De Novo Peptide Sequencing Reveals a Vast Cyclopeptidome in Human Gut and Other Environments

Bahar Behsaz[1], Hosein Mohimani[2], Alexey Gurevich[3], Andrey Prjibelski[3], Mark F. Fisher[4], Larry Smarr[2,5,6], Pieter C. Dorrestein[6,7], Joshua S. Mylne[4], and Pavel A. Pevzner[2,3,6]

[1] Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, USA

[2] Department of Computer Science and Engineering, University of California San Diego, La Jolla, USA

[3] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St Petersburg, Russia

[4] The University of Western Australia, School of Molecular Sciences and The ARC Centre of Excellence in Plant Energy Biology, Crawley, Australia

[5] California Institute for Telecommunications and Information Technology, University of California San Diego, La Jolla, USA[6]

[6] Center for Microbiome Innovation, University of California at San Diego, La Jolla, USA

[7] Department of Pharmacology, University of California at San Diego, La Jolla, USA

ABSTRACT

Cyclic and branch cyclic peptides (cyclopeptides) represent an important class of bioactive natural products that include many antibiotics and anti-tumor compounds. However, little is known about cyclopeptides in the human gut, despite the fact that humans are constantly exposed to them. To address this bottleneck, we developed the CycloNovo algorithm for *de novo* cyclopeptide sequencing that employs de Bruijn graphs, the workhorse of DNA sequencing algorithms. CycloNovo reconstructed many new cyclopeptides that we validated with transcriptome, metagenome, and genome mining analyses. Our benchmarking revealed a vast hidden cyclopeptidome in the human gut and other environments and suggested that CycloNovo offers a much-needed step-change for cyclopeptide discovery. Furthermore, CycloNovo revealed a wealth of anti-microbial cyclopeptides from food that survive the complete human gastrointestinal tract, raising the question of how these cyclopeptides might affect the human microbiome.

SIGNIFICANCE

The golden age of antibiotics was followed by a decline in the pace of antibiotics discovery in the 1990s. The key prerequisite for the resurgence of antibiotics research is the development of a computational discovery pipeline for antibiotics sequencing. We describe such pipeline for cyclic and branch cyclic peptides (cyclopeptides) that represent an important class of bioactive natural products such as antibiotics and anti-tumor compounds. Our CycloNovo algorithm for cyclopeptide sequencing reconstructed many new cyclopeptides that we validated with transcriptome, metagenome, and genome mining analyses. CycloNovo revealed a wealth of anti-microbial cyclopeptides from food that survive the complete human gastrointestinal tract, raising the question of how these cyclopeptides might affect the human microbiome.

INTRODUCTION

The golden age of antibiotics was followed by a decline in the pace of antibiotics discovery in the 1990s. However, antibiotics and other natural products are again at the center of attention as exemplified by the recent discovery of teixobactin[1]. The key prerequisite for the resurgence of antibiotics research is the development of computational discovery pipelines[2] such as the Global Natural Products Social (GNPS) molecular networking[3], Dereplicator[4], and VarQuest[5]. The GNPS project already accumulated over a billion mass spectra, an untapped resource for discovery of new antibiotics. However, new algorithms are needed for the GNPS project to realize its promise for antibiotics discovery. Currently, the GNPS network is mainly used for identification of previously discovered natural products and their analogs, emphasizing the need for algorithms for discovery of novel natural products.

This study focuses on *cyclopeptides*, an important class of bioactive natural products with an unparalleled track record in pharmacology: many antibiotics as well as anti-tumor agents, immunosuppressors, and toxins are cyclopeptides. Cyclopeptide sequencing from tandem mass spectra is challenging as their propensity to break at all pairs of points in their cyclic backbone gives a far more complex series of ions than in linear peptides. Cyclopeptides are divided into cyclic *Non-Ribosomal Peptides* (*NRPs*) and cyclic *Ribosomally synthesized and Posttranslationally modified Peptides* (*RiPPs*). NRPs are built from 300 different naturally occurring amino acids according to the complex *non-ribosomal code*[6] rather than the genetic code. RiPPs are encoded using the genetic code and so built from the twenty proteinogenic amino acids, which however are subjected to numerous post-translational modifications.

The discovery of the cyclopeptide gramicidin S in 1942 (first antibiotic used for treating soldiers during the World War II) led to two Nobel prizes and has been followed by the discovery of ≈400 families of cyclopeptides (*cyclofamilies*) in the last 75 years[5]. A relatively small number of known cyclofamilies reflects the experimental and computational challenges in cyclopeptide discovery. Moreover, the question of how many cyclofamilies remained below the radar of previous studies (even though their spectra have already been deposited to public databases!) remains open.

To answer this question, we consider the problem of recognizing *cyclospectra* (tandem mass spectra that originated from cyclopeptides) that can be matched against biosynthetic genes using various genome mining and peptidogenomics tools[7,8]. These tools typically generate a huge database of putative cyclopeptides, making it prohibitively time-consuming to search large spectral datasets against such databases. Fast algorithms for recognizing cyclospectra are critical for genome mining as they greatly reduce the set of spectra that need to be matched against databases of putative cyclopeptides.

Bandeira et al.[9] introduced the concept of *spectral networks* that reveal the spectra of related peptides without knowing their amino acid sequences. Nodes in a spectral network correspond to spectra while edges connect *spectral pairs*, i.e. spectra of peptides differing by a single modification or a mutation (see Supplementary Note "Surugamide spectral network.") Ideally, each connected component of a spectral network corresponds to a cyclofamily[10] representing a set of similar cyclopeptides. Although spectral networks of various GNPS datasets have become the workhorse of the cyclopeptide studies, they typically contain false positive edges that make analysis of cyclofamilies challenging[3]. Moreover, constructing the spectral network of all GNPS spectra remains an open algorithmic problem.

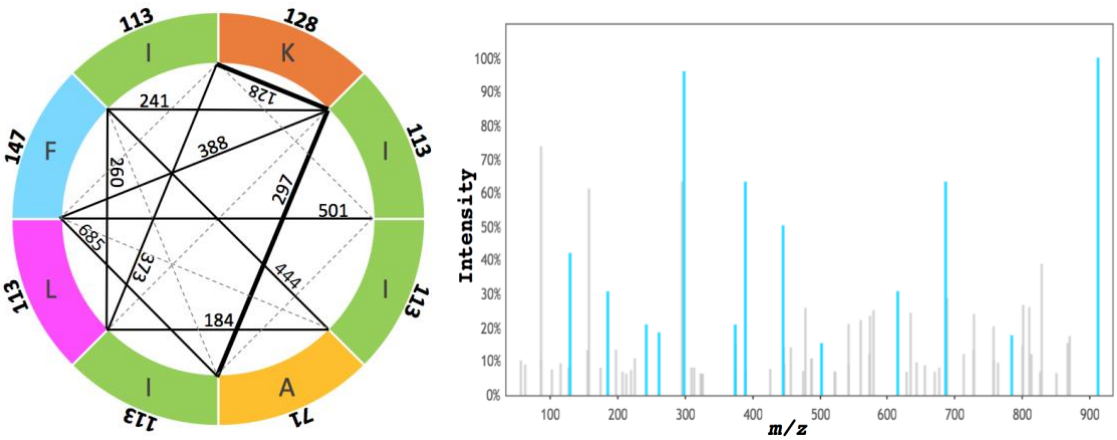
Recognition of cyclospectra creates a possibility to construct a small *cyclospectral sub-network* of the entire GNPS network and to evaluate the number of cyclofamilies in the GNPS network as the number of connected components in this sub-network. This analysis revealed that many cyclopeptides evaded detection in previous studies and that the known cyclopeptides represent the tip of the iceberg of cyclopeptides that are waiting to be decoded from the GNPS network.

We distinguish between *cyclopeptide identification* (identifying cyclopeptides by matching their spectra against databases of known cyclopeptides[11]) and *de novo cyclopeptide sequencing* (determining the cyclopeptide sequence from a spectrum alone). Although recent studies have made progress towards cyclopeptide identification[4,5,12,13], the previously developed cyclopeptide sequencing algorithms[14–17] are rarely used[11] because they are rather inaccurate, too slow for analyzing high-throughput spectral datasets, and cannot distinguish cyclopeptides from other compounds.

Here we describe CycloNovo, a fast cyclopeptide sequencing algorithm based on the concept of the *de Bruijn graph* of a spectrum, a compact representation of putative *k*-mers (strings formed by *k* consecutive amino acids) in an unknown cyclopeptide (see Supplementary Note "An example of the de Bruijn graph.") Although de Bruijn graphs represent the workhorse of DNA sequencing[18], they have not previously been applied to cyclopeptide sequencing. We demonstrate that CycloNovo enables high-throughput analysis of cyclopeptides in large spectral datasets and sequence many cyclopeptides in diverse samples that include marine, soil, and human gut bacterial communities.

RESULTS

To illustrate how CycloNovo works, we used a spectrum of the cyclopeptide surugamide A[19] (referred to as surugamide hereon) with the amino acid sequence AIIKIFLI (Figure 1).



TheoreticalSpectrum (*Surugamide*)	71 113 **128** 147 **184** 226 **241** **260** **297** 354 **373** **388** 410 425 **444** 467 486 **501** 523 538 557 **614** 651 670 **685** 727 764 **783** 798 840 **911**
Spectrum$_{Surugamide}$	55 61 85 85 85 101 114 126 **128** 128 154 156 173 **184** 196 205 211 218 224 **241** **260** 294 294 **297** 297 308 312 321 323 325 373 **373** 373 **388** 388 425 **444** 445 445 455 473 477 486 486 **501** 520 521 541 541 559 572 573 578 578 **614** 627 629 633 642 654 669 675 **685** 687 712 726 727 756 756 763 **783** 799 800 809 812 825 828 849 866 869 **911**

Figure 1. Theoretical and experimental spectra of surugamide. (Top left) Diagram of the surugamide from a marine *Streptomyces* CNQ329 (mass 911.62 Da). Each color represents an amino acid (the numbers on the outer edge are the nominal masses of amino acids in Daltons).

Each chord corresponds to a fragment of surugamide. The solid chords represent the fragments in *TheoreticalSpectrum*(*Surugamide*) whose masses match masses in the experimental spectrum $Spectrum_{Surugamide}$. The numbers on solid chords show the nominal masses of the corresponding fragment, e.g., the chord labeled 297 corresponds to the fragment Ile-Ile-Ala of mass 297 Da. Each chord corresponds either to a single mass *x* or two masses *x* and *mass(Spectrum)-x* (in the latter case the chord is shown in bold). Given a set of fragments with the same mass, we show one of them (arbitrarily chosen) by a solid chord and the others by dashed chords. For example, one of two fragments with the same integer mass 241 (Ile-Lys and Lys-Ile in clockwise order), is shown by a solid chord and another by a dashed chord. (Top right) The experimental spectrum of surugamide ($Spectrum_{Surugamide}$) with 82 peaks (GNPS ID MSV000078839). The *y*-axis in the $Spectrum_{Surugamide}$ shows the ion intensities as the percentage of the intensity of the highest intensity peak. Blue peaks represent masses shared with *TheoreticalSpectrum*(*Surugamide*) for the error threshold $\varepsilon = 0.015$ Da. (Bottom) The theoretical and (pre-processed) experimental spectra for surugamide rounded to the nearest integer (this rounding results in the repetitive integers in the list). Masses in the pre-processed experimental spectrum are reduced by the mass of hydrogen $m_H \approx 1.0078$ Da. A mass in the theoretical spectrum is shared with a mass in the experimental spectrum if they are within the error threshold. The numbers in bold represent 13 shared masses.

Theoretical and experimental spectra. Given an amino acid string, its *mass* is defined as the sum of masses of its amino acids. Given a cyclopeptide *Peptide*, its *theoretical spectrum TheoreticalSpectrum*(*Peptide*) is the set of masses of all substrings of *Peptide* (Figure 1). For example, *TheoreticalSpectrum*(AGCD) contains masses of A, G, C, D, AG, GC, CD, DA, AGC, GCD, CDA, DAG, and AGCD. Note that if multiple fragments have the same mass, they contribute a single mass to the theoretical spectrum.

An *experimental spectrum* is a list of *peak*s, where each peak is characterized by its *intensity* and *m/z* (*m* and *z* represent the mass and the charge of the ion corresponding to the peak). For simplicity, we will represent a pre-processed spectrum as an increasing sequence of numbers $Spectrum=\{s_1, \ldots, s_n\}$, assuming that all peaks in the spectrum have charge 1 and ignoring intensities (see Supplementary Note: "Preprocessing spectra.") We estimate the *PeptideMass* of the cyclopeptide that generated *Spectrum* based on the precursor mass and the charge of *Spectrum*. We define the *symmetric* version of *Spectrum* (denoted *Spectrum**) as a spectrum that, in addition to all masses in *Spectrum*, contains *PeptideMass-s* for each mass *s* in *Spectrum*.

Scoring Peptide-Spectrum Matches. A mass *s* in a (pre-processed) experimental spectrum *Spectrum* matches a mass *s'* in *TheoreticalSpectrum*(*Peptide*) if *s* is "equal" to *s'*. By "equal" we mean "approximately equal" with error below the *error threshold* ε (all default values are specified in the Supplementary Note "CycloNovo parameters.") The score between *Peptide* and *Spectrum* (denoted *score*(*Peptide, Spectrum*) is defined as the number of matches between masses in *Spectrum* and masses in *TheoreticalSpectrum*(*Peptide*). Although CycloNovo uses accurate masses, examples below use nominal masses for simplicity.

Figure 1 illustrates that *score*(*Surugamide*, $Spectrum_{Surugamide}$)=13. For a linear peptide *Peptide*, *score*(*Peptide, Spectrum*) is the number of matches between masses of all linear substrings of *Peptide* and all masses in *Spectrum*. For example, *score*(ILFIK, $Spectrum_{Surugamide}$)=7 because the theoretical spectrum of the linear peptide ILFIK has 7 shared masses with $Spectrum_{Surugamide}$ corresponding to 7 chords within the ILFIK segment in Figure 1. These chords correspond to the following substrings: K (nominal mass 128), IL (226), IK (241), LF (260), LFI (373), LFIK (501), and ILFIK (614).

Cyclopeptidic amino acids. Gurevich et al.[5] combined all currently known peptidic natural products into a single *PNPDatabase*. This database contains 1,257 cyclopeptides (387 cyclofamilies) that we refer to as *CyclopeptideDatabase*. We formed the set of *cyclopeptidic amino acids* (i.e., amino acids that occur in many cyclopeptides) by considering 25 most frequent amino acids in cyclopeptides from *CyclopeptideDatabase* and extending this list to include all proteinogenic amino acids and some common amino acids in RiPPs (see Supplementary Note "Cyclopeptidic amino acids"). 255 out of 1,257 cyclopeptides in the *CyclopeptideDatabase* include only cyclopeptidic amino acids.

Spectral convolution. The *convolution* of a spectrum is the set of all pairwise differences between its masses[15]. Given a mass *a*, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is defined as the number of masses in the convolution equal to *a* (with error up to ε). As shown by Ng et al.[15], the value *convolution*(*Spectrum, a*) is expected to be high if *a* is the mass of an amino acid in a cyclopeptide that gave rise to *Spectrum*. Thus, offsets with high convolutions reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum.

To account for measurement errors, we cluster the masses in the convolution using *single linkage clustering* by combining pairs of masses in a cluster if they are less than ε apart. We define the *cluster mass* as the median mass of its members, and *cluster multiplicity* as the number of elements in the cluster. We call a cluster *cyclopeptidic* if one of its elements is within ε of the mass of a cyclopeptidic amino acid. Since high-multiplicity clusters reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids in an unknown cyclopeptide[15]. See Supplementary Note "Analyzing spectral convolution" for more details.

CycloNovo outline. Given an experimental spectrum *Spectrum*, the Cyclopeptide Sequencing Problem refers to finding a cyclopeptide *Peptide* that maximizes *score*(*Peptide, Spectrum*). Figure 2 illustrates the CycloNovo pipeline for solving this problem:

- ***Recognizing cyclospectra.*** Natural product researchers use *Marfey's analysis* for inferring the amino acid composition of an unknown peptide. However, since Marfey's analysis requires a purified peptide and has a number of limitations[20], we describe its *in silico* alternative for deriving an *approximate* amino acid composition of a cyclopeptide that gave rise to a given spectrum (see Methods section). If applying this approach reveals that a spectrum originated from a cyclopepeptide, we classify it as a cyclospectrum.

- ***Predicting amino acids in a cyclopeptide.*** For each cyclospectrum, CycloNovo predicts the set of putative amino acids in a cyclopeptide that gave rise to this spectrum. CycloNovo considers each cyclopeptidic cluster with multiplicity exceeding the *cyclopeptidic aa threshold* and classifies the cyclopeptidic amino acids corresponding to this cluster as a *putative amino acid* of the cyclopeptide that generated the cyclospectrum. Figure 2 illustrates that CycloNovo classifies amino acids **A, I/L, F, K,** T, W, R, and G as putative amino acids for $Spectrum_{Surugamide}$ (amino acids occurring in suragamide are shown in bold).

