2023

Published online: 02 January 2024

Check for updates

Mihir Mongia[1,6], Tyler M. Yasaka[1,6], Yudong Liu[1,6], Mustafa Guler[1], Liang Lu[1], Aditya Bhagwat[1], Bahar Behsaz[1,2], Mingxun Wang[3], Pieter C. Dorrestein[4,5] & Hosein Mohimani[1]

The throughput of mass spectrometers and the amount of publicly available metabolomics data are growing rapidly, but analysis tools such as molecular networking and Mass Spectrometry Search Tool do not scale to searching and clustering billions of mass spectral data in metabolomics repositories. To address this limitation, we designed MASST+ and Networking+, which can process datasets that are up to three orders of magnitude larger than those processed by state-of-the-art tools.

During the past decade, the amount of mass spectral data collected in the fields of natural products, exposomics and metabolomics has grown exponentially[1–3]. In accordance with advances in mass spectrometry technology, multiple computational methods have been developed for analyzing these massive datasets. Recently, Mass Spectrometry Search Tool (MASST) was introduced as a search engine for finding analogs of a query spectrum in mass spectrometry repositories[4]. MASST has demonstrated utility in the annotation of a wide variety of unidentified metabolites, including clinically important molecules in patient cohorts[5–9], toxins or pesticides in environmental samples[10], fungal metabolites[11] and metabolites from pathogenic microorganisms[12–15]. Moreover, molecular networking has been introduced for clustering spectral datasets into families of related molecules[16,17]. Molecular networking has yielded a systematic view of the chemical space in different ecosystems and helped determine the structure of many compounds[18–25].

MASST and molecular networking are based on a naive approach for scoring two tandem mass spectra. MASST compares the query spectrum against all reference spectra one by one and computes a similarity score based on the relative intensities of shared and shifted peaks. Therefore, the runtime of MASST grows linearly with the repository size. Molecular networking first uses MS-Clustering[16] to cluster identical spectra by calculating a dot-product score (ExactScore, Fig. 1a(i)) between the spectra. Then, spectral networking[17] is used to calculate a dot-product score accounting for peaks that are shared or shifted (ShiftedScore, Fig. 1a(ii)) between all pairs of clusters to find groups of related molecules. This latter procedure grows quadratically with the number of clusters. Current trends show that the size of public mass spectral repositories doubles every two to three years (Supplementary Fig. 1). Therefore, the current implementations of MASST and molecular networking will not be able to scale with the growth of future repositories. A MASST search for a single spectrum against the clustered global natural product social (GNPS) database (~83 million clusters) currently takes about an hour on a single thread, and a MASST search against the entire GNPS (717 million spectra) does not complete after being run for 3 days. Currently, molecular networking analysis of a million spectra takes a few hours, whereas molecular networking of ~20 million spectra does not yield results after running for a week. Similar to the area of computational genomics, handling the exponential growth of repositories requires the development of more efficient and scalable search algorithms.

In this work, we introduce a fast dot-product algorithm that preprocesses a set of spectra into an indexing table. This indexing table maps all possible precursor m/z and fragment ion m/z pairs to the spectra that contain them. Using this indexing, given a query spectrum, the dot product with respect to all spectra can be computed efficiently by iterating through each query peak and using the indexing table to retrieve spectra with similar peaks (Fig. 1b). As mass spectra are sparse, only a small fraction of spectra and peaks are retrieved for each query. The ability to leverage this sparsity requires only a small fraction of the computation used by naive scoring methods, because the vast majority of the tandem mass spectra in the index are never touched during the

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Chemia Biosciences Inc., Pittsburgh, PA, USA. [3]Computer Science and Engineering, University of California Riverside, Riverside, CA, USA. [4]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, USA. [5]Department of Pharmacology and Pediatrics, University of California San Diego, San Diego, CA, USA. [6]These authors contributed equally: Mihir Mongia, Tyler M. Yasaka, Yudong Liu. e-mail: hoseinm@andrew.cmu.edu

