5-28-2021

Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery.

Bahar Behsaz

Edna Bode

Alexey Gurevich

Yan-Ni Shi

Florian Grundmann

See next page for additional authors

Authors

Bahar Behsaz, Edna Bode, Alexey Gurevich, Yan-Ni Shi, Florian Grundmann, Deepa Acharya, Andrés Mauricio Caraballo-Rodríguez, Amina Bouslimani, Morgan Panitchpakdi, Annabell Linck, Changhui Guan, Julia Oh, Pieter C Dorrestein, Helge B Bode, Pavel A Pevzner, and Hosein Mohimani

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23502-4 OPEN

Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery

Bahar Behsaz[1,2,3], Edna Bode[4], Alexey Gurevich[5], Yan-Ni Shi[4], Florian Grundmann[4], Deepa Acharya[6], Andrés Mauricio Caraballo-Rodríguez[7], Amina Bouslimani[7], Morgan Panitchpakdi[7], Annabell Linck[4], Changhui Guan[8], Julia Oh[8], Pieter C. Dorrestein[2,7], Helge B. Bode[4,9,10]✉, Pavel A. Pevzner[3,11]✉ & Hosein Mohimani[3]✉

Non-Ribosomal Peptides (NRPs) represent a biomedically important class of natural products that include a multitude of antibiotics and other clinically used drugs. NRPs are not directly encoded in the genome but are instead produced by metabolic pathways encoded by *biosynthetic gene clusters* (BGCs). Since the existing genome mining tools predict many putative NRPs synthesized by a given BGC, it remains unclear which of these putative NRPs are correct and how to identify post-assembly modifications of amino acids in these NRPs in a blind mode, without knowing which modifications exist in the sample. To address this challenge, here we report NRPminer, a modification-tolerant tool for NRP discovery from large (meta)genomic and mass spectrometry datasets. We show that NRPminer is able to identify many NRPs from different environments, including four previously unreported NRP families from soil-associated microbes and NRPs from human microbiota. Furthermore, in this work we demonstrate the anti-parasitic activities and the structure of two of these NRP families using direct bioactivity screening and nuclear magnetic resonance spectrometry, illustrating the power of NRPminer for discovering bioactive NRPs.

[1] Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA. [2] Center for Microbiome Innovation, University of California at San Diego, La Jolla, CA, USA. [3] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [4] Molecular Biotechnology, Department of Biosciences, Goethe University Frankfurt, Frankfurt am Main, Germany. [5] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St Petersburg, Russia. [6] Tiny Earth Chemistry Hub, University of Wisconsin–Madison, Madison, WI, USA. [7] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [8] The Jackson Laboratory of Medical Genomics, Farmington, CT, USA. [9] Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt & Senckenberg Research Institute, Frankfurt am Main, Germany. [10] Max-Planck-Institute for Terrestrial Microbiology, Department for Natural Products in Organismic Interactions, Marburg, Germany. [11] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. ✉email: helge.bode@mpi-marburg.mpg.de; ppevzner@ucsd.edu; hoseinm@andrew.cmu.edu

Microbial natural products represent a major source of bioactive compounds for drug discovery[1]. Among these molecules, non-ribosomal peptides (NRPs) represent a diverse class of natural products that include antibiotics, immunosuppressants, anticancer agents, toxins, siderophores, pigments, and cytostatics[1–4]. NRPs have been reported in various habitats, from marine environments[5] to soil[3] and even human microbiome[6–9]. However, the discovery of NRPs remains a slow and laborious process because NRPs are not directly encoded in the genome and are instead assembled by non-ribosomal peptide synthetases (NRPSs).

NRPSs are multi-modular proteins that are encoded by a set of chromosomally adjacent genes called biosynthetic gene clusters (BGCs)[10,11]. Each NRP-producing BGC encodes for one or more genes composed of NRPS modules. Together the NRPS modules synthesize the core NRP in an assembly line fashion, with each module responsible for adding one amino acid to the growing NRP. Each NRPS module contains an Adenylation domain (A-domain) that is responsible for recognition and activation of the specific amino acid[12] that can be incorporated by that module through the non-ribosomal code[10] (as opposed to the genetic code). At minimum, each NRPS module also includes a Thiolation domain (T-domain) and a Condensation domain (C-domain) that are responsible for loading and elongation of the NRP scaffold, respectively. Additionally, an NRPS module may include additional domains such as Epimerization domain (E-domain) or dual-function Condensation/Epimerization domain (C/E domain). An “NRPS assembly line” refers to a sequence of NRPS modules that together assemble a core NRP. The core NRP often undergoes post-assembly modifications (PAMs) that transform it into a mature NRP. The order of the modules in an NRPS assembly line can be different from the order of NRPS modules encoded in the BGC through iterative use of NRPS modules[13,14].

In the past decade, genome mining methods have been developed for predicting the NRP sequences from their BGC sequences[15,16]. Genome mining tools, such as antiSMASH[17], start by identifying the NRPS BGCs in a microbial genome using Hidden Markov Models (HMMs). Afterwards, they identify NRPS modules and predict the amino acids incorporated by the A-domain in each module using the substrate prediction algorithms (such as NRPSpredictor2 (ref. [15]) or SANDPUMA[18]) that are based on machine learning techniques trained on a set of A-domains with known specificities[16,18]. For each observed A-domain, these algorithms predict a set of amino acids potentially recruited by that A-domain, along with the specificity score reflecting confidence of each amino acid prediction. The use of genome mining is becoming increasingly popular for discovering NRPs over the past decade[19–21], demonstrating the potential of large-scale (meta)genomic projects for NRP discovery.