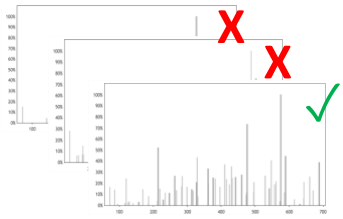
- ***Predicting amino acid composition of a cyclopeptide.*** For each cyclospectrum, CycloNovo uses dynamic programming to find all combinations of putative amino acids with total mass matching the precursor mass of the spectrum. We refer to

each such combination as *putative composition Composition(Spectrum)*, which may include the same amino acid multiple times. Figure 2 illustrates that CycloNovo predicts the following putative compositions for $Spectrum_{Surugamide}$: **$A^1I/L^5K^1F^1$ ($71^1113^5128^1147^1$)**, $I/L^4F^1R^2$ ($113^4147^1156^2$), $A^2T^1K^4R^1$ ($71^2101^1128^4156^1$), and G^1T^1I/L^1K^5 ($57^1101^1113^1128^5$). The putative composition of surugamide with one A, five I/L, one K, and one F is represented as **$A^1I/L^5K^1F^1$** and is shown in bold.

- ***Predicting k-mers in a cyclopeptide.*** For each *Composition(Spectrum*), it analyzes all linear *k*-mers formed by amino acids in this composition (the default value *k*=5) and scores them against *Spectrum** using linear scoring. It assumes that if a Peptide-Spectrum Match has a high score *score*(*Peptide,Spectrum*) (a condition that usually holds for well-fragmented spectra), then each linear *k*-mer in *Peptide* also has high score (for an appropriately chosen *k*). High-scoring *k*-mers (defined as *k*-mers with scores exceeding the *k-mer score threshold*) represent putative *k*-mers in an unknown cyclopeptide. For example, for *Composition*=$71^1113^5128^1147^1$, there exist 4^5=1024 5-mers and CycloNovo identifies 524 of them as high-scoring 5-mers. We refer to the set of high-scoring *k*-mers as $Kmers_{Composition,k}$(*Spectrum*). Figure 2 illustrates that three out of six highest scoring 5-mers for $Spectrum_{Surugamide}$ are correct, i.e., represent 5-mers from surugamide. CycloNovo computes the *k-merScore*, the score of the highest-scoring *k*-mer.

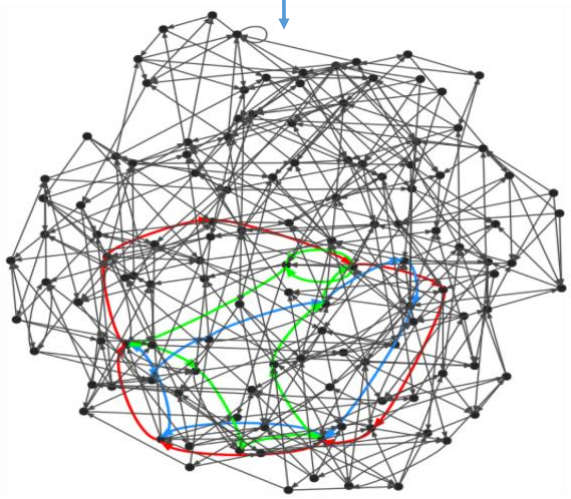
- ***Constructing the de Bruijn graph of a spectrum.*** Given a set *Kmers*= $Kmers_{Composition,k}$(*Spectrum*), CycloNovo constructs the *de Bruijn graph* DB_{Kmers}(*Spectrum*)[18]. Nodes in DB_{Kmers}(*Spectrum*) correspond to all (*k-1*)-mers from *Kmers* and each directed edge corresponds to a *k*-mer from *Kmers* and connects its first (*k-1*)-mer with its last (*k-1*)-mer. Each cycle in DB_{Kmers}(*Spectrum*) spells out a cyclic amino acid sequence. Figure 2 presents the *pruned de Bruijn graph* for the putative composition $71^1113^5128^1147^1$ that is obtained by iterative removal of tips (nodes without outgoing or incoming edges), and single isolated edges from the de Bruijn graph. The composition $113^5128^171^1147^1$ results in a de Bruijn graph with 202 vertices and 524 edges and the pruned de Bruijn graph with 126 vertices and 392 edges (Figure 2).

- ***Generating cyclopeptide reconstructions.*** A cycle in the de Bruijn graph of a spectrum is *feasible* if it spells a cyclopeptide with the mass matching the precursor mass of the spectrum. Using the breadth-first search algorithm, CycloNovo finds all feasible cycles in the de Bruijn graph with length equal to the number of amino acids in *Composition* (a cycle may traverse the same edge multiple times). Each such cycle spells a putative cyclopeptide and CycloNovo scores each of them against *Spectrum*. Finally, it reports the highest scoring cyclopeptides along with the P-values of their Peptide-Spectrum Matches (PSMs) computed using MS-DPR[21].



putative amino acids	113 (I/L)	128 (K)	186 (W)	71 (A)	147 (F)	156 (R)	57 (G)	101 (T)

putative compositions	top-scoring putative *k*-mers		#putative *k*-mers
$113^4156^2147^1$	(156 156 113 113 147) (156 113 147 113 156) (113 147 113 113 113)	(156 113 147 113 113) (113 113 147 156 113) (113 147 113 156 113)	57
$128^471^2156^1101^1$	(101 128 128 128 156) (128 128 101 128 71) (128 128 128 128 71)	(128 128 128 101 71) (128 113 71 113 113) (127 101 128 71 128)	76
$113^5128^171^1147^1$	(113 113 128 113 113) (113 147 113 113 128) (147 113 128 113 71)	(113 147 113 128 113) (113 113 71 113 113) (128 113 71 113 113)	524
$128^5113^1101^157^1$	(57 128 113 128 128) (101 128 113 128 128) (128 113 128 101 57)	(128 128 128 113 57) (128 128 101 128 113) (128 57 128 113 128)	180



sequence of masses of amino acids in a cyclopeptide								P-value
128	113	147	113	113	71	113	113	3.9×10^{-23}
128	113	113	147	113	113	71	113	3.9×10^{-23}
128	113	113	147	113	71	113	113	7.1×10^{-21}

Figure 2. CycloNovo outline illustrated using *Spectrum*$_{Surugamide}$. CycloNovo includes six steps: (1) recognizing cyclospectra, (2) predicting amino acids in a cyclopeptide, (3) predicting amino acid composition of a cyclopeptide, (4), predicting *k*-mers in a cyclopeptide, (5) constructing the de Bruijn graph of a spectrum, and (6) generating cyclopeptide reconstructions. Only six top-scoring putative *k*-mers for each putative amino acid composition are shown. Masses of amino acids occurring in surugamide are shown in red and *k*-mers occurring in surugamide are underlined. To simplify the de Bruijn graph (corresponding to the composition $71^1113^5128^1147^1$), all tips and isolated edges in the graph were removed. Red, blue and green feasible cycles in the graph spell out three cyclopeptides shown in the bottom table along with their P-values. The red cycle spells out surugamide.

In the case of *Spectrum*$_{Surugamide}$, CycloNovo found three similar cyclopeptides (Figure 2) spelled by feasible cycles in the de Bruijn graph with a putative composition

$113^5128^171^1147^1$ (the highest-scoring one corresponds to surugamide). The remaining three putative compositions do not yield feasible cycles in their de Bruijn graphs (Supplementary Note "De Bruijn Graphs for $Spectrum_{Surugamide}$.") CycloNovo sequenced $Spectrum_{Surugamide}$ in ≈3 seconds on a laptop with a single 2.5GHz processor (see Supplementary Note "CycloNovo running time").

Datasets. We analyzed various spectral datasets obtained from diverse bacterial communities (Table 1 and Supplementary Note "Information about spectral datasets"). To benchmark CycloNovo, we also analyzed a plant spectral dataset that had a paired RNA-seq dataset, thus enabling us to validate the CycloNovo reconstructions by matching them against the transcriptome.

The CYCLOLIBRARY dataset contains 81 spectra from 81 distinct cyclopeptides (forming 41 cyclofamilies) that were identified by Dereplicator[4] after searching the GNPS network against *CyclopeptideDatabase*[5].

The S.VULGARIS dataset is generated from a single sample collected from seeds of the plant *Senecio vulgaris* (both medicinal and poisonous)[22] from the *Asteraceae* family. We also analyzed the RNA-Seq reads from the same sample (~74 million 100 bp long Illumina reads)[22], assembled them using rnaSPAdes[23], and used the assembled transcripts (61.9 Mb total length) and prior knowledge of cyclopeptide processing[24–26] to validate the reconstructed cyclopeptides.

The HUMANSTOOL dataset is generated from 65 stool samples of a single person (L.S., co-author of this paper and a contributor to the "Quantified self" initiative) collected over a course of four years. This dataset is accompanied by the detailed medical and food metadata[27] as well as metagenomics reads generated from the same samples (project ID PRJEB24161).

The GNPS dataset is formed by combining forty datasets from GNPS[3]. The $GNPS_{CYANO}$, $GNPS_{PSEUDO}$, and $GNPS_{ACTI}$ datasets represent sub-datasets of the GNPS dataset corresponding to three phyla with extensively analyzed cyclopeptides (*Cyanobacteria*, *Pseudomonas* and *Actinomyces*).

dataset	#spectra	#spectra after preprocessing	#cyclospectra	#distinct cyclopeptides/ cyclofamilies	#known cyclopeptides/ cyclofamilies
CYCLOLIBRARY	81	81	45	45/27	45/27
S.VULGARIS	667	212	23	12/9	4/4
HUMANSTOOL	1,242,178	451,962	703	79/69	7/5
GNPS	51,220,679	27,883,895	12,004	512/213	67/37
$GNPS_{ACTI}$	5,903,921	4,435,893	1,478	116/56	38/24
$GNPS_{CYANO}$	23,582,408	12,118,482	317	74/35	5/4
$GNPS_{PSEUDO}$	697,812	581,012	2,076	120/39	5/2

Table 1. Information about various high-resolution spectral datasets analyzed by CycloNovo. The number of distinct cyclopeptides and cyclofamilies was estimated using MS-Cluster[28] and SpecNets[3], respectively. The last column shows the number of known cyclopeptides/cyclofamilies (identified by Dereplicator) in each dataset. For each identified cyclopeptide in the CYCLOLIBRARY dataset, we selected the PSM with the minimum P-value (among all PSMs for that cyclopeptide), resulting in a spectral dataset CYCLOLIBRARY with 81 spectra.

Analyzing the CYCLOLIBRARY dataset. As the cyclopeptides that gave rise to the spectra in the CYCLOLIBRARY dataset are known, we used this dataset to benchmark CycloNovo. We considered a cyclopeptide/spectrum as correctly sequenced if the sequence of the cyclopeptide appeared among reconstructions with three highest-scores. CycloNovo recognized 45 spectra in the CYCLOLIBRARY dataset as cyclospectra and correctly sequenced 38 of these cyclospectra (see Supplementary Note "CycloNovo analysis of the CYCLOLIBRARY dataset").

Analyzing the S.VULGARIS dataset. 23 recognized cyclospectra in this dataset correspond to twelve distinct cyclopeptides. CycloNovo sequenced ten of them with P-values below 10^{-15} (Table 2). Nine of ten reconstructed cyclopeptides match the assembled transcriptome. One reconstructed cyclopeptide (with the highest-scoring reconstruction AFLLADV and score 22), does not match the assembled transcriptome but a suboptimal ALFLGLD reconstruction with score 20 does (see Supplementary Note "Cyclopeptide-encoding transcripts in the S.VULGARIS dataset").

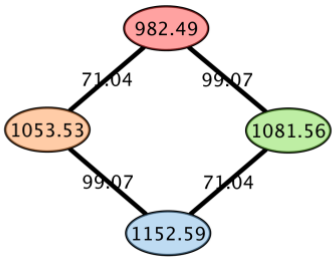
The ten reconstructed cyclopeptides (nine highest-scoring reconstruction and one suboptimal reconstruction) match 11 transcripts (some transcripts encode multiple cyclopeptides and some cyclopeptides are encoded by multiple transcripts) that belong to cyclopeptide-encoding *PawS1-Like* genes in various *Asteraceae* species[22,24]. While three out of eleven identified *PawL1* ORFs and the four cyclopeptides encoded by them (PLP-12 through PLP-15) have been extensively analyzed in recent studies[22,24], the remaining eight ORFs represent previously unknown cyclopeptide-encoding genes in *S. vulgaris*. See Supplementary Note "Cyclopeptide-encoding transcripts in the S.VULGARIS dataset."

precursor mass	sequence matching transcripts	PSM score	highest score	#reconstructions with score ≥ PSM score	P-value	peptide ID	gene
899.36	DNFVDTTGYDRLSDN	24	24	1	1.4×10^{-47}	PLP-14	*Sv_PawL1c*
811.37	DNFVGGTSFDRLSDN	14	14	2	2.4×10^{-24}	PLP-12	*Sv_PawL1c*
803.42	DNTFGVVIADRLSEN	30	30	1	1.2×10^{-61}	PLP-13	*Sv_PawL1b*
762.32	DNGFHGTFDGLDN	13	13	1	3.2×10^{-23}	**PLP-47**	*Sv_PawL1e*
730.41	DNALFLGLDGLDN	20	22	12	2.2×10^{-39}	**PLP-48**	*Sv_PawL1f*
702.38	DNAIFGVVDGLDN	20	20	1	5.6×10^{-36}	**PLP-49**	*Sv_PawL1j*
688.36	DNFVGGVIDGLDN	21	21	1	1.0×10^{-40}	**PLP-50**	*Sv_PawL1g*
674.35	DNGVVVGFDGLDN	14	14	5	1.1×10^{-25}	**PLP-51**	*Sv_PawL1l*
668.40	DNALVVGLDGLDN	14	14	1	1.9×10^{-27}	PLP-15	*Sv_PawL1d* *Sv_PawL1g*
654.39	DNALLGIADGLDN	18	18	5	6.9×10^{-34}	**PLP-52**	Sv_PawL1i

Table 2. Cyclopeptides reconstructed in the S.VULGARIS dataset. Ten reconstructed cyclopeptides (highlighted in yellow) along with their flanking sequences in transcripts translated into amino acids. For each of these cyclopeptides (reconstructed with P-values below 10^{-15}), we selected one representative spectrum with the highest score. The conserved flanking amino acids in the transcripts on the left and right sides of the highlighted cyclopeptides (preceding and succeeding motifs) are shown in red and green, respectively. For nine out of ten cyclopeptides, the reconstruction with the highest score matches one of the transcripts. For the cyclopeptide with mass 730.41 (highlighted in pink), the highest scoring reconstruction AFLLADV (score 22), does not match the assembled transcriptome but a suboptimal ALFLGLD reconstruction (score 20) does. The novel cyclopeptides discovered by CycloNovo are shown with bold IDs and named PLP-47 through PLP-52. For this dataset we used the error threshold ε=0.015 Da as recommended in Fisher *et al* [24].

Analyzing the HUMANSTOOL dataset. A Dereplicator search of the HUMANSTOOL dataset against *CyclopeptideDatabase* identified seven PSMs at 0% False Discovery Rate (FDR), namely an antimicrobial orbitide citrusin V found in various *Citrus* species[29,30] and cyclolinopeptides A[31], B[32], C[33], D[33], H[33], and E[33]. Cyclolinopeptides are bioactive flaxseed orbitides from *Linum usitatissimum*. The individual who provided the HUMANSTOOL sample frequently ate flaxseeds because they contain α-linolenic acids. CycloNovo sequenced six flaxseed cyclopeptides from the *CyclopeptideDatabase* as well cyclolinopeptide P (a recently discovered cyclopeptide[34] that has not been added to *CyclopeptideDatabase* yet) as the highest-scoring reconstructions (Supplementary Note "Cyclopeptides in the HUMANSTOOL dataset.")