https://doi.org/10.1038/s41587-023-01985-4

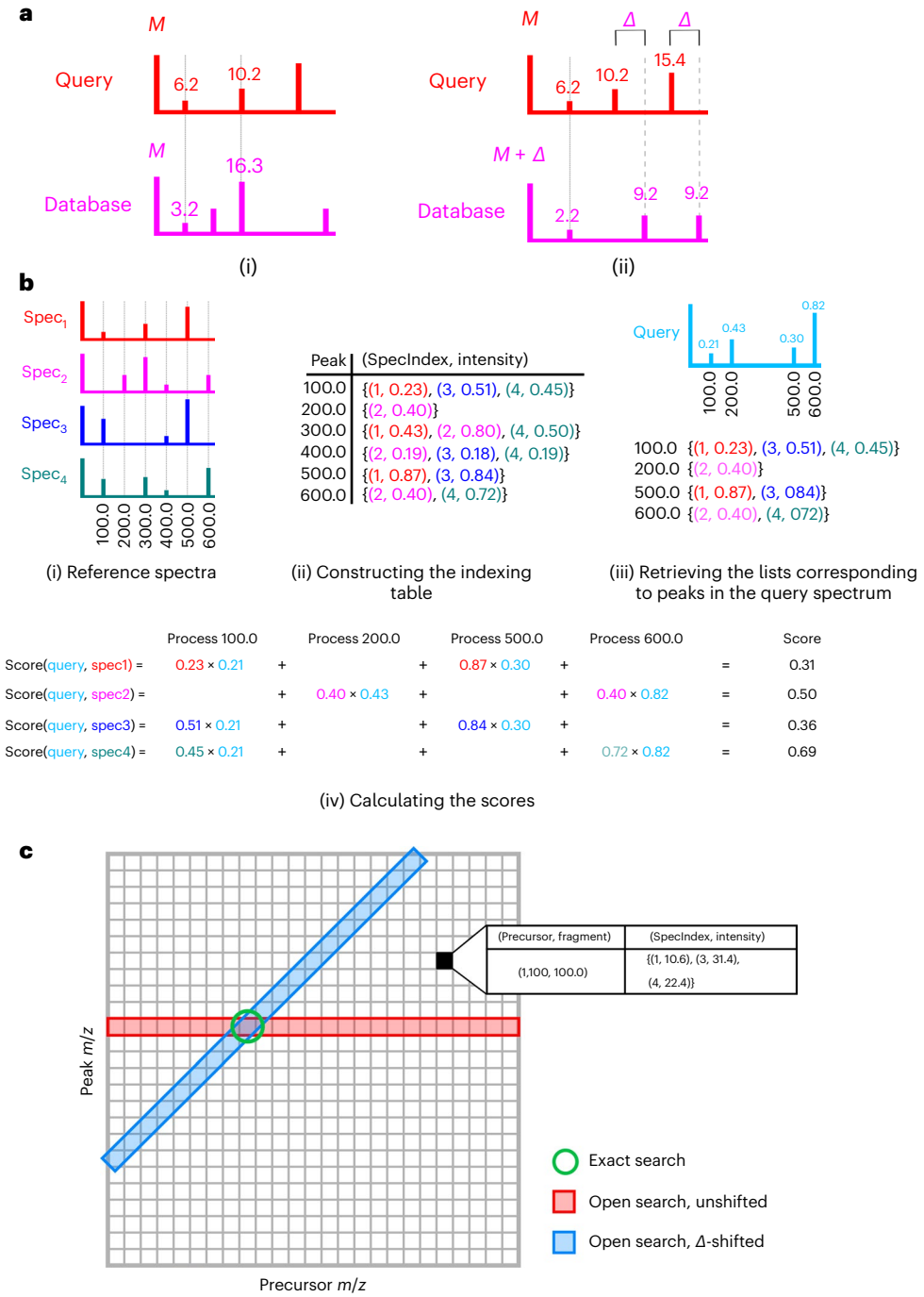

Fig. 1 | Fast scoring with indexing. a, Similarity score. (i) In exact search, MASST searches a query spectrum against all database spectra with similar precursor masses and computes the ExactScore, a sum of multiplications between intensities of peaks shared by the query and database spectrum (shown in solid gray). In this case, the score is $6.2 \times 3.2 + 10.2 \times 16.3 = 186.1$. (ii) In the case of analog search, MASST searches the query spectrum against all database spectra within a specific precursor mass range (for example, 300 Da) and computes the ShiftedScore, a sum of multiplications between intensities of peaks that are shared and Δ-shifted between the query and database spectrum. Here, there is one shared (solid gray) and two Δ-shifted (dashed gray) peaks, yielding a total score of $6.2 \times 2.2 + 10.2 \times 9.2 + 15.4 \times 9.2 = 249.16$. **b**, Fast dot product. (i) Given a database of spectra, the fast dot procedure starts with (ii) construction of an indexing table, where each row corresponds to a fragment peak mass and contains a list of tuples of spectra indices that contain the peak, along with the intensity of the peak in these spectra. (iii) Given a query spectrum, all lists corresponding to peaks present in the query are retrieved. Then (iv), for each list, and for each tuple in the list, the product of the intensity of the corresponding query peak and database peak is added to the total dot-product score of query and database spectra. For simplicity, in this illustration all the spectra have the same precursor mass. **c**, Fast dot-product indexing. The fast dot-product indexing table corresponds to a two-dimensional grid, with precursor mass on the *x* axis and peak mass on the *y* axis. Each database peak is inserted into a list corresponding to a specific location in the grid, determined by the peak mass and the precursor mass. In exact search, for each query peak only the list in a single cell will be retrieved (highlighted with green circle). For analog search, red cells (corresponding to shared peaks) and blue cells (corresponding to Δ-shifted peaks) are retrieved. spec, spectrum.

query process. By integrating this indexing approach into the scoring subroutines of MASST and molecular networking, we develop two computational tools, MASST+ and Networking+, which are two to three orders of magnitude faster than state-of-the-art tools on large datasets. Further, the indexing approach supports online growth, that is, the insertion of new spectra without the need for recalculation from scratch. This enables both MASST+ and Networking+ to efficiently handle the dynamic growth of reference spectra. Currently, MASST+ is available as a web service from https://masst.ucsd.edu/masstplus/. GNPS supports stand-alone MASST+ (Supplementary Fig. 2) and integration with molecular networking (Supplementary Fig. 3).

