nature biotechnology

Brief Communication

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

Fast mass spectrometry search and clustering of untargeted metabolomics data

In the format provided by the authors and unedited

Supplementary Information: Fast Mass Spectrometry Search and Clustering of Untargeted Metabolomics Data

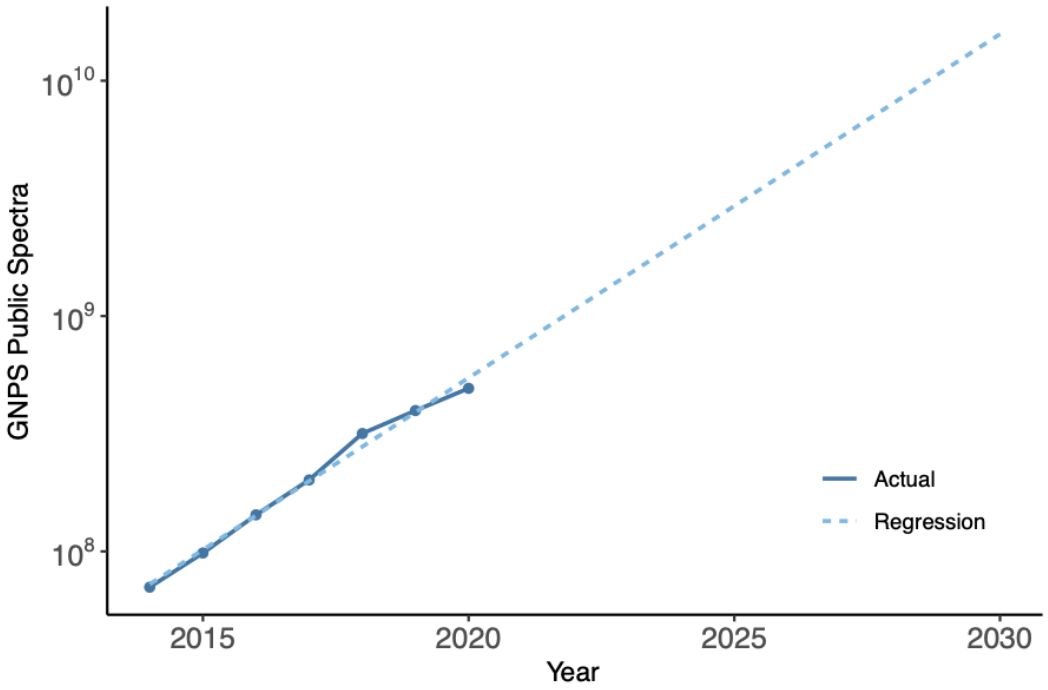
Mihir Mongia[*1], Tyler M. Yasaka[*1], Yudong Liu[*1], Mustafa Guler[1], Liang Lu[1], Aditya Bhagwat[1], Bahar Behsaz[1], Mingxun Wang[2], Pieter C. Dorrestein[3,4], and Hosein Mohimani[1]

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

2 Computer Science and Engineering, University of California Riverside

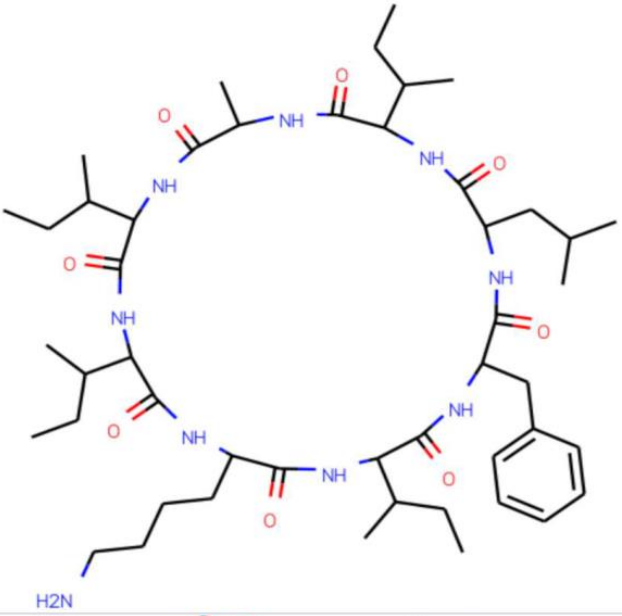
[3]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego

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



Supplementary Figure S1. 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.

(a)

Spectrum ID	CCMSLIB00000579271
Compound Name	Surugamide A
Structure	
NPAtlas/MIBiG	
External Chemical Resources	View All External Resources
Metabolomics Spectrum Resolver	mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00000579271
Precursor M/Z	912.628
Exact Mass	912.628
Charge	1
Adduct	M+H
Ion Source	LC-ESI
Instrument	Orbitrap
Ion Mode	Positive
Provenance LCMS File Viewer	View Provenance Full LCMS File
Spectrum Provenance Input File	Download Provenance File
Download Library Spectrum Peaks	Download Spectrum Peaks
MASST Spectrum Links	MASST+
Library Membership	GNPS-LIBRARY

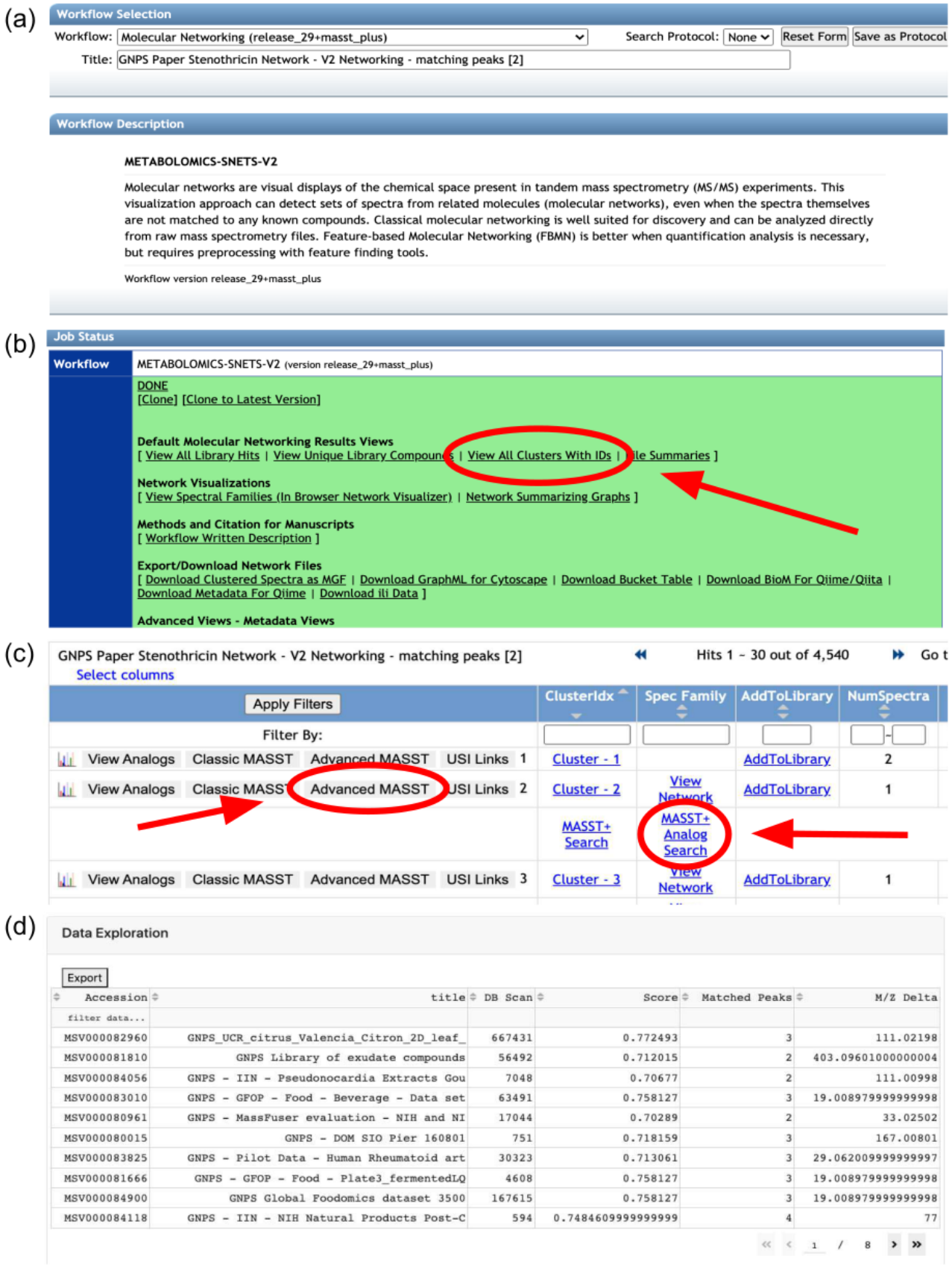
(b)

Data Exploration

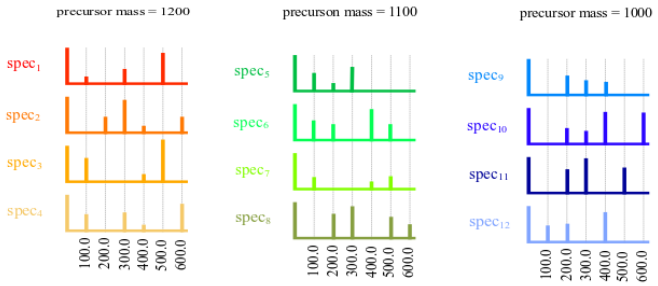
Export

Accession	title	DB Scan	Score	Matched Peaks	M/Z Delta
filter data...					
MSV000080559	GNPS - Amycolatopsis Extracts	10770	0.903961	29	0.00103
MSV000079519	GNPS-Streptomyces albus J1074	3368	0.7338279999999999	33	0.01702
MSV000079502	GNPS-Streptomyces albus J1074	980	0.828615	32	0.00001
MSV000079517	GNPS-Streptomyces albus J1074_13C6-label	963	0.806639	35	0.00099

Supplementary Figure S2. MASST+ search of spectra in the GNPS spectral library. (a) For each spectrum in the GNPS spectral library, users can search against the entire MassIVE repository by clicking the "MASST+" link. (b) Within a few seconds, MASST+ retrieves the search results. Surugamide is shown to be produced by *Streptomyces albus* [12].



Supplementary Figure S3. MASST+ has been integrated into molecular networking and clusters identified by molecular networking can be searched against the entire MassIVE library using MASST+. (a) Spectral dataset is analyzed using molecular networking. (b) All clustered spectra are retrieved. (c) Each cluster can be searched using MASST+ in exact or analog search mode. (d) MASST+ results are retrieved in a few seconds.



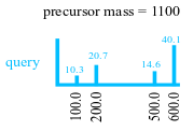
(a) database spectra

(precursor,fragment)	(SpecIndex, Intensity)
(1000,100.0)	{(1,10.6),(3,31.4),(4,22.4)}
(1000,200.0)	{(2,21.6)}
(1000,300.0)	{(1,20.1),(2,42.7),(4,24.6)}
(1000,400.0)	{(2,10.3),(3,11.2),(4,9.2)}
(1000,500.0)	{(1,40.2),(3,52.4)}
(1000,600.0)	{(2,21.6),(4,35.4)}

(precursor, fragment)	(SpecIndex, Intensity)
(1100,100.0)	{(5,31.4),(6,31.4),(7,10.3)}
(1100,200.0)	{(5,10.2),(6,29.4),(8,38.1)}
(1100,300.0)	{(5,38.1),(8,45.1)}
(1100,400.0)	{(6,45.1),(7,9.2)}
(1100,500.0)	{(6,20.1),(7,10.3),(8,18.8)}
(1100,600.0)	{(8,10.3)}

(precursor,fragment)	(SpecIndex, Intensity)
(1200,100.0)	{(9,20.4)}
(1200,200.0)	{(9,30.6),(10,20.1),(11,35.1),(12,22,4)}
(1200,300.0)	{(9,17.1),(10,15.1),(11,50.2)}
(1200,400.0)	{(9,12.3),(10,49.2),(12,35.3)}
(1200,500.0)	{(11,36.4)}
(1200,600.0)	{(10,46.6)}

(b) constructing the index table for each precursor mass



Δ = 1200.0 - 1100.0

100.0 + Δ {(2,21.6)}
200.0 + Δ {(1,20.1),(2,42.7),(4,24.6)}
500.0 + Δ {(2,21.6),(4,35.4)}

Δ = 1100.0 - 1100.0

100.0 + Δ {(5,31.4),(6,31.4),(7,10.3)}
200.0 + Δ {(5,10.2),(6,29.4),(8,38.1)}
500.0 + Δ {(6,20.1),(7,10.3),(8,18.8)}
600.0 + Δ {(8,10.3)}

Δ = 1000.0 - 1100.0

200.0 + Δ {(9,20.4)}
500.0 + Δ {(9,12.3),(10,49.2),(12,35.3)}
600.0 + Δ {(11,36.4)}

(c) retrieve rows from indexing table corresponding to shared and shifted peaks of query spectrum (only shifted peaks shown)

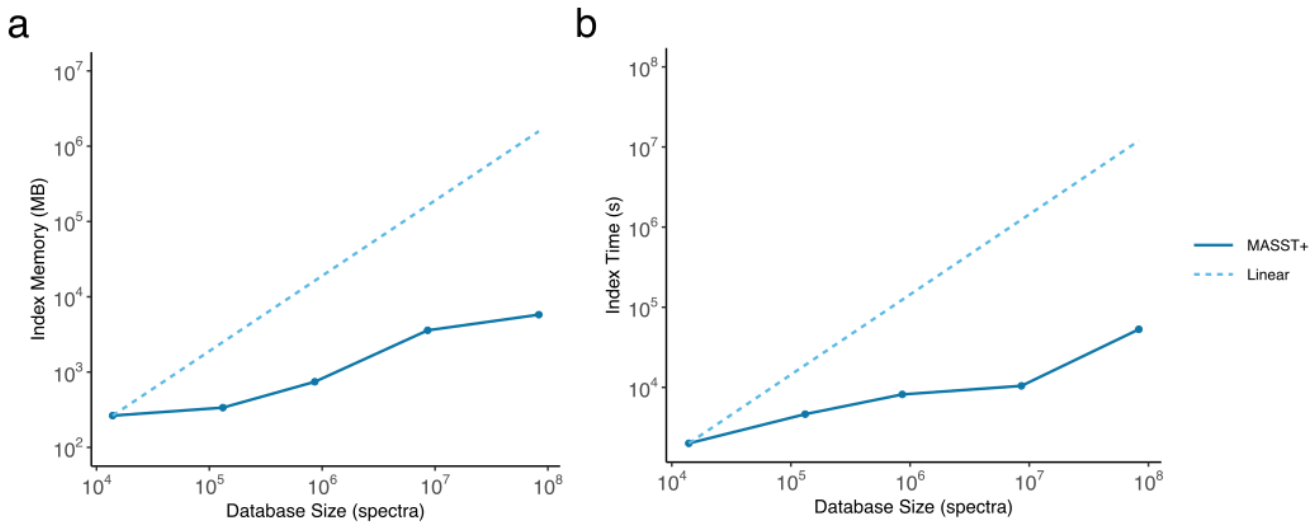
	process 100.0 + Δ		process 200.0 + Δ		process 500.0 + Δ		process 600.0 + Δ		score
ShiftedScore(query,spec1) =		+	20.1 * 20.7	+		+		=	416.07
ShiftedScore(query,spec2) =	21.6 * 10.3	+	42.7 * 20.7	+	21.6 * 14.6	+		=	1419.57
ShiftedScore(query,spec3) =		+		+		+		=	0
ShiftedScore(query,spec4) =		+	24.6 * 20.7	+	35.4 * 14.6	+		=	1025.06

	process 100.0 + Δ		process 200.0 + Δ		process 500.0 + Δ		process 600.0 + Δ		
ShiftedScore(query,spec6) =	31.4 * 10.3	+	29.4 * 20.7	+	20.1 * 14.6	+		=	1225.46
ShiftedScore(query,spec5) =	31.4 * 10.3	+	10.2 * 20.7	+		+		=	543.56
ShiftedScore(query,spec7) =	10.3 * 10.3	+		+	10.3 * 14.6	+		=	256.47
ShiftedScore(query,spec8) =		+	38.1 * 20.7	+	10.3 * 14.6	+	10.3 * 40.1	=	1352.08

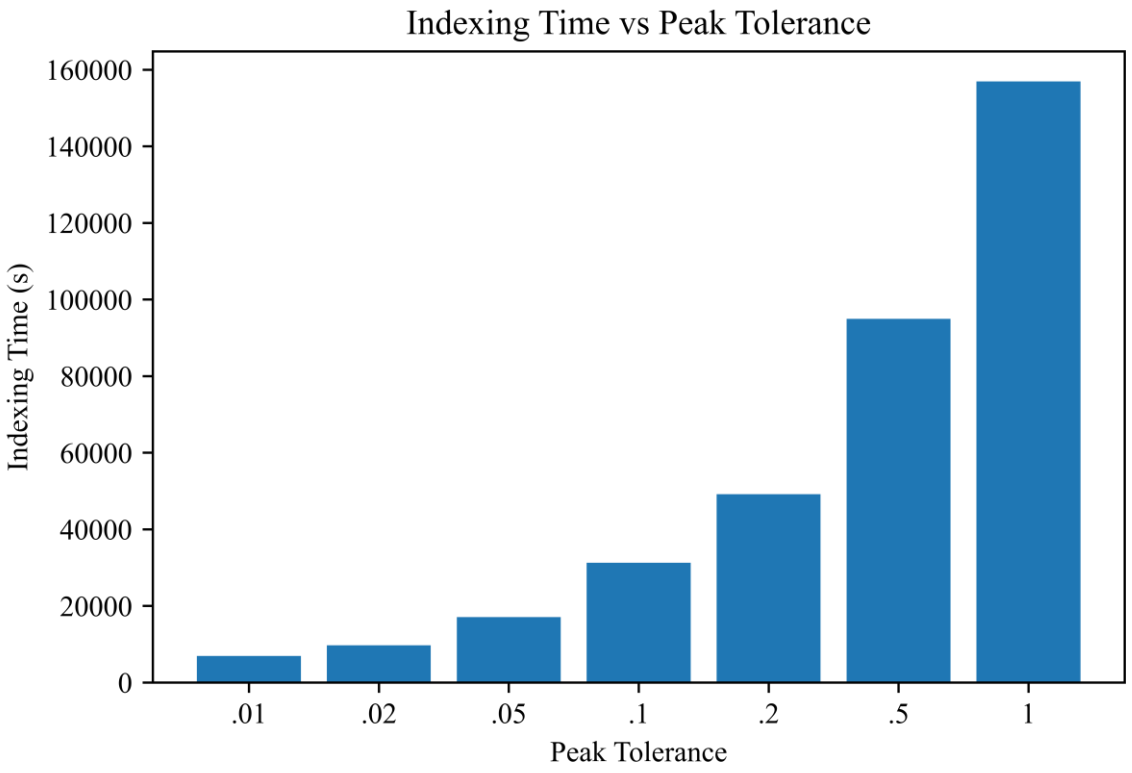
	process 200.0 + Δ		process 500.0 + Δ		process 600.0 + Δ		
ShiftedScore(query,spec9) =	20.4 * 20.7	+	12.3 * 14.6	+	0	=	601.86
ShiftedScore(query,spec10) =	0	+	49.2 * 14.6	+	0	=	718.32
ShiftedScore(query,spec11) =	0	+	0	+	36.4 * 40.1	=	1459.64
ShiftedScore(query,spec12) =	0	+	35.3 * 14.6	+	0	=	515.38

(d) Computing shared and shifted scores (only shifted scores shown)

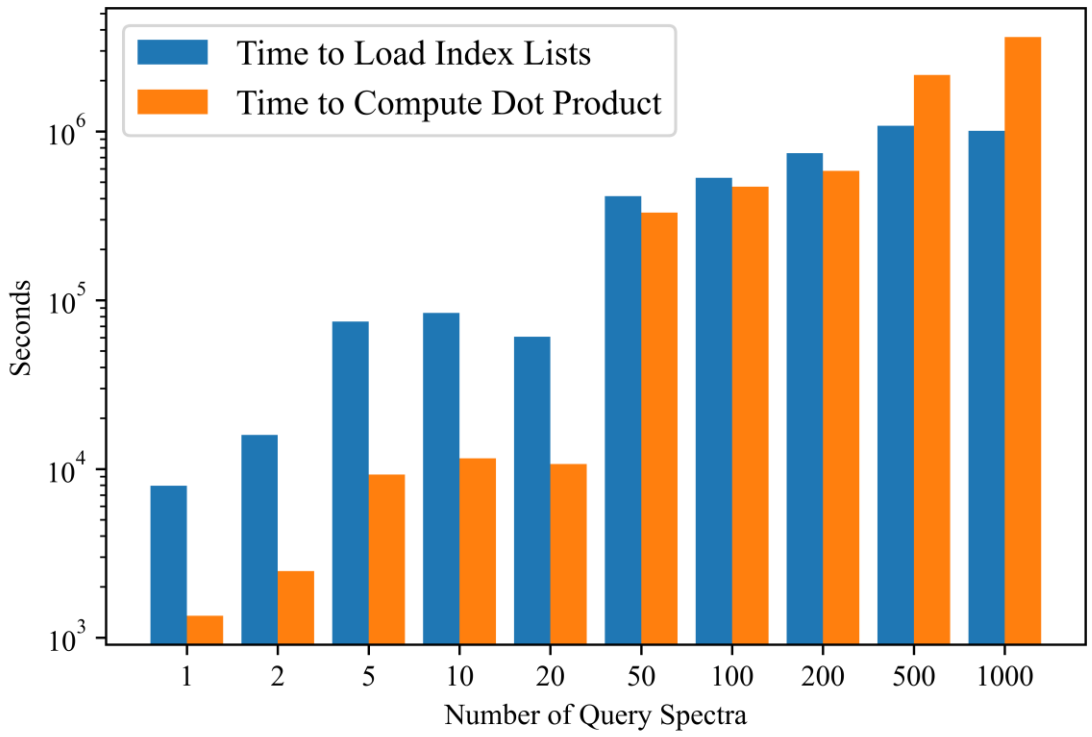
Supplementary Figure S4: MASST+ analog search. Given (a) a reference mass spectral database, MASST+ (b) constructs multiple indexing tables, each corresponding to spectra with a certain precursor mass. Each indexing table contains rows corresponding to a specific precursor mass, and each row contains a list of spectra in which a specific fragment peak is present, along with the intensity of the fragment peak in those spectra. Then (c) given a query spectrum, MASST+ iterates through each indexing table, and retrieves rows correspond to query peaks and Δ-shifted version of query peaks, where Δ is the precursor mass difference between the query spectrum and the indexing table. Then (d) MASST+ computes the ShiftedScore by adding up the product of the intensity of query peaks and database peaks.



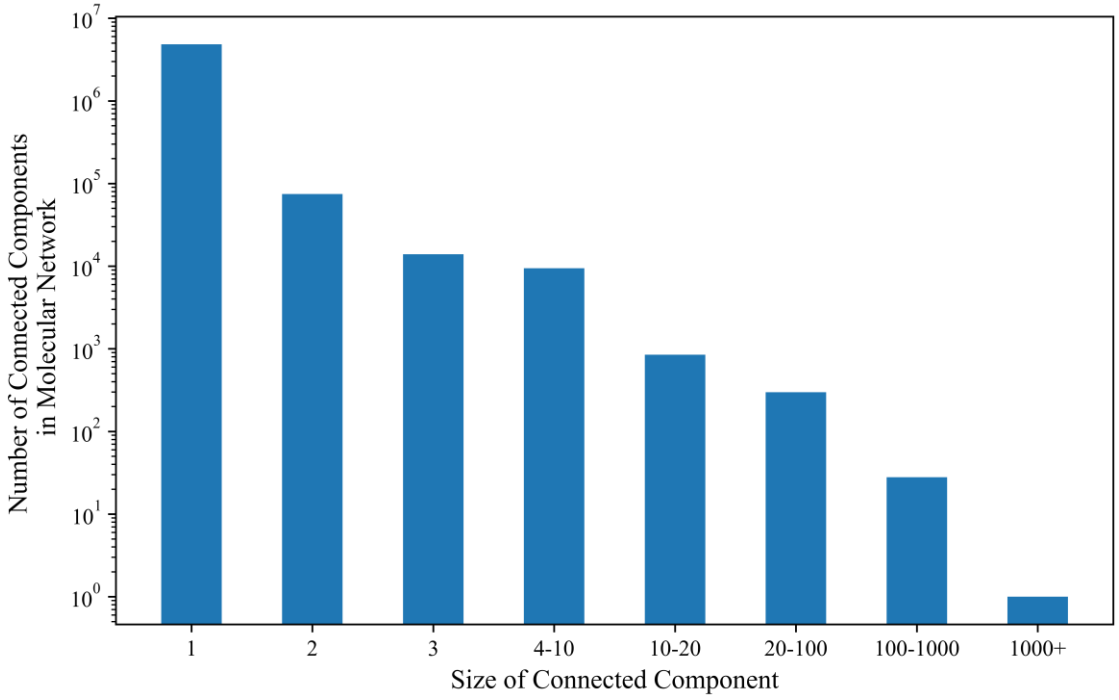
Supplementary Figure S5. MASST + indexing memory (a) and run time (b) as database size grows. Both runtime and memory grow sub-linearly (linear growth shown on dashed line). On the clustered GNPS, MASST+ requires eight hours of and eight gigabytes of memory. Note that indexing needs only to be performed once for each database.



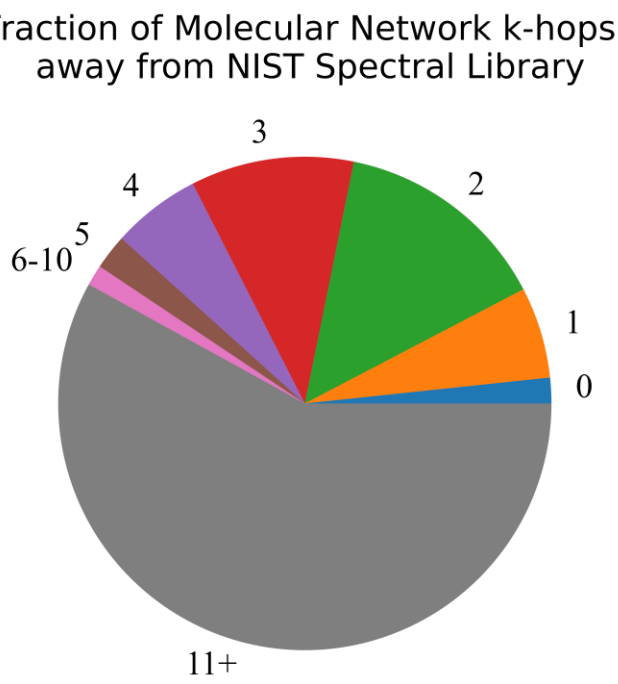
Supplementary Figure S6. Time required for indexing 1 million spectra for various values of peak tolerance. Indexing time increases monotonically with respect to peak tolerance.



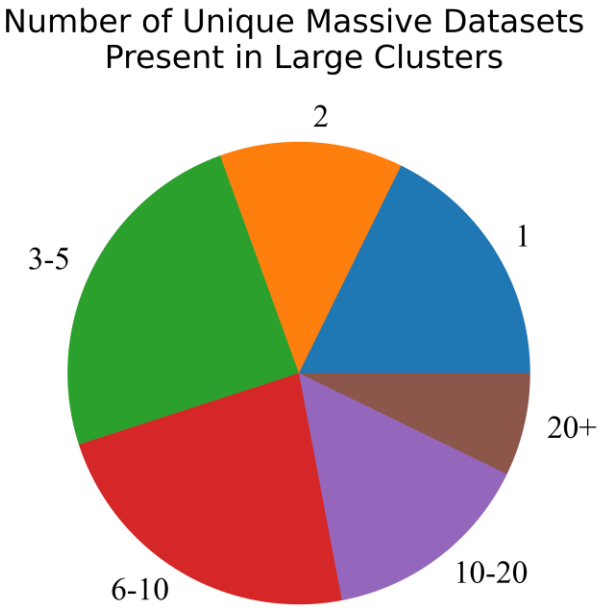
Supplementary Figure S7. Time required to load lists from index and compute dot product for various numbers of query spectra. Fraction of time devoted to loading decreases for larger numbers of query spectra.



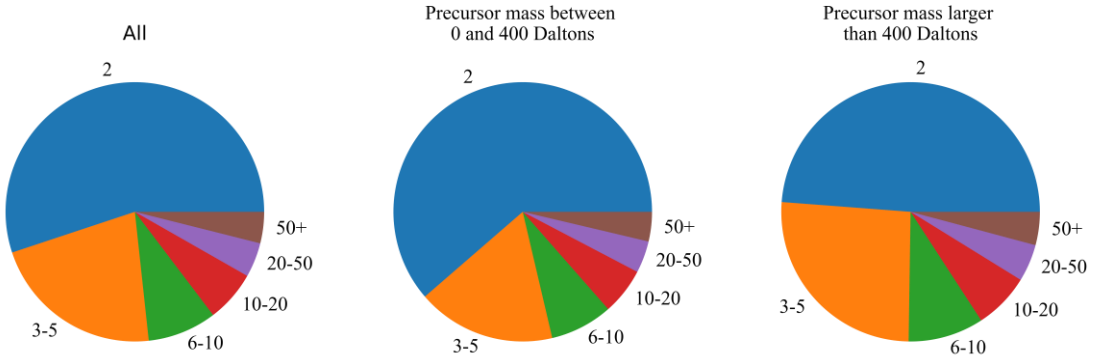
Supplementary Figure S8. Number of size-k connected components existing in the molecular network for all values of k. Largest connected component contains 4656 spectra/clusters.



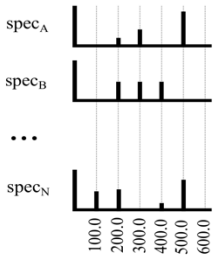
Supplementary Figure S9. Fraction of nodes/clusters in the molecular network that 0,1,2 … 11+ edges away from clusters that are exact matches to spectra in the NIST spectral library. More than one-third of the nodes are within 10 edges of exact NIST spectral library matches.



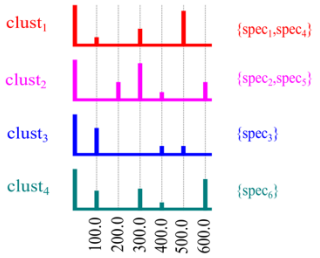
Supplementary Figure S10. Fraction of large clusters (consisting of more than 20 GNPS spectra), that contain spectra coming from 1,2,3-5,6-10,10-20, and 20+ unique MASSIVE datasets.



Supplementary Figure S11. 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.



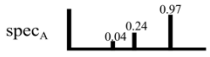
(a) Spectra



(b) Clusters and cluster centers

Peak	(SpecIndex, Intensity)
100.0	{(1,0.23),(3,0.89),(4,0.45)}
200.0	{(2,0.40)}
300.0	{(1,0.43),(2,0.80),(4,0.50)}
400.0	{(2,0.19),(3,0.32),(4,0.19)}
500.0	{(1,0.87),(3,0.32)}
600.0	{(2,0.40),(4,0.72)}

(c) Indexing table

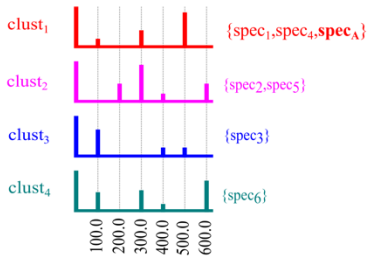


200.0 {(2,0.40)}
300.0 {(1,0.43),(2,0.80),(4,0.50)}
500.0 {(1,0.87),(3,0.32)}

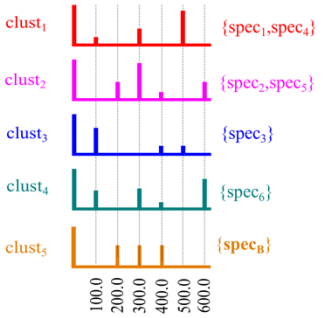
(d) Retrieving lists corresponding to each peak.

	process 200.0		process 300.0		process 500.0		score
Score(spec$_7$,clust$_1$) =	0	+	0.43 * 0.24	+	0.87 * 0.97	=	0.95
Score(spec$_7$,clust$_2$) =	0.40 * 0.04	+	0.80 * 0.24	+	0	=	0.21
Score(spec$_7$,clust$_3$) =	0	+	0	+	0.32 * 0.97	=	0.31
Score(spec$_7$,clust$_4$) =	0	+	0.50 * 0.24	+	0	=	0.12

(e) Calculate ExactScore



(f) Adding spectrum to existing cluster.