In addition to the eight reconstructed orbitides, CycloNovo reconstructed 32 cyclopeptides in the HUMANSTOOL dataset with P-values below 10^{-15} forming 26 cyclofamilies (see Supplementary Notes "Cyclopeptides in the HUMANSTOOL dataset"). Figure 3 shows a connected component in the spectral network formed by four novel cyclopeptides in the HUMANSTOOL dataset and illustrates that CycloNovo reconstructions are consistent with the spectral network.



precursor mass	peptide	PSM score	#reconstructions with score ≥ PSM score	P-value	dates
982.49	SVTFEAPLH	24	1	2.6×10^{-37}	07.14.2014 07.19.2015
1053.53	SVTFEAPLAH	25	1	8.6×10^{-38}	07.14.2014 07.19.2015
1081.56	SVVTFEAPLH	21	1	3.0×10^{-36}	07.14.2014 07.19.2015
1152.59	SVVTFEAPLAH	19	1	2.3×10^{-27}	07.14.2014

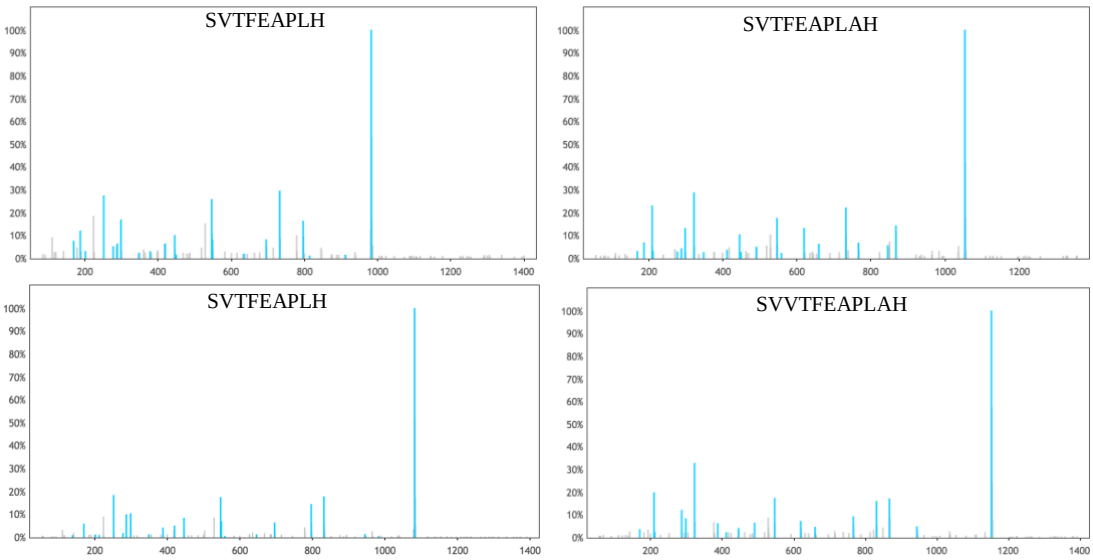


Figure 3. A novel cyclofamily reconstructed by CycloNovo in the HUMANSTOOL dataset. (Top) Four cyclopeptides reconstructed by CycloNovo form a cyclofamily represented by a connected component in the spectral network of the HUMANSTOOL dataset (label "L" stands for one of amino acids L and I). Each node represents a spectrum and two nodes are connected by an edge if their spectral similarity[3] exceeds 0.8. The numbers on the edges show the mass shifts between the corresponding spectra. (Middle) The *de novo* reconstructions corresponding to the four spectra forming the spectral network. For each cyclopeptide, the cyclic sequence of the highest-scoring reconstruction along with their scores, the number of reconstructions with scores larger or equal to the PSM score (column "#reconstructions with score ≥ PSM score"), and P-values are listed. The "dates" column shows the dates when the corresponding samples were taken. Note that the cyclopeptides in this cyclofamily appear on the same dates. (Bottom) The annotated spectra of the four cyclopeptides based on the CycloNovo reconstructions.

The Dereplicator search of all 703 cyclospectra in the HUMANSTOOL dataset against *PNPDatabase* resulted in a single hit and identified a cyclic lipopeptide massetolide F[35] with P-value 7.5×10^{-22}. As this compound includes lipid chains not included in the set of cyclopeptidic amino acids, CycloNovo was not able to generate its full-length reconstructions, but correctly reconstructed its partial amino acid sequence (see Supplementary Note "Cyclopeptides in the HUMANSTOOL dataset.")

Massetolides are non-ribosomal lipopeptides produced by *Pseudomonas fluoresences*, an indigenous member of human and plant microbiota[36,37]. Analysis of the metagenome assembly of reads paired with the HUMANSTOOL dataset confirmed that *P. fluoresences* is present in the stool samples where massetolide F was detected. Therefore, massetolide F most likely originated from *P. fluoresences* in the human microbiome (see Supplementary Note "Cyclopeptides in the HUMANSTOOL dataset" for information about the assemblies).

Analyzing the GNPS dataset. We analyzed all cyclospectra in the GNPS dataset using MS-Cluster[28] and SpecNets[38] with the goal of estimating the number of still unknown cyclopeptides and cyclofamilies originating from spectra already deposited into GNPS. To provide a conservative estimate for the number of cyclopeptides and cyclofamilies, we limited the analysis to clusters with at least three spectra. 12,004 cyclospectra in the GNPS dataset originated from 512 cyclopeptides and 213 cyclofamilies. Dereplicator search of these cyclospectra against *CyclopeptideDatabase* identified only 67 cyclopeptides from 37 cyclofamilies (see Supplementary Note "Cyclopeptides in the GNPS dataset"). For each putative cyclopeptide, we selected a representative spectrum with the highest *k-merScore*, resulting in 512 spectra corresponding to the 512 cyclopeptides. CycloNovo *de novo* sequenced 94 cyclopeptides with P-values below 10^{-15} in this set of 512 cyclospectra (see Supplementary File).

Comparing CycloNovo and Dereplicator. Figure 4 compares the number of distinct cyclopeptides, including some branch-cyclic peptides, (see Supplementary Note "Cyclopeptides in the HUMANSTOOL dataset") and cyclofamilies revealed by CycloNovo and identified by Dereplicator in searches against the *PNPDatabase*. As Figure 4 illustrates, even for the extensively studied phyla of *Cyanobacteria* and *Pseudomonas*, only a small fraction of cyclopeptides and cyclofamilies revealed by CycloNovo are currently known. Moreover, CycloNovo revealed many novel cyclopeptides in known cyclofamilies. For example, CycloNovo reconstructed six novel variants of surugamide by analyzing the $\text{GNPS}_{\text{ACTI}}$ dataset and revealed the widespread proliferation of the recently described *A-domain skipping* phenomenon[5,39], suggesting that it is more prevalent than was previously thought (each A-domain encodes a single amino acid in an NRP according to non-ribosomal code). Genome mining efforts typically rule

out such events due to the consecutive arrangements of A-domains in NRP synthetases (see Supplementary Note "Surugamide spectral network.").

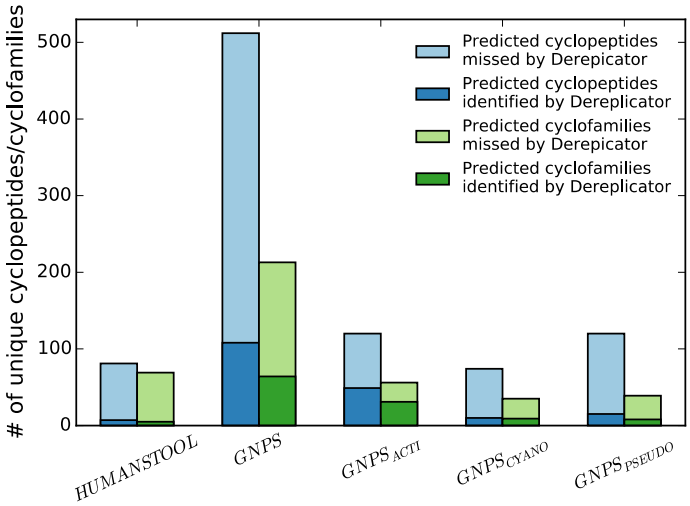


Figure 4. Number of cyclopeptides (blue bars) and cyclofamilies (green bars) predicted by CycloNovo and identified/missed by Dereplicator in various spectral datasets. Missed cyclopeptides/cyclofamilies are not present in *PNPDatabase*.

DISCUSSION

Although the advent of the GNPS molecular network has created a new resource for natural product discovery, there exists a large body of still unknown bioactive compounds represented by various spectra in the GNPS network[4] (less than one percent of GNPS spectra have been identified so far). As the existing database search approaches are limited to identifying known cyclopeptides and their variants, *de novo* cyclopeptide sequencing is needed to reveal the "dark matter of cyclopeptidomics."

Charlop-Powers et al.[40], recently demonstrated that New York urban parks (rather than exotic areas like rainforests or coral reefs) represent an untapped resource for discovery of clinically important non-ribosomal peptides. This surprising discovery illustrated the still unexplored biosynthetic potential of environmental metagenomes but has not revealed the chemical compounds that these metagenomes encode. As Nothias et al.[41], wrote in the follow-up commentary *Antibiotic discovery is a walk in the park*, Charlop-Powers et al., revealed the potential for discovering antibiotics in our backyards but did not answer the question what these molecules are and whether they are actually produced by the microbes. To address these questions, we analyzed the GNPS molecular network (our "digital version" of the New York Central Park that contains mass spectra from many environmental samples) and demonstrated that it contains spectra that originated from hundreds of still unknown cyclopeptides.

Only 81 out of 1,257 known cyclopeptides (42 out of 387 known cyclofamilies) have been identified in the GNPS network[5]. CycloNovo revealed over 400 unknown cyclopeptides from 176 novel cyclofamilies by analyzing only ≈51 million GNPS spectra, illustrating that the currently known cyclopeptides represent just a small fraction of cyclopeptides whose spectra have been already deposited into the GNPS network. CycloNovo correctly sequenced many known cyclopeptides in a blind mode and reconstructed many novel cyclopeptides that were validated using transcriptomics data.

Our analysis of the HUMANSTOOL dataset demonstrates that numerous bioactive cyclopeptides from consumed plants remain stable throughout the proteolytic, absorptive and microbial ecosystem provided by the gastrointestinal system and thus interact with human microbiome. It also found cyclospectra originating from the branch cyclic peptide massetolide produced by an indigenous member of the human microbiota and confirmed by metagenomics analysis. In addition, it revealed a large number of still unknown cyclopeptides in the human gut that are either a part of the human diet or are products of the human gut microbiome.

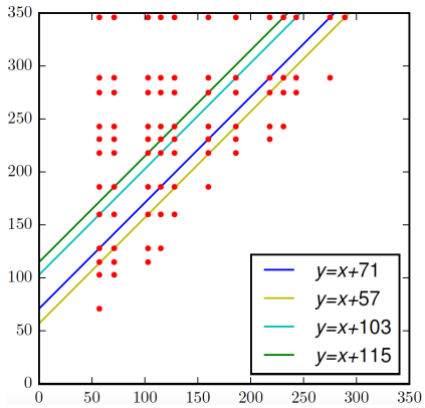
The cyclolinopeptides constitutes the largest identified component of the spectral network of all cyclospectra in the HUMANSTOOL dataset. Controlled diets have demonstrated beneficial effects of flaxseed consumption[32] and revealed that flaxseed is an effective chemo-preventive agent[42]. However, it remains unknown how flaxseed cyclopeptides affect the human microbiome. Previous studies of flaxseed focused on α-linolenic acid[43] and other bioactive compounds[44] affecting the digestive system. However, none of the flaxseed studies have identified what specific flaxseed ingredients are associated with the observed biological outcomes. As discussed in Shim et al.[43], attributing a certain bioactivity to specific flaxseed compounds is a difficult task as multiple compounds are present in various flaxseed fractions.

It is remarkable that cyclolinopeptides remain stable in a proteolytic environment of the human gut and are not degraded. Although the immunosuppressive potential of cyclolinopeptides was established two decades ago[32], little was known about their antimicrobial potential. Recent studies demonstrated significant antimicrobial activities of flaxseed, but it remains unclear which specific compounds are responsible for these activities[45]. As our analysis revealed that cyclolinopeptides survive the human digestive tract, we propose that the antimicrobial activities of flaxseeds might be caused by cyclolinopeptides, complementing their known anti-fungal[46] and anti-malarial[47] activities. Finding antimicrobial cyclopeptides in human stool raises the question of how these bioactive antimicrobial cyclopeptides might affect the human microbiome.

METHODS

Spectral convolution. We represent each spectrum $Spectrum=\{s_1, \ldots, s_n\}$ as its *spectral diagram*, the set of $n\times(n-1)/2$ 2-dimensional points (s_i, s_j) for $1 \le i < j \le n$. Given a mass a, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is equivalently defined as the number of points in the diagonal (45°) band $y \approx x+a$ in the spectral diagram. Figure 5 presents the spectral diagram of *TheoreticalSpectrum*(AGCD) and reveals that bands corresponding to its amino acids (71, 57, 103, and 115 Da) are the most *populous* (contain a large number of points as compared to other bands), i.e., *convolution*(*Spectrum,a*) is high when a is the mass of amino acids A, G, C, or D. For example, *TheoreticalSpectrum*(AGCD) includes five pairs of fragment masses ((G,AG), (D,AD), (AGC,GC), (CD,CDA), and (GCD,AGCD)) that are located on the "blue" diagonal $y = x+mass(A)$ in Figure 5.

The spectral diagrams for *TheoreticalSpectrum*(*Surugamide*) and experimental $Spectrum_{Surugamide}$ highlight four populous diagonal bands $y \approx x+a$, where a is the mass of one of four amino acids in surugamide with integer masses 71, 113, 128, and 147 (Supplementary Note "Analyzing spectral convolution.") These populous bands in the spectral diagram reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum.



cluster mass	cluster multiplicity	cluster distance
113.07	**12**	**0.006 (I/L)**
128.08	**7**	**0.006 (K)**
56.02	6	-
186.10	**6**	**0.029 (W)**
71.04	**5**	**0.004 (A)**
147.07	**4**	**0.001 (F)**
156.10	**4**	**0.002 (R)**
184.11	4	-
98.08	4	-
169.11	4	-
57.05	4	0.027 (G)

Figure 5. The spectral diagram of *TheoreticalSpectrum*(AGCD) (left) and the list of clusters in the convolutions of *Spectrum*$_{Surugamide}$ (right). (Left) The highlighted lines with slope 1 correspond to the masses of the amino acids, A, G, C, and D and contain 5, 9, 5, and 6 points, respectively. (Right) Clusters in the convolutions of *Spectrum*$_{Surugamide}$ in the decreasing order of their multiplicities. Only clusters with masses between 55 and 190 Da and multiplicity exceeding 3 are shown. Cyclopeptidic clusters are shown in bold and cyclopeptidic clusters with masses similar to the masses of amino acids in surugamide are shown in red. *Cluster distance* is defined as the distance between the cluster mass and a closest mass of a cyclopeptidic amino acid.

Figure 5 lists high-multiplicity clusters for *Spectrum*$_{Surugamide}$ (see Supplementary Note "Analyzing spectral convolution") and shows that many of them have masses that are similar to the masses of amino acids in surugamide. Since populous diagonals (high-multiplicity clusters) in the spectral diagram reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids[18].

Recognizing cyclospectra. A cluster in the spectral convolution is called *frequent* if its multiplicity exceeds the *cluster multiplicity threshold* (the default threshold for *Spectrum*$_{Surugamide}$ is 7). CycloNovo classifies a spectrum as a cyclospectrum if the number of frequent cyclopeptidic clusters in its spectral convolution is at least *minNumberFrequentClusters* (the default value *minNumberFrequentClusters*=2). Since there exist two frequent cyclopeptidic clusters for *Spectrum*$_{Surugamide}$ (corresponding to amino acids I/L and K), it is classified as cyclopeptidic (Figure 5). In addition to *Spectrum*$_{Surugamide}$, out of 938 spectra passing the preprocessing step in the small spectral dataset for *Streptomyces CNQ329* that contains *Spectrum*$_{Surugamide}$, CycloNovo recognized only one cyclospectrum, also originated from surugamide. See Supplementary Note "Distinguishing cyclospectra from spectra of linear peptides and polymers" for selecting CycloNovo parameters.

Estimating the number of distinct cyclopeptides and cyclofamilies. Spectral datasets often contain multiple spectra originating from the same compound. CycloNovo clusters similar cyclospectra using MS-Cluster[28] and estimates for the number of distinct cyclopeptides as the number of constructed clusters. It further constructs the spectral network of cyclospectra using SpecNets[3] and estimates for the number of distinct cyclofamilies as the number of connected components in this network.

Acknowledgements. We thank Ben Pullman, Sergey Nurk, Alexey Melnik, and Louis-Felix Nothias for fruitful discussions.

Availability. CycloNovo is available as both a stand-alone tool (https://github.com/bbehsaz/cyclonovo) and a web application (http://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp). All described datasets are available through the corresponding public repositories.

Funding. The work of B.B., H.M. and P.A.P. was supported by the US National Institutes of Health (grant 2-P41-GM103484). The work of B.B. was also supported by the Natural Science and Engineering Research Council of Canada. The work of A.G. and A.P. was supported by the Russian Science Foundation (grant 14-50-00069). J.S.M. was supported in part by an Australian Research Council Future Fellowship (FT120100013). M.F.F. was supported by the Australian Government's Research Training Program and a Bruce and Betty Green Postgraduate Research Scholarship.

Pavel Pevzner is a co-founder, has an equity interest and receives income from Digital Proteomics, LLC. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Contributions. P.A.P. and B.B. designed the CycloNovo algorithm and B.B. implemented it. B.B. did the benchmarking and spectral network analysis for all datasets included in this study. M.F. and J.M. generated the S.VULGARIS spectral dataset and helped with the biological interpretation and validation of the identified cyclopeptides. A.P. assembled RNA-seq data for S.VULGARIS. L.S. contributed the HUMANSTOOL sample set for analysis and commented on the results. H.M and A.G. performed the Dereplicator search for all datasets. P.C.D., H.M. and P.A.P. directed the work. B.B. and P.A.P. wrote the manuscript with contributions from all co-authors.