Results

Given a query spectrum, MASST+ efficiently searches a database of reference spectra to find similar entries by creation of an indexing table—a data structure that allows rapid retrieval of similar spectra based on the peaks present in the query spectrum. For each precursor mass M and each peak mass p, a list of indices of spectra with precursor M and peak p are stored, along with the intensity of the peaks. In the case of exact search, MASST+ iterates through the peaks in the query spectrum and retrieves the lists associated with a query peak and the query's precursor mass. The ExactScore is calculated by multiplying and adding the intensities of each peak in the query spectrum and reference spectra (Fig. 1b). In the case of analog search (Supplementary Fig. 4), MASST+ uses a much larger precursor mass tolerance (for instance, 300 Da) and computes a ShiftedScore that takes into account both shared and Δ-shifted peaks (peaks in reference spectra that are Δ Da larger than peaks in the query), where Δ is the mass difference between the precursors of the query and reference spectra (Fig. 1c).

Networking+ clusters spectral datasets into families of related molecules by first putting spectra from identical molecules into the same clusters (Clustering+), then forming the centers of each cluster by taking their consensus and then connecting the clusters that are predicted to be generated from related molecules (Pairing+). Clustering+ iterates over all spectra and associates each spectrum with a cluster that is highly similar. It uses a strategy similar to MASST+ exact search for efficiently calculating the SharedScore between the spectrum and each cluster center. Pairing+ uses a shared and Δ-shifted dot product as a similarity measure for identifying related spectra. It uses a strategy similar to MASST+ analog search to find all pairs of clusters with high ShiftedScore.

We have benchmarked MASST+ (Supplementary Table 1) on various GNPS datasets, including the MSV000078787 dataset collected on *Streptomyces* cultures (5,433 spectra), clustered GNPS (83,131,248 spectra) and entire GNPS (717,395,473 spectra). Supplementary Data 1 lists the accession identifiers of all GNPS datasets used in our study. While MASST and MASST+ reported identical hits, MASST+ was two orders of magnitude faster and more memory efficient (Supplementary Table 1). For small datasets, we only achieved a threefold increase in speed; however, this was magnified when larger datasets were searched. In the case of the clustered GNPS, MASST+ performed analog search in 15 s, whereas MASST took 49 min, a 196-fold increase. In the case of the entire GNPS, MASST+ performed analog search in under 2 h on average, whereas MASST search did not finish within 3 days on the GNPS server, making it practically not possible to routinely perform such a search.

Figure 2a illustrates the runtime and memory consumption of MASST+ in exact and analog mode for various subsets of the clustered GNPS. Indexing time and memory consumption grew linearly with the size of datasets (Supplementary Fig. 5), and indexing time increased for larger values of peak mass tolerance (Supplementary Fig. 6). MASST+ took 8 h of computational time and 8 GB of memory to index ~83 million spectra from the clustered GNPS, and 72 h of compute time and 9 GB of memory to index 717 million spectra contained in GNPS. Supplementary Fig. 7 breaks down MASST+ runtime into two different steps, loading peak lists and computing dot products, for various numbers of query spectra. Loading peak lists consumed about half of the total runtime when the number of query spectra was greater than 100.

Figure 2b and Supplementary Tables 2–5 benchmark Networking+ against molecular networking for various data sizes with runtime less than 24 h. In 24 h, Clustering+ could process 300 million spectra on a single CPU, whereas MS-Clustering could process 20 million spectra. Moreover, in this timeline, Pairing+ could process 2 million spectra, whereas spectral networking could handle 0.2 million spectra. Clustering+ and Pairing+ were two orders of magnitude faster than their counterparts, MS-Clustering[16] and spectral networking[17]. The clusters and networks reported by Clustering+ and Pairing+ were identical to those obtained with MS-Clustering and spectral networks. As noted by Bittremieux et al.[26], it was not previously possible to directly create a molecular network from all the GNPS spectra; here, we show that this is now possible with Networking+ with minimal computer memory requirements.

