Fast Mass Spectrometry Search and Clustering of Untargeted Metabolomics Data

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

[1]Computational Biology Department, School of Computer Science Carnegie Mellon University

[2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego

3 Department of Pharmacology and Pediatrics, University of California San Diego

4 Computer Science and Engineering, University of California Riverside

Abstract: The throughput rate of mass spectrometers and the size of publicly available metabolomics data are growing rapidly. Illuminating the molecules present in untargeted mass spectrometry data that cannot be identified by existing approaches (the dark matter of metabolomics) remains a challenging task. In the past decade, molecular networking and MASST were introduced to organize and query untargeted mass spectrometry data. While useful for single datasets, these methods cannot scale to searching and clustering billions of mass spectral data in metabolomics repositories, e.g. the Global Natural Product Social (GNPS) molecular networking infrastructure. To address this shortcoming, we developed an efficient strategy for the computation of dot-product between mass spectra, where the relevant information from spectral datasets is stored in an indexing table. Based on this strategy, we designed MASST+ and Networking+, scalable approaches for querying and clustering mass spectra that can process datasets that are up to three orders of magnitude larger than the state-of-the-art. For example, MASST+ can query against 717 million spectra from the GNPS public data in less than an hour and Networking+ is able to map the chemical diversity of all GNPS public data in days.

Introduction

During the past decade, the size of mass spectral data collected in the fields of natural products, exposomics, and metabolomics has grown exponentially[9,16,18]. In accordance with the advances in mass spectrometry technology, multiple computational methods were developed for analyzing this massive data. Recently Mass Spectrometry Search Tool (MASST) was introduced as a search engine for finding analogs of a query spectrum in mass spectrometry repositories[19]. MASST has demonstrated utility in the annotation of a wide variety of unidentified metabolites, including clinically important molecules in patient cohorts[15,3,6], toxins/pesticides in environmental samples[14], fungal metabolites[10], and metabolites from pathogenic microorganisms[4,11,5]. Moreover, molecular networking was introduced for clustering spectral datasets into families of related molecules[29,28]. Molecular Networking has yielded a systematic view of the chemical space in different ecosystems and helped determine the structure of many compounds[20,21,22,23,24,25,27,26].

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[28] to cluster identical spectra by calculating a dot-product score (ExactScore, Figure 1a) between the spectra. Then Spectral Networking[29] is used to calculate a dot product score that accounts for peaks that are shared or shifted (ShiftedScore, Figure 1b) between all pairs of clusters in order 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 three days. Currently, molecular networking analysis of a million spectra takes a few hours, while 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 paper, 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 (Figure 2). Since mass spectra are sparse, only a small fraction of spectra/peaks are retrieved for each query. The ability to leverage this sparsity requires only a small fraction of the compute used by naive scoring methods because the vast majority of the MS/MS spectra in the index are never touched during the query process. By integrating this indexing approach into the scoring subroutines of MASST and Molecular Networking, we develop two new computational tools, MASST+ and Networking+, that are two to three orders of magnitude faster than state-of-the-art on large datasets. Further, this indexing approach supports on-line growth, that is, the insertion of new spectra without the need for recalculation from scratch. The 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).

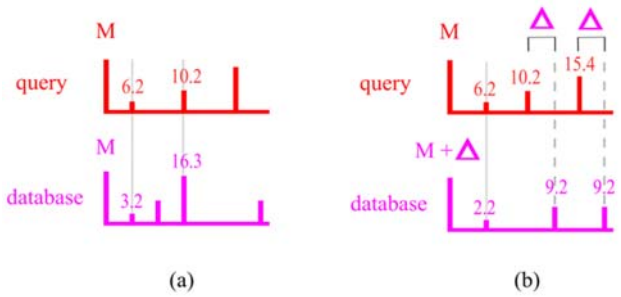

Figure 1. Similarity score. (a) In case of exact search, MASST searches a query spectrum against all database spectra with similar precursor masses, and computes the ExactScore, a sum of multiplications between intensities of peaks shared by the query and database spectrum (shown in solid grey). In this case the score is 6.2 * 3.2 + 10.2 * 16.3 = 186.1. (b) In the case of analog search, MASST searches the query spectrum against all database spectra within a specific precursor mass range (e.g. 300 Da) and computes the ShiftedScore, a sum of multiplications between intensities of peaks that are shared and Δ-shifted between query and database spectrum. Here there is one shared (solid grey) and two Δ-shifted (dashed grey) peaks, yielding a total score of 6.2 * 2.2 + 10.2 * 9.2 + 15.4 * 9.2 = 249.16. Δ denotes the precursor mass difference between query and database spectra.