REFERENCES

1. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517,** 455–459 (2015).
2. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11,** 639–648 (2015).
3. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34,** 828–837 (2016).
4. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13,** 30–37 (2017).
5. Gurevich, A. *et al.* Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **3,** 319–327 (2018).
6. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem. Rev.* **97,** 2651–2674 (1997).
7. Mohimani, H. *et al.* NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *J. Nat. Prod.* **77,** 1902–1909 (2014).
8. Mohimani, H. *et al.* Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9,** 1545–1551 (2014).
9. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 6140–6145 (2007).
10. Mohimani, H., Wei-Ting Liu, Y.-L. Y., Susana P. Gaudêncio, W. F., Dorrestein, P. C. & Pevzner, P. A. Multiplex de novo sequencing of peptide antibiotics. *J. Comput. Biol.* **18,** 1371–1381 (2011).

11. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Nat. Prod. Rep.* **33,** 73–86 (2016).
12. Mohimani, H. *et al.* Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases. *J. Proteome Res.* **10,** 4505–4512 (2011).
13. Ibrahim, A. *et al.* Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 19196–19201 (2012).
14. Mohimani, H. *et al.* Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **11,** 3642–3650 (2011).
15. Ng, J. *et al.* Dereplication and de novo sequencing of nonribosomal peptides. *Nat Methods* **6,** 596–599 (2009).
16. Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V. CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. *Journal of the American Society for Mass Spectrometry* **24,** 1177–1184 (2013).
17. Townsend, C. *et al.* CycLS: Accurate, whole-librarysequencing of cyclic peptides using tandem mass spectrometry. *Bioorg. Med. Chem.* **26,** 1232–1238 (2018).
18. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29,** 987–991 (2011).
19. Takada, K. *et al.* Surugamides A − E, Cyclic Octapeptides with Four D-Amino Acid Residues, from a Marine Streptomyces sp.: LC-MS-Aided Inspection of Partial Hydrolysates for the Distinction of D- and L-Amino Acid Residues in the Sequence. **50,** 3–7 (2013).
20. Bhushan, R. & Bruckner, H. Use of Marfey's reagent and analogs for chiral amino acid analysis: Assessment and applications to natural products and biological systems. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **879,** 3148–3161 (2011).
21. Mohimani, H., Kim, S. & Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research* **12,** 1560–1568 (2013).
22. Jayasena, A. S. *et al.* Stepwise Evolution of a Buried Inhibitor Peptide over 45 My. *Mol. Biol. Evol.* **34,** 1505–1516 (2017).
23. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
24. Fisher, M. F. *et al.* A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch. *Plant Direct* **2,** e00042 (2018).
25. Mylne, J. S. *et al.* Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nat. Chem. Biol.* **7,** 257–259 (2011).
26. Elliott, A. G. *et al.* Evolutionary Origins of a Bioactive Peptide Buried within Preproalbumin. *Plant Cell* **26,** 981–995 (2014).
27. Yazdani, M. *et al.* Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016* 1272–1280 (2016).
28. Frank, A. M. *et al.* Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8,** 587–591 (2011).
29. Noh, H. J. *et al.* Anti-inflammatory activity of a new cyclic peptide, citrusin XI, isolated from the fruits of Citrus unshiu. *J. Ethnopharmacol.* **163,** 106–112 (2015).
30. Belknap, W. R. *et al.* A family of small cyclic amphipathic peptides (SCAmpPs) genes in citrus. *BMC Genomics* **16,** 303 (2015).
31. Kaufmann, H. P. & Tobschirbel, A. Über ein oligopeptid aus leinsamen. *Eur. J.*

Inorg. Chem. **92,** 2805–2809 (1959).
32. Morita, H. *et al.* A new immunosuppressive cyclic nonapeptide, cycloinopeptide B from Linum usitatissimum. *Bioorganic Med. Chem. Lett.* **7,** 1269–1272 (1997).
33. Morita, H., Shishido, A., Matsumoto, T., Itokawa, H. & Takeya, K. Cyclolinopeptides B-E, new cyclic peptides from Linum usitatissimum. *Tetrahedron* **55,** 967–976 (1999).
34. Okinyo-Owiti, D. P., Young, L., Burnett, P. G. G. & Reaney, M. J. T. New flaxseed orbitides: Detection, sequencing, and15N incorporation. *Biopolym. - Pept. Sci. Sect.* **102,** 168–175 (2014).
35. Gerard, J. *et al.* Massetolides A-H, antimycobacterial cyclic depsipeptides produced by two pseudomonads isolated from marine habitats. *J. Nat. Prod.* **60,** 223–229 (1997).
36. Scales, B. S., Dickson, R. P., Lipuma, J. J. & Huffnagle, G. B. Microbiology, genomics, and clinical significance of the Pseudomonas fluorescens species complex, an unappreciated colonizer of humans. *Clin. Microbiol. Rev.* **27,** 927–948 (2014).
37. O'Sullivan, D. J. & O'Gara, F. Traits of fluorescent Pseudomonas spp. involved in suppression of plant root pathogens. *Microbiol. Rev.* **56,** 662–676 (1992).
38. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E1743–E1752 (2012).
39. Nguyen, D. D. *et al.* Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2,** (2016).
40. Charlop-Powers, Z. *et al.* Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc. Natl. Acad. Sci.* (2016).
41. Nothias, L.-F., Knight, R. & Dorrestein, P. C. Antibiotic discovery is a walk in the park. *Proc. Natl. Acad. Sci.* (2016).
42. Bommareddy, A. *et al.* Effects of dietary flaxseed on intestinal tumorigenesis in Apc Min mouse. *Nutr. Cancer* **61,** 276–283 (2009).
43. Shim, Y. Y., Gui, B., Arnison, P. G., Wang, Y. & Reaney, M. J. T. Flaxseed (Linum usitatissimum L.) bioactive compounds and peptide nomenclature: Areview. *Trends in Food Science and Technology* **38,** 5–20 (2014).
44. Adolphe, J. L., Whiting, S. J., Juurlink, B. H. J., Thorpe, L. U. & Alcorn, J. Health effects with consumption of the flax lignan secoisolariciresinol diglucoside. *Br. J. Nutr.* **103,** 929 (2010).
45. Son, H.-J. & Song, K. Bin. Antimicrobial Activity of Flaxseed Meal Extract against Escherichia coli O157: H7 and Staphylococcus aureus Inoculated on Red Mustard. *J. Microbiol. Biotechnol.* **27,** 67--71 (2017).
46. Tian, J. *et al.* Antifungal cyclic peptides from Psammosilene tunicoides. *J. Nat. Prod.* **73,** 1987–1992 (2010).
47. Pinto, M. E. F. *et al.* Ribifolin, an orbitide from jatropha ribifolia, and its potential antimalarial activity. *J. Nat. Prod.* **78,** 374–380 (2015).
48. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** W362-367 (2011).
49. Mingxun, W. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst. Press*
50. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta* **627,** 71–81 (2008).
51. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30,** 108–160 (2013).

SUPPLEMENTARY NOTES

SURUGAMIDE SPECTRAL NETWORK 20

AN EXAMPLE OF THE DE BRUIJN GRAPH 22

PREPROCESSING SPECTRA 22

CYCLONOVO PARAMETERS 22

DISTINGUISHING CYCLOSPECTRA FROM SPECTRA OF LINEAR PEPTIDES AND POLYMERS 23

CYCLOPEPTIDIC AMINO ACIDS 26

ANALYZING SPECTRAL CONVOLUTION 27

DE BRUIJN GRAPHS FOR $SPECTRUM_{SURUGAMIDE}$ 29

CYCLONOVO RUNNING TIME 29

INFORMATION ABOUT SPECTRAL DATASETS 30

CYCLONOVO ANALYSIS OF THE CYCLOLIBRARY DATASET 33

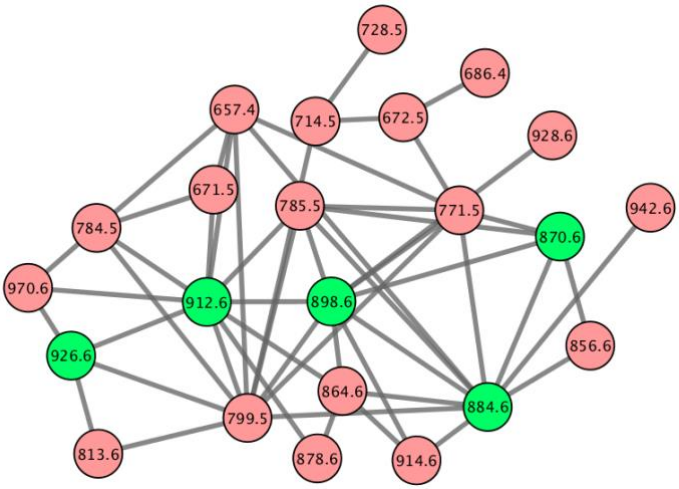
CYCLOPEPTIDE-ENCODING TRANSCRIPTS IN THE S.VULGARIS DATASET 35

CYCLOPEPTIDES IN THE HUMANSTOOL DATASET 36

CYCLOPEPTIDES IN THE GNPS DATASET 38

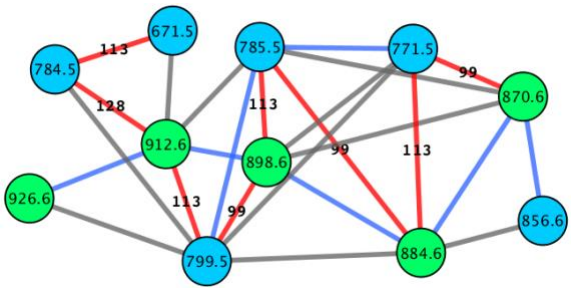
Supplementary Note: Surugamide spectral network

Supplementary Figure S1 shows a connected component in the spectral network containing known and novel surugamide variants (spectral dataset $GNPS_{ACTI}$ generated from samples collected from various *Actinomyces* species).



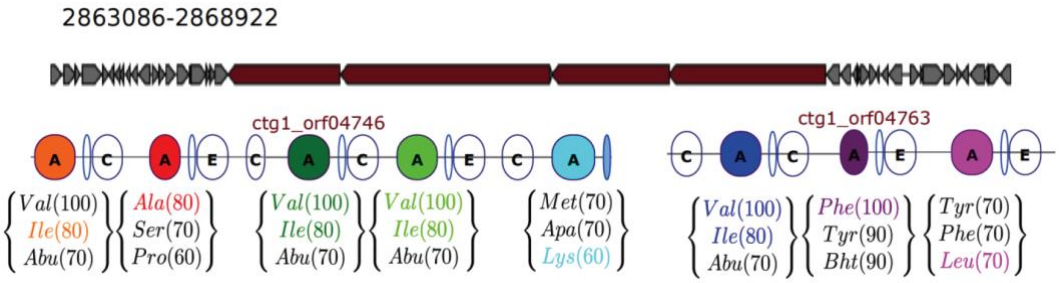
Supplementary Figure S1. A connected component in the spectral network that contains various surugamide variants. Each node in the network is labeled by the precursor mass of a spectrum and each edge connects spectral pairs that reveal related cyclopeptides. The five green nodes are the known surugamide variants[1]. The pink nodes represent unknown cyclopeptides. The spectral network was constructed based on all cyclospectra in the $GNPS_{ACTI}$ dataset.

Supplementary Figure S2 shows a subgraph of the spectral network shown in Supplementary Figure S1 that includes only known surugamides and six novel variants reconstructed by CycloNovo. Supplementary Figure S3 illustrates that five of these novel variants differ from known surugamides by deletions of some amino acids.



Supplementary Figure S2. A subgraph of the suragamide connected component in the spectral network of all cyclospectra from the $GNPS_{ACTI}$ dataset showing only the known and novel surugamide variants sequenced by CycloNovo. The green nodes correspond to known surugamides and the blue nodes represent the novel surugamide variants reconstructed by CycloNovo. The numbers on edges represent the nominal mass shift between the corresponding spectra. The red edges highlight the mass shifts that suggest loss/addition of an amino acid in the peptide and the blue edges connects peptides that differ from each other by a single Ile ➔ Val or Val ➔ Ile substitution (resulting in a nominal offset 14 Da). Although the 14 Da offset can also correspond to methylation, the substitutions represent the more likely explanations in this case. The grey edges show mass shifts that represent combinations of those mass shifts. Supplementary Figure S1 presents the entire connected component.

(a)



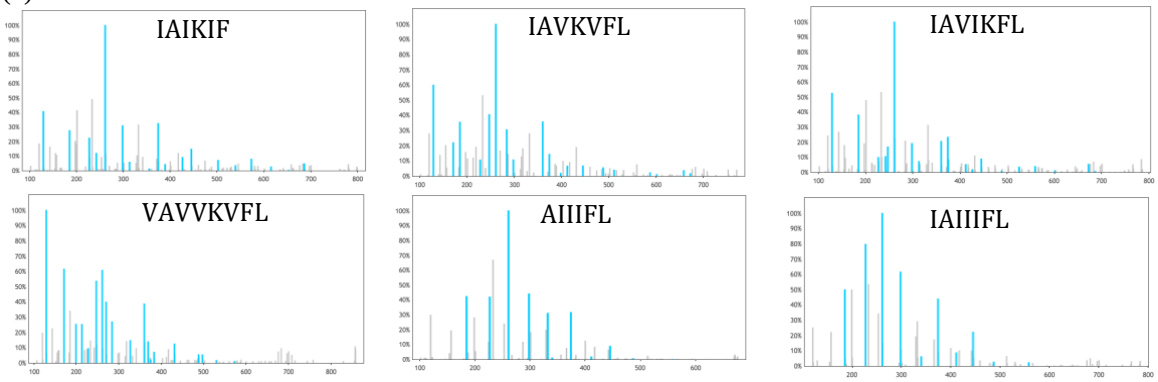
(b)

870.6	Ile	Ala	Val	Val	Lys	Val	Phe	Leu
884.6	Ile	Ala	Val	Ile	Lys	Val	Phe	Leu
898.6	Ile	Ala	Val	Ile	Lys	Ile	Phe	Leu
912.6	Ile	Ala	Ile	Ile	Lys	Ile	Phe	Leu
926.6	Ile	Ala	Ile	Ile	Lys+14	Ile	Phe	Leu
799.5	Ile	Ala	-	Ile	Lys	Ile	Phe	Leu
785.5	Ile	Ala	Val	Ile	Lys	-	Phe	Leu
771.5	Ile	Ala	Val	-	Lys	Val	Phe	Leu
856.5	Val	Ala	Val	Val	Lys	Val	Phe	Leu
784.5	Ile	Ala	Ile	Ile	-	Ile	Phe	Leu
671.5	-	Ala	Ile	Ile	-	Ile	Phe	Leu

(c)

precursor mass	sequence	PSM score/ highest score	P-value	#reconstructions with score ≥ PSM score	reconstruction with the highest score	78604	78787	78936	78937	79516
799.5	IA-IKIFL	19/19	6.2×10^{-37}	1		6				
785.5	IAVIK-FL	22/22	1.6×10^{-38}	4		5	1	4	3	
771.5	IAV-KVFL	20/22	9.4×10^{-39}	3	IAVKV**LF**	2	2	2	1	
856.6	VAVVKVFL	20/22	4.3×10^{-36}	5	VVV**A**KVFL	2				1
784.5	IAII-IFL	10/12	3.4×10^{-19}	2	**AI**IIIFL	5				
671.5	-AII-IFL	12/12	5.0×10^{-21}	1		5				

(d)

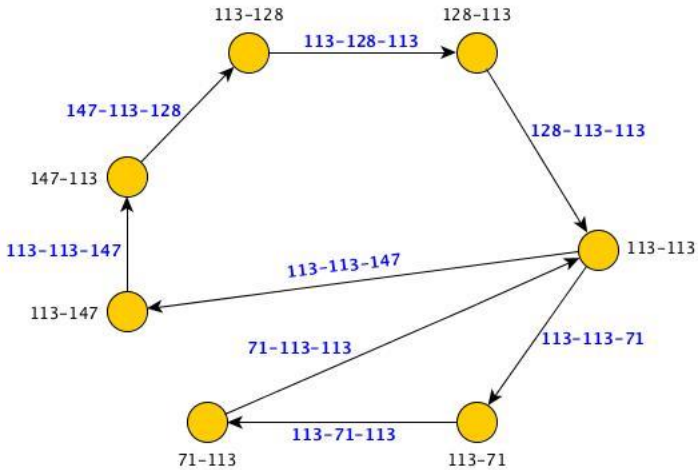


Supplementary Figure S3. Known and novel surugamide variants. (a) Surugamide gene cluster in *Streptomyces albus* along with the three most likely amino acids for each A-domain and their scores predicted by NRPSpredictor2[2]. See Mohimani et al, 2017[1] for more details on this representation. (b) Five known (first five rows) and six novel (last six rows) surugamide variants. Each column is color-coded based on the color of the A-domain they represent in the top figure. The dash symbols indicate a violation of the non-ribosomal code (A-domain skipping) when an A-domain in the surugamide gene cluster does not

add an amino acid to a cyclopeptide. (c) *De novo* reconstructions of the novel surugamide variants. The column 'PSM score/highest score' shows the score of the cyclopeptide and the highest score observed for that spectrum among all CycloNovo reconstructions. The "P-value" column presents the P-value of the PSM (for each cyclopeptide, the spectrum that yielded the lowest P-value is reflected). The column "#reconstructions with score ≥ PSM score" shows the number of reconstructions with score greater or equal to the PSM score. The column "reconstruction with the highest score" shows a highest-scoring reconstruction for the cases when the PSM score is below the highest score. The number of spectra corresponding to each novel surugamide variant in the five GNPS datasets are presented in the columns '78604', '78787', '78936', '78937', and '79516', representing the GNPS sub-datasets MSV000078604, MSV000078787, MSV000078936, MSV000078937, and MSV000079516, respectively. Finding the same surugamide variants in different studies makes it unlikely that they represent artifacts. (d) Annotated spectra of six novel surugamide variants.

Supplementary Note: An example of the de Bruijn graph

Figure S4 presents an example of the de Bruijn graph.



Supplementary Figure S4. The de Bruijn graph constructed from eight 3-mers of surugamide. The sequence of nominal masses of amino acids in surugamide is represented as 71-113-113-147-113-128-113-113. Nodes (edges) in the de Bruijn graph correspond to seven 2-mers (eight 3-mers) in surugamide. Each edge (3-mer) connects the node corresponding to its initial 2-mer to the node corresponding to its final 2-mer. A traversal of edges of the graph spells out the sequence of masses of amino acids in surugamide.

Supplementary Note: Preprocessing spectra

Similar to pre-processing practices in proteomics[3], CycloNovo filters out low-intensity peaks in each spectrum by retaining at most 5 peaks with the highest intensities in each 50 Da window. CycloNovo further filters out all peaks that are less than 0.05 Da apart from another peak with higher intensity. It further removes spectra with a small number of peaks (less than 20) and spectra with a small precursor mass (less than 500 Da). We subtract the mass of a hydrogen atom from all masses in the spectrum (for simplicity, we assume that each ion is protonated with a single proton).

Supplementary Note: CycloNovo parameters

Table S1 specifies the default values of CycloNovo parameters.

Universal Parameter	**default value**	**command line argument**
error threshold (Da)	0.02	--precursor_ion_thresh
Parameters used for recognizing cyclospectra		

α	0.07	--alpha
β	-1	--beta
cycloIntensity threshold	60%	--cyclointensity
k-merScore threshold	4	--kmer_score
minNumberFrequentClusters threshold	2	--num_frequent_clusters
Parameters used for *de novo* sequencing		
cyclopeptidic aa threshold	1	--aminoacid_multiplicity
k-mer threshold	2	--kmer_threshold
k-mer size	5	--kmer_size

Supplementary Table S1. The default values of CycloNovo parameters along with the command-line arguments to specify their values. The parameters introduced in the main text are shown in green rows and the parameters introduced in the Supplementary Note "Distinguishing cyclospectra from spectra of linear peptides and polymers" are shown in pink rows. For all datasets in this paper, all cyclospectra were recognized with the default parameters $\alpha = 0.007$ and $\beta = -1$, except for cyclospectra in the GNPS datasets that were recognized with even more conservative parameters $\alpha = 0.008$ and $\beta = 0$.

Supplementary Note: Distinguishing cyclospectra from spectra of linear peptides and polymers

The challenge of distinguishing cyclospectra from spectra of linear peptides and polymers. Fragmentation of linear peptides typically results in *prefix* (e.g., b-ions) and *suffix* (e.g., y-ions) ions and rarely generates *internal* ions. However, spectra of some linear peptides feature a substantial number of internal ions, leading to a possibility to erroneously classify them as cyclospectra. Another source of a potential misclassification of some spectra as cyclospectra are polymers that represent a common source of contamination in mass spectral datasets. Since polymers are made up of repeated units, the spectral convolution of a polymer spectrum typically has high-multiplicity clusters (for clusters corresponding to masses of the repeat units). In some cases, the adducts of these repeat units form high multiplicity clusters with masses equal to the masses of a cyclopeptidic amino acid, triggering a possibility to misclassify a polymer spectrum as a cyclospectrum.

LINEARLIBRARY and POLYMERLIBRARY datasets. To ensure that CycloNovo does not misclassify spectra of linear peptides and polymers as cyclospectra, we analyzed two spectral datasets described below:

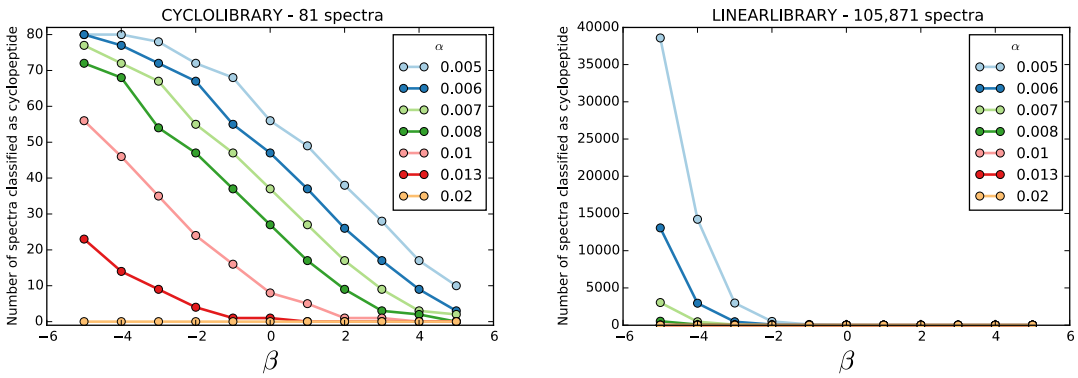
- LINEARLIBRARY is a set of 105,871 Collision-Induced Dissociation (CID) tandem mass spectra of distinct linear peptides from the Massive Knowledge-Based spectral library[4] of linear peptides distilled from all human proteomics data in the MassIVE database.
- POLYMERLIBARY is a set of 448 tandem spectra generated from polyethylene glycol (MSV000081544).

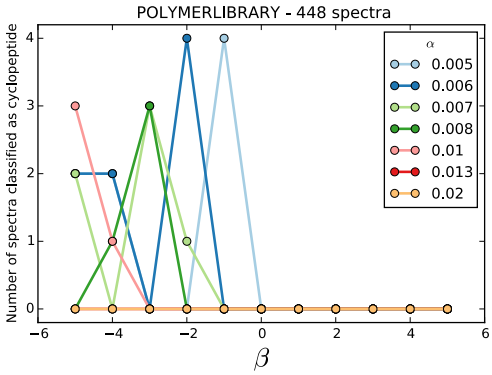
These spectral datasets have spectra with the precursor masses varying between 500 Da and 2000 Da and the charges at most 2.

Additional tests for recognizing cyclospectra. To distinguish cyclospectra from spectra of linear peptides and polymers, CycloNovo only classifies a spectrum as cyclopeptidic if it passes additional tests described below.

- **High multiplicity cyclopeptidic clusters test (distinguishing cyclospectra from spectra of linear peptides).** As described in the main text, CycloNovo first selects a spectrum for further analysis if its spectral convolution has at least *minNumberFrequentClusters* frequent cyclopeptidic clusters, i.e., clusters with multiplicities exceeding the cluster multiplicity threshold. Since the cluster multiplicities typically increase with the increase in the length of a peptide, this threshold increases with the increase in the peptide mass. We thus defined the *cluster multiplicity threshold* as $\alpha \times precursorMass + \beta$ (see below for selecting parameters α and β).
- **Polymer test (distinguishing cyclospectra from polymer spectra).** For each cyclospectrum *Spectrum*, CycloNovo analyzes clusters with masses of repeat units observed in background contamination from polyethylene glycol, NaCl, polypropylene glycol, and trimethylsiloxane (44.03, 57.96, 58.04, and 72.04 Da, respectively). We refer to these masses as *polymeric units*[5] and refer to clusters with masses equal to polymeric units as *polymer-clusters*. CycloNovo classifies a spectrum as polymeric if there exist at least *minNumberFrequentClusters* polymer-clusters with multiplicities at least the *cluster multiplicity threshold*. Polymeric spectra are filtered out from the set of found cyclospectra.
- ***cycloIntensity* test.** For each cyclospectrum, CycloNovo considers all frequent cyclopeptidic clusters. For each such cluster of mass a, we consider all pairs of masses x and y in the spectrum contributing to this cluster, i.e., satisfying the condition $y \approx x + a$. The *cyclointensity* of the spectrum, referred to as *cycloIntensity*, is defined as the total intensity of all such peaks (across all frequent cyclopeptidic clusters) divided by the total intensity of all peaks in *Spectrum*. Spectra with cyclointensity below the *cycloiIntensity threshold* are filtered out.
- ***k-merScore* test.** CycloNovo computes the *k-merScore*, the score of the highest-scoring *k*-mer that contributes to the de Bruijn graph of the spectrum and filters out cyclospectra with *k-merScore* below the *k-merScore threshold*.

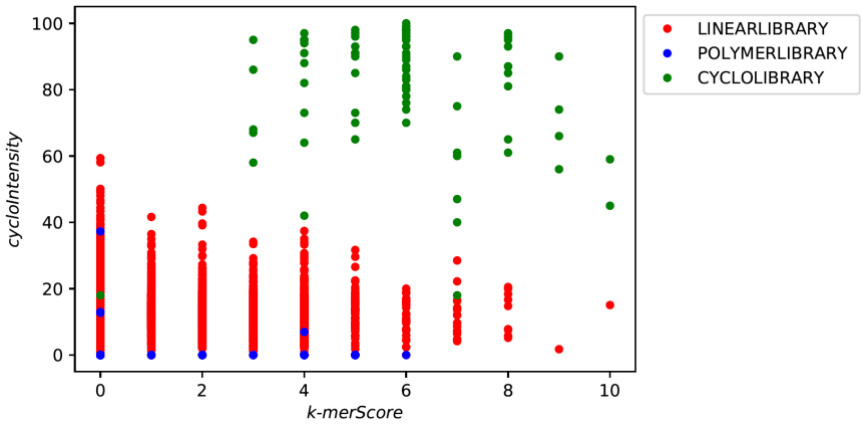
Selecting thresholds for recognizing cyclospectra. To select the default value of *cluster multiplicity threshold*$=\alpha \times precursorMass + \beta$, we varied parameters α (from 0.005 to 0.02) and β (from -5 to +5) and analyzed all found cyclospectra in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBARY datasets (Supplementary Figure S5). Despite its smaller size, CYCLOLIBRARY is the only dataset where CycloNovo recognizes cyclospectra for all analyzed values of α and β. Since α=0.07 and β=-1 yielded the largest number of recognized cyclospectra in CYCLOLIBRARY (46 out of 81) and no cyclospectra in the LINEARLIBRARY and POLYMERLIBRARY datasets (Supplementary Figure S5), we selected these values as the default parameters.





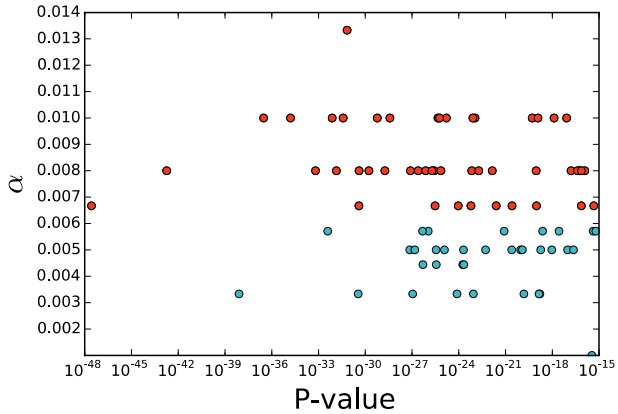
Supplementary Figure S5. Number of spectra passing both the "high multiplicity cyclopeptidic cluster" and the "polymer" tests in the CYCLOLIBRARY, LINERALIBRARY, and POLYMERLiBRARY datasets (for various values of parameters *α* and *β*).

Supplementary Figure S6 presents the values of *cycloIntensity* and *k-merScore* for each spectrum in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBRARY datasets and reveals a separation between the former and the two latter datasets with respect to these two parameters. CycloNovo thus classifies a spectrum as a cyclospectrum if its *cycloIntensity* exceeds the *cycloIntensity* threshold (60%) and its *k-merScore* exceeds the *k-merScore* threshold (5). 45 spectra in the CYCLOLIBRARY datasets, that pass all four tests described above, are classified as cyclospectra.



Supplementary Figure S6. Values of *cycloIntensity* and *k-merScore* for all spectra in the CYCLOLIBRARY, POLYMERS, and LINEARLIBRARY datasets (for *k*=5).

We also investigated how CycloNovo's ability to recognize a cyclospectrum is affected by the fragmentation quality of the corresponding PSM (measured by the P-value of this PSM). For each spectrum in the CYCLOLIBRARY dataset, we identified the minimum value of the parameter *α* that leads to classifying this spectrum as cyclopeptidic (for *β*=-1). Supplementary Figure S7 illustrates that well-fragmented spectra can be recognized even with more restrictive threshold values (larger values of *α*).



Supplementary Figure S7. Dependence between the P-value of each spectrum in the CYCLOLIBRARY dataset and the minimum value of the parameter α that leads to classifying this spectrum as a cyclospectrum (for $\beta = -1$). Each point represents a spectrum in the CYCLOLIBRARY dataset. The *x*-axis shows the P-value of the PSM for that spectrum and the *y*-axis shows the minimum value of the parameter α that leads to classifying this spectrum as a cyclospectrum. The points corresponding to cyclospectra recognized with the default parameter α=0.07 are shown in red.

Supplementary Note: Cyclopeptidic amino acids

Table S2 lists all cyclopeptidic amino acids.

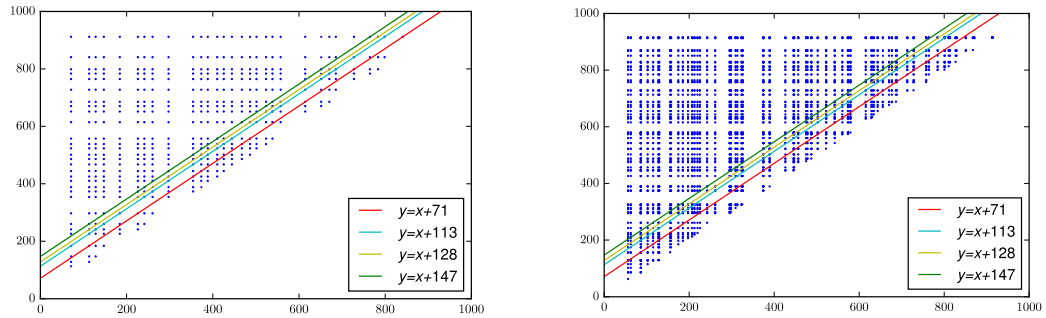
amino acid	elemental composition	monoisotopic mass (Da)	% of cyclopeptides in *CyclopeptideDatabase* containing the amino acid
isoleucine/leuscine	$C_6H_{11}ON$	113.084	55.8
valine	C_5H_9ON	99.068	38.2
proline	C_5H_7ON	97.053	36.9
alanine	C_3H_5ON	71.037	36.4
phenylalanine	C_9H_9ON	147.068	24.6
methyl-isoleucine/leucine	$C_7H_{13}ON$	127.101	18.7
glycine	C_2H_3ON	57.022	17
threonine	$C_4H_7O_2N$	101.048	16.5
serine	$C_3H_5O_2N$	87.032	15.8
ornithine	$C_5H_{10}ON^2$	114.068	15
methyl-alanine	C_4H_7ON	85.054	14.6
tyrosine	$C_9H_9O_2N$	163.063	10.9
glutamine	$C_5H_7O_3N$	129.043	10.1
asparagine	$C_4H_6O_2N_2$	114.043	9.5
methyl-phenylalanine	$C_{10}H_{11}ON$	161.085	9.5
aspartic acid	$C_4H_5O_3N$	115.027	7.6
glutamic acid	$C_5H_8O_2N_2$	128.059	7.6
arginine	$C_6H_{12}ON_4$	156.101	7.2
tryptophan	$C_{11}H_{10}ON_2$	186.079	7.2
methyl-oxazoline	C_4H_7NO	85.104	2.9
lysine	$C_6H_{12}ON_2$	128.095	2.9
oxazoline	C_3H_5NO	71.078	2.5
methionine	C_5H_9ONS	131.040	2.4
oxazole	C_3H_3NO	69.062	1.3
methionine-oxide	$C_5H_9O_2NS$	147.040	1.2
histidine	$C_6H_7ON_3$	137.059	1
methionine-dioxide	$C_5H_9O_3NS$	163.040	0.3
methyl-oxazole	C_4H_5NO	83.089	<0.1
thiazole	C_3H_3NS	85.128	<0.1
thiazoline	C_3H_5NS	87.143	<0.1
cysteine	C_3H_5ONS	103.009	<0.1

Supplementary Table S2. The list of 33 cyclopeptidic amino acids (corresponding to 31 unique amino acid masses). Proteinogenic amino acids are shown in blue, common amino acids in RiPPs[6] are shown in black, and the remaining amino acids that appeared in the top 25 most frequent residues in *CyclopeptideDatabase* are shown in red.

Supplementary Note: Analyzing spectral convolution

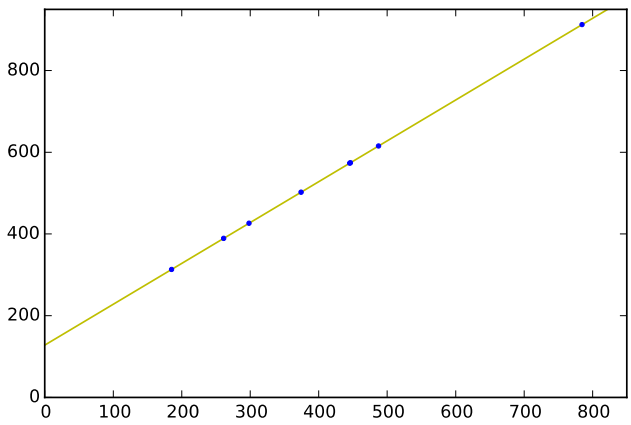
Supplementary Figure S8 illustrates that each amino acid in surugamide results in a populous diagonal in the spectral diagram of $Spectrum_{Surugamide}$. For each constructed cluster (diagonal band in the spectral diagram), we consider all pairs of masses in *Spectrum* that contributed to this cluster and form a *band* as the set of these *k* pairs.

We define the *cluster diameter* as the difference between its maximum and minimum elements. Supplementary Figure S9 presents the band for the cluster with multiplicity 8 and mass 128.09 (diameter 0.03) in the spectral convolution of $Spectrum_{Surugamide}$ and reveals that the 8 elements of this band can be partitioned into 7 groups of closely located points. We are interested in the number of such groups (rather than the raw cluster multiplicities) since experimental spectra often contain *satellite masses* resulting from *neutral losses* and *isotopic peaks*. For example, in addition to the integer mass 242 Da corresponding to the peptide IK, $Spectrum_{Surugamide}$ also contains the integer mass 225 Da corresponding to the loss of NH_3 from this peptide.



Supplementary Figure S8. The spectral diagrams of the *TheoreticalSpectrum(Surugamide)* (left) and $Spectrum_{Surugamide}$ (right). The highlighted lines with slope +1 have *y*-intercepts equal to the masses of the constituent amino acids of surugamide (A, L/I, K, and F). Amino acids A, L/I, K, and F correspond to populous diagonals containing 11, 23, 11, and 11 points (left figure) and 5, 14, 8, and 4 points (right figure), respectively.

$y \approx x + 128$	(185.13,313.22) (261.17,389.26) (298.21,426.28) (374.24,502.32) **(446.30,574.38) (445.28,573.35)** (487.35,615.42) (784.52,912.62)



Supplementary Figure S9. A band with multiplicity eight in *Spectrum*$_{Surugamide}$ (cluster with mass 128.09 and diameter 0.03). (Top) Coordinates of the points in the band. Since the difference between the *x*-coordinates and *y*-coordinates of the two points shown in bold match the mass of hydrogen, these two points are clustered together in this band. (Bottom) The same band in the spectral diagram for *Spectrum*$_{Surugamide}$. The points of the band can be partitioned into seven groups of closely located points: six singleton groups and one group with two elements.

Since satellite masses artificially inflate cluster multiplicities, there is a need to reduce biases caused by these masses. We thus define the set of common satellite offsets (1 Da (H), 18 Da (H_2O), 17 Da (NH_3), and 28 Da (CO)) and perform additional single linkage clustering in each populous band by combining pairs of masses in a single cluster if both their *x*-coordinates and *y*-coordinates differ by a satellite offset. We redefine the concept of cluster multiplicity as the number of the resulting clusters in the band (Supplementary Table S3).

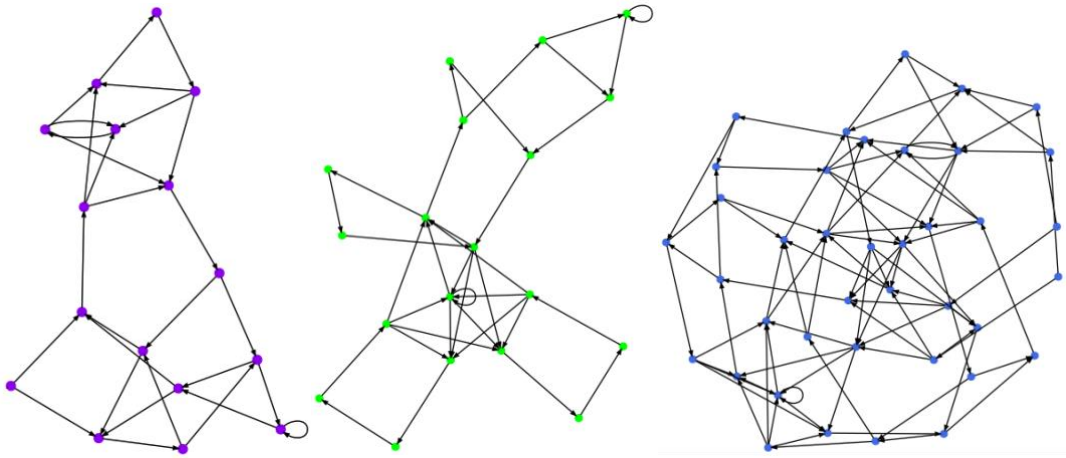
cluster mass	multiplicity after satellite removal	multiplicity before satellite removal	cluster diameter	cluster distance
113.078	**12**	**14**	**0.106**	**0.006 (I/L)**
128.089	**7**	**8**	**0.032**	**0.006 (K)**
56.025	6	6	0.033	-
186.108	**6**	**6**	**0.077**	**0.029 (W)**
71.041	**5**	**5**	**0.022**	**0.004 (A)**
147.069	**4**	**4**	**0.027**	**0.001 (F)**
156.099	**4**	**5**	**0.036**	**0.002 (R)**
184.106	4	5	0.032	-
98.079	4	4	0.031	-
169.11	4	5	0.01	-
57.049	4	4	0.02	-
70.042	3	3	0.01	-
132.058	3	3	0.006	-
133.584	3	3	0.013	-
183.116	3	3	0.024	-
168.159	3	3	0.023	-
52.049	3	3	0.005	-
96.035	3	3	0.017	-
165.115	3	3	0.004	-
76.041	3	4	0.013	-
73.08	3	3	0.017	-
101.053	**3**	**3**	**0.011**	**0.005 (T)**
167.886	3	4	0.024	-
189.105	2	3	0.009	-
57.018	**2**	**2**	**0.006**	**0.004 (G)**

114.09	2	2	0.01	-
45.041	2	2	0.012	-

Supplementary Table S3. List of clusters in the spectral convolution of *Spectrum*$_{Surugamide}$. Clusters are shown in the decreasing order of their multiplicities (only clusters with multiplicity at least 2 are shown). Cyclopeptidic clusters are shown in bold and cyclopeptidic clusters with masses similar to masses of amino acids in surugamide are shown in red.

Supplementary Note: De Bruijn graphs for *Spectrum*$_{Surugamide}$

Supplementary Figure S10 shows the pruned *de Bruijn* graphs of three compositions of *Spectrum*$_{Surugamide}$ that do not contain feasible cycles.



Supplementary Figure S10. The pruned *de Bruijn* graphs of the compositions of *Spectrum*$_{Surugamide}$ that do not contain feasible cycles. (Left) The composition $113^4156^2147^1$ results in a de Bruijn graph with 40 vertices and 57 edges and a pruned de Bruijn graph with 18 vertices and 40 edges. (Middle) The composition $128^471^2156^1101^1$ results in a de Bruijn graph with 52 vertices and 76 edges and a pruned de Bruijn graph with 20 vertices and 42 edges. (Right) The composition $128^5113^1101^157^1$ results in a de Bruijn graph with 94 vertices and 180 edges and a pruned de Bruijn graph with 40 vertices and 92 edges.

Supplementary Note: CycloNovo running time

CycloNovo recognizes cyclospectra by constructing their spectral convolutions ($O(n^2)$ running time for a spectrum with n peaks) and further sequences all found cyclospectra. Using a single 2.5GHz processor, CycloNovo recognized all cyclospectra in the HUMANSTOOL and GNPS datasets in ≈35 minutes and ≈31 hours, respectively.

The sequencing step is only applied to a small fraction of all spectra in spectral datasets, e.g., CycloNovo recognizes only ≈0.05% of all spectra in the GNPS dataset as cyclospectra. The running time for the sequencing step varies widely between spectra and depends on the number of putative amino acid compositions, the number of putative *k*-mers, and the number feasible cycles in the de Bruijn graphs.

We were not able to benchmark CycloNovo against CYCLONE[7] since CYCLONE failed to reconstruct most spectra in the CYCLOLIBRARY dataset. For example, CycloNovo took ≈3 seconds to sequence *Spectrum*$_{Surugamide}$. In contrast, CYCLONE[7] failed to sequence *Spectrum*$_{Surugamide}$ and was not even able to infer alanine and lysine as amino acids in suragamide.

We thus compared CycloNovo with a brute force sequencing algorithm by generating and scoring all possible permutations of all amino acid compositions resulting from the putative amino acids for $Spectrum_{Surugamide}$. The brute-force approach took ≈27 seconds to sequence $Spectrum_{Surugamide}$ as the highest scoring reconstruction among all 113 generated cyclopeptides. In a more difficult example, we analyzed the spectrum with precursor mass 899.36 and reconstructed the orbitide FVDTTGYD in the S.VULGARIS dataset (Table 2). CycloNovo sequenced this spectrum in 56 seconds while the brute force approach took 58 minutes. For this spectrum, there exist 321 putative amino acid compositions yielding over 400,000 candidate sequences. Since the majority of those ~400,000 sequences do not produce high-scoring 5-mers, they yield the relatively small de Bruijn graphs and hence small numbers of feasible cycles and candidate cyclopeptides. By only exploring the feasible cycles in the de Bruijn graphs, CycloNovo reduced the number of candidate sequences to 2,491.

These examples illustrate that the running time of CycloNovo varies by orders of magnitude depending on analyzed cyclospectra. By only exploring the sequences spelled by the feasible cycles in the de Bruijn graphs, CycloNovo greatly reduces the search space compared to the brute force approach. For example, the brute force approach failed after 1000 hours on just two cyclospectra from the CYCLOLIBRARY dataset, while the de Bruijn graph approach finished analysis of all spectra in this library in ≈48 hours. Supplementary Note "CycloNovo analysis of the CYCLOLIBRARY dataset" lists all CycloNovo reconstructions for this dataset. CycloNovo analysis of the HUMANSTOOL and GNPS datasets took ~34 and ~149 hours, respectively.

Supplementary Note: Information about spectral datasets

Information about *CyclopeptideDatabase*. The *PNPDatabase*[8] combines all known peptidic natural products from various databases. Many peptides in this database are *lipopeptides* containing a lipid chain, e.g., surfactin is a cyclopeptide containing a fatty acid side chain connected to a fully peptidic part via a peptide bond.

We classify a peptide in the *PNPdatabase* as a cyclopeptide if its backbone could be represented as a circular graph (cycle) with nodes corresponding to either a single amino acid or a single lipid tail (i.e. monomers) and edges corresponding to the amide bonds in the peptide structure. 1,257 out of 5,021 peptides in the *PNPDatabase* represent cyclopeptides and form *CyclopeptideDatabase* (note that the *CyclopeptideDatabase* database contains lipopeptides).

Information about the CYCLOLIBRARY dataset. We searched ~130 million GNPS spectra against the *CyclopeptideDatabase* using Dereplicator[1] and identified 81 distinct cyclopeptides (41 cyclofamilies) corresponding to PSMs with FDR=0% and P-value below 10^{-15}. For each identified cyclopeptide, we selected the PSM with the minimum P-value (among all PSMs identified for this cyclopeptide), resulting in a set of 81 PSMs and hence created a spectral dataset CYCLOLIBRARY with 81 spectra (Table S4). CYCLOLIBRARY includes only 13 cyclopeptides (6 cyclofamilies) that are made up entirely of cyclopeptidic amino acids (Table S4). 34 peptides (25 cyclofamilies) in the CYCLOLIBRARY dataset contain lipid tails and 34 peptides (14 cyclofamilies) contain non-cyclopeptidic amino acids.

peptide ID	**peptide mass**	**compound type**	**P-value**	**GNPS ID**	***k-merScore***	***cycloIntensity***
Antibiotic_FR_901459	609.9	peptide	1.8×10^{-24}	MSV000079098	10	0.59

Arthrofactin	1354.8	lipopeptide	1.2×10^{-16}	MSV000079772	8	0.96
Bacillomycin_D2	1031.5	lipopeptide	6.0×10^{-24}	MSV000078635	4	0.95
Bacillomycin_D3	1045.6	peptide	3.9×10^{-31}	MSV000079450	4	0.94
Bacillomycin_D5	1059.6	peptide	2.5×10^{-21}	MSV000078635	5	0.85
Bacillopeptin_B	1035.5	peptide	1.2×10^{-19}	MSV000079054	4	0.88
Bacillus_amyloliquefaciens_Surfactin_1	1036.7	lipopeptide	1.3×10^{-18}	MSV000080116	6	0.98
Bacillus_amyloliquefaciens_Surfactin_22	1022.7	lipopeptide	3.0×10^{-26}	MSV000078936	6	0.87
BK_10 _101A-form	1021.7	lipopeptide	2.5×10^{-22}	MSV000078688	6	0.78
BK_10 _101C	1035.7	lipopeptide	2.6×10^{-21}	MSV000078937	6	0.96
Champacyclin	898.6	peptide	5.8×10^{-26}	MSV000078936	6	0.98
Cyclolinopeptide_A	1040.7	peptide	4.0×10^{-31}	MSV000080050	6	0.93
Cyclolinopeptide_B	1058.6	peptide	1.8×10^{-29}	MSV000080050	8	0.87
Cyclolinopeptide_B_S-Oxide	1074.6	peptide	4.6×10^{-26}	MSV000080050	6	0.81
Cyclolinopeptide_D	1064.6	peptide	9.2×10^{-20}	MSV000080050	6	0.86
Cyclolinopeptide_E	977.6	peptide	2.6×10^{-26}	MSV000079777	4	0.73
Cyclolinopeptide_H	1082.5	peptide	5.1×10^{-20}	MSV000080050	6	0.80
Cyclosporin_B	1188.8	peptide	3.8×10^{-29}	MSV000079098	8	0.65
Cyclosporin_C	1218.8	peptide	1.4×10^{-32}	MSV000079581	6	0.81
Cyclosporin_E	1188.8	peptide	4.7×10^{-27}	MSV000079098	8	0.93
Cyclosporin_L	1188.7	peptide	1.7×10^{-30}	MSV000079098	8	0.81
Cyclosporin_P	1204.8	peptide	4.6×10^{-16}	MSV000079777	6	0.74
Cyclosporin_U	594.9	peptide	3.5×10^{-26}	MSV000079098	9	0.66
Cyclosporin_Y	601.9	peptide	2.0×10^{-24}	MSV000079098	10	0.45
Cyclosporin,_9CI_4	1188.8	peptide	1.5×10^{-27}	MSV000079098	8	0.61
Cyclosporin,_9CI_9	1202.8	peptide	1.8×10^{-43}	MSV000079098	8	0.95
Cyclosporin,_9CI_Deoxy	1186.9	peptide	1.6×10^{-35}	MSV000079098	8	0.85
Cyclosporin,_9CI_N9-De-Me	1188.8	peptide	3.8×10^{-32}	MSV000079098	9	0.74
[8'-Hydroxy-MeBmf]1-cyclosporin	1218.8	peptide	3.0×10^{-37}	MSV000079581	8	0.97
Daitocidin_B2	1064.7	lipopeptide	1.4×10^{-22}	MSV000078937	6	0.90
Daitocidin_Pumilacidin_F	1050.7	lipopeptide	7.5×10^{-27}	MSV000078936	6	0.76
Dolastatin_1_11-N-Me	999.6	lipopeptide	9.6×10^{-18}	MSV000078568	3	0.58
Dolastatin_1_15-Epimer,_31-methyl,_11-N-Me	1013.6	lipopeptide	3.5×10^{-16}	MSV000079050	0	0.18
Dolastatin_1_31	492.3	lipopeptide	1.4×10^{-19}	MSV000078568	4	0.42
Dolastatin_12	969.6	lipopeptide	4.2×10^{-16}	MSV000078568	5	0.65
Dolastatin_14_Dolastatin_14	1089.7	lipopeptide	1.5×10^{-20}	MSV000078568	7	0.47
g-Hydroxy-Meleu4-cyclosporin	609.9	peptide	3.6×10^{-26}	MSV000079581	9	0.56
Ilamycin_B1	1012.6	peptide	7.7×10^{-25}	MSV000078937	5	0.73
Ilamycin_B2	1028.6	peptide	1.6×10^{-19}	MSV000078936	3	0.67
Isocyclosporin_D	1216.9	peptide	5.0×10^{-27}	MSV000079098	7	0.40
Laxaphycin_A	1196.7	peptide	2.7×10^{-48}	MSV000079050	6	0.95
Laxaphycin_B	1395.9	peptide	3.9×10^{-33}	MSV000079050	5	0.97
Laxaphycin_B_32-Epimer,_53-deoxy	690.4	peptide	1.1×10^{-27}	MSV000079050	7	0.18
Laxaphycin_D	1367.8	peptide	7.1×10^{-28}	MSV000079050	3	0.86
Laxaphycin_E	1224.8	peptide	7.9×10^{-39}	MSV000079050	6	0.89
Lichenysin_A	1007.7	lipopeptide	1.2×10^{-20}	MSV000079481	3	0.95
Lichenysin-G1a	993.7	lipopeptide	1.8×10^{-19}	MSV000078936	5	0.70
Lichenysin-G3	1007.7	lipopeptide	8.1×10^{-22}	MSV000078936	6	0.93
Lichenysin-G5b	1021.7	lipopeptide	9.7×10^{-20}	MSV000078936	6	0.99
Lipodepsipeptides_KMM_A	1036.7	lipopeptide	9.2×10^{-25}	MSV000078635	5	0.98
Lipodepsipeptides_KMM_E	1064.7	lipopeptide	6.2×10^{-16}	MSV000078937	6	0.96
Lipodepsipeptides_KMM_F	1078.8	lipopeptide	2.4×10^{-19}	MSV000078936	6	0.97
Lipopeptide_NO	994.6	lipopeptide	4.1×10^{-17}	MSV000078688	6	0.98
Majusculamide_C	985.6	lipopeptide	1.1×10^{-23}	MSV000078892	5	0.91
Majusculamide_C_Demethoxy	955.6	lipopeptide	7.3×10^{-17}	MSV000078568	6	0.95
Nocardiamide_A	687.5	peptide	8.0×10^{-24}	MSV000078936	6	0.97
NVA2-g-hydroxy-Meleu4-cyclosporin	1232.9	peptide	5.5×10^{-17}	MSV000079777	9	0.90
Peptidolipin_NA	964.7	lipopeptide	1.6×10^{-17}	MSV000078937	4	0.64
Pitipeptolide_E	794.5	peptide	9.3×10^{-19}	MSV000078568	5	0.90
Pitiprolamide	905.5	peptide	7.3×10^{-17}	MSV000078568	4	0.95
Precarriebowmide	865.5	lipopeptide	2.1×10^{-24}	MSV000079050	7	0.75
Precarriebowmide_S-Oxide	881.5	lipopeptide	9.3×10^{-21}	MSV000079050	6	0.70
Puwainaphycin_A	1235.7	peptide	7.1×10^{-26}	MSV000078982	4	0.97
Puwainaphycin_B	1233.7	peptide	6.4×10^{-34}	MSV000078982	5	0.91
Puwainaphycin_C	1227.7	peptide	7.9×10^{-28}	MSV000078982	4	0.91
Sch_378167_5'-Amide	569.3	peptide	3.5×10^{-31}	MSV000079098	7	0.61
SCH-378161	1123.6	peptide	2.5×10^{-27}	MSV000079098	6	0.99
Streptocidin_C	649.9	peptide	8.6×10^{-24}	MSV000079598	7	0.60

Surfactin_ A1	1008.7	lipopeptide	1.1×10^{-26}	MSV000078936	6	0.83
Surfactin_7-L-Valine_analogue	1022.7	lipopeptide	1.9×10^{-26}	MSV000078936	5	0.93
Surfactin_B1	1022.7	lipopeptide	5.2×10^{-23}	MSV000078937	3	0.68
Surfactin_C1	1036.7	lipopeptide	1.2×10^{-25}	MSV000078688	6	0.84
[Ile2,Val7]-Surfactin_C14i	1008.7	lipopeptide	2.3×10^{-17}	MSV000079450	4	0.82
[Val7]-Surfactin_C13ai	994.7	lipopeptide	1.9×10^{-23}	MSV000078936	6	0.80
Surfactin_D	1050.7	lipopeptide	6.8×10^{-24}	MSV000078937	6	0.96
Surugamide_A	912.6	peptide	1.6×10^{-25}	MSV000078936	6	0.91
Surugamide_B	898.6	peptide	7.5×10^{-33}	MSV000079519	7	0.90
Surugamide_C	898.6	peptide	6.8×10^{-32}	MSV000079519	6	1.00
Surugamide_D	898.6	peptide	5.9×10^{-30}	MSV000078937	6	0.98
Viequeamide_B	808.5	lipopeptide	2.7×10^{-18}	MSV000078568	5	0.96
[Dihydro-MeBmt]1-[g-hydroxy-Meleu]4	1220.9	peptide	8.3×10^{-18}	MSV000079777	8	0.87

Supplementary Table S4. Cyclopeptides in the CYCLOLIBRARY dataset. The peptides that gave rise to 81 spectra in the CYCLOLIBRARY dataset with their corresponding peptide mass, P-value, the GNPS ID of the dataset a spectrum belongs to, *k-merScore*, and *cycloIntensity*. The column "compound type" specifies whether the compound is fully peptidic or represents a lipopeptide. The blue rows show the 13 cyclopeptides that are made up entirely of cyclopeptidic amino acids.

Information about the GNPS dataset. The GNPS dataset is formed by 40 MassIVE datasets that were selected from 120 datasets analyzed in Gurevich et al.[8] to exclude potentially miscalibrated spectral datasets. Since miscalibrated datasets typically do not result in any cyclopeptide identifications, we searched each of these 120 datasets with Dereplicator and excluded datasets that did not result in any identifications (with 0% FDR and P-value below 10^{-15}) from further analysis, leaving us with 40 datasets (Supplementary Table S5).

GNPS ID	#spectra	#spectra after pre-processing	#cyclo spectra	#putative cyclo-peptides/ cyclo-families found by CycloNovo	#identified cyclo-peptides/ cyclo-families identified by Dereplicator (among cyclo-spectra)	#identified cyclo-peptides/ cyclo-families identified by Dereplicator (among all spectra)	#identified branch-cyclic . peptides/ branch-cyclic families identified by Dereplicator (among cyclo-spectra)
MSV000078567	730582	316993	4	2/1	2/1	4/2	0/0
MSV000078568	23582408	12118472	317	74/35	9/8	15/10	1/1
MSV000078584	680906	263160	0	0/0	0/0	0/0	0/0
MSV000078604	311617	281617	606	56/25	6/3	6/3	3/3
MSV000078606	289170	237988	122	32/12	1/1	1/1	7/6
MSV000078635	680168	569316	2388	124/40	9/4	12/5	10/7
MSV000078656	2844	1023	88	14/1	10/5	11/6	3/3
MSV000078710	1469076	689912	6	1/1	2/2	3/3	0/0
MSV000078787	1767830	1281235	208	58/31	25/13	25/13	8/7
MSV000078839	717600	504350	1	1/1	1/1	1/1	0/0
MSV000078847	167917	115603	19	7/5	1/1	1/1	1/1
MSV000078892	847114	461769	27	8/4	3/2	4/3	0/0
MSV000078936	2059306	1538683	526	58/30	25/13	30/15	5/5
MSV000078937	1694918	1303349	256	52/26	26/14	33/19	11/9
MSV000078982	984	727	32	4/2	3/1	3/1	2/2
MSV000079044	576282	270860	2	1/1	1/1	2/1	0/0
MSV000079050	1241328	683124	207	24/8	3/1	7/3	3/3
MSV000079054	702020	364382	112	16/7	13/6	13/6	3/3
MSV000079069	847145	215229	1066	23/2	0/0	1/1	0/0
MSV000079140	607488	443147	1118	25/7	14/6	15/7	0/0
MSV000079274	5433248	3457806	14	2/2	1/1	1/1	0/0
MSV000079312	54806	25354	1112	15/3	0/0	1/1	0/0
MSV000079450	697812	581012	2245	120/39	6/2	6/2	9/6

MSV000079471	22379	16138	68	14/4	10/4	10/4	0/0
MSV000079481	45742	7692	48	2/1	0/0	2/2	0/0
MSV000079502	47450	3167	14	3/1	3/1	4/2	0/0
MSV000079516	120154	19113	138	22/6	4/2	5/2	0/0
MSV000079517	22516	2911	37	7/3	3/1	4/2	0/0
MSV000079519	76289	13985	112	15/6	3/1	5/2	0/0
MSV000079568	224645	10273	0	0/0	0/0	4/2	0/0
MSV000079581	129012	41779	74	4/3	2/2	5/5	0/0
MSV000079598	919494	286130	9	3/1	4/1	6/3	0/0
MSV000079651	81818	5239	0	0/0	0/0	2/1	0/0
MSV000079679	595244	300682	109	24/14	12/7	14/7	2/2
MSV000079772	75916	13870	40	7/5	2/2	2/2	0/0
MSV000079778	1242178	451962	265	50/44	6/1	7/2	0/0
MSV000079813	578683	170990	23	6/4	2/2	3/2	0/0
MSV000079888	238820	74317	170	17/7	8/4	9/5	1/1
MSV000080115	1567520	709527	400	33/9	12/6	13/7	9/6
MSV000080116	70250	31009	19	9/5	6/3	6/3	0/0
TOTAL	51220679	27883895	12004	512/213	61/37	91/51	41/27

Supplementary Table S5. Information about the GNPS dataset. The last row shows the total number of spectra and unique cyclopeptides/cyclofamiles across all datasets. The datasets marked in red, blue, and green form $GNPS_{CYANO}$, $GNPS_{PSEUDO}$, and $GNPS_{ACTI}$ subsets of the GNPS dataset, respectively.

Supplementary Note: CycloNovo analysis of the CYCLOLIBRARY dataset

CycloNovo recognized 45 out of 81 spectra in the CYCLOLIBRARY dataset as cyclospectra. It classified 12 out of 13 cyclopeptides built from cyclopeptidic amino acids as cyclospectra and *de novo* sequenced them with one of the top three highest scores.

CycloNovo is unable to sequence most spectra in the CYCLOLIBRARY dataset since 68 of them originated from lipopeptides or peptides containing non-cyclopeptidic amino acids. To evaluate how CycloNovo performs on 45 cyclospectra in this dataset, we extended the set of cyclopeptidic amino acids to include the mass of the lipid chain and/or the masses of non-cyclopeptidic amino acids for each spectrum. Using this admittedly imperfect benchmarking approach, CycloNovo sequenced 22 of 45 cyclospectra as a highest-scoring *de novo* reconstructions and an additional 16 spectra with one of the three highest scores. Supplementary Table S6 lists the highest-scoring reconstruction for these spectra and illustrates that the highest-scoring reconstruction is similar to the correct amino acid sequence for all these spectra.

peptide ID	**sequence of aa masses in the peptide vs. sequence of aa masses in the highest-scoring reconstruction (if PSM score ≠ max score)**	**PSM score**	**max score**	**# reconstructions with score ≥ PSM score**
BK 101C	113 113 115 99 113 113 128 240	21	21	6
nocardiamide A	113 113 99 99 99 163	19	19	1
cyclolinopeptide A	113 113 113 147 147 97 97 99 113	30	30	1
cyclolinopeptide B	113 99 147 147 97 97 113 113 131	24	24	1
bacillopeptin B	239 101 87 129 87 114 163 114	17	17	1
daitocidin_Pumilacidin F	254 99 113 115 113 113 113 129	24	24	4
BK 101A	113 113 115 99 113 113 128 226	19	19	6
cyclolinopeptide H	113 186 147 147 97 131 113 147	16	16	1
cyclosporin 9CI_Deoxy	167 113 127 127 71 71 127 99 127 71 85	29	29	6
cyclosporin B	183 113 127 127 71 71 127 99 127 71 71	30	30	4

laxaphycin A	57 113 113 113 113 147 101 113 83 101 141	44	44	1
surfactin 2	240 113 113 115 99 113 99 129	21	21	8
cyclolinopeptide D	113 186 147 147 97 113 113 147	20	20	2
cyclolinopeptide E	113 147 113 97 147 99 113 147	23	23	1
lipodepsipeptide KMM 1364A	240 99 113 115 113 113 113 129	20	20	8
Lipodepsipeptide KMM 1364E	268 99 113 115 113 113 113 129	20	20	2
cyclolinopeptide C	113 99 147 147 97 97 113 113 147	24	24	1
bacillomycin D2	97 114 163 114 225 101 87 129	22	22	1
bacillomycin D3	97 114 163 114 239 101 87 129	21	21	2
SCH-378161	113 57 97 147 114 143 99 97 113 142	29	29	2
[Val7]-Surfactin C13ai	99 113 113 129 212 99 113 115	21	21	3
lipopeptide_NO	99 113 113 129 198 113 113 115	17	17	15
cyclosporin,9CI 9	183 113 127 127 71 71 127 99 127 71 85 183 113 127 127 71 71 127 **113 99 85** 85	32	33	10
surfactin C1	240 113 113 115 99 113 113 129 240 113 113 **99 113 115** 113 129	20	21	15
cyclosporin C	183 113 127 127 71 71 127 99 127 71 101 183 113 127 **85** 71 113 **128** 71 99 99 **128**	33	34	49
surfactin 1	254 113 113 115 99 113 99 129 254 113 113 **99** 99 **129** 99 129	23	24	16
puwainaphycin_B	325 97 128 115 57 128 99 101 83 99 325 97 **99** 128 **57 115** 128 101 83 99	26	27	6
surugamide A	128 113 113 71 113 113 147 113 128 113 113 71 113 113 **113 147**	23	24	5
surugamide B	128 99 113 71 113 113 147 113 128 99 113 71 113 **147 113** 113	26	27	5
surugamide D	128 113 99 71 113 113 147 113 128 113 **113** 71 **99** 113 147 113	28	30	7
lichenysin G5b	99 113 115 99 113 113 128 240 99 113 **99 115** 113 113 128 240	20	21	6
pitiprolamide	100 97 99 142 97 175 97 97 100 97 **142 99** 97 175 97 97	17	18	6
surfactin 7-L-Valine	240 99 113 115 99 113 113 129 240 99 **99 129** 99 113 113 129	22	24	15
surfactin D	254 113 113 115 99 113 113 129 254 113 113 113 **115 113 99** 129	22	25	14
majusculamide C Demethoxy	57 114 113 71 141 161 113 57 127 57 **113 114** 71 **161 141** 113 57 127	20	22	69
cyclosporin E	183 99 127 127 71 71 127 99 127 71 85 183 99 127 127 **99** 71 71 127 127 71 85	24	26	53
champacyclin	128 99 113 147 113 113 71 113 128 99 **71** 113 147 113 113 113	21	23	21
surugamide C	128 113 113 71 113 113 147 99 128 113 113 71 113 **99 147 113**	29	31	10

Supplementary Table S6. 38 cyclopeptides reconstructed by CycloNovo from 45 cyclospectra in the CYCLOLIBRARY dataset. The PSM score represents the score of the PSM in the CYCLOLIBRARY dataset. The "max score" represents the score of the top-scoring reconstruction. For 22 cyclopeptides, the correct sequence of the cyclopeptide has the highest-scoring reconstruction. For the remaining 16 cyclopeptides, a highest-scoring reconstruction is listed below the correct sequence of the cyclopeptide in blue (differently arranged amino acid masses in the reconstructed cyclopeptide are shown in bold blue). Only in one case (cyclosporin C), CycloNovo predicted the wrong amino acids (shown in red) for the top-scoring reconstruction. CycloNovo failed to sequences 45-38=7 cyclospectra in the CYCLOLIBRARY dataset since it was not able to predict all their amino acids.

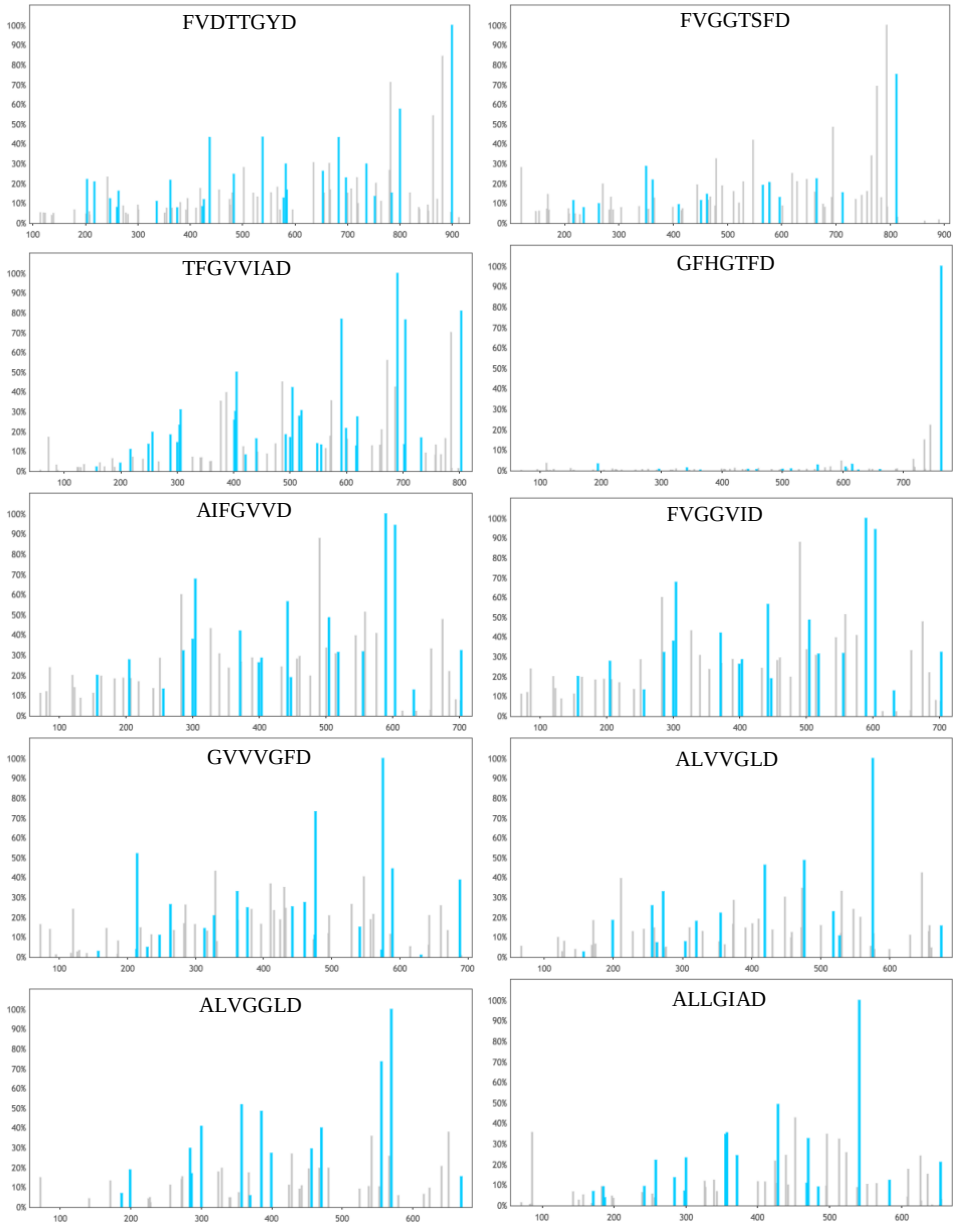
Supplementary Note: Cyclopeptide-encoding transcripts in the S.VULGARIS dataset

Supplementary Table S7 lists ORFs (translated into amino acid sequences) in the orbitide-encoding transcripts. The PawL1 proteins have dual fates; they encode an albumin as well as a cyclopeptide(s). An enzyme asparaginyl endopeptidase (targets Asp, Asn) matures both the albumin and the cyclopeptide.

gene	ORF sequence
Sv_PawL1b	AKLIVVVFAFAVIVAFAEVSAYKTTITTTTVEDNFVGGTSFDRLSENFMYGTPVDRL SDNRGSQKQCHRQIP
Sv_PawL1c	AKLIVVVFAFAVIVAFAEVSAYKTTITTTTVEDNTFGVVIADRLSDNFVDTTGYDRL SDNRGSQKQCHRQIP
Sv_PawL1d	ITTVEDNALVVGLDGLDNPITTTVEDNYFAGLIDGLDNPITTTVEDNGVFLGLDGLD NPSGSTYQCRRQIQGQQLNHCQMHIIQQGRSLVE
Sv_PawL1e	FVAIVAFSEQVSAYKTTIPTTVEDNALLVALDGLDNGFHGTFDGLDNGFHGTFDGLD NPSGSTYQCRRQIQ*
Sv_PawL1f	TTVEDNALFLGLDGLDNPSGSTYQCRRQIQGQQLNHCQMHITQQGRSLMENPRQQQL LQMCCNQLRQVEEECQCE*
Sv_PawL1g	ITTTVEDNALVVGLDGLDNPITTTVEDNFVGGVIDGLDNFVGGVIDGLDNPSGSTYK CRRQIQGQQLNHCQMHITQQGRSLVE
Sv_PawL1h	MTKVSAIVVLAFVAIVAFSEQVSAYKTTITTPVEDNAIFLGVDGLDNPI*
Sv_PawL1i	LDGLDNALLGIADGLDNPSGSTYQCRMQIQGQQLNHCQMHIIQQGRSLVENPRQQQQ LQMCCNQLR*
Sv_PawL1j	SEQVSAYKTTITTTVEDNAIFGVVDGLDNPSGSTYQCRKQIQGQQ*
Sv_PawL1k	AIVAFSEQVSAYKTTITTTVEDNAIFLGVDGLDNPITTTVEDNGVSDFFDDGLDKPS GSTYQCRRQIQGQQLNHCQMHISQQGRSLVENPRQQQQLQM*
Sv_PawL1l	FVAIVAFSEQVSAYKTTITTPVEDNGVVVGFDGLDNPSGSTYQCRKQIQGQQ*

Supplementary Table S7. ORFs in the cyclopeptide-encoding transcripts. All identified ORFs originate from various *PawS1-Like* genes. The sequences are color-coded based on the subunits they belong to: endoplasmic reticulum signal sequence (pink), the reconstructed cyclopeptide (blue), 2S albumin small subunit (lime green), and 2S albumin large subunit (orange). While the first three sequences (*Sv_PawL1b*, *Sv_PawL1c*, and *Sv_PawL1d*) are known *PawS1-Like* genes in *S. vulgaris*, the other eight sequences (named *Sv_PawL1e* through *Sv_PawL1l*) are novel *PawS1-Like* genes that were identified by searching for novel cyclopeptides.

Supplementary Figure S11 shows cyclospectra in the S.VULGARIS dataset, annotated using their CycloNovo reconstructions.



Supplementary Figure S11. Annotated cyclospectra of the ten reconstructed cyclopeptides in the S.VULGARIS dataset. The *x*-axis shows the *m/z* ratios and the *y*-axis shows the percentage of the peak intensity compared to the intensity of the largest peak in that spectrum.

Supplementary Note: Cyclopeptides in the HUMANSTOOL dataset

Identification (Dereplicator) and *de novo* reconstruction (CycloNovo) of peptides in the HUMANSTOOL dataset. Supplementary Table S8 lists cyclopeptides identified by Dereplicator in the HUMANSTOOL datasets. Supplementary Table S9 lists CycloNovo reconstructions of 31 cyclopeptides in the HUMANSTOOL dataset.

precursor mass	peptide	PSM score	#reconstructions with score ≥ PSM score	P-value	peptide ID
1040.66	ILVPPFFLI	31	1	1.2×10^{-54}	cyclolinopeptide A
1058.61	MLIPPFFVI	24	1	2.3×10^{-42}	cyclolinopeptide B
1074.62	M^{+16}LIPPFFVI	16	1	9.7×10^{-18}	cyclolinopeptide C
1064.57	M^{+16}LLPFFWI	20	2	1.5×10^{-33}	cyclolinopeptide D
1082.52	M^{+16}LMPFFWI	19	1	1.2×10^{-31}	cyclolinopeptide H

977.56	M^{+16}LVFPLFI	25	1	1.6×10^{-43}	cyclolinopeptide E
961.55	MLVFPLFI	25	10	3.1×10^{-42}	cyclolinopeptide P
567.36	GIVIPS	11	1	2.1×10^{-17}	citrusin V

Supplementary Table S8. Cyclopeptides identified by Dereplicator in the HUMANSTOOL dataset. The correct sequence of all reconstructed cyclopeptides has the highest score among all reconstructions. For each cyclopeptide, the score of the correct cyclopeptide (column "PSM score"), the number of reconstructions with scores larger or equal to the PSM score (column "#reconstructions score ≥ PSM score"), and P-values are listed.

peptide mass	precursor mass	sequence of amino acid masses	score	P-value
1151.52	1152.53	87 99 99 101 147 129 71 97 113 71 137	19	2.3×10^{-27}
1098.49	549.75	147 71 87 137 57 99 71 129 163 137	23	6.6×10^{-24}
1085.53	543.27	99 99 137 57 113 115 186 71 137 71	20	6.2×10^{-30}
1081.51	1082.52	147 113 131 97 147 147 186 113	19	1.2×10^{-31}
1080.55	1081.56	87 99 99 101 147 129 71 97 113 137	22	3.0×10^{-36}
1073.61	1074.62	147 113 113 97 97 147 147 99 113	16	9.7×10^{-18}
1063.56	1064.57	147 113 113 97 147 147 186 113	20	1.5×10^{-33}
1060.75	530.74	147 163 57 99 57 137 71 129 71 129	22	2.7×10^{-29}
1057.61	1058.62	131 113 113 97 97 147 147 99 113	24	2.3×10^{-42}
1052.52	1053.53	87 99 101 147 129 71 97 113 71 137	25	8.6×10^{-38}
1039.65	1040.66	113 113 99 97 97 147 147 113 113	31	1.2×10^{-54}
1003.54	1004.55	71 97 147 99 147 97 147 99 99	13	4.0×10^{-17}
981.493	982.50	101 147 129 71 97 137 113 186	24	2.6×10^{-37}
978.543	979.55	128 113 147 87 113 57 99 87 147	18	6.1×10^{-25}
976.553	977.56	147 113 99 147 97 113 147 113	25	1.6×10^{-43}
960.553	961.56	131 113 99 147 97 113 147 113	25	3.1×10^{-42}
948.493	949.50	147 128 99 87 71 147 57 99 113	17	2.3×10^{-23}
891.463	892.47	113 147 101 71 57 57 99 99 147	15	3.4×10^{-19}
889.453	890.46	128 101 103 71 113 147 129 97	14	2.0×10^{-20}
888.493	889.50	57 97 147 113 113 99 99 163	15	9.7×10^{-21}
877.453	878.46	113 99 57 71 147 163 113 57 57	20	1.1×10^{-27}
873.403	874.41	71 97 115 71 97 101 87 97 137	17	2.7×10^{-23}
872.453	873.46	57 71 87 113 71 99 186 101 87	18	1.7×10^{-24}
871.453	872.46	147 101 71 57 99 114 97 57 128	19	2.5×10^{-26}
856.463	857.47	99 99 147 97 147 71 97 99	21	3.2×10^{-37}
841.433	842.44	97 147 101 57 97 113 128 101	20	2.2×10^{-29}
829.433	830.44	57 113 57 97 71 147 101 99 87	13	3.7×10^{-17}
826.393	827.40	147 99 137 71 71 99 115 87	17	5.8×10^{-23}
812.493	813.50	87 99 57 99 99 71 128 71 101	15	6.6×10^{-19}
811.413	812.42	147 113 57 57 57 97 99 97 87	16	2.0×10^{-20}
801.383	802.39	97 87 57 129 99 57 71 57 147	17	1.9×10^{-23}
695.273	696.28	71 57 163 129 57 71 147	18	6.1×10^{-27}

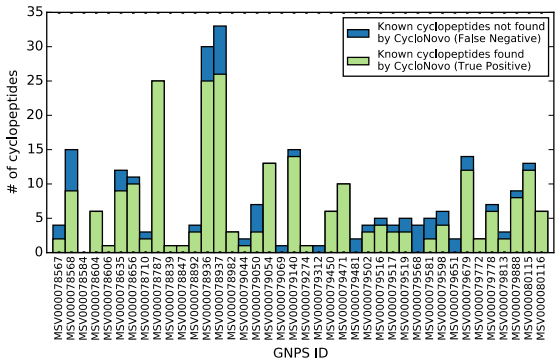
Supplementary Table S9. *De novo* reconstructions of 31 cyclopeptides in the HUMANSTOOL dataset. For each spectrum, its precursor mass, the *de novo* reconstruction (shown as a sequence of nominal masses of amino acids), the score, and the P-value are shown. *De novo* reconstructions are ordered in the decreasing order of their precursor masses. Precursor masses of spectra identified by Dereplicator are highlighted in blue. Cyclopeptides highlighted in green represent a novel cyclofamily described in the Supplementary Note "Novel cyclofamily in the HUMANSTOOL dataset."

Cyclospectra of branch-cyclic peptides in the HUMANSTOOL dataset. We classify a peptide as branch-cyclic if its backbone includes a cycle (with all monomers connected via amide bonds) and a side chain that includes at least one additional amide bond not included in the cycle. Although CycloNovo classify spectra of some branch cyclic peptides as cyclospectra (see Supplementary Note "Information about spectral datasets"), it is unable to *de novo* sequence them. Nevertheless, CycloNovo provides information about substrings of branch-cyclic peptides made of cyclopeptidic amino acids. For example, CycloNovo classified the spectrum of massetolide F in the HUMANSTOOL dataset as a cyclospectrum. The lipopetide massetolide F consists of the cycle TILSLSLV and a branch EL (along with a fatty acid chain tail with nominal mass 171 Da) connected to the cycle via an amide bond between T and E. We represent this branch cyclic peptides as a concatenate between the sequence of nominal masses of the cyclic and branch region separated by "*" sign, i.e., massetolide F is represented as 100, 113, 87, 113, 87, 113, 99 * 129, 113, 171. CycloNovo found five cyclopeptidic amino acids in massetolide F (S, I, L, V, T, and E) and missed the lipid chain (171 Da).

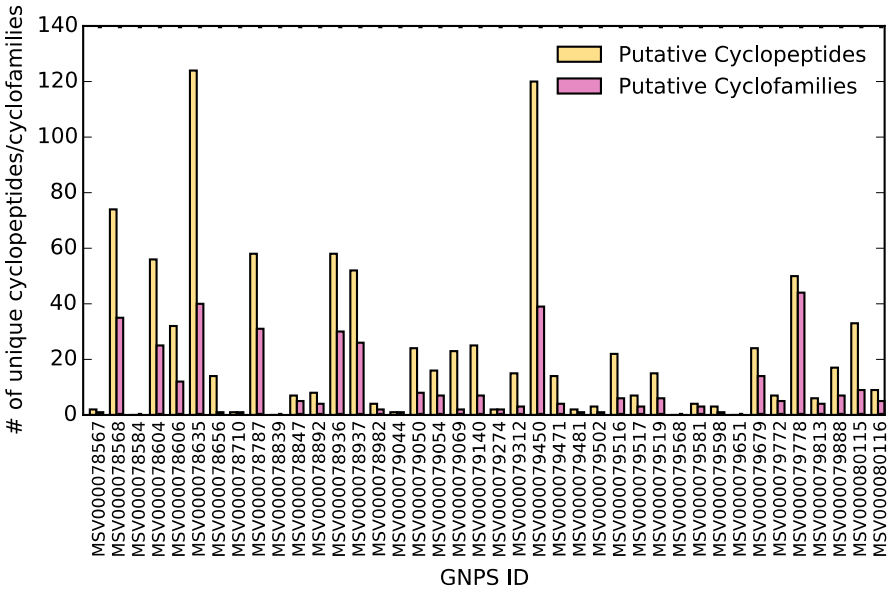
Assembly of the human stool sample where massetolide F was detected. We used metaSPAdes[10] to assemble the metagenomic dataset, generated from the stool sample (dated by 6/16/2014) where massetolide F was detected, This dataset includes 34.5 million paired reads which are assembled into 81 thousand scaffolds of lengths longer than 500 bp amounting to 407 Mb total assembly length.

Supplementary Note: Cyclopeptides in the GNPS dataset

Supplementary Figure S13 shows the number of identified cyclopeptides across all GNPS sub-datasets. Supplementary Figure S14 shows the number of cyclopeptides and cyclofamilies that gave rise to cyclospectra found by CycloNovo across all GNPS sub-datasets.



Supplementary Figure S13. Number of cyclopeptides identified by Dereplicator across all GNPS sub-dataset. Dereplicator identified 81 cyclopeptides in the GNPS dataset. Since some cyclopeptides are identified in multiple sub-datasets, the total numbers of identified cyclopeptides across all GNPS sub-datasets (180) exceeds 81. The green (blue) part of each bar represent spectra that were (were not) classified by CycloNovo as cyclospectra.



Supplementary Figure S14. Number of cyclopeptides (yellow) and cyclofamilies (pink) found by CycloNovo across all GNPS dataset.

SUPPLEMENTARY REFERENCES

1. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13,** 30–37 (2017).
2. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** W362-367 (2011).
3. Frank, A. M. *et al.* Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8,** 587–591 (2011).
4. Mingxun, W. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst. Press*
5. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta* **627,** 71–81 (2008).
6. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30,** 108–160 (2013).
7. Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V. CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. *Journal of the American Society for Mass Spectrometry* **24,** 1177–1184 (2013).
8. Gurevich, A. *et al.* Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **3,** 319–327 (2018).