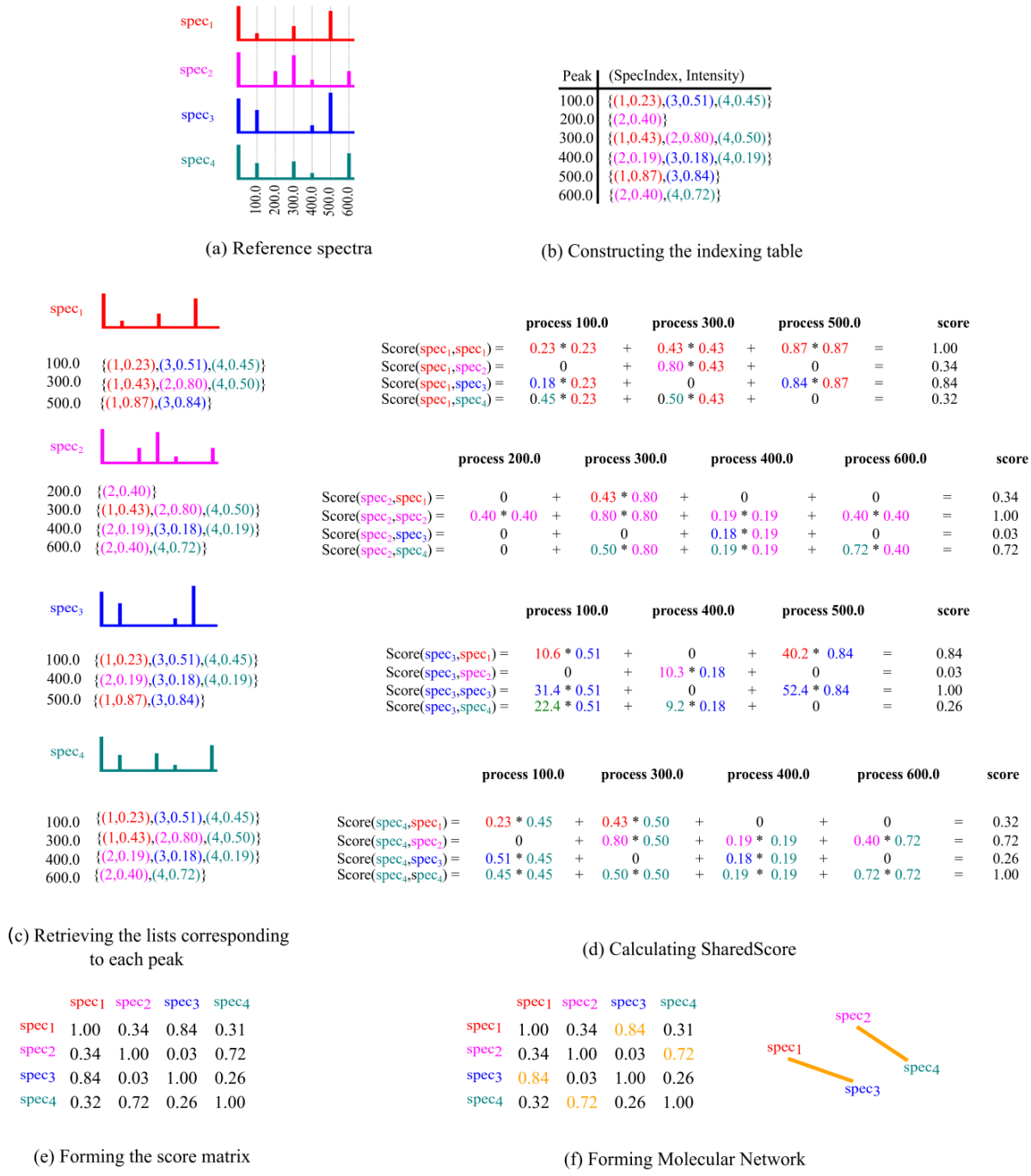


(g) If no similar cluster found, create new cluster. Center of new cluster is set to spectrum.

Peak	(SpecIndex, Intensity)
100.0	{(1,0.23),(3,0.89),(4,0.45)}
200.0	{(2,0.40),(5,0.57
300.0	{(1,0.43),(2,0.80),(4,0.50),(5,.057)}
400.0	{(2,0.19),(3,0.32),(4,0.19),(5,.057)}
500.0	{(1,0.87),(3,0.32)}
600.0	{(2,0.40),(4,0.72)}

(h) Updating the indexing table.

Supplementary Figure S12. One iteration of Clustering+. Given (a) spectral dataset of N spectra, Clustering+ performs the following steps N times. Starting from (b) the current set of clusters and their centers, (c) an indexing table is constructed. Then, (d) a spectrum that has yet to be associated with a cluster is chosen and lists corresponding to each of the peaks are retrieved from the indexing table. Then (e), the ExactScore between query spectra and each cluster center is calculated by adding up the product of intensity of peaks in query and the cluster. (f) If the maximum score is higher than a threshold, the spectrum is added to the highest scoring cluster. (g) Otherwise, a new cluster is created, where the spectrum is set to be the new cluster's center. Lastly, (h) the indexing table is updated.



Supplementary Figure S13. Pairing+. Starting from (a) reference spectra, Pairing+ first (b) constructs the indexing table. Then (c) retrieves the lists corresponding to each peak from the indexing table, and (d) pairwise ShiftedScore between each pair of spectra are calculated. Here, only shared peaks are highlighted for simplicity. Then (e) a similarity matrix is formed between all pairs of spectra, and (f) molecular network is formed by retaining all the pairs with ShiftedScore exceeding a threshold (e. g. 0.7).

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

* Bold font indicates the proposed methods.

Supplementary Table S1. 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	Dataset (Size)	clustering time	Clustering memory	#clusters	networking time	Networking memory

Molecular Networking	MSV000078787 (219,915)	321 sec	662Mb	5,288	8 sec	1224Kb
Networking +	**MSV000078787 (219,915)**	**27 sec**	**992Kb**	**5,288**	**.25 sec**	**996Kb**
Molecular Networking	Entire GNPS	N/A	N/A	N/A	N/A	N/A
Networking +	**Entire GNPS**	**60 hours**	**93Gb**	**8,453,822**	**97 hours**	**23Gb**

* Bold font indicates the proposed methods.

Supplementary Table S2. 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.

Dataset size	Clustering+	MS-Clustering
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

Supplementary Table S3. Benchmarking Clustering+ and MS-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	Pairing+	Spectral Networking
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

Supplementary Table S4. Benchmarking 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	Networking+	Molecular Networking
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

Supplementary Table S5. 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.

MassIVE ID	**#strain**	**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

Supplementary Table S6. List of MassIVE datasets mined for lanthipeptides.

Tag length	#nodes	overlap
6	32449	4008
7	20474	2472
8	12957	1544
9	8337	1013
10	5400	647
11	3607	441
12	2353	285
13	1560	193
14	1077	135
15	717	86
16	506	60
17	344	39

Supplementary Table S7. The number of nodes across different sequence tag lengths in the network of microbial datasets with known genomes from GNPS. The second column shows the total number of nodes, while the third columns show the nodes that possess H2O, NH3 and amino acid mass difference network motifs.

Organism	name	Sequence	score	p-value	Mass	Walker et al.	reference
Streptomyces rimosus NRRL WC-3904	CHM-1793	DT-18GHCS-18GVCT-18VLVCT-18VAVC	21	2.50E-36	1793.77	YN	This paper
Streptomyces albus NRRL F-5917	CHM-1731	YS-18QVCS-18IVVCNT-18VVICG	19	5.80E-33	1731.81	YN	This paper
Streptomyces lavenduligriseus NRRL ISP-5487	SapT	YT-18QGCS-18GLCT-18IVICAT-18VVICG	18	1.40E-32	2030.95	YN	Kodani et al.[41]
Streptomyces species NRRL S-	CHM-1911	S-18TAGCS-18GLCT-18IIVCAT-18VVICA	17	5.20E-31	1911.91	YN	This paper

340							
Streptomyces pathocidini NRRL B-24287	CHM-2168	IT-18S-18IS-18YCT-18PGCT-18SDGGGS-18GCS-18HCC	16	1.60E-26	2168.76	YY	This paper
Streptomyces moroccanus NRRL B-24548	CHM-2182	IT-18S-18IS-18YCT-18PGCT-18SEGGGS-18GCS-18HCC	15	2.00E-25	2182.78	YY	This paper
Streptomyces cinerochromogenes NBRC 13822	CHM-1974	YT-18EGCS-18GLCT-18ILVCAT-18VVIC	13	9.10E-24	1974.91	NN	This paper
Streptomyces hygroscopicus NRRL ISP-5087	CHM-1354	MT-18QVCPVT-18SWHC	13	3.60E-23	1354.56	YN	This paper
Streptomyces rimosus NRRL WC-3874	CHM-1831	PSRSSSPGSFPPGST-18PS-18APS-18	14	1.60E-21	1831.85	NN	This paper
Streptomyces albus NBRC 13041	CHM-1775	YS-18QVCS-18IVICNT-18VVICS	11	5.50E-20	1775.84	NN	This paper
Streptomyces kanamyceticus NBRC 13414	CHM-1748	IS-18GEES-18CFRT-18CT-18TCS-18LC	12	3.40E-19	1748.68	YN	This paper
Streptomyces sulphureus NRRL B-2195	CHM-2229	TEGGGGDS-18SGCS-18GVCT-18IVVCT-18VIVC	9	1.10E-17	2229.95	YN	This paper
Streptomyces anulatus NBRC 12853	AmfS	T-18GS-18QVS-18LLVCEYSS-18LSVVLCTP	11	2.10E-17	2212.09	YN	Ueda et al.[42]
Streptomyces anulatus NBRC 13369	CHM-1669	C-34LPEPFP+16TATT-18RVGCD	11	9.50E-17	1669.78	YN	This paper
Streptomyces paludis JCM 33019	CHM-1635	S-18GEES-18CFRT-18CT-18T-18CSLC	11	2.30E-16	1635.59	YN	This paper

Streptomyces anulatus NBRC 12861	CHM-2433	CRPPSASLCIT-18SDRS-18S-18TGRYLSM	11	3.10E-16	2433.14	NN	This paper
Streptomyces brasiliensis NBRC 101283	Amfs analog	TGS-18QVS-18VLVCEYS-18S-18LSVVLCTP	11	7.10E-16	2198.08	YN	Ueda et al.[42]

Supplementary Table S8. 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[46]. This is a one-sided p-value, where adjustment was made for multiple comparisons.