We clustered the entire GNPS (717 million scans) using Clustering+ and formed a network using Pairing+. This resulted in 8,453,822 million clusters and 4,947,928 connected components with a total of 17,533,386 edges (available from https://github.com/mohimanilab/MASSTplus). Among the 4,948,146 connected components in the network, 98% (4,849,047 components) consisted of a single node, whereas 1.5%, 0.3%, 0.2% and 0.02% (74530, 13957, 9239 and 1152 components) had 2, 3, 4–9 and 10+ nodes, respectively (Supplementary Fig. 8). Among 7,986,356 clusters in the network, 1.7% (134,198 clusters) matched reference spectra from the NIST library, 6% (477,721 clusters) were a neighbor of a cluster-matched NIST library, 14% (1,130,092 clusters) were a neighbor of a neighbor, and 78% (5,390,554 clusters) were three or more hops away from any cluster-matching NIST library (Supplementary Fig. 9). Of the 307,709 clusters consisting of 20 or more spectra, for 18% (54,518 clusters) all spectra came from a single MassIVE dataset, whereas for 13% and 69% (39,428 and 213,763 clusters) spectra came from 2 or 3+ MassIVE datasets, respectively (Supplementary Fig. 10). About 61% of the clusters with precursor mass between 0 and 400 Da consisted of only two GNPS spectra, whereas fewer than half the clusters with precursor mass above 400 Da consisted of only two GNPS spectra (Supplementary Fig. 11). Networking+ took 6 days to finish this task on one CPU. This task was not feasible using previous approaches.

The indexing strategies proposed here are applicable to all classes of small molecules. Here, we illustrate the application of these methods in the case of lanthipeptide natural products. Currently, methods for high-throughput discovery of lanthipeptides through computational analysis of genomics and metabolomics data have various limitations, especially at repository scale. Lanthipeptides are a biologically important class of natural products that include antibiotics[27], antifungals[28], antivirals[29] and antinociceptives[30]. Lanthipeptides are structurally defined by the thioether amino acids lanthionine, methyllanthionine and labionin. Lanthionine and methyllanthionine are introduced by dehydration of a serine or threonine (to generate a dehydroalanine or dehydrobutyrine) and addition of a cysteine thiol, catalyzed by a dehydratase and a cyclase, respectively[31]. During lanthipeptide biosynthesis, a precursor gene lanA is translated by the ribosome to yield a precursor peptide LanA that consists of an amino-terminal leader peptide and a carboxy-terminal core peptide sequence. The core peptide is post-translationally modified by the lanthionine biosynthetic machinery and other enzymes. It is then proteolytically cleaved from the leader peptide to yield the mature lanthipeptide and exported out of the cell by transporters.

Lanthipeptides usually possess network motifs that enable their mining in spectral networks. These motifs include mass shifts of −18.01 Da (H_2O mass) that correspond to the varying number of dehydrations and mass shifts equal to amino acid masses that correspond to promiscuity in N-terminal leader processing. We formed a spectral network using Networking+ for a subset of 500 *Streptomyces*

https://doi.org/10.1038/s41587-023-01985-4

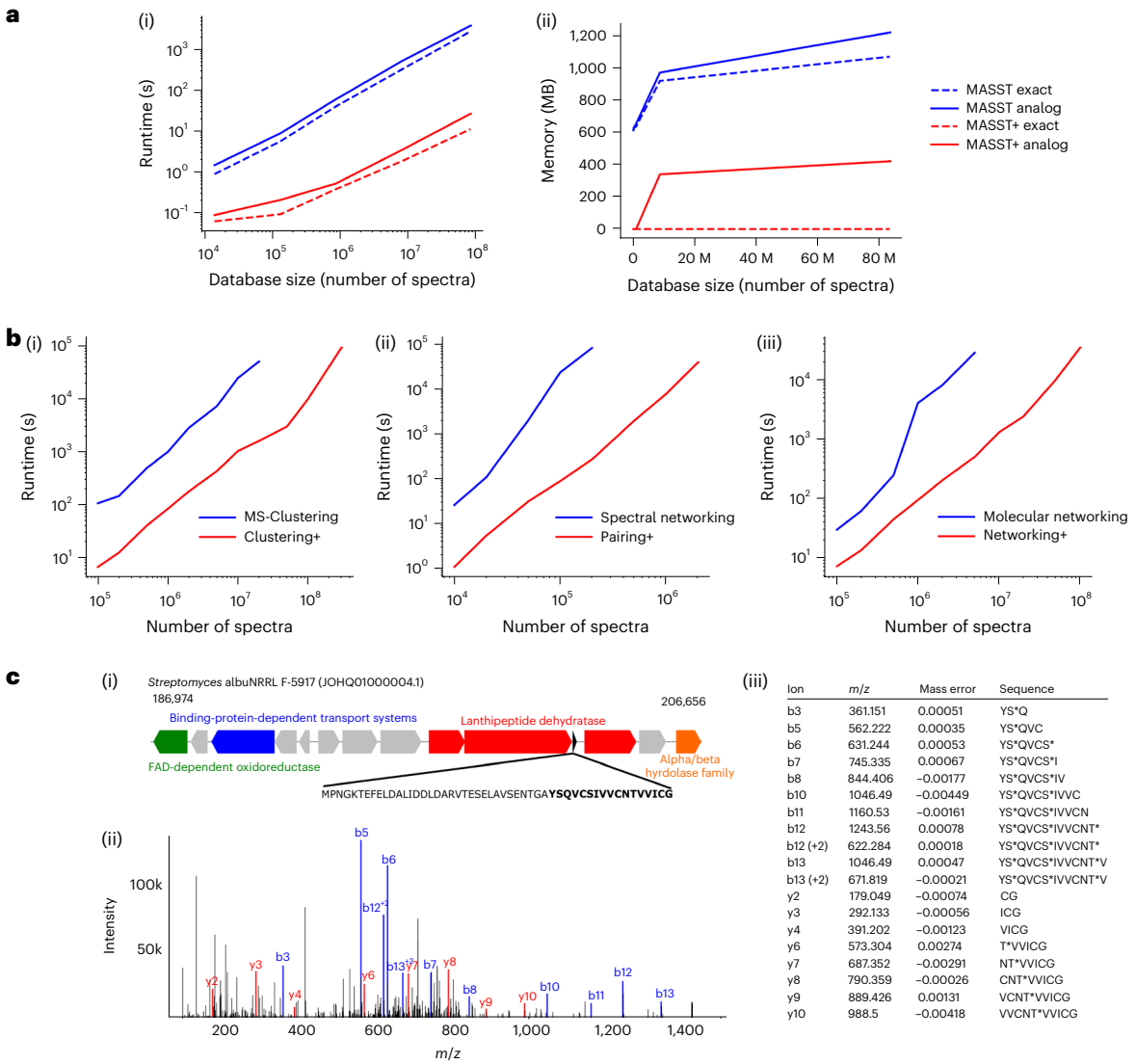

Fig. 2 | MASST+, Clustering+ and Networking+ enable lanthipeptide discovery. a, MASST+ performance. (i) MASST+ is two orders of magnitude faster than MASST in exact and analog search for various database sizes. (ii) MASST+ outperforms MASST in terms of memory efficiency. M, million. **b**, Clustering+ and Networking+ performance. (i) Clustering+ runtime versus MS-Clustering. (ii) Pairing+ runtime versus spectral networking. (iii) Networking+ runtime versus molecular networking. Clustering+, Pairing+ and Networking+ are two orders of magnitude faster than the state-of-the-art methods when processing large datasets. **c**, Lanthipeptides. (i) Biosynthetic gene cluster of CHM-1731. Genes with different functions are highlighted with different colors. (ii) Annotation of peaks in mass spectrum representing CHM-1731. b-ions (prefix fragmentations) are shown in blue, and y-ions (suffix fragmentations) are shown in red. k, thousand. (iii) Mass error of annotations shown in parts per million. Asterisks indicate dehydrated serine/threonine.

cultures with known genomes (Supplementary Table 6). The dataset contained 9,410,802 scans, which were clustered into 354,401 nodes, 6,032 connected components and 1,265,311 edges. Molecular networking crashes on this dataset after 8 days of processing. We further only retained 29,639 nodes that possess the network motif by filtering for edges with mass differences equal to a loss of H_2O, NH_3 or an amino acid mass. Then, we filtered for nodes with long amino acid sequence tags of various lengths using PepNovo[32] (Supplementary Table 7). There were a total of 2,353 nodes with sequence tags of length 12 or longer, and 285 of these nodes were connected to an edge with a mass difference equal to the mass of one H_2O or an amino acid loss. We further inspected these nodes using our in-house software algorithm, Seq2RiPP (https://github.com/mohimanilab/seq2ripp). Given a lanthipeptide precursor, Seq2Ripp generates all molecular structures of all possible candidate molecules by considering different cores and various modifications and then searches the candidate molecular structures against mass spectra using Dereplicator[15]. This strategy identified three known and 14 new lanthipeptides with P values below 1×10^{-15} (Supplementary Table 8). Among them, the precursor of 13 lanthipeptides (76%) overlapped with reports using the genome mining strategy introduced by Walker et al.[33–35]. However, the core peptides predicted were consistent with predictions by Walker et al. for only two lanthipeptides (11%). Note that in contrast to our approach, the strategy used by Walker et al. was based solely on genomics and did not use metabolomics data for identifying the start of the core peptide. This demonstrates that MASST+ and Molecular Networking+ can be used to gain insight into previously uncharacterized molecules. One of the new peptides (CHM-1731 from *Streptomyces albus*) is further described in Fig. 2c.

Discussion

MASST and molecular networking have become powerful strategies for analysis of data based on liquid chromatography coupled with tandem mass spectrometry, with a broad range of users in the research community[9,18,36–41]. However, these tools do not scale to searching and

clustering large spectral repositories with hundreds of millions of spectra. As the size of mass spectral repositories doubles every 2–3 years, the current implementations of MASST and molecular networking will soon not be able to meet the needs of biologists and clinicians. Thus, new solutions are urgently needed.

Recent advances have enabled the determination of molecular formula[42] and chemical class[43,44] for a large portion of spectra in GNPS. Despite these efforts, it is challenging to assign a chemical structure to the majority of spectra in GNPS. MASST+ and Networking+ provide efficient ways to annotate this dark matter by elucidating known molecules and their novel variants in repositories as they grow to billions of mass spectra. MASST+ currently searches query spectra against the clustered GNPS in a few seconds (in comparison with an hour for MASST), enabling instant analysis of the query mass spectrum of interest. Further, MASST+ can search the entire GNPS, which contains hundreds of millions of spectra, in less than 2 h, a task that is currently impossible with MASST. MASST+ can be parallelized by splitting a set of query spectra among several computational nodes or threads. Each thread then can run a separate MASST+ search job that uses the same index stored on disk.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01985-4.

References

1. Kale, N. S. et al. MetaboLights: an analog-access database repository for metabolomics data. *Curr. Protoc. Bioinformatics* **53**, 14–13 (2016).
2. Sud, M. et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
3. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
4. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
5. Courraud, J., Ernst, M., Svane Laursen, S., Hougaard, D. M. & Cohen, A. S. Studying autism using untargeted metabolomics in newborn screening samples. *J. Mol. Neurosci.* **71**, 1378–1393 (2021).
6. Ernst, M. et al. Gestational age-dependent development of the neonatal metabolome. *Pediatr. Res.* **89**, 1396–1404 (2021).
7. Frank, A. M. et al. Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
8. Jarmusch, A. K. et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
9. Quinn, R. A. et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
10. Petras, D. et al. Non-targeted metabolomics enables the prioritization and tracking of anthropogenic pollutants in coastal seawater. *Chemosphere* **271** (2020).
11. Kuo, T.-H., Yang, C.-T., Chang, H.-Y., Hsueh, Y.-P. & Hsu, C.-C. Nematode-trapping fungi produce diverse metabolites during predator–prey interaction. *Metabolites* **10**, 117 (2020).
12. Depke, T., Thöming, J. G., Kordes, A., Häussler, S. & Brönstrup, M. Untargeted LC-MS metabolomics differentiates between virulent and avirulent clinical strains of *Pseudomonas aeruginosa*. *Biomolecules* **10**, 1041 (2020).
13. Eberhard, F. E., Klimpel, S., Guarneri, A. A. & Tobias, N. J. Metabolites as predictive biomarkers for *Trypanosoma cruzi* exposure in triatomine bugs. *Comput. Struct. Biotechnol. J.* **19**, 3051–3057 (2021).
14. Lybbert, A. C., Williams, J. L., Raghuvanshi, R., Jones, A. D. & Quinn, R. A. Mining public mass spectrometry data to characterize the diversity and ubiquity of *P. aeruginosa* specialized metabolites. *Metabolites* **10**, 445 (2020).
15. Mohimani, H. et al. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
16. Frank, A. M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**, 587–591 (2011).
17. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proc. Natl Acad. Sci. USA* **104**, 6140–6145 (2007).
18. Ramos, A. E. F., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Prod. Rep.* **36**, 960–980 (2019).
19. Kalinski, J.-C. J. et al. Molecular networking reveals two distinct chemotypes in pyrroloiminoquinone-producing *Tsitsikamma favus* sponges. *Marine Drugs* **17**, 60 (2019).
20. Raheem, D. J., Tawfike, A. F., Abdelmohsen, U. R., Edrada-Ebel, R. & Fitzsimmons-Thoss, V. Application of metabolomics and molecular networking in investigating the chemical profile and antitrypanosomal activity of British bluebells (*Hyacinthoides non-scripta*). *Sci. Rep.* **9**, 2547 (2019).
21. Trautman, E. P., Healy, A. R., Shine, E. E., Herzon, S. B. & Crawford, J. M. Domain-targeted metabolomics delineates the heterocycle assembly steps of colibactin biosynthesis. *J. Am. Chem. Soc.* **139**, 4195–4201 (2017).
22. Vizcaino, M. I., Engel, P., Trautman, E. & Crawford, J. M. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *J. Am. Chem. Soc.* **136**, 9244–9247 (2014).
23. Nguyen, D. D. et al. Indexing the *Pseudomonas* specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2**, 16197 (2016).
24. Woo, S., Kang, K. B., Kim, J. & Sung, S. H. Molecular networking reveals the chemical diversity of selaginellin derivatives, natural phosphodiesterase-4 inhibitors from *Selaginella tamariscina*. *J. Nat. Prod.* **82**, 1820–1830 (2019).
25. Reginaldo, F. P. S. et al. Molecular networking discloses the chemical diversity of flavonoids and selaginellins in *Selaginella convoluta*. *Planta Med.* **87**, 113–123 (2021).
26. Bittremieux, W. et al. Analog access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.15.490691 (2022).
27. Schnell, N. et al. Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. *Nature* **333**, 276–278 (1988).
28. Mohr, K. I. et al. Pinensins: the first antifungal lantibiotics. *Angew. Chem. Int. Ed.* **54**, 11254–11258 (2015).
29. Férir, G. et al. The lantibiotic peptide labyrinthopeptin A1 demonstrates broad anti-HIV and anti-HSV activity with potential for microbicidal applications. *PLoS ONE* **8**, e64010 (2013).
30. Iorio, M. et al. A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity. *ACS Chem. Biol.* **9**, 398–404 (2014).
31. Arnison, P. G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).

32. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
33. Walker, M. C. et al. Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* **21**, 387 (2020).
34. Kodani, S., Lodato, M. A., Durrant, M. C., Picart, F. & Willey, J. M. SapT, a lanthionine-containing peptide involved in aerial hyphae formation in the streptomycetes. *Mol. Microbiol.* **58**, 1368–1380 (2005).
35. Ueda, K. et al. AmfS, an extracellular peptidic morphogen in *Streptomyces griseus*. *J. Bacteriol.* **184**, 1488–1492 (2002).
36. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
37. Aron, A. T. et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
38. Nothias, L.-F. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
39. van Der Hooft, J. J. et al. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).
40. Yang, J. Y. et al. Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).
41. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752 (2012).
42. Ludwig, M., Fleischauer, M., Dührkop, K., Hoffmann, M. A. & Böcker, S. De novo molecular formula annotation and structure elucidation using SIRIUS 4. *Methods Mol. Biol.* **2104**, 185–207 (2020).
43. Dührkop, K. et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol* **39**, 462–471 (2021).
44. Mohimani, H., Kim, S. and Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *J. Proteome Res.* **12**, 1560–1568 (2013).

Methods

Overview of MASST algorithm

In exact search mode, MASST performs the exact search by retrieving the spectra in the database that have the same precursor mass as the query and computing a SharedScore between each retrieved spectrum and the query. Analog search is conducted by retrieving all spectra within a large precursor mass tolerance (for example, 300 Da) of the query precursor mass and computing the ShiftedScore (Fig. 1a(ii)). To compute these scores, MASST iterates over all the peaks in the query spectrum; for each peak, it explores whether a peak with similar or shifted m/z is present in each database spectrum. Whenever such a peak is present, MASST increments the score between the query and that database spectrum by the product of the intensity of peaks in the query and the database spectrum.

MASST+ exact search

Given a query spectrum, MASST+ efficiently searches a database of reference spectra to find similar spectra using the fast dot-product algorithm (Fig. 1b). For each precursor mass M and each peak mass p, a list of indices of all spectra with precursor mass M and peaks with mass within a tolerance threshold of p are stored, along with the intensities of the peaks. In the case of exact search, given a query spectrum with precursor mass M, MASST+ iterates through the peaks in the query spectrum and retrieves lists corresponding to the peaks and M. As each list is stored on disk, it can be retrieved in $O(1)$ time. The SharedScore is then calculated by multiplying and adding up the intensity of each peak in the query spectrum and reference spectra (Fig. 1b(iv)).

MASST+ analog search

In the case of analog search, MASST+ uses a large precursor mass tolerance (for example, 300 Da) and computes a ShiftedScore (Fig. 1a(ii)). The ShiftedScore takes into account both shared and Δ-shifted peaks. In analog mode, all reference spectra are processed into lists as in MASST+ exact search. Given a query spectrum, MASST+ analog search iterates through each peak p in the query spectrum with precursor mass M and scan lists (M', p') where either $p = p'$ (shared peak) or $M - p = M' - p'$ (shifted peak). The ShiftedScore between the query and each reference spectrum is calculated by multiplying and adding the intensities of shared and shifted peaks in the two spectra (Supplementary Fig. 4). Note that MASST+ analog search is a variant of the fast dot-product algorithm (Fig. 1b), as both methods rely on similarly structured index tables. Rather than just retrieving one list for each query spectrum peak, however, MASST+ analog search retrieves two lists.

MASST+ indexing

To handle continuous values of peak masses, we bin peak masses into discrete values. Depending on the bin size and product mass tolerance, one or more bins must be retrieved when processing each query peak during search. We use a bin size of 0.01 Da, which can handle both high-resolution (0.01 Da accuracy) and low-resolution (0.5 Da accuracy) data.

Overview of molecular networking

To find structurally related families of small molecules, the existing molecular networking method first clusters spectra from identical molecules using MS-Clustering[16]. It then connects clusters of related molecules using spectral networking[17]. MS-Clustering puts two spectra in the same cluster if their precursor mass difference is below a threshold (usually 2 Da) and their cosine dot product (a normalized SharedScore) is above a certain threshold (usually 0.7). Then, for each cluster, a consensus spectrum is constructed using the approach introduced by Frank et al.[16]. In spectral networking, two consensus spectra are connected to each other if the shared-shifted cosine score (normalized ShiftedScore) is above a threshold (the default is 0.7).

Networking+ algorithm

Networking+ consists of two modules, Clustering+ and Pairing+. Clustering+ is implemented using a greedy procedure (Supplementary Fig. 12). Given a dataset of N spectra, Clustering+ creates an initial cluster whose center is set to be the first spectrum in the dataset. Then, in the following $N - 1$ iterations, the similarity score between each remaining spectrum and all the existing cluster centers is calculated. To efficiently calculate the similarity score between a spectrum and all cluster centers, an indexing table similar to MASST+ exact search is constructed and iteratively updated. For each precursor mass M and peak mass p, the indexing table stores the list of all clusters that have centers with a specific precursor mass M and a peak mass p. At each iteration, whenever the highest score between the spectrum and cluster centers is greater than a threshold (the default is 0.7), the spectrum is added to the highest-scoring cluster, and the center of the cluster is updated. If the highest score is below the threshold, then a new cluster is created, and the current spectrum is set as the center of the cluster. This procedure continues until all the spectra have been clustered.

To maintain efficiency, whenever a new spectrum is added, the center is updated only when the cluster size doubles (for example, after the addition of the first, second, fourth, eighth and 16th spectrum to the cluster). Similar to the method of Frank et al.[16], the center is computed by adding peaks that are present in the majority of the members of the cluster. The intensity of each peak is calculated as the average of the intensity of the corresponding peaks in members. All spectra are initially normalized.

Pairing+ computes a score similar to that used in MASST+ analog search (Supplementary Fig. 4), which accounts for Δ-shifted and shared peaks for all pairs of input spectra (for example, cluster centers from Clustering+). To do this, it constructs an indexing table similar to that used in MASST+ analog search. Then, the table is used to efficiently compute the scores between all pairs of spectra (Supplementary Fig. 13).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets analyzed are available at gnps.ucsd.edu. Accession codes related to the lanthipeptides part of the study are MSV000090476, MSV000090473, MSV000090472, MSV000090471, MSV000090457, MSV000089818, MSV000089817, MSV000089816, MSV000089815, MSV000089813, MSV000088816, MSV000088801, MSV000088800, MSV000088764 and MSV000088763. For comparing MASST+ and Networking+ against previous state-of-the-art tools, datasets MSV000078787, clustered GNPS, and unclustered GNPS were used. The accession codes for clustered GNPS and unclustered GNPS are available in Supplementary Data 1.

Code availability

MASST+ and Networking+ are available at https://github.com/mohimanilab/MASSTplus. Other custom software used in this work includes Seq2Ripp (https://github.com/mohimanilab/seq2ripp), PepNovo (https://github.com/jmchilton/pepnovo) and Dereplicator (https://ccms-ucsd.github.io/GNPSDocumentation/dereplicator/).

Acknowledgements

The work of T.M.Y., M.M., Y.L., B.B. and H.M. was supported by National Institutes of Health New Innovator Award DP2GM137413, US Department of Energy award DE-SC0021340, National Science Foundation award DBI-2117640 and National Institute of General Medicine Sciences of the National Institutes of Health award R43GM150301 (B.B. only). The work of P.C.D. and M.W. was supported by R03OD034493, U24DK133658 and R01GM107550 (P.C.D. only).

Author contributions

M.M., T.M.Y., Y.L., M.G., L.L. and A.B. implemented the algorithms. M.M., T.M.Y. and Y.L. performed the analysis. M.W. designed and implemented the GNPS web service for MASST+. B.B., P.C.D. and H.M. designed and directed the work. M.M. and H.M. wrote the manuscript, and all authors contributed to its revision.

Competing interests

H.M. and B.B. are cofounders of and have equity interests in Chemia.ai, LLC. P.C.D. is an advisor of and holds equity in Cybele, consulted for MSD Animal Health in 2023 and is a cofounder of, holds equity in and is scientific advisor for Ometa Labs, Arome and Enveda with prior approval by the University of California San Diego. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01985-4.

Correspondence and requests for materials should be addressed to Hosein Mohimani.

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

nature portfolio

Corresponding author(s): Hosein Mohimani

Last updated by author(s): Aug 30, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	MASST+ and Networking+ presented in the paper are available at https://github.com/mohimanilab/MASSTplus. Additionally, Seq2Ripp (https://github.com/mohimanilab/seq2ripp), PepNovo (https://github.com/jmchilton/pepnovo), and Dereplicator (https://ccms-ucsd.github.io/GNPSDocumentation/dereplicator/) were used for the data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets analyzed are available at gnps.ucsd.edu. Accession codes related to Lanthepeptide portion of manuscript are MSV000090476 ,MSV000090473 ,MSV000090472 ,MSV000090471 ,MSV000090457 ,MSV000089818 ,MSV000089817 ,MSV000089816 ,MSV000089815 ,MSV0000

89813 ,MSV000088816 ,MSV000088801 ,MSV000088800 ,MSV000088764 ,MSV000088763 . For comparing MASST+ and Networking+ against previous state of the art, datasets MSV000078787 , Clustered GNPS, and Unclustered GNPS were used. The Accession codes for Clustered GNPS and Unclustered GNPS are in Supplementary_Data_1.xlsx

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	MASST+ and Networking+ are general software tools designed to process tandem mass spectra. As part of this project, no data was collected.
Data exclusions	No data was excluded in any of the experiments.
Replication	MASST+ and Networking+ are general software tools designed to process tandem mass spectra. As part of this project, no data was collected.
Randomization	MASST+ and Networking+ are general software tools designed to process tandem mass spectra. As part of this project, no data was collected.
Blinding	MASST+ and Networking+ are general software tools designed to process tandem mass spectra. As part of this project, no data was collected.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging