Efficient Mass Spectrometry Searching and Clustering of Untargeted Metabolomics Data

Yudong Liu

CMU-CS-23-145
December 2023

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania

Thesis Committee:
Hosein Mohimani (Chair)
Carl Kingsford

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science.

Keywords: mass spectrometry, computational biology, bioinformatics, data mining, clustering

Abstract

Efficiently searching and dereplicating know entities from raw databases of biological extracts has been one of the major difficulties in natural product discoveries. Due to the wide usage of high-throughput mass spectrometry technique (MS) for building NP databases, there has been a pressing demand for an efficient infrastructure capable of organizing community-wide available MS libraries into solid datasets that allows cross-referencing between different MS spectral data of the same molecules. While 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 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. Our method enables querying against 717 million spectra from the GNPS public data in less than an hour and mapping the chemical diversity of all GNPS public data in days.

Acknowledgments

I would like to thank my advisor, Professor Hosein Mohimani, for his unwavering support throughout my last two years of research. I would like to thank Professor Carl Kingsford for attending my thesis committee. I would like to thank Mihir Mongia for his guidance and encouragement during my most struggling and difficult times in the project. I would also like to thank my collaborators Tyler Yasaka, Mustafa Guler and Bahar Behsaz for offering me valuable advice in research, as well as Liang Lu, Aditya Bhagwat, Mingxun Wang and Professor Pieter Dorrestein who have been great collaborators and mentors

I would also like to thank Professor Phillip Gibbons and Professor David Anderson for their valuable academic and research advice. I would also like to thank my room mates Zihao Deng, Runxuan Wang, and Zhengze Gong for their support in my academic journey, as well as those who inspired me with new ideas and knowledge through conversations and projects: Vivswan Shah, Fan Pu Zeng, Yiwen Song, and hopefully many more to come.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

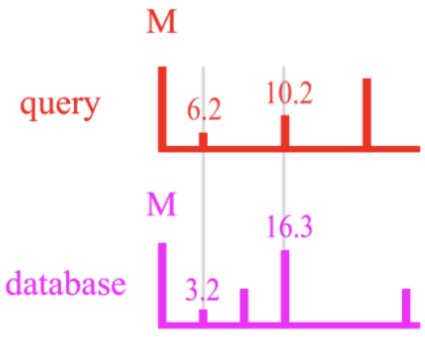
Identifying and discoverying new molecules from a collected sample of mixed products has always been an important task in all areas of research in life science. For example, to understand the function of uncharacterized genomic sequences, proteomics scientists need to identify all relevant final proteome product of the gene that's inherently more complex and closer to function than the gene itself [25, 60]. Metablomics scientists, on the other hand, need to discover small molecule metabolites to improve the understanding metabolic mechanism of numerous diseases, and improve the ability for monitoring various metabolic changes in clinical settings [51]. Scientists working on NP-drug discovery also need to determine all known molecules in their samples before identifying a bioactive 'hit' extract that's further fractionated to isolate the active Natural Products [5]. Mass spectrometry (MS) is a commonly used, high-throughput tool for identifying proteins and small molecules from mixtures [59]. Specifically, mass spectrometry breaks the molecules into smaller pieces and measures the mass of each fragment to determine the unique mass spectrometry fingerprints of molecules. The fingerprints of each molecule consists of a mass to charge ratio vector (representing masses of molecular fragments) and the intensity vector (representing the abundance of each fragment). An approach to identify all known molecules in the sample is to search mass spectra fingerprints of molecules collected from the samples against those of all existing molecules in the database with the probabilistic measure between a query spectra and a reference spectra in the database marked by the dot-product between their fingerprint vectors.

During the past decade, the size of mass spectral data collected in the fields of natural products, exposomics, and metabolomics has grown exponentially [30, 62, 70]. How to design algorithms and systems that can perform efficient searching and analysis across large number mass spectral datasets has been an important open problem to tackle. In accordance with the advances in mass spectrometry technology , multiple computational methods were developed for analyzing this massive data. For example, in order to determine whether a spectrum shares the same identity with the ones in the dataset, a naive approach is to brute-force compute the probabilistic measure between the query spectrum and each reference in the dataset of known molecules. However, the runtime of this approach grows linearly with the size of the reference database, which is proven to be really expensive when searching across very large public spectra datasets such as the Global Natural Product Social molecular networking infrastructure (GNPS) dataset [70] which contains hundreds of millions of mass spectrometry data. For example,

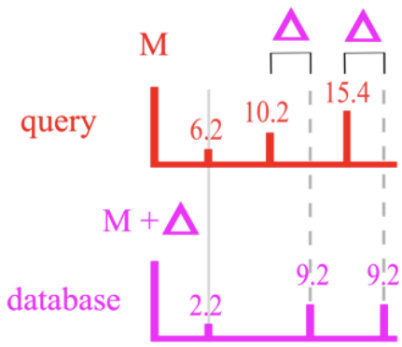
searching a single query spectra against all reference spectra in GNPS using this naive approach can take more than a whole day on a single CPU. In the case of unrestricted search allowing for a modification in the query spectrum in relation to the reference, the runtime increases to multiple months per query spectrum.

Recently Mass Spectrometry Search Tool (MASST) was introduced as a search engine for finding analogs of a query spectrum in mass spectrometry repositories [71]. MASST has demonstrated utility in the annotation of a wide variety of unidentified metabolites, including clinically important molecules in patient cohorts [11, 18, 53] toxins/pesticides in environmental samples [50] fungal metabolites [35], and metabolites from pathogenic microorganisms [14, 17, 38]. Moreover, molecular networking was introduced for clustering spectral datasets into families of related molecules [6, 22]. Molecular Networking has yielded a systematic view of the chemical space in different ecosystems and helped determine the structure of many compounds [31, 44, 54, 55, 64, 67, 74]. 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 [22] to cluster identical spectra by calculating a dot-product score (Figure 1.1a) between the spectra. Then Spectral Networking [6] is used to calculate a dot product score that accounts for peaks that are shared or shifted (Figure 1.1b) 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 (Figure 1.2). 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 54 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 thesis, 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 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.



(a) peak matching when calculating products in exact search



(b) peak matching when calculating products in analog search

Figure 1.1: In case of exact search, MASST searches a query spectrum against all database spectra with similar precursor masses, and computes the ExactScore, a sum multiplications between intensities of peaks shared by the query and database spectrum (shown in solid grey). In this case the score is $6.2 \times 3.2 + 10.2 \times 16.3 = 186.1$. 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 $6.2 \times 2.2 + 10.2 \times 9.2 + 15.4 \times 9.2 = 249.16$. Δ denotes the precursor mass difference between query and database spectra

In the following sections of the thesis, we present the outline of our algorithm and the performances evaluated on publicly available MS datasets. The contributions of our research consist of the following aspects:

1. We introduced a fast indexing-based algorithm for calculating the dot-product similarity score between two mass-spectras.
2. Based on the indexing we designed and implemented an efficient computational method, **MASST+**, for searching a query spectra against an untargeted mass-spectrometry dataset that's three magnitudes faster than the current state-of-the-art approach.
3. Similar to MASST+, we designed and implemented Networking+ for clustering spectral datasets into families of related molecules also three magnitudes faster than the state-of-the-art method. Our **Networking+** tool consists of **Clustering+** and **Pairing+** for clustering identical spectra from the same molecule and finding groups of related molecules respectively using dot-product scores respectively. Both of these methods turns out to be about two or three magnitudes faster than existing approaches (MS-Clustering and Spectral Networking).

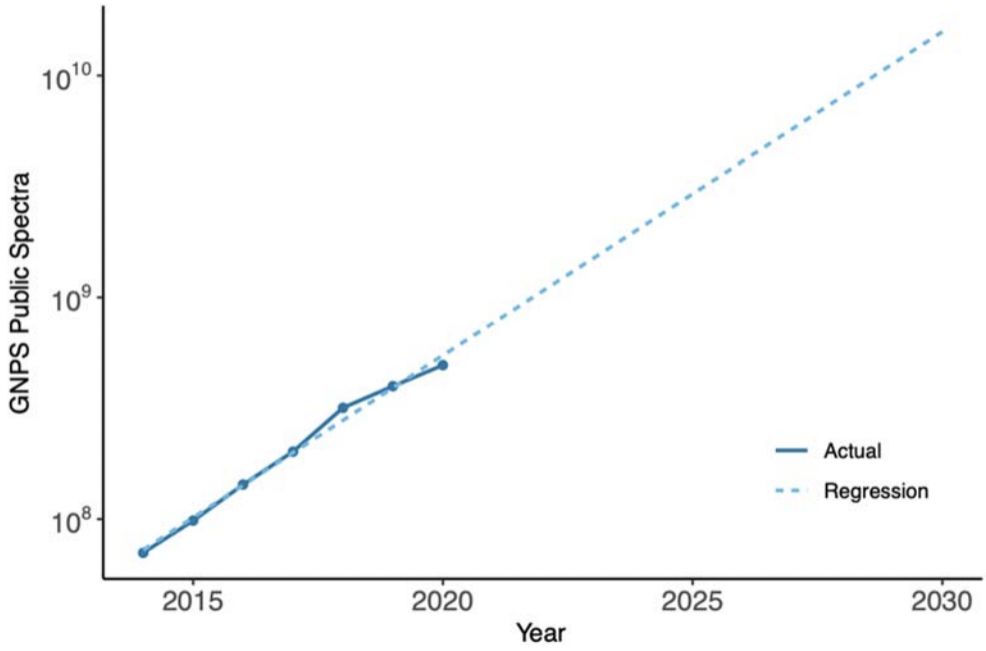


Figure 1.2: Growth of the GNPS database size since 2015. The size of the public GNPS database is projected to contain a billion spectra by the year 2026.

4. Our methods allows searching and networking analysis on the whole publicly available GNPS dataset (currently 717 million MS-spectras), a task that's not achievable by any existing approaches. The high efficiency of our algorithm allows the democratization of large mass-spectral dataset searching, clustering and networking tasks.
5. The high efficiency of our algorithm enables forming spectral networks on large-scale peptide datasets for discovery of new molecules.
6. We posted the networking analysis results of GNPS dataset using our method, providing solid stepping-stone for future dereplication and searching of public available mass-spectrometry datasets.

Currently MASST+ is available as a web service from `https://masst.ucsd.edu/masstplus/`. GNPS supports stand-alone MASST+ and integration with molecular networking.

Chapter 2

Related Work

In this section, we provide a brief outline of previous work done in mass-spectrometry analysis, spectral library search and spectral molecular networking.

2.1 Mass Spectrometry Analysis

Mass spectrometry (MS) techniques are increasingly used in the computational biology community due to its suitability for high-throughput characterization of NP [70]. Due to rapid growth of mass spectral datasets collected in the fields of natural products, exposomics, and metabolomics [30, 62, 70], there has been a pressing need for an efficient infrastructure for sharing and curation of crowdsourced MS datasets [70]. Over the past decade, there has been an increase in biological research that take advantage of publicly available MS datasets. For example, past works using mass spectrometry data involves discovery of new molecules [9, 10, 56], molecular structure identifications [19, 28, 39] or metabolomics studies using spectral data analysis [6, 22, 50].

Of particular relevance to our study is the literature on applied algorithms for MS dataset analysis. To address the issue that NP databases is not searchable with raw MS data, Global Natural Products Social Molecular Networking (GNPS) [70] is introduced as a publicly available infrastructure that incorporates MS data from . Further research in the field involves database searching [24, 34, 61, 69, 76], dereplication [33, 41, 42, 75] and clustering of raw Mass spectrometry data [1, 7, 26, 63]. Most of these tools focused on a only a narrow range of mass spectrometry data, such as specific types of proteins [72] or small molecules [34], or specialized for certain research purposes, such as providing fine-grained toolkit for compound identification for NIST spectral libraries [61], metabolite profiling [72], protein isoforms detection [2] or specific types of protein profiling and identifications [63, 69]

2.2 Spectral Library Search

Spectral library search has become a mature computational method for identifying tandem mass spectra in proteomics studies [77]. Spectral library search engines use spectral libraries of identified, generally experimental MS/MS spectra to identify observed MS/MS spectra to match raw spectral information with molecules in the databases [78]. Over the past two decades, several

spectral library search engines, including SpectraST [36] from trans-proteomics pipeline (TPP), Bibliospec [23] from MacCross lab, XHunter [12] from Global Proteome Machine (GPM) project, have been developed and many resources started to provide every growing, reliable spectral libraries. However, none of the above tools enables finding specific MS/MS spectra of interest, including unannotated spectra or structural analogs, in public repositories of metabolomics MS data and natural product MS data. With the development of publicaly avaiable infrastructures such as GNPS/MassIVE knowledge base [70] and adoption of universal, non-vendor-specific MS data formats [32] in multiple publicly available MS datasets, MASST[71] was developed as a web-based system that enables searching a single MS/MS spectrum for identical or analogous MS/MS spectra against public data in repositories, including unknown molecules.

Almost all of these search engines use a dot-product based similarity scoring approach to combine the experimental spectra against the library spectra, which treats each spectrum as a vector of the ordered peak intensities and measures the cosine of the angle between the spectra using the product of matched peak values. The naive dot product approaches for calculating similarity score had several limitations. For example, it fails to take into account the fact that matching peaks to fragments from peptide bonds is more important than matching peaks from other ions; it is also over-dependent of the resulting dot-product to very high peaks, and fails to take into account the discrepancy between the m/z values of the peaks in dot-product. Latest search tools like MASST [71] manage to alleviate the influence of very-high peaks by square-rooting the peak intensity values and performing normalization prior to calculating the dot product. MASST [71] performs peak-filtering and merging to get rid of noisy peaks and performs matching between peaks using an adjustable m/z threshold. MASST also incorporates user-defined parameters of minimum number of ions to match, precursor (parent) and product (fragment) ion tolerances, and performs both exact similarity and analog similarity searches based on non-identical precursor masses [73].

2.3 Spectral Molecular Networking

Molecular networking is a key method to visualize and annotate the chemical space in non-targeted mass spectrometry data first introduced in 2012 [73]. Unlike spectra searching algorithms, molecular networking goes beyond spectral matching against reference spectra, by aligning experimental spectra against one another and connecting related molecules by their spectral similarity. Idealy, in a molecular network, related molecules are referred to as a 'molecular family', differing by simple transformations such as glycosylation, alkylation and oxidation/reduction. [46]. Molecular networking was first publicly released as a part of the GNPS platform [70], and has become an essential bioinformatics tool for non-targeted mass spectrometry (MS) visualization and annotation since then. It's widely used in fields such as drug discoveries [13, 45, 52], genome mining [16, 48, 65] and metablomics studies [47, 49, 57] due to its potential of deciphering the "dark matter" of metabolomics through publicly available MS datasets and showing cross-associations between the chemistries of seemingly unrelated biological systems.

Similar to most MS searching algorithms, Molecular Networking uses a vector-based computational algorithm to compare the degree of spectral similarity between every MS/MS spectra

in a dataset [27]. It serves as is a graph-based workflow that aims to organize large MS datasets by mining spectral similarity between the MS/MS fragmentation patterns of different, but structurally related precursor ions. It first performs clustering to merge spectra with the identical precursor ion mass-to-charge ratio (m/z) and high dot exact dot product similarity score into a single consensus spectrum. Additionally, it removes low-intensity fragment ions to simplify the MS dataset and reduces the downstream computational load for the spectral similarity algorithm [21, 22]. After forming the consensus spectrum (cluster centers) into a vector of m/z peak values, it calculates a cosine score (normalized dot product) between every possible pair of consensus MS/MS spectra , which allows the determination of the degree of spectral similarity between them. The molecular networking tool released with GNPS allows customized precursor ion mass tolerance for the consensus spectrum and the fragment ion mass tolerance for clustering. It also allows adjustments to the minimum-match fragment ions to meet the specificity of the fragmentation behavior of the analyzed molecules. [52].

Chapter 3

Methods

In this section, we describe the algorithm design of MASST+ and Networking+ and provide theoretical justifications for its high efficiency that we described in Section 1.

In Section 3.1, we provide the outline of our algorithm and justifications for its high performance. In Section 3.2, we describe the data processing techniques and implementation details for MASST+. In Section 3.3, we describe the data processing techniques and implementation details for Molecular Networking+. We provide detailed outline of the algorithms in Section 6.3 of the Appendix.

3.1 Indexing-based dot-product score

In this subsection, we describe and provide justification for an indexing based fast dot product algorithm for calculating similarity scores between spectras that serves as the fundamental cause for the high time-efficiency of our methods. Researchers usually evaluate the affinity between a pair of spectra S_1 and S_2 using cosine similarity score that's calculated through the dot-product between the two spectras. When calculating the dot product, each spectra is treated as a vector of m/z peak intensity pairs (m_i, p_i) such that m represents the mass divided by charge number (m/z value) of the peak in the mass spectrum, whereas p_i represents the normalized intensity value of the peak. The dot product between

$$\begin{aligned} S_1 &= \{(m_1^{(1)}, p_1^{(1)}), (m_2^{(1)}, p_2^{(1)}), (m_3^{(1)}, p_3^{(1)}), \cdots, (m_M^{(1)}, p_M^{(1)})\}, \\ S_2 &= \{(m_1^{(2)}, p_1^{(2)}), (m_2^{(2)}, p_2^{(2)}), (m_3^{(2)}, p_3^{(2)}), \cdots, (m_N^{(2)}, p_N^{(2)})\}. \end{aligned} \tag{3.1}$$

can be written as the matched peak intensity products sum divided by the product of spectral norm values. We consider a pair of peak to be matched if they have m/z values within a certain threshold value tol (such as 0.01 Da)

$$P(S_1, S_2) = \sum_{|m_i^{(1)} - m_j^{(2)}| < tol} \frac{(p_i^{(1)} * p_j^{(2)})}{||S_1||_2 * ||S_2||_2} \tag{3.2}$$

$$||S||_2 = \sqrt{\sum_{(m_i, p_i) \in S} p_i^2} \tag{3.3}$$

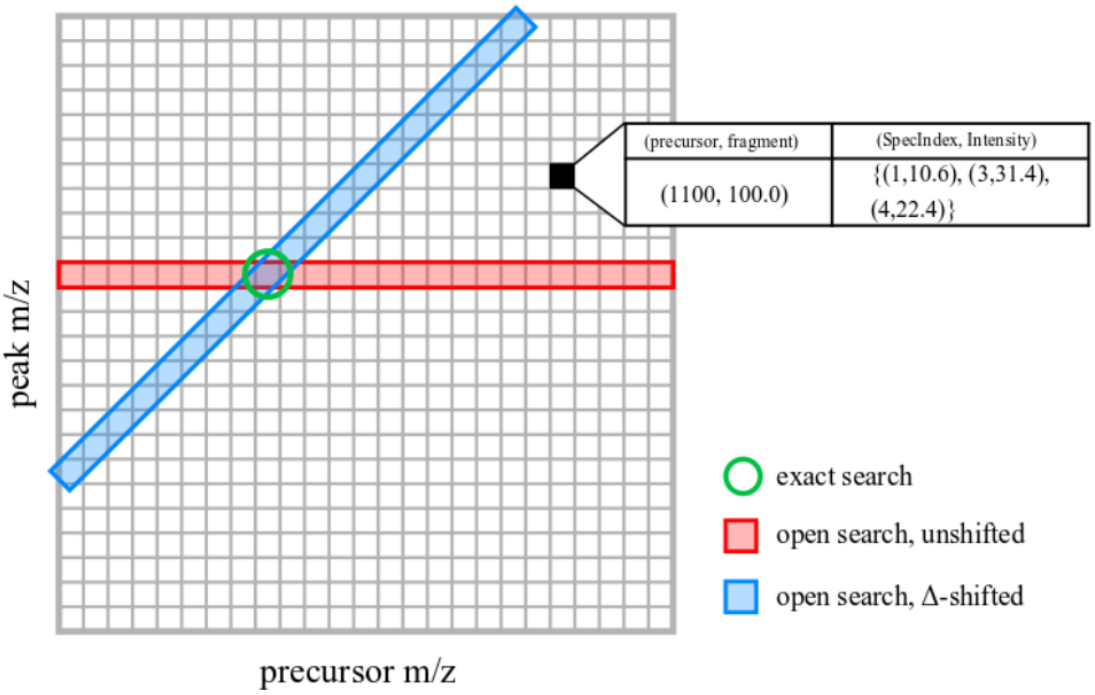


Figure 3.1: **Fast Dot Product Indexing:** The fast dot product indexing table corresponds to a two-dimensional grid, with precursor mass on the x-axis and peak mass on the y-axis. Each database peak is inserted into a list corresponding to a specific location in the grid, determined by the peak mass 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.

where Equation 3.3 represents the L2-norm of the spectra vector

The main underlying intuition of the algorithm is to preprocessing a set of spectra into an indexing table that maps all possible precursor m/z and fragment ion m/z pairs to the spectra that contain them.

In algorithm 1 of Section 6.3 we provide the brief outline of the non-indexing algorithm for performing exact search of a query spectra against all spectras in the library through calculating the one-versus-all exact dot-products as the similarity scores and only preserving the ones above a certain threshold. Specifically, or each target spectra inside the library s, for each peak inside the query spectra $P = (m, p)$, we search through the peaks of s and try to find peaks in the same location as P. After finding all the matched pairs, we perform sorting and selection on the matches to make sure each peak from each spectra is matched only once when calculating pair-wise dot product similarity score. According to Algorithm 1, it's clear that we have a linear growth of runtime with respect to the library size of L for searching matched peaks between the

query and Library spectra.

We provide an example for performing indexing-based one-versus-all peak matching for exact search in Algorithm 3 of Section 6.3. Specifically, we assume that we've already constructed the indexing database D such that the t-th index contains peaks of all spectras from the library L that have a m/z value between $t * tol$ and $(t + 1) * tol$. Although the size of this database grows linearly with the number of the spectra in the library, It only needs to be constructed once and takes only $O(1)$ time for adding peaks from a new library spectra to the D, offering a high scalability.

Since each peak $P = (m, p)$ of the query spectra could only match with peaks from library spectras that are within a m/z difference tolerance range of tol. Therefore, we only need to search inside the index bin m **div** tol and its neighboring bins to find the peaks from the entire library that matches with the query peak. Based on the fact that most queries spectra have sparsed distribution of peaks, we only need to consider the library peaks from a few m/z indexes when performing one-versus-all dot-product calculation between the query and the library spectras. In fact, the peaks from most library spectra won't even be accessed during the calculation. Therefore, the complexity of searching on a query spectra is only proportional with the number of peaks inside the indexed-bins we need to look at, offering us a two or three magnitudes of speedup when searching large open libraries.

3.2 MASST+

Given a query spectrum, MASST+ efficiently searches a database of reference spectra to find similar spectra by using the fast dot product algorithm. The backbone of **MASST+** software implementation consists of two parts, constructing indexing-based database and performing searching of a query spectra against all spectra inside the database through dot-product calculations. **MASST+** conducts searching based on exact similarity score and analog similarity score.

When conducting exact search, 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 tol 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.

In the case of analog search, MASST+ uses a large precursor mass tolerance (e.g. 300Da) and computes ShiftedScore that 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 (mz, p) in the query spectrum with precursor mass M, and scans lists of shared (a_i, p_i) and shifted peaks (b_j, p_j) from a library spectra with precursor mass M' such that either $|a_i - mz| < tol$ or $|(b_j - M') - (mz - M)| < tol$. 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. Note that **MASST+** analog search is a variant of the fast dot product algorithm 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.

During the preprocessing step of indexing, We perform filtering using a certain window size to preserve the top-K peaks with the highest intensity for each window. We then perform L2-normalization on peaks of each spectra to make sure the square sum of all peaks is equal to 1. We construct the shifted and unshifted indexing database according to procesures in algorithm 2 and then split the indexes into different ranges and stores peak from each range in a group of binary files

When conducting exact search on a given set of query spectra, we ues Algorithm 3 to find the exact matched peaks between the query and library spectra that has at least one unshifted matched peak with the query. to calculate the dot-product similarity score between query spectra and library spectra according to Algorithm 5 in Section 6.3. When conducting analog search, we use Algorithm 4 in Section 6.3 to perform shifted or exact peak matching between the query and library spectra. We calculate the dot-product similarity score between query spectra and library spectra according to Algorithm 5 ensuring each peak is only matched once for calculating the pairwise dot-product.

3.3 Networking+

The Networking+ software consists primarily of two phases, **Clustering+** and **Pairing+**. In the **Clustering+** phase, we cluster the input spectral library as a number of groups such that members within each group has a exact dot-product score of larger than a certain threshold (0.9). In the **Pairing+**, we select a candidate from each cluster as the cluster center and calculate all-versus-all pairwise exact or analog dot-product similarity score. We treat each cluster as a node and preserve the similarity scores above a certain threshold as an edge between the two clusters. We're then able to perform extensive metabolomics analysis or signature small molecule discoveries on the connected components of in the graph

3.3.1 Clustering+

The clustering+ part is aimed to gather untargeted spectra that are likely to belong to the same molecular structure by comparing the exact similarity scores between their mass-spectrometry fingerprints. In order for the fingerprints of two spectra to be matched, they should have almost identical precursor masses. Therefore, in the first step of clustering, we preprocess the spectra into different precursor mass bins separated according to a certain precursor mass tolerance. Each bin correspond to a small precursor mass range. We wouldn't consider clustering mass spectra with a precursor mass difference larger than the precursor mass tolerance value.

For each precursor mass bin, we perform clustering using a greedy algorithm by iterating through all spectra in the bin and compare them with candidates of existing clusters. If the spectra has an above threshold exact similarity score with the candidate spectra of at least one existing clusters, we add the spectra to the cluster. Otherwise if the spectra is not similar to any of the existing spectras, we create a new cluster for the spectra and use the spectra as the candidate for the new cluster.

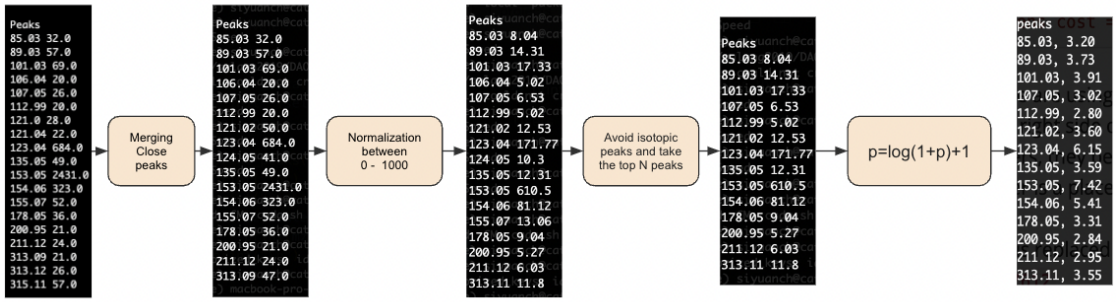


Figure 3.2: **Preprocessing pipeline of Clustering+**

The naive approach of directly comparing the dot-product score between normalized spectra can yield misleading clustering results in certain cases. Since a lot of spectra in the GNPS library contains one dominant peak that takes up to over 80% of the normalized intensity value, it would only take two spectra to have an overlapping dominant peak in order to have a high cosine similarity score rather than having a exact matched mass-spec fingerprint. The existence of isotopes also generates neighboring peaks that are $+1$ or $+2$ units away in m/z values with a random ratio between the peak intensities directly influenced by the isotopic abundance ratio of the elements in the samples. The error produced by noise signal of mass-spectrometer is also likely to cause deviation of mass-distribution, yielding peaks in additional locations or breaking a larger peak into two smaller ones with close m/z values. These factors would greatly influence the dot-product similarity score, either increase the chance of spectra from heterogenous molecules to be clustered or reduce the cosine similarity score between untargeted spectra of the same molecule. We inherited the preprocessing steps (Figure 3.2) from the source code of **MS-Clustering** [22].

During preprocessing step of each spectra, we first merge the peaks close to each other in consideration of random errors produced by mass-spectrometers using a scanning window. We then rescale the peak values so that they sum up to 1000 to avoid explosion of dominent peak when performing the L2-normalization. We also further search peaks off by one or two units in m/z value (peaks with a difference of $0.99 - 1.01$ or $1.99 - 2.01$ in m/z value) to reduce additional peaks caused by isotopic mass differences.

We've provided an outline of the **Clustering+** procedures below. The first step involves generating indexing databases according to Algorithm 2. However, contrary to the indexing step of **MASST+**, the indexing database for **clustering+** is hierarchical. The first layer indexes based on precursor mass ranges, while the second layer indexes based on m/z value of peaks.

Indexing for Clustering+ Variables:

L the spectral library to be clustered

$mzTol$ the m/z tolerance for two peaks to be matched

$massTol$ the precursor mass tolerance for two spectra to be clustered

1: **procedure** INDEXING($L, mzTol, massTol$)

```
    Initialize Database D = {}
    D.mzTol = mzTol
    D.massTol = massTol
    for spectra s ∈ L do
        massBin = s.precurMass div massTol    ▷ calculate the precursor mass bin of s
        Store spectra s in the precursor mass bin D[massBin]
        for peak ∈ s.peaks do                  ▷ indexing peaks accoding to Algorithm 2
            Add peak to the m/z Indexed Structure inside the precursor mass bin
        end for
    end for
    return D
end procedure
```

After performing indexing, we perform greedy clustering to iterate over spectra of each presursor mass bin. For each spectra q in the bin, we calculate the exact cosine dot-product similarity score between the spectra and candidate spectra of all clusters inside the same presursor mass bin based on the one-versus-all dot product similarity score Algorithm 3 and 5. If the spectra q does not fit with the candidate of any cluster, we make a new cluster for q and mark it as the candidate for the new cluster.

Variables:

D the indexed database generated by Indexing step

$thresh$ the minimum threshold for exact dot product similarity score to cluster two spectra

Clustering+ procedure based on Indexed Database

```
procedure CLUSTERING+(D)
    AllClusters = {}
    for massBin ∈ D do
        Initialize empty ClusterCenters to store candidate spectras of clusters
        Initialize empty CandidateIndex to store peaks of candidate spectras
        for q ∈ D[massBin] do
            Preprocess q according to Figure 3.2
            Based on Algorithm 5,
            calculate exact scores between q and ClusterCenters using CandidateIndex
            if at least one similarity score greater than thresh then
                Add q to the cluster with the highest candidate similarity score with q
            else
                Create new cluster C in AllClusters
                Add q to new cluster C
                Add q to ClusterCenters
                Add peaks of q to CandidateIndex using Indexing functions of Algorithm 2
            end if
        end for
    end for
    return AllClusters
end procedure
```

3.3.2 Pairing+

Pairing+ computes a score similar to MASST+ analog search that accounts for Δ-shifted and shared peaks for all pairs of input spectra (e.g. candidates as cluster center 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.

We implemented **Pairing+** as a more time-efficient approach for spectral network generation than the baseline proposed in Molecular Networking implementation [27]. Based on the assumption that each cluster generated by **MS-Clustering** contains spectra from the same molecular structure, **Pairing+** is designed to establish connections between the molecular clusters to form a graph that visualizes the similarity between molecules. The edges of the graph were determined based on whether the exact/analog dot-product similarity score between two cluster candidates are above a certain threshold (which is usually set to 0.7 or 0.9 depending on the task we're running). In order to determine all possible pairwise connections between clusters, we need to calculate the all-versus-all pairwise dot product scores above the threshold. We apply the same intuition as **MASST+** and **Clustering+** by preprocessing the cluster candidates into an indexed-bin based storage structure (If we were using results directly generated by **Clustering+** then we can skip this step by taking the indexing database storing the cluster center peaks generated during clustering). We then take each candidate as the query, and calculate its product with all other candidates through the same methodology as **MASST+**. After constructing all possible connections between molecular clusters, we eliminate noise by removing singletons from the graph. We then perform further analysis by calculating the connected components using Breath-first search algorithm. We provide a brief outline for the procedure of performing **Pairing+** based on analog similarity score below

Variables:

D the database containing all candidate spectra

I the indexing database containing peaks from all candidates

$thresh$ the minimum exact/analog similarity score for forming an edge between candidates

Clustering+ procedure based on Indexed Database

procedure PAIRING+(D, I)
  Edges = $\{\}$
  for $q \in D$ **do**
    Initialize empty **Product** vector to store dot-product scores with other candidates
    Calculate one versus all product score using query **q** and store in **Product**
    for $(specIdx, score) \in$ **Product do**
      if $score > thresh$ **then**
        Add $(q.id, specIdx, score)$ to **Edges**
      end if
    end for
  end for
  Construct V as the indexes of candidates appear as with endpoint of at least one edge
  Perform **BFS** on $G = (V, E)$ to generate all connected components information S
  return S
end procedure

When performing testing of the above draft implementation on large scale MS-datasets, we realized that our **Pairing+** implementation had a low time and memory-efficiency. This is likely due to the redundant memory utilization for storing matched peaks between low-similarity spectra pairs since memory allocation procedure in C++ is very expensive. In order to further optimize our implementation, we take advantage of a prefiltering technique based on an estimated upperbound for the dot-product similarity score between two spectral peak vectors. Given spectral peaks $v_1 = \{(p_i.mz, p_i.int)\}_{i=0}^{M}$ and $v_2 = \{(q_i.mz, q_i.int)\}_{i=0}^{N}$ such that v_1, v_2 contains the exact and shifted peaks of two spectra, and a tolerance for peak matching of ϵ,

$$Prod(v_1, v_2) \leq \sum_{||q_i.mz - p_j.mz|| < \epsilon} p_i.int \times q_j.int$$

The above inequality holds since when performing dot-product calculation between two spectra, each peak index from a spectra can only be matched to one peak index from the other spectra, therefore, we need sorting in Algorithm 4 to take the optimal set of matching between peaks from two spectra. The in-place update rule for calculating the product upperbound equation 3.3.2 requires only Constant amount of memory, and thus over two magnitudes more time-efficient than directly calculating the actual product score. Additionally, due to the sparsity of spectral peaks, majority of the pairs in the database are expected to have a low product similarity score. Therefore, our algorithm can be greatly optimized by filtering out the product pairs with a product upperbound below the threshold

in addition to the above optimization, we also take similar approaches as the preprocessing steps of **Clustering+** to ensure the spectral fingerprint of two molecules match as much as possible. For example, we take log transformation when normalizing the peak intensity values and require two matched candidates to have at least 6 matched peaks in additional to having a high dot-product score. These steps alleviate the influence of the most dominate peak in the spectra for dot-product score calculation and ensures two matched candidates have spectral fingerprint that are similar to a reasonable extent.

Chapter 4

Results

In this section, we describe the results of our main results and supplemental analysis. We include the comparison between our method and baseline in both time and memory efficiency metric when testing on different scale of sampled spectral datasets. We benchmark our **MASST+** and **Networking+** toolkit on the whole GNPS containing over 717 million spectra, a dataset that couldn't be handled by the baseline on commodity computational resources. We report our time and memory utilization for performing searching and networking on the dataset along with the connected-component graph generated by **Networking+**. Additionally, we provide example of using **Networking+** for identification of certain types of small molecules such as lanthipeptides

4.1 Benchmarking MASST+

When benchmarking **MASST+**, we use a bin size of $0.01Da$, which can handle both high-resolution ($0.01Da$ accuracy) and low-resolution ($0.5Da$ accuracy) data. we use a dot-product similarity score threshold of 0.7 for searching matched spectra.

We have benchmarked MASST+ (Table 4.1) 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. 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.

Table 4.1 illustrates the runtime and memory consumption of MASST+ versus MASST in exact and analog mode for various subsets of the clustered GNPS. According to 4.1, the indexing time and memory consumption grows linearly with the size of datasets. According to 4.2a, the 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. Figure 4.2b breaks down MASST+ runtime into two different steps, loading

Table 4.1: **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.

Method	Mode	Dataset(size)	Search Time	Search Memory	Matched 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 sec**	**159MB**	**2,175**
MASST	exact	Whole GNPS (717M)	N/B	N/B	N/B
MASST+	**exact**	Whole GNPS (717M)	**43 min**	**21GB**	**171**
MASST	analog	Whole GNPS (717M)	N/B	N/B	N/B
MASST+	**analog**	**Whole GNPS (717M)**	**115 min**	**35GB**	**265,958**

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.

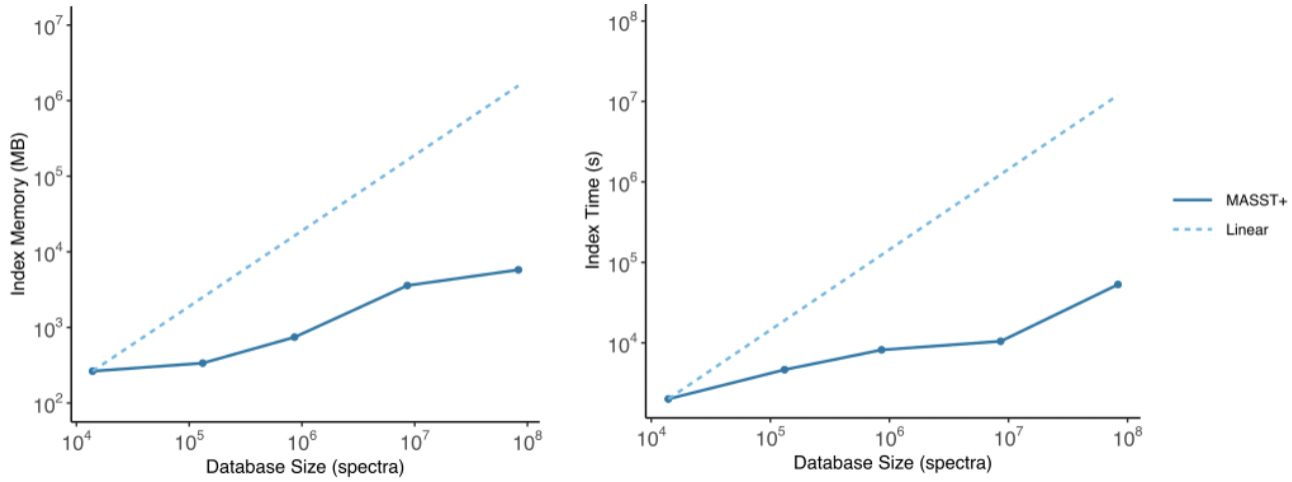
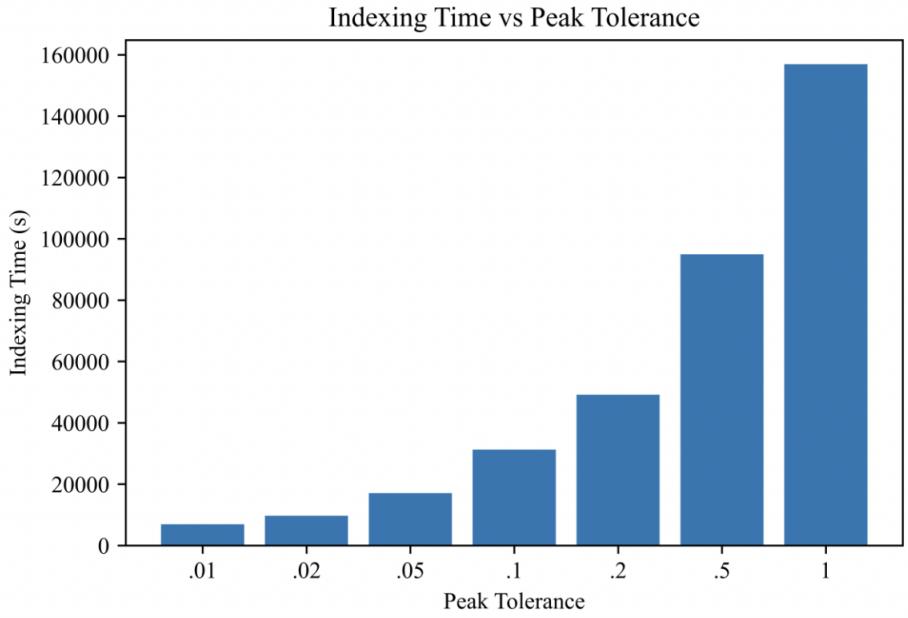
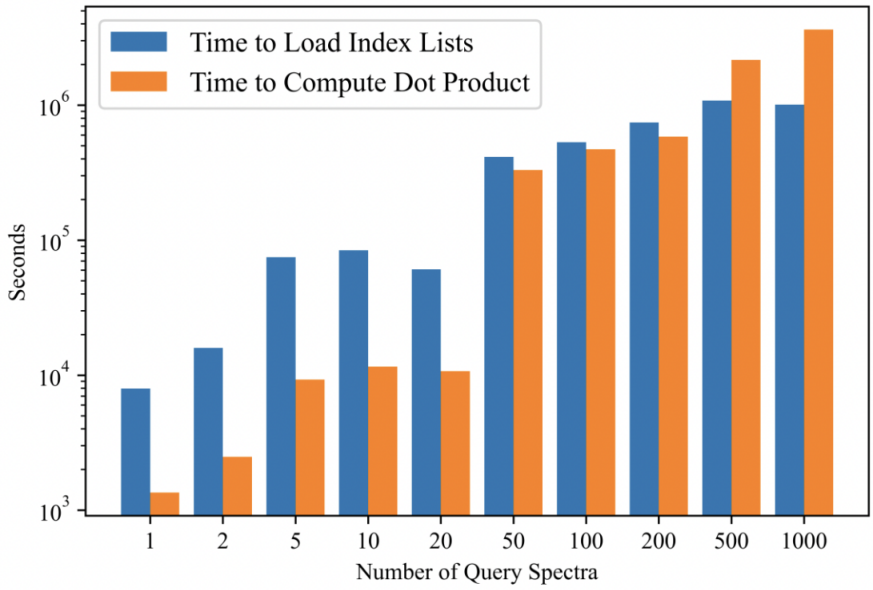


Figure 4.1: MASST + indexing memory (left) and run time (right) as database size grows. Both runtime and memory grow sub-linearly (linear growth shown on dashed line). On the clustered GNPS, MASST+ requires eight hours of and eight gigabytes of memory. Note that indexing needs only to be performed once for each database.



(a) Time required for indexing 1 million spectra for various values of peak tolerance. Indexing time increases monotonically with respect to peak tolerance.



(b) Time required to load lists from index and compute dot product for various numbers of query spectra. Fraction of time devoted to loading decreases for larger numbers of query spectra.

Figure 4.2: **Indexing time changes as peak tolerance and number of query spectra grows**

4.2 Benchmarking Networking+

We compare our implementation of **Networking+** with Molecular Networking under the same parameter values. In order to find structurally related families of small molecules, the existing Molecular Networking method first clusters spectra from identical molecules using MS-Clustering [6]. It then connects clusters of related molecules using Spectral Networking [22]. MS-Clustering puts two spectra in the same cluster if their precursor mass difference is below a threshold (usually $2Da$) 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 [22]. 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).

We provide benchmarking result for **Networking+** against molecular networking on various data sizes for which runtime is less than 24 hours in Table 4.2. We also provide a benchmark of **Clustering+** in Table 4.4 and **Pairing+** in Table 4.5 separately. 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 and Spectral Networking . The clusters and networks reported by Clustering+ and Pairing+ are identical to MSClustering and spectral networks. As previously noted in Bittremieux et al.[8], 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.

Table 4.2: **Benchmarking Networking+** Comparison of Molecular Networking and Molecular Networking+ runtimes for various sizes of spectral datasets (runtimes are shown in seconds). The cases where the search did not yield results within 24 hours are shown with N/A

Dataset size	Networking+ Runtime (sec)	Molecular Networking Runtime (sec)
100,000	7	30
200,000	13	62
500,00	44	247
1,000,000	94	4041
2,000,000	202	8067
5,000,000	500	28117
10,000,000	1296	N/A
20,000,000	2400	N/A
50,000,000	9931	N/A
100,000,000	34359	N/A

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. As shown in 4.3, 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. 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 as shown in Figure 4.4a. As for the networking results, 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 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 as shown in Figure 4.4b. Networking+ took 6 days to finish this task on 1 CPU. Currently, this task is not feasible using existing approaches.

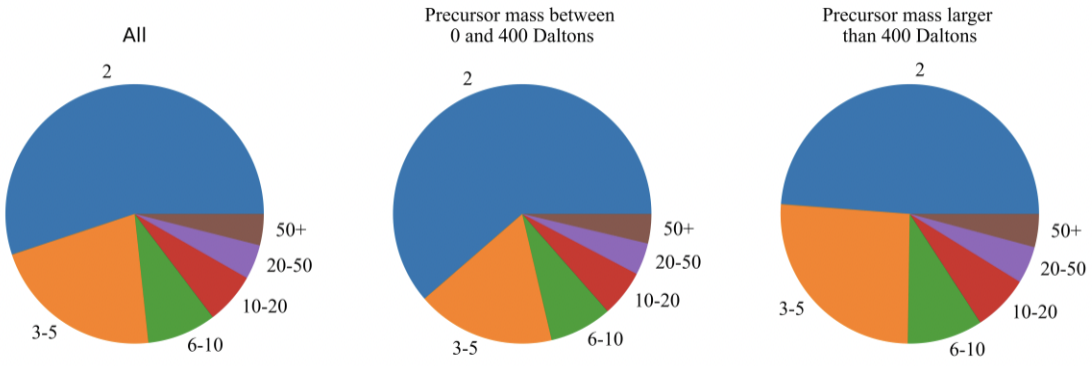
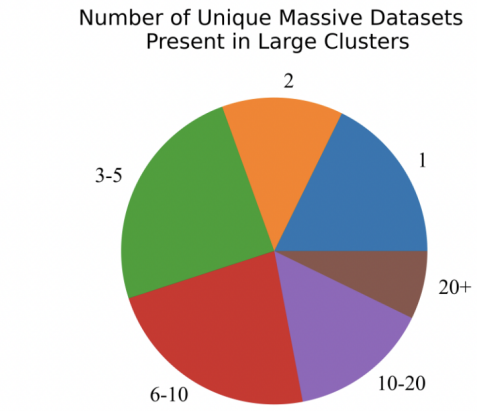


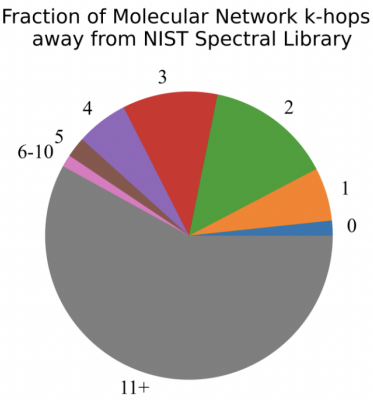
Figure 4.3: Portion of clusters containing 2, 3-5, 6-10, 10-20, 20-50, and 50+ spectra for clusters of varying mass ranges. For precursor mass ranges of 0Da-400Da, a significantly larger fraction of clusters contain 2 spectra compared to clusters with precursor mass larger than 400Da.

Table 4.3: **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

Method	Dataset(size)	Clustering Time	Clustering Mem	Clusters count	networking time	Networking Mem
Molecular-Networking	MSV000078787 (219,915)	321 sec	662Mb	5,288	8 sec	1224Kb
Networking+	**MSV000078787 (219,915)**	**27 sec**	**992Kb**	**5,288**	**0.25 sec**	**996Kb**
Molecular-Networking	Whole GNPS (717M)	N/A	N/A	N/A	N/A	N/A
Networking+	**Whole GNPS (717M)**	**60 hours**	**93Gb**	**8,453,822**	**97 hours**	**23Gb**



(a) Fraction of large clusters (consisting of more than 20 GNPS spectra), that contain spectra coming from 1,2,3-5,6-10,10-20, and 20+ unique MASSIVE datasets



(b) Fraction of nodes/clusters in the molecular network that $0, 1, 2 \ldots 11+$ edges away from clusters that are exact matches to spectra in the NIST spectral library.

Table 4.4: **Benchmarking Clustering+** Comparison of **Clustering+** and Molecular Clustering+ runtimes for various sizes of spectral datasets (runtimes are shown in seconds). The cases where the search did not yield results within 24 hours are shown with N/A

Dataset size	Clustering+ Runtime (sec)	Molecular Clustering Runtime (sec)
100,000	7	110
200,000	13	151
500,00	42	506
1,000,000	87	1009
2,000,000	185	2867
5,000,000	444	7273
10,000,000	1043	24596
20,000,000	1620	50557
50,000,000	3005	N/A
100,000,000	9933	N/A
300,000,000	91729	N/A

4.3 Identification of lanthipeptides using Networking+

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 [58], antifungals [43], antiviral [20], and antinociceptives [29]. Lanthipeptides are structurally defined by the thioether amino acids lanthionine, methyllanthionine and labionin. Lanthionine and

Table 4.5: **Benchmarking Pairing+** Comparison for Pairing+ and Spectral Networking runtimes for various sizes of spectral datasets (runtimes are shown in seconds). The cases where the search did not yield results within 24 hours are shown with N/A

Dataset size	Pairing+ Runtime (sec)	Spectral Networking Runtime (sec)
10,000	1.14	27
20,000	5.62	111
50,000	32.5	2072
100,000	91.8	23808
200,000	278.3	83101
500,000	2018.8	N/A
1,000,000	7900.2	N/A
2,000,000	39737	N/A

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 [3]. 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 Nterminal leader processing.

We formed the spectral network using **Networking+** for a subset of 500 Streptomyces cultures with known genomes (Table 4.6). 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 PepNovo35. 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. 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 [41]. This strategy identified three known and 14 novel lanthipeptides with p-values below 1e-15 (Table 4.7). Among them, the precursor of 13 lanthipeptides (76%) overlaps with reports by the genome mining strategy introduced by Walker et al. [68]. 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.

Table 4.6: List of MassIVE datasets mined for lanthipeptides

MassIVE ID	number of strains	media
MSV000090476	60	ISP-2
MSV000090473	60	ISP-4
MSV000090472	60	NSG
MSV000090471	60	TSA
MSV000090457	60	Czapek
MSV000089818	264	ISP-4
MSV000089817	264	TSA
MSV000089816	264	Czapek
MSV000089815	264	NSG
MSV000089813	264	ISP-2
MSV000088816	176	ISP-4
MSV000088801	176	TSA
MSV000088800	176	Czapek
MSV000088764	176	NSG
MSV000088763	176	ISP-2

Table 4.7: 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. The p-values were computed using Markov Chain Monte Carlo approach [40]. This is a one-sided p-value, where adjustment was made for multiple comparisons.

Organism	name	Sequence	score	p-value	Mass	Walker et al.
Streptomyces rimosus NRRL WC-3904	CHM-1793	DT-18GHCS-18GVCT18VLVCT-18VAVC	21	2.50E-36	1793.77	YN
Streptomyces albus NRRL F5917	CHM-1731	YS-18QVCS-18IVVCNT18VVICG	19	5.80E-33	1731.81	YN
Streptomyces lavenduligriseus NRRL ISP-5487	SapT	YT-18QGCS-18GLCT18IVICAT-18VVICG	18	1.40E-32	2030.95	YN
Streptomyces species NRRL S-240	CHM-1911	S-18TAGCS-18GLCT-18IIVCAT18VVICA	17	5.20E-31	1911.91	YN
Streptomyces pathocidini NRRL B-24287	CHM-2168	IT-18S-18IS-18YCT-18PGCT18SDGGGS-18GCS-18HCC	16	1.60E-26	2168.76	YY
Streptomyces moroccanus NRRL B-24548	CHM-2182	IT-18S-18IS-18YCT-18PGCT18SEGGGS-18GCS-18HCC	15	2.00E-25	2182.78	YY
Streptomyces cinerochromoge nes NBRC 13822	CHM-1974	YT-18EGCS-18GLCT18ILVCAT-18VVIC	13	9.10E-24	1974.91	NN
Streptomyces hygroscopicus NRRL ISP-5087	CHM-1354	MT-18QVCPVT-18SWHC	13	3.60E-23	1354.56	YN
Streptomyces rimosus NRRL WC-3874	CHM-1831	PSRSSSPGSFPPGST-18PS18APS-18	14	1.60E-21	1831.85	NN
Streptomyces albus NBRC 13041	CHM-1775	YS-18QVCS-18IVICNT18VVICS	11	5.50E-20	1775.84	NN
Streptomyces kanamyceticus NBRC 13414	CHM-1748	IS-18GEES-18CFRT-18CT18TCS-18LC	12	3.40E-19	1748.68	YN
Streptomyces sulphureus NRRL B-2195	CHM-2229	TEGGGGDS-18SGCS-18GVCT18IVVCT-18VIVC	9	1.10E-17	2229.95	YN
Streptomyces anulatus NBRC 12853	AmfS	T-18GS-18QVS-18LLVCEYSS18LSVVLCTP	11	2.10E-17	2212.09	YN
Streptomyces anulatus NBRC 13369	CHM-1669	C-34LPEPFP+16TATT18RVGCD	11	9.50E-17	1669.78	YN
Streptomyces paludis JCM 33019	CHM-1635	S-18GEES-18CFRT-18CT-18T18CSLC	11	2.30E-16	1635.59	YN
Streptomyces anulatus NBRC 12861	CHM-2433	CRPPSASLCIT-18SDRS-18S18TGRYLSM	11	3.10E-16	2433.14	NN
Streptomyces brasiliensis NBRC 101283	Amfs analog	TGS-18QVS-18VLVCEYS-18S18LSVVLCTP	11	7.10E-16	2198.08	YN

Chapter 5

Conclusion and Discussion

5.1 Conclusion

In this thesis, we present the design and results of **MASST+** and **Networking+**, two efficient softwares for mass spectrometry searching and analysis over large datasets. We included thorough comparison with baselines including MASST, Molecular Networking, Molecular Clustering and Spectral Networking and proved the huge improvement of our method over existing approaches in terms of both time and memory efficiencies. By taking advantage of indexing calculation for dot-product similarity scores, we were able to make **MASST+** over two magnitudes faster than MASST, and **Networking+** over two magnitudes faster than Molecular Networking. We also proved the effectiveness of **Networking+** for natural product discovery of small molecules by performing identification of lanthipeptide.

5.2 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 [4, 46, 53, 55, 66, 73, 75]. 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 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 [37] and chemical class [15] 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.

Chapter 6

Appendix

6.1 Code Availability

We provide the code for the algorithm and software tools mentioned in the thesis below:

1. **MASST+** and **Networking+** https://github.com/mohimanilab/MASSTplus
2. Seq2Ripp (https://github.com/mohimanilab/seq2ripp
3. PepNovo (https://github.com/mohimanilab/seq2ripp
4. Dereplicator (https://ccmsucsd.github.io/GNPSDocumentation/dereplicator/)

6.2 Data Availability

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, MSV000089813, 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.

6.3 Algorithms Outline

We provide detailed outline for all the algorithms mentioned in the thesis, including the non-indexing algorithm for dot-product similarity search (Algorithm 1), indexing procedures for **MASST+** and **Networking+** (Algorithm 2), peak matching process for one-versus-all exact (Algorithm 3)or analog (Algorithm 4) search, and calculating dot-product scores from matched peaks (Algorithm 5)

Algorithm 1 A Naive algorithm for query againset all dot-product calculation

Input

A query spectra **q** containing peaks $\{(m_1^{(q)}, p_1^{(q)}), \cdots, (m_K^{(q)}, p_K^{(q)})\}$
A spectra library $\mathbf{L} = \{s_1, s_2, \cdots, s_N\}$ such that each spectra contains a list of m/z values
Similarity dot-product score threshold $thresh$ (default set to be 0.7 or 0.9)
m/z tolerance tol for two peaks to be matched (default set to be 0.01 or 0.02)

Output

A mapping $Res = \{(i_1, v_1), (i_2, v_2), \cdots, (i_n, v_n)\}$ containing idx of matched library spectra and their similarity product with the query

REQUIRE

peak intensities in q and all library spectra s are L2-normalized
To normalize spectra $s = \{(m_k, p_k)\}$, we perform $p_k = \frac{p_k}{\sqrt{\sum_j p_j^2}}$ for each peak in the list

function NONINDEXSEARCH($L, q, thresh = 0.9, tol = 0.02$)
  $Res \leftarrow \{\}$
  for $s = \{(m_1^{(s)}, p_1^{(s)}), (m_2^{(s)}, p_2^{(s)}), \cdots)\} \in L$ **do**
    $Prod \leftarrow 0$ ▷ Dot product similarity score between s and q
    $M \leftarrow \{\}$ ▷ M contains all the matched spectra peaks between s and q
    for $(m_j^{(s)}, p_j^{(s)}) \in s$ **do**
      $L \leftarrow$ **Search all peaks** $(m_k^{(q)}, p_k^{(q)}) \in q$ **such that** $|m_k^{(q)} - m_j^{(s)}| < tol$
      for each matched peak $(m_k^{(q)}, p_k^{(q)})$ **append** $(k, j, p_k^{(q)} p_j^{(s)})$ to M
    end for
    Sort $M = \{(i, j, v)\}$ **based on descending order of** v
    $Q \leftarrow \{\}$ ▷ A set containing indexes of all peaks from q for the product
    $S \leftarrow \{\}$ ▷ A set containing indexes of all peaks from s for the product
    for $(i, j, v) \in M$ **do**
      if $i \notin Q \ \ \& \ \ j \notin S$ **then**
        $Q.\mathbf{add}(i)$
        $S.\mathbf{add}(j)$
        $Prod \leftarrow Prod + v$
      end if
    end for
    if $Prod \geq thresh$ **then**
      $idx \leftarrow$ **Index of** s **inside dataset** S
      $Res[idx] \leftarrow Prod$
    end if
  end for
  return Res
end function

Algorithm 2 Storing peaks in the indexing-based structure

Input

A spectra library $\mathbf{L} = \{s_1, s_2, \cdots, s_N\}$ such that each spectra contains a list of m/z values m/z tolerance tol for each indexed bin)

Output

An indexing database $D = \{(idx, i, m, p)\}$ such that idx is the index of spectra in L, i is the index of the peak in the spectra, m is the m/z value, p is the normalized peak intensity

Helper Function generating indexing structure for unshifted peaks

procedure UNSHIFTEDINDEXING($L, tol = 0.02$)
    Initialize empty indexing database $D = []$ **in storage**
    $idx \leftarrow 0$
    for $L[idx] \in L$ **do**
        $i \leftarrow 0$
        for $(m_i, p_i) \in L[idx].peaks$ **do**
            $j \leftarrow \mathbf{Round}(m_i \text{ div } tol)$
            if $j \geq D.size()$ **then**
                $D.\mathbf{resize}(j + 1)$
            end if
            $D[j].\mathbf{add}((idx, i, m_i, p_i))$
        end for
    end for
end procedure

Helper Function generating indexing database for shifted peaks

procedure SHIFTEDINDEXING($L, tol = 0.02$)
    Initialize empty indexing structure $D = []$ **in storage**
    $idx \leftarrow 0$
    for $L[idx] \in L$ **do**
        $i \leftarrow 0$
        for $(m_i, p_i) \in L[idx].peaks$ **do**
            $j \leftarrow \mathbf{Round}((m_i - L[idx].precursorMass) \text{ div } tol)$
            if $j \geq D.size()$ **then**
                $D.resize(j + 1)$
            end if
            $D[j].\mathbf{add}((idx, i, m_i, p_i))$
        end for
    end for
end procedure

Algorithm 3 Generating exact matched peaks between query spectra and library spectra

Input

A query spectra $\mathbf{q}$ containing peaks $\{(m_1^{(q)}, p_1^{(q)}), \cdots, (m_K^{(q)}, p_K^{(q)})\}$

A spectra library $\mathbf{L} = \{s_1, s_2, \cdots, s_N\}$

The address of the unshifted indexing database $\mathbf{D}$, generated by Algorithm 2

Exact similarity dot-product score threshold $thresh$ (default set to be 0.7 or 0.9)

m/z tolerance tol for two peaks to be matched (default set to be 0.01 or 0.02)

Output

A dictionary M containing matched peaks between library spectra and the query

function EXACTMATCH($q, L, D, thresh = 0.9, tol = 0.02$)
  $M \leftarrow \{\}$ ▷ A set containing matched peaks for each spectra
  for $(m_i^{(q)}, p_i^{(q)}) \in q$ **do**
    $bin \leftarrow \mathbf{Round}(m_i^{(q)}$ div $tol)$
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D[bin]$ **do**
      $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
    end for
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D[bin-1] \bigcup D[bin+1]$ **do**
      if $|p_k^{(idx)} - p_i^{(q)}| \leq tol$ **then**
        $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
      end if
    end for
  end for
  return M
end function

Algorithm 4 Generating analog matched peaks between query spectra and library spectra

Input

A query spectra $\mathbf{q}$ containing peaks $\{(m_1^{(q)}, p_1^{(q)}), \cdots, (m_K^{(q)}, p_K^{(q)})\}$

A spectra library $\mathbf{L} = \{s_1, s_2, \cdots, s_N\}$

The addresses of unshifted indexing database $\mathbf{D_1}$ and shifted indexing database $\mathbf{D_2}$

Analog similarity dot-product score threshold $thresh$ (default set to be 0.7 or 0.9)

m/z tolerance tol for two peaks to be matched (default set to be 0.01 or 0.02)

Output

A dictionary M containing unshifted and shifted matched peaks between library spectra and the query

function ANALOGMATCH($q, L, D_1, D_2, thresh = 0.9, tol = 0.02$)
  $M \leftarrow \{\}$ ▷ A set containing matched peaks for each spectra
  for $(m_i^{(q)}, p_i^{(q)}) \in q$ **do**
    $bin_1 \leftarrow \mathbf{Round}(m_i^{(q)}$ div $tol)$
    $bin_2 \leftarrow \mathbf{Round}((m_i^{(q)} - q.precursorMass)$ div $tol)$
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D_1[bin]$ **do**
      $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
    end for
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D_2[bin]$ **do**
      $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
    end for
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D_1[bin_1 - 1] \bigcup D_1[bin_1 + 1]$ **do**
      if $|p_k^{(idx)} - p_i^{(q)}| \leq tol$ **then**
        $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
      end if
    end for
    for $(idx, k, m_k^{(idx)}, p_k^{(idx)}) \in D_2[bin_2 - 1] \bigcup D_2[bin_2 + 1]$ **do**
      if $|p_k^{(idx)} - p_i^{(q)}| \leq tol$ **then**
        $M[idx].add(i, k, p_i^{(q)}, p_k^{(idx)})$
      end if
    end for
  end for
  return M
end function

Algorithm 5 Similarity scores between query and library spectra based on matched peaks

Input

A spectra library $\mathbf{L} = \{s_1, s_2, \cdots, s_N\}$

The dictionary M output by Algorithm 3 or Algorithm 4

Output

A mapping Res between idx of matched lib spectra and exact score with the query

function SCORES(M, L)
  $S \leftarrow \{\}$
  for $s = L[idx]$ **do**
    Sort $M[idx] = \{(i, j, v)\}$ **based on descending order of** v
    $Q, S \leftarrow \{\}$ ▷ Sets containing indexes of all peaks from q, s for the product
    for $(i, j, v) \in M$ **do**
      if $i \notin Q$ & $j \notin S$ **then**
        $Q.\mathbf{add}(i), \quad S.\mathbf{add}(j)$
        $Prod \leftarrow Prod + v$
      end if
    end for
    if $Prod \geq thresh$ **then**
      $Res[idx] \leftarrow Prod$
    end if
  end for
  return Res
end function

Bibliography

[1] Theodore Alexandrov and Jan Hendrik Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27 (13):i230–i238, 2011. 2.1

[2] Rikard Alm, Peter Johansson, Karin Hjernø, Cecilia Emanuelsson, Markus Ringner, and Jari Häkkinen. Detection and identification of protein isoforms using cluster analysis of maldi- ms mass spectra. *Journal of Proteome Research*, 5(4):785–792, 2006. 2.1

[3] Paul G Arnison, Mervyn J Bibb, Gabriele Bierbaum, Albert A Bowers, Tim S Bugni, Grzegorz Bulaj, Julio A Camarero, Dominic J Campopiano, Gregory L Challis, Jon Clardy, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports*, 30 (1):108–160, 2013. 4.3

[4] Allegra T Aron, Emily C Gentry, Kerry L McPhail, Louis-Félix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, Julia M Gauglitz, Nicole Sikora, Fernando Vargas, et al. Reproducible molecular networking of untargeted mass spectrometry data using gnps. *Nature protocols*, 15(6):1954–1991, 2020. 5.2

[5] Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: Advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021. 1

[6] Nuno Bandeira, Dekel Tsur, Ari Frank, and Pavel A Pevzner. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences*, 104(15): 6140–6145, 2007. 1, 2.1, 4.2

[7] Ilan Beer, Eilon Barnea, Tamar Ziv, and Arie Admon. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, 4(4):950–960, 2004. 2.1

[8] Wout Bittremieux, Nicole E Avalon, Sydney P Thomas, Sarvar A Kakhkhorov, Alexander A Aksenov, Paulo Wender P Gomes, Christine M Aceves, Andrés Mauricio Caraballo Rodríguez, Julia M Gauglitz, William H Gerwick, et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *BioRxiv*, pages 2022–05, 2022. 4.2

[9] Liu Cao, Mustafa Guler, Azat Tagirdzhanov, Yi-Yuan Lee, Alexey Gurevich, and Hosein Mohimani. Moldiscovery: Learning mass spectrometry fragmentation of small molecules. *Nature communications*, 12(1):3718, 2021. 2.1

[10] Fengju Chen, Darshan S Chandrashekar, Sooryanarayana Varambally, and Chad J

Creighton. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nature communications*, 10(1):5679, 2019. 2.1

[11] Julie Courraud, Madeleine Ernst, Susan Svane Laursen, David M Hougaard, and Arieh S Cohen. Studying autism using untargeted metabolomics in newborn screening samples. *Journal of Molecular Neuroscience*, 71:1378–1393, 2021. 1

[12] Robertson Craig, JC Cortens, David Fenyo, and Ronald C Beavis. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research*, 5(8):1843–1849, 2006. 2.2

[13] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013. 2.3

[14] Tobias Depke, Janne Gesine Thöming, Adrian Kordes, Susanne Häussler, and Mark Brönstrup. Untargeted lc-ms metabolomics differentiates between virulent and avirulent clinical strains of pseudomonas aeruginosa. *Biomolecules*, 10(7):1041, 2020. 1

[15] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4):462–471, 2021. 5.2

[16] Katherine R Duncan, Max Crüsemann, Anna Lechner, Anindita Sarkar, Jie Li, Nadine Ziemert, Mingxun Wang, Nuno Bandeira, Bradley S Moore, Pieter C Dorrestein, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from salinispora species. *Chemistry & biology*, 22 (4):460–471, 2015. 2.3

[17] Fanny E Eberhard, Sven Klimpel, Alessandra A Guarneri, and Nicholas J Tobias. Metabolites as predictive biomarkers for trypanosoma cruzi exposure in triatomine bugs. *Computational and Structural Biotechnology Journal*, 19:3051–3057, 2021. 1

[18] Madeleine Ernst, Simon Rogers, Ulrik Lausten-Thomsen, Anders Björkbom, Susan Svane Laursen, Julie Courraud, Anders Børglum, Merete Nordentoft, Thomas Werge, Preben Bo Mortensen, et al. Gestational age-dependent development of the neonatal metabolome. *Pediatric Research*, 89(6):1396–1404, 2021. 1

[19] Rong Feng, Yasuo Konishi, and Alexander W Bell. High accuracy molecular weight determination and variation characterization of proteins up to 80 ku by ionspray mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 2(5):387–401, 1991. 2.1

[20] Geoffrey Férir, Mariya I Petrova, Graciela Andrei, Dana Huskens, Bart Hoorelbeke, Robert Snoeck, Jos Vanderleyden, Jan Balzarini, Stefan Bartoschek, Mark Brönstrup, et al. The lantibiotic peptide labyrinthopeptin a1 demonstrates broad anti-hiv and anti-hsv activity with potential for microbicidal applications. *PloS one*, 8(5):e64010, 2013. 4.3

[21] Ari M Frank, Nuno Bandeira, Zhouxin Shen, Stephen Tanner, Steven P Briggs, Richard D

Smith, and Pavel A Pevzner. Clustering millions of tandem mass spectra. *Journal of proteome research*, 7(01):113–122, 2008. 2.3

[22] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, Nuno Bandeira, Ronald J Moore, Gordon A Anderson, Richard D Smith, and Pavel A Pevzner. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature methods*, 8(7):587–591, 2011. 1, 2.1, 2.3, 3.3.1, 4.2

[23] Barbara E Frewen, Gennifer E Merrihew, Christine C Wu, William Stafford Noble, and Michael J MacCoss. Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry*, 78(16):5678–5684, 2006. 2.2

[24] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–964, 2004. 2.1

[25] Paul R Graves and Timothy AJ Haystead. Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews*, 66(1):39–63, 2002. 1

[26] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Giuseppe Tradigo, and Pierangelo Veltri. A time series approach for clustering mass spectrometry data. *Journal of Computational Science*, 3(5):344–355, 2012. 2.1

[27] Adrian Guthals, Jeramie D Watrous, Pieter C Dorrestein, and Nuno Bandeira. The spectral networks paradigm in high throughput mass spectrometry. *Molecular bioSystems*, 8(10): 2535–2544, 2012. 2.3, 3.3.2

[28] Franziska Hufsky and Sebastian Böcker. Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. *Mass spectrometry reviews*, 36(5):624–633, 2017. 2.1

[29] Marianna Iorio, Oscar Sasso, Sonia I Maffioli, Rosalia Bertorelli, Paolo Monciardini, Margherita Sosio, Fabiola Bonezzi, Maria Summa, Cristina Brunati, Roberta Bordoni, et al. A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity. *ACS chemical biology*, 9(2):398–404, 2014. 4.3

[30] Namrata S Kale, Kenneth Haug, Pablo Conesa, Kalaivani Jayseelan, Pablo Moreno, Philippe Rocca-Serra, Venkata Chandrasekhar Nainala, Rachel A Spicer, Mark Williams, Xuefei Li, et al. Metabolights: an open-access database repository for metabolomics data. *Current protocols in bioinformatics*, 53(1):14–13, 2016. 1, 2.1

[31] Jarmo-Charles J Kalinski, Samantha C Waterworth, Xavier Siwe Noundou, Meesbah Jiwaji, Shirley Parker-Nance, Rui WM Krause, Kerry L McPhail, and Rosemary A Dorrington. Molecular networking reveals two distinct chemotypes in pyrroloiminoquinone-producing tsitsikamma favus sponges. *Marine Drugs*, 17(1):60, 2019. 1

[32] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24 (21):2534–2536, 2008. 2.2

[33] Tobias Kind and Oliver Fiehn. Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry. *Phytochemistry letters*, 21:313–319, 2017. 2.1

[34] Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R Showalter, Masanori Arita, et al. Identification of small molecules using accurate mass ms/ms search. *Mass spectrometry reviews*, 37(4):513–532, 2018. 2.1

[35] Ting-Hao Kuo, Ching-Ting Yang, Hsin-Yuan Chang, Yen-Ping Hsueh, and Cheng-Chih Hsu. Nematode-trapping fungi produce diverse metabolites during predator–prey interaction. *Metabolites*, 10(3):117, 2020. 1

[36] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, Nichole King, Stephen E Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–667, 2007. 2.2

[37] Marcus Ludwig, Markus Fleischauer, Kai Dührkop, Martin A Hoffmann, and Sebastian Böcker. De novo molecular formula annotation and structure elucidation using sirius 4. *Computational Methods and Data Analysis for Metabolomics*, pages 185–207, 2020. 5.2

[38] Andrew C Lybbert, Justin L Williams, Ruma Raghuvanshi, A Daniel Jones, and Robert A Quinn. Mining public mass spectrometry data to characterize the diversity and ubiquity of p. aeruginosa specialized metabolites. *Metabolites*, 10(11):445, 2020. 1

[39] Matthias Mann, Peter Højrup, and Peter Roepstorff. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological mass spectrometry*, 22(6):338–345, 1993. 2.1

[40] Hosein Mohimani, Sangtae Kim, and Pavel A Pevzner. A new approach to evaluating statistical significance of spectral identifications. *Journal of proteome research*, 12(4):1560–1568, 2013. (document), 4.7

[41] Hosein Mohimani, Alexey Gurevich, Alla Mikheenko, Neha Garg, Louis-Felix Nothias, Akihiro Ninomiya, Kentaro Takada, Pieter C Dorrestein, and Pavel A Pevzner. Dereplication of peptidic natural products through database search of mass spectra. *Nature chemical biology*, 13(1):30–37, 2017. 2.1, 4.3

[42] Hosein Mohimani, Alexey Gurevich, Alexander Shlemov, Alla Mikheenko, Anton Korobeynikov, Liu Cao, Egor Shcherbin, Louis-Felix Nothias, Pieter C Dorrestein, and Pavel A Pevzner. Dereplication of microbial metabolites through database search of mass spectra. *Nature communications*, 9(1):4035, 2018. 2.1

[43] Kathrin I Mohr, Carsten Volz, Rolf Jansen, Victor Wray, Judith Hoffmann, Steffen Bernecker, Joachim Wink, Klaus Gerth, Marc Stadler, and Rolf Müller. Pinensins: the first antifungal lantibiotics. *Angewandte Chemie International Edition*, 54(38):11254–11258, 2015. 4.3

[44] Don D Nguyen, Alexey V Melnik, Nobuhiro Koyama, Xiaowen Lu, Michelle Schorn, Jinshu Fang, Kristen Aguinaldo, Tommie L Lincecum, Maarten GK Ghequire, Victor J Carrion, et al. Indexing the pseudomonas specialized metabolome enabled the discovery of poaeamide b and the bananamides. *Nature microbiology*, 2(1):1–10, 2016. 1

[45] Louis-Félix Nothias, Mélissa Nothias-Esposito, Ricardo Da Silva, Mingxun Wang, Ivan Protsyuk, Zheng Zhang, Abi Sarvepalli, Pieter Leyssen, David Touboul, Jean Costa, et al.

Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *Journal of natural products*, 81(4):758–767, 2018. 2.3

[46] Louis-Félix Nothias, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, Madeleine Ernst, Hiroshi Tsugawa, Markus Fleischauer, et al. Feature-based molecular networking in the gnps analysis environment. *Nature methods*, 17 (9):905–908, 2020. 2.3, 5.2

[47] Florent Olivon, Fanny Roussi, Marc Litaudon, and David Touboul. Optimized experimental workflow for tandem mass spectrometry molecular networking in metabolomics. *Analytical and bioanalytical chemistry*, 409:5767–5778, 2017. 2.3

[48] Bruno S Paulo, Renata Sigrist, Célio FF Angolini, and Luciana G De Oliveira. New cyclodepsipeptide derivatives revealed by genome mining and molecular networking. *ChemistrySelect*, 4(27):7785–7790, 2019. 2.3

[49] Leonardo Perez De Souza, Saleh Alseekh, Yariv Brotman, and Alisdair R Fernie. Network-based strategies in metabolomics data analysis and interpretation: From molecular networking to biological interpretation. *Expert Review of Proteomics*, 17(4):243–255, 2020. 2.3

[50] Daniel Petras, Jeremiah J Minich, Emily Kunselman, Mingxun Wang, Margot E White, Eric E Allen, Lihini I Aluwihare, and Pieter C Dorrestein. Non-targeted metabolomics enables the prioritization and tracking of anthropogenic pollutants in coastal seawater. 2020. 1, 2.1

[51] Shi Qiu, Ying Cai, Hong Yao, Chunsheng Lin, Yiqiang Xie, Songqi Tang, and Aihua Zhang. Small molecule metabolites: discovery of biomarkers and therapeutic targets. *Signal Transduction and Targeted Therapy*, 8(1):132, 2023. 1

[52] Robert A Quinn, Louis-Felix Nothias, Oliver Vining, Michael Meehan, Eduardo Esquenazi, and Pieter C Dorrestein. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends in pharmacological sciences*, 38(2):143–154, 2017. 2.3

[53] Robert A Quinn, Alexey V Melnik, Alison Vrbanac, Ting Fu, Kathryn A Patras, Mitchell P Christy, Zsolt Bodai, Pedro Belda-Ferre, Anupriya Tripathi, Lawton K Chung, et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*, 579(7797): 123–129, 2020. 1, 5.2

[54] Dotsha J Raheem, Ahmed F Tawfike, Usama R Abdelmohsen, RuAngelie Edrada-Ebel, and Vera Fitzsimmons-Thoss. Application of metabolomics and molecular networking in investigating the chemical profile and antitrypanosomal activity of british bluebells (hyacinthoides non-scripta). *Scientific reports*, 9(1):2547, 2019. 1

[55] Alexander E Fox Ramos, Laurent Evanno, Erwan Poupon, Pierre Champy, and Mehdi A Beniddir. Natural products targeting strategies involving molecular networking: Different manners, one goal. *Natural product reports*, 36(7):960–980, 2019. 1, 5.2

[56] Lukas Reiter, Manfred Claassen, Sabine P Schrimpf, Marko Jovanovic, Alexander Schmidt, Joachim M Buhmann, Michael O Hengartner, and Ruedi Aebersold. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & Cellular Proteomics*, 8(11):2405–2417, 2009. 2.1

[57] Robin Schmid, Daniel Petras, Louis-Félix Nothias, Mingxun Wang, Allegra T Aron, Annika Jagels, Hiroshi Tsugawa, Johannes Rainer, Mar Garcia-Aloy, Kai Dührkop, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the gnps environment. *Nature communications*, 12(1):3832, 2021. 2.3

[58] Norbert Schnell, Karl-Dieter Entian, Ursula Schneider, Friedrich Götz, Hans Zähner, Roland Kellner, and Günther Jung. Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. *Nature*, 333(6170):276–278, 1988. 4.3

[59] ThermoFisher Scientific. Overview of mass spectrometry for protein analysis, 2016. 1

[60] Andrej Shevchenko, Ole N Jensen, Alexandre V Podtelejnikov, Francis Sagliocco, Matthias Wilm, Ole Vorm, Peter Mortensen, Anna Shevchenko, Helian Boucherie, and Matthias Mann. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proceedings of the National Academy of Sciences*, 93(25):14440–14445, 1996. 1

[61] Stephen E Stein and Donald R Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994. 2.1

[62] Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, Oliver Fiehn, Richard Higashi, K Sreekumaran Nair, 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 (D1):D463–D470, 2016. 1, 2.1

[63] Kelly Tilleman, Katrien Van Beneden, Aline Dhondt, Ilse Hoffman, Filip De Keyser, Eric Veys, Dirk Elewaut, and Dieter Deforce. Chronically inflamed synovium from spondyloarthropathy and rheumatoid arthritis investigated by protein expression profiling followed by tandem mass spectrometry. *Proteomics*, 5(8):2247–2257, 2005. 2.1

[64] Eric P Trautman, Alan R Healy, Emilee E Shine, Seth B Herzon, and Jason M Crawford. Domain-targeted metabolomics delineates the heterocycle assembly steps of colibactin biosynthesis. *Journal of the American Chemical Society*, 139(11):4195–4201, 2017. 1

[65] Daniela BB Trivella and Rafael de Felicio. The tripod for bacterial natural product discovery: genome mining, silent pathway induction, and mass spectrometry-based molecular networking. *MSystems*, 3(2):10–1128, 2018. 2.3

[66] Justin JJ van Der Hooft, Hosein Mohimani, Anelize Bauermeister, Pieter C Dorrestein, Katherine R Duncan, and Marnix H Medema. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews*, 49(11):3297–3314, 2020. 5.2

[67] Maria I Vizcaino, Philipp Engel, Eric Trautman, and Jason M Crawford. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *Journal of the American Chemical Society*, 136(26):9244–9247, 2014. 1

[68] Mark C Walker, Sara M Eslami, Kenton J Hetrick, Sarah E Ackenhusen, Douglas A Mitchell, and Wilfred A Van Der Donk. Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC genomics*, 21:1–17, 2020. 4.3

[69] Le-heng Wang, De-Quan Li, Yan Fu, Hai-Peng Wang, Jing-Fen Zhang, Zuo-Fei Yuan, Rui-Xiang Sun, Rong Zeng, Si-Min He, and Wen Gao. pfind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 21(18):2985–2991, 2007. 2.1

[70] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016. 1, 2.1, 2.2, 2.3

[71] Mingxun Wang, Alan K Jarmusch, Fernando Vargas, Alexander A Aksenov, Julia M Gauglitz, Kelly Weldon, Daniel Petras, Ricardo da Silva, Robert Quinn, Alexey V Melnik, et al. Mass spectrometry searches using masst. *Nature biotechnology*, 38(1):23–26, 2020. 1, 2.2

[72] Elizabeth J Want, Grace O'Maille, Colin A Smith, Theodore R Brandon, Wilasinee Uritboonthai, Chuan Qin, Sunia A Trauger, and Gary Siuzdak. Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Analytical chemistry*, 78(3):743–752, 2006. 2.1

[73] Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S Heath, Jane Y Yang, Roland D Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M Raaijmakers, et al. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26):E1743–E1752, 2012. 2.2, 2.3, 5.2

[74] Sunmin Woo, Kyo Bin Kang, Jinwoong Kim, and Sang Hyun Sung. Molecular networking reveals the chemical diversity of selaginellin derivatives, natural phosphodiesterase-4 inhibitors from selaginella tamariscina. *Journal of natural products*, 82(7):1820–1830, 2019. 1

[75] Jane Y Yang, Laura M Sanchez, Christopher M Rath, Xueting Liu, Paul D Boudreau, Nicole Bruns, Evgenia Glukhov, Anne Wodtke, Rafael De Felicio, Amanda Fenner, et al. Molecular networking as a dereplication strategy. *Journal of natural products*, 76(9):1686–1699, 2013. 2.1, 5.2

[76] John R Yates III. Database searching using mass spectrometry data. *Electrophoresis*, 19 (6):893–900, 1998. 2.1

[77] Ding Ye, Yan Fu, Rui-Xiang Sun, Hai-Peng Wang, Zuo-Fei Yuan, Hao Chi, and Si-Min He. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics*, 26(12):i399–i406, 06 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq185. URL `https://doi.org/10.1093/bioinformatics/btq185`. 2.2

[78] Ding Ye, Yan Fu, Rui-Xiang Sun, Hai-Peng Wang, Zuo-Fei Yuan, Hao Chi, and Si-Min

He. Open ms/ms spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics*, 26(12):i399–i406, 2010. 2.2