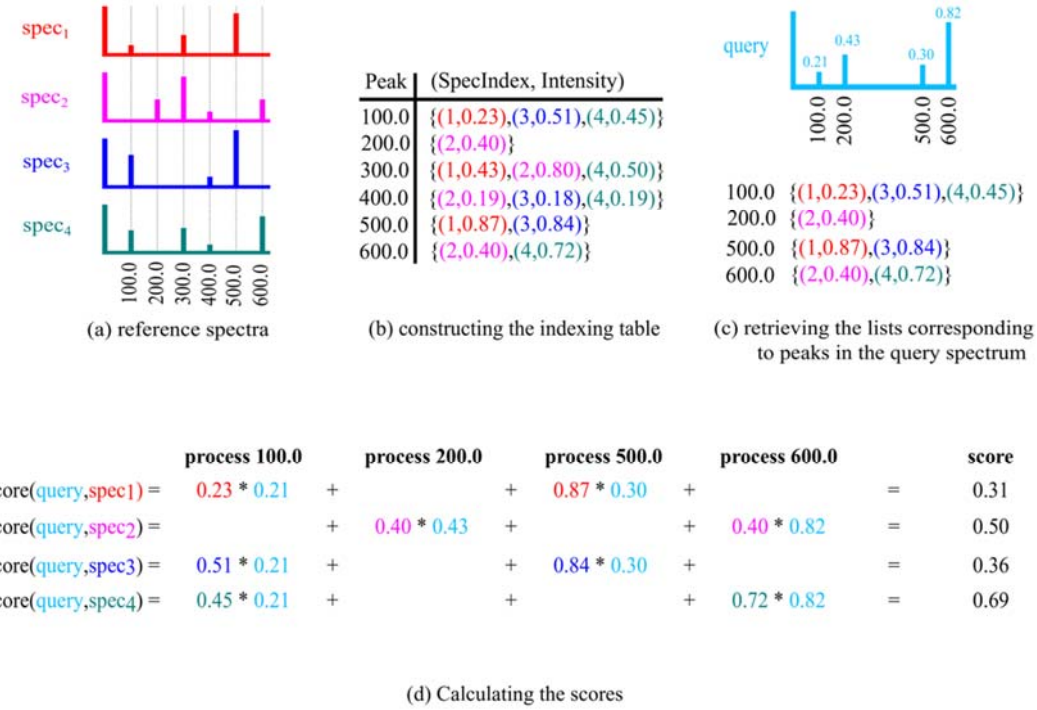

Figure 2: Fast Dot Product. (a) Given a database of spectra the fast dot procedure starts with (b) constructing an indexing table, where each row corresponds to a fragment peak mass, and contains a list of tuples of spectra indices that contain the peak, along with the intensity of the peak in these spectra. (c) Given a query spectrum, all lists corresponding to peaks present in the query are retrieved. Then, (d) 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.

Results.

Outline of MASST+ algorithm. 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 which 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 case of exact search, MASST+ iterates through the peaks in the query spectrum and retrieves the lists associated with a query peak and query's precursor mass. The ExactScore is calculated by multiplying and adding up the intensity of each peak in query spectrum and reference spectra (Figure 2). In case of analog search (Supplementary Fig. 4), MASST+ uses a much larger precursor mass tolerance (e. g. 300Da) and computes ShiftedScore that takes into account both shared and Δ-shifted peaks (peaks in reference spectra that are Δ Da larger than peaks in query), where Δ is the mass difference between the precursor of query and reference spectra (Figure 3).

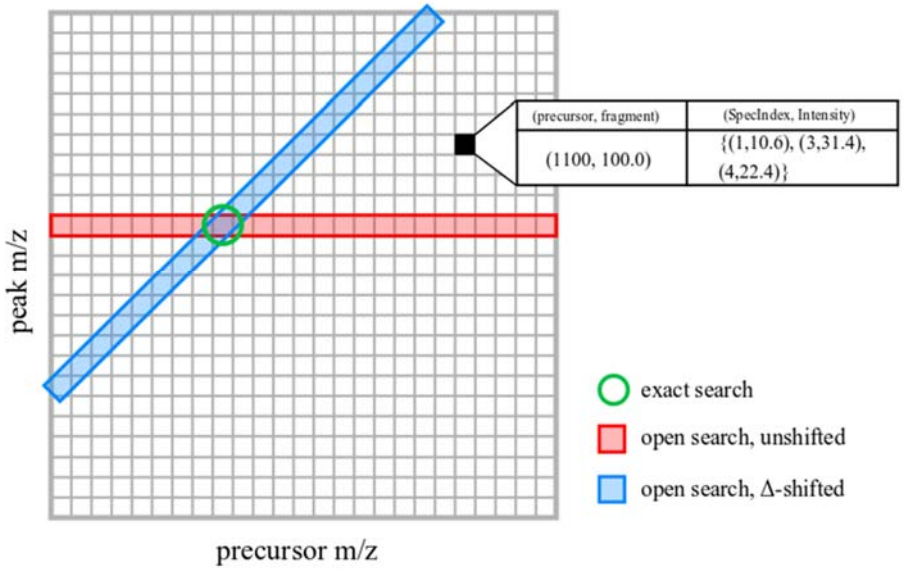

Figure 3. Fast Dot Product Indexing. The fast dot product indexing table corresponds to a two-dimensional grid, with precursor mass on the x-axis and peak mass on the y-axis. Each database peak is inserted into a list corresponding to a specific location in the grid, determined by the peak mass and the 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 the shared peaks) and blue cells (corresponding to Δ-shifted peaks) are retrieved.

Outline of Networking+ algorithm. 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.

Benchmarking MASST+. We have benchmarked MASST+ (Table 1) on various GNPS datasets

including MSV000078787 dataset collected on *Streptomyces* cultures (5,433 spectra), clustered GNPS (83,131,248 spectra), and entire GNPS (717,395,473 spectra). While MASST and MASST+ report identical hits, MASST+ is two orders of magnitude faster and more memory efficient (Table 1). For small data sets we only get a 3-fold increase in speed. This becomes magnified when the data set that is searched becomes larger. In case of the clustered GNPS, MASST+ performs analog search in 15 seconds while MASST takes 49 min, a 196-fold increase. In case of the entire GNPS, MASST+ performs analog search in under two hours on average, while MASST search does not finish within three days on the GNPS server making it practically not possible to routinely perform such a search.

Figure 4 illustrates the runtime and memory consumption of MASST+ in exact and analog mode for various subsets of the clustered GNPS. Supplementary Fig. 5 illustrates that indexing time and memory consumption grows linearly with the size of datasets and Supplementary Fig. 6 shows indexing time increases for larger values of peak mass tolerance. MASST+ takes eight hours of compute time and eight gigabytes of memory to index ~83 million spectra from the clustered GNPS and 72 hours of compute time and 9 gigabytes of memory to index 717 million spectra contained in GNPS. Supplementary Fig. 7 breaks down MASST+ runtime into two different steps, loading peaks lists and computing dot product, for various numbers of query spectra. Loading peak lists consumes about half of the total runtime when the number of query spectra is greater than 100.

Method	Mode	Dataset (Size)	Search Time	Search Memory	# IDs
MASST	exact	MSV000078787 (195K)	0.41 sec	50Mb	10
MASST+	**exact**	**MSV000078787 (195K)**	**0.13 sec**	**0Kb**	**10**
MASST	analog	MSV000078787 (195K)	0.61 sec	40Mb	16
MASST+	**analog**	**MSV000078787 (195K)**	**0.14 sec**	**0Kb**	**16**
MASST	exact	Clustered GNPS (83M)	34 min	952Mb	49
MASST+	**exact**	**Clustered GNPS (83M)**	**8.6 sec**	**24Mb**	**49**
MASST	analog	Clustered GNPS (83M)	49 min	1.1Gb	2,175
MASST+	**analog**	**Clustered GNPS (83M)**	**15.0 sec**	**159Mb**	**2,175**
MASST	exact	Entire GNPS (717M)	N/A	N/A	N/A
MASST+	**exact**	**Entire GNPS (717M)**	**43 min**	**21Gb**	**171**
MASST	analog	Entire GNPS (717M)	N/A	N/A	N/A
MASST+	**analog**	**Entire GNPS (717M)**	**115 min**	**35Gb**	**265,958**

Table 1. Benchmarking MASST+ search. MSV000078787 (~195K spectra), clustered GNPS (~83M spectra), or entire GNPS (~717M spectra) are used as the reference database. Search time, search memory consumption, and number of identifications resulting from searching queries are shown. For MSV000078787, clustered GNPS, and entire GNPS, MASST+ is two orders of magnitude faster than MASST while consuming the same or less memory. MASST search did not yield results for entire GNPS in a reasonable time frame (three days threshold). MASST+ reports are identical to MASST.

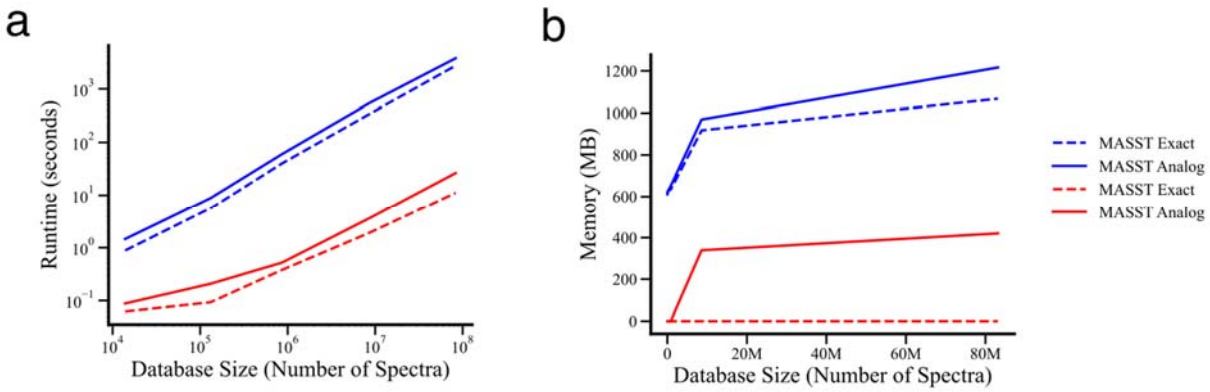

Figure 4. (a) MASST+ is two orders of magnitudes faster than MASST in exact and analog search for various database sizes. (b) MASST+ outperforms MASST in memory efficiency.

Benchmarking networking+. Figure 5, Table 2 and Supplementary Tables 1-3 benchmark Networking+ against molecular networking on various data sizes for which runtime is less than 24 hours. In 24 hours Clustering+ can process 300 million spectra on a single CPU, while MS-Clustering can process 20 million spectra. Moreover, in this timeline, Pairing+ can process 2 million spectra, while spectral networking can handle 0.2 million spectra. Clustering+ and Pairing+ are two orders of magnitude faster than their counterparts, MS-Clustering[29] and Spectral Networking[28]. The clusters and networks reported by Clustering+ and Pairing+ are identical to MS-Clustering and spectral networks. As previously noted in Bittremieux et al.[43], it was not 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.

Method	Dataset (Size)	clustering time	Clustering memory	#clusters	networking time	Networking memory
Molecular Networking	MSV000078787 (219,915)	321 sec	662Mb	5,288	8 sec	1224Kb
Networking +	**MSV000078787 (219,915)**	**27 sec**	**992Kb**	**5,288**	**.25 sec**	**996Kb**
Molecular Networking	Entire GNPS	N/A	N/A	N/A	N/A	N/A
Networking +	**Entire GNPS**	**25 hours**	**93Gb**	**8,453,822**	**97 hours**	**23Gb**

Table 2. Benchmarking Molecular Networking and Networking+. MSV000078787 (~195K spectra), entire GNPS (~717M spectra) are used as spectral datasets. Clustering time, clustering memory, number of clusters, networking time and networking memory are shown. Networking+ clusters and networks the entire GNPS in 25 and 97 hours respectively while Molecular Networking does not complete clustering in 14 days.

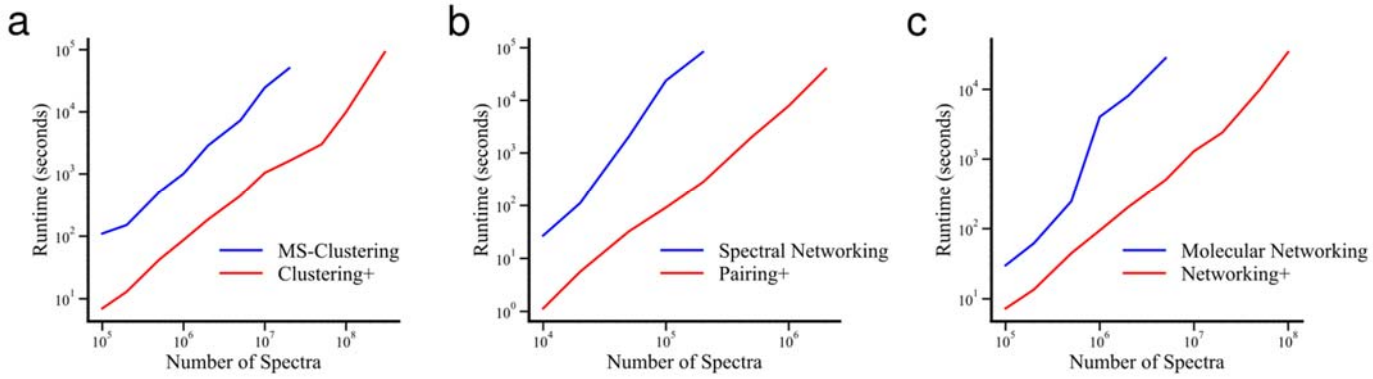

Figure 5. a) Clustering+ runtime versus MS-Clustering. b) Pairing+ runtime versus spectral networking. c) Networking+ runtime versus Molecular Networking. Clustering+, Pairing+ and Networking+ are two order of magnitudes faster than the state-of-the-art methods when processing large datasets.

Networking the entire GNPS. We clustered the entire GNPS (717 million scans) using Clustering+ and formed the 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 4,948,146 connected components in the network, 98% (4,849,047 components) consist of a single node, while 1.5%, 0.3%, 0.2% and 0.02% (74530, 13957, 9239, and 1152 components) had 2, 3, and 4-9 and 10+ nodes (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 307,709 clusters consisting of 20 or more spectra, for 18% (54,518 clusters) all spectra came from a single MassIVE dataset, while for 13% and 69% (39,428 and 213,763 clusters) spectra came from 2 or 3+ MassIVE datasets (Supplementary Fig. 10). About 61 percent of the clusters with precursor mass between 0 and 400 Daltons consisted of only two GNPS spectra whereas less than half the clusters with precursor mass above 400 Daltons consisted of only two GNPS spectra (Supplementary Fig. 11). Networking+ took 5 days to finish this task on 1 CPU. Currently, this task is not feasible using existing approaches.

Applying Networking+ for Identification of novel lanthipeptides. 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 suffer from various limitations, especially at repository scale. Lanthipeptides are a biologically important class of natural products that include antibiotics[30], antifungals[31], antivirals[32], and antinociceptives[33]. 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[34]. During lanthipeptide biosynthesis, a precursor gene lanA is translated by the ribosome to yield a precursor peptide LanA that consists of a N-terminal leader peptide and a C-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 mining them in spectral networks. These motifs include mass shifts of -18.01Da (H2O 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 the spectral network using Networking+ for a subset of 500 *Streptomyces* cultures with known genomes (Supplementary Table 4). The dataset contains 9,410,802 scans, which are clustered into 354,401 nodes, 6,032 connected components, and 1,265,311 edges. Currently, Molecular Networking crashes on this dataset after eight 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 H2O, NH3, or an amino acid mass. Then we filtered for nodes with long amino acid sequence tags of various lengths using PepNovo[35] (Supplementary Table 5). There are a total of 2,353 nodes with sequence tags of length 12 or longer, and 285 of these nodes are connected to an edge with a mass difference equal to the mass of one H2O 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[36]. This strategy identified three known and 14 novel lanthipeptides with p-values below 1e-15 (Table 3). Among them, the precursor of 13 lanthipeptides (76%) overlaps with reports by the genome mining strategy introduced by Walker et al.[40]. However, only for two lanthipeptides, the core peptides predicted are consistent with predictions from Walker et al. (11%). Note that in contrast to our approach, Walker et al. is based solely on genomics, and it does not use metabolomics data for identifying the start of core peptide. This demonstrates that MASST+ and Molecular Networking+ can be used to gain insight into previously uncharacterized molecules. One of the novel peptides (CHM-1731 from *Streptomyces albus*) is further described in Figure 6.

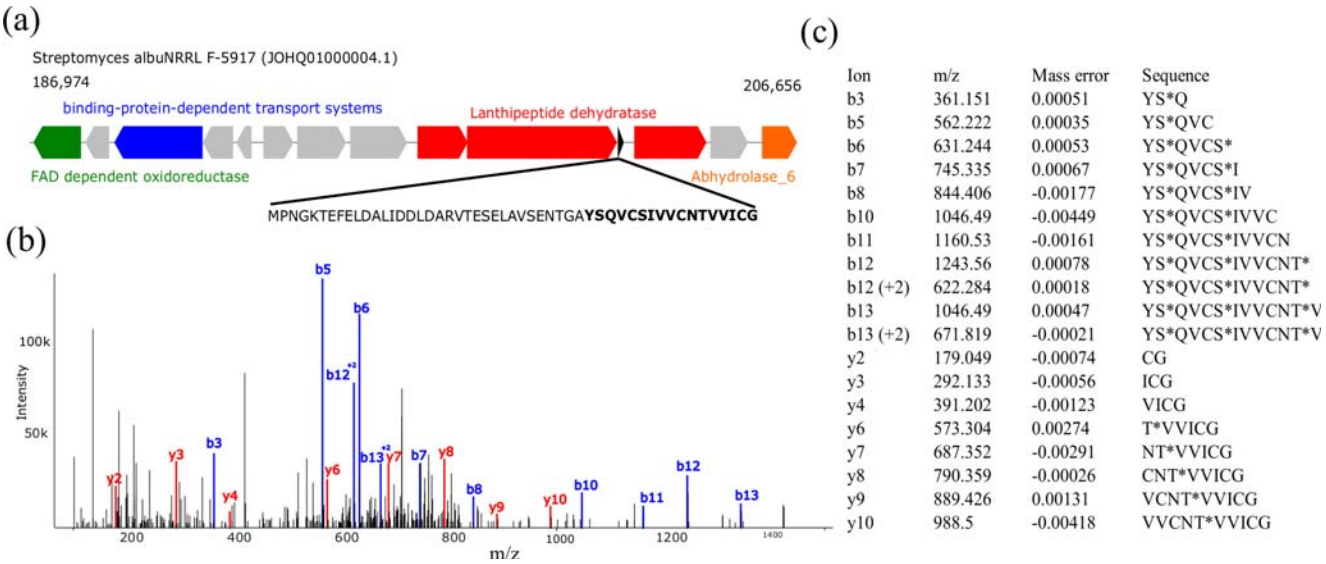

Figure 6. (a) Biosynthetic gene cluster of CHM-1731. Genes with different functions are highlighted with different colors. (b) Annotation of peaks in mass spectrum representing CHM-1731. B-ions (prefix fragmentations) are shown in blue, and y-ions (suffix fragmentations) are shown in red. (c) Mass error of annotations are shown in parts per million (ppm). Stars stand for dehydrated serine / threonine.

Organism	name	Sequence	score	p-value	Mass	Walker et al.	reference
Streptomyces rimosus NRRL WC-3904	CHM-1793	DT-18GHCS-18GVCT-18VLVCT-18VAVC	21	2.50E-36	1793.77	YN	This paper

Streptomyces albus NRRL F-5917	CHM-1731	YS-18QVCS-18IVVCNT-18VVICG	19	5.80E-33	1731.81	YN	This paper
Streptomyces lavenduligriseus NRRL ISP-5487	SapT	YT-18QGCS-18GLCT-18IVICAT-18VVICG	18	1.40E-32	2030.95	YN	Kodani et al.[41]
Streptomyces species NRRL S-340	CHM-1911	S-18TAGCS-18GLCT-18IIVCAT-18VVICA	17	5.20E-31	1911.91	YN	This paper
Streptomyces pathocidini NRRL B-24287	CHM-2168	IT-18S-18IS-18YCT-18PGCT-18SDGGGS-18GCS-18HCC	16	1.60E-26	2168.76	YY	This paper
Streptomyces moroccanus NRRL B-24548	CHM-2182	IT-18S-18IS-18YCT-18PGCT-18SEGGGS-18GCS-18HCC	15	2.00E-25	2182.78	YY	This paper
Streptomyces cinerochromogenes NBRC 13822	CHM-1974	YT-18EGCS-18GLCT-18ILVCAT-18VVIC	13	9.10E-24	1974.91	NN	This paper
Streptomyces hygroscopicus NRRL ISP-5087	CHM-1354	MT-18QVCPVT-18SWHC	13	3.60E-23	1354.56	YN	This paper
Streptomyces rimosus NRRL WC-3874	CHM-1831	PSRSSSPGSFPPGST-18PS-18APS-18	14	1.60E-21	1831.85	NN	This paper
Streptomyces albus NBRC 13041	CHM-1775	YS-18QVCS-18IVICNT-18VVICS	11	5.50E-20	1775.84	NN	This paper
Streptomyces kanamyceticus NBRC 13414	CHM-1748	IS-18GEES-18CFRT-18CT-18TCS-18LC	12	3.40E-19	1748.68	YN	This paper
Streptomyces sulphureus NRRL B-2195	CHM-2229	TEGGGGDS-18SGCS-18GVCT-18IVVCT-18VIVC	9	1.10E-17	2229.95	YN	This paper
Streptomyces anulatus NBRC 12853	AmfS	T-18GS-18QVS-18LLVCEYSS-18LSVVLCTP	11	2.10E-17	2212.09	YN	Ueda et al.[42]
Streptomyces anulatus NBRC 13369	CHM-1669	C-34LPEPFP+16TATT-18RVGCD	11	9.50E-17	1669.78	YN	This paper
Streptomyces paludis JCM 33019	CHM-1635	S-18GEES-18CFRT-18CT-18T-18CSLC	11	2.30E-16	1635.59	YN	This paper
Streptomyces anulatus NBRC 12861	CHM-2433	CRPPSASLCIT-18SDRS-18S-18TGRYLSM	11	3.10E-16	2433.14	NN	This paper
Streptomyces brasiliensis NBRC 101283	Amfs analog	TGS-18QVS-18VLVCEYS-18S-18LSVVLCTP	11	7.10E-16	2198.08	YN	Ueda et al.[42]

Table 3. Novel and known lanthipeptides discovered by network motif mining. The producer organism, name, sequence, Dereplicator score, and p-value, mass and references are shown. Moreover, it is also indicated whether the precursor genes and core peptides are identified by Walker et al. YY means both precursor gene and core peptide are predicted by Walker et al. YN means the precursor gene is predicted by Walker et al., but the core peptide is inconsistent. NN means the precursor gene is not predicted by Walker et al.

Discussion.

The mass spectrometry search tool (MASST) and molecular networking have become powerful strategies to analyze LC-MS/MS based data to a broad range of users in the research community[2,13,15,17,37,38,39]. However, these tools can not scale to searching and clustering large spectral repositories with hundreds of millions of spectra. As the size of mass spectral repositories doubles every two to three years, the current implementation of MASST and Molecular Networking will soon

not be able to meet the needs of biologists and clinicians and thus new solutions are urgently needed.

Recent advances have enabled the determination of molecular formula[44] and chemical class[45] 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 to an hour for MASST), hence 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 two hours, a task that is currently impossible with MASST. MASST+ can be parallelized by splitting a set of query spectra among several computational nodes/threads. Each thread then can run a separate MASST+ search job that utilizes the same index stored on disk.

Methods

Overview of MASST algorithm. In exact search mode, MASST performs exact search by retrieving the spectra in the database that have the same precursor mass as the query and computing SharedScore between each retrieved spectrum and the query. Analog search is conducted by retrieving all spectra within a large precursor mass tolerance (e.g. 300 Da) of the query precursor mass, and computing the ShiftedScore (Figure 1). To compute these scores, MASST iterates over all the peaks in the query spectrum, and 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 by using the fast dot product algorithm (Figure 2). For each precursor mass M and each peak mass p , a list of indices of all spectra with precursor M and peak within a tolerance threshold of p are stored, along with intensity of peaks. In case of exact search, given a query spectrum with precursor mass M, MASST+ iterates through the peaks in the query spectrum and retrieves the lists corresponding to the peaks and precursor mass M. As each list is stored on disk, each list 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 (Figure 1).

MASST+ analog search. In the case of analog search, MASST+ uses a large precursor mass tolerance (e. g. 300Da) and computes ShiftedScore (Figure 1). ShiftedScore takes into account both shared and Δ-shifted peaks, where Δ is the mass difference between the query and each reference spectrum. 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 □ in the query spectrum with precursor mass □, and scans lists (□′,□′) where either □=□′(shared peak) or □−□=□′−□′ (shifted peak). The ShiftedScore between the query and each reference spectrum is calculated by multiplying and adding up the intensity 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 (Figure 2) 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 0.01Da, which can handle both high-resolution (0.01Da accuracy) and low-resolution (0.5Da accuracy) data.

Overview of Molecular Networking. In order to find structurally related families of small molecules, the existing molecular networking method first clusters spectra from identical molecules using MS-Clustering[29]. It then connects clusters of related molecules using spectral networking[28]. 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[28]. In spectral networking, two consensus spectra are connected to each other if the shared-shifted cosine score (normalized ShiftedScore) is above a threshold (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 spectra and all the existing cluster centers is calculated. To efficiently calculate the similarity score between the 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 □. At each iteration, whenever the highest score between the spectrum and cluster centers is greater than a threshold (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 are clustered.

To maintain efficiency, whenever a new spectrum is added, the center is updated only when the cluster size doubles (e.g. after the addition of the first, second, fourth, eighth, sixteenth, etc. spectrum to the cluster). Similar to Frank et al [28], 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 MASST+ analog search (Supplementary Fig. 4) that accounts for Δ-shifted and shared peaks for all pairs of input spectra (e.g. cluster centers from clustering+). To do this, it constructs an indexing table similar to MASST+ analog search. Then the table is used to efficiently compute the score between all pairs of spectra (Supplementary Fig. 13).

Data Availability. The datasets analyzed are available at gnps.ucsd.edu. Accession codes to all the analyzed datasets are available in the supplementary material.

Code Availability. MASST+, Clustering+, and Networking+ are available at https://github.com/mohimanilab/MASSTplus.

References

1.da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* **112**, 12549–12550 (2015).

2.Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature protocols* **15**, 1954–1991 (2020).

3.Courraud, J., Ernst, M., Svane Laursen, S., Hougaard, D. M. & Cohen, A. S. Studying autism using untargeted metabolomics in newborn screening samples. *Journal of Molecular Neuroscience* **71**, 1378–1393 (2021).

4.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).

5.Eberhard, F. E., Klimpel, S., Guarneri, A. A. & Tobias, N. J. Metabolites as predictive biomarkers for Trypanosoma cruzi exposure in triatomine bugs. *Computational and structural biotechnology journal* **19**, 3051–3057 (2021).

6.Ernst, M. *et al.* Gestational age-dependent development of the neonatal metabolome. *Pediatric Research* **89**, 1396–1404 (2021).

7.Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *Journal of proteome research* **7**, 113–122 (2008).

8.Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nature methods* **17**, 901–904 (2020).

9.Kale, N. S. *et al.* MetaboLights: an analog-access database repository for metabolomics data. *Current protocols in bioinformatics* **53**, 14–13 (2016).

10.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).

11.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).

12.Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nature chemical biology* **13**, 30–37 (2017).

13.Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nature methods* **17**, 905–908 (2020).

14.Petras, D. *et al.* Non-Targeted Metabolomics Enables the Prioritization and Tracking of Anthropogenic Pollutants in Coastal Seawater. (2020).

15.Quinn, R. A. *et al.* Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).

16.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 research* **44**, D463–D470 (2016).

17.van Der Hooft, J. J. *et al.* Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews* **49**, 3297–3314 (2020).

18.Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology* **34**, 828–837 (2016).

19.Wang, M. *et al.* Mass spectrometry searches using MASST. *Nature biotechnology* **38**, 23–26 (2020).

20.Ramos, A. E. F., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Natural product reports* **36**, 960–980 (2019).

21.Kalinski, J.-C. J. *et al.* Molecular networking reveals two distinct chemotypes in pyrroloiminoquinone-producing Tsitsikamma favus sponges. *Marine drugs* **17**, 60 (2019).

22.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). *Scientific reports* **9**, 1–13 (2019).

23.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. *Journal of the American Chemical Society* **139**, 4195–4201 (2017).

24.Vizcaino, M. I., Engel, P., Trautman, E. & Crawford, J. M. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *Journal of the American Chemical Society* **136**, 9244–9247 (2014).

25.Nguyen, D. D. *et al.* Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nature microbiology* **2**, 1–10 (2016).

26.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. *Journal of natural products* **82**, 1820–1830 (2019).

27.Reginaldo, F. P. S. *et al.* Molecular Networking Discloses the Chemical Diversity of Flavonoids and Selaginellins in Selaginella convoluta. *Planta Medica* **87**, 113–123 (2021).

28.Frank, A. M. *et al.* Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature methods* **8**, 587–591 (2011).

29.Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences* **104**, 6140–6145 (2007).

30.Schnell, N. *et al.* Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. *Nature* **333**, 276–278 (1988).

31.Mohr, K. I. *et al.* Pinensins: the first antifungal lantibiotics. *Angewandte Chemie International Edition* **54**, 11254–11258 (2015).

32.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).

33.Iorio, M. *et al.* A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity. *ACS chemical biology* **9**, 398–404 (2014).

34.Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports* **30**, 108–160 (2013).

35.Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry* **77**, 964–973 (2005).

36.Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nature chemical biology* **13**, 30–37 (2017).

37.Yang, J. Y. *et al.* Molecular networking as a dereplication strategy. *Journal of natural products* **76**, 1686–1699 (2013).

38.Ramos, A. E. F., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Natural product reports* **36**, 960–980 (2019).

39.Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences* **109**, E1743–E1752 (2012).

40.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**, 1–17 (2020).

41.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. *Molecular microbiology* **58**, 1368–1380 (2005).

42.Ueda, K. *et al.* AmfS, an extracellular peptidic morphogen in Streptomyces griseus. *Journal of Bacteriology* **184**, 1488–1492 (2002).

43. Bittremieux et al. Analog Access Repository-Scale Propagated Nearest Neighbor Suspect Spectral Library for Untargeted Metabolomics. BioRxiv, [Preprint] (2022) Available from: https://doi.org/10.1101/2022.05.15.490691

44. Ludwig, M., Fleischauer, M., Dührkop, K., Hoffmann, M. A. & Böcker, S. De novo molecular formula annotation and structure elucidation using SIRIUS 4. *Computational Methods and Data Analysis for Metabolomics* 185–207 (2020).

45. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **39**, 462–471 (2021).

Acknowledgements

The work of T.Y., M.M., Y.D. and H.M. was supported by a research fellowship from the Alfred P. Sloan Foundation, a National Institutes of Health New Innovator Award DP2GM137413, and a U.S. Department

of Energy award DE- SC0021340, P.C.D M.W. were supported by R03OD034493, U24DK133658, and R01GM107550 (P.C.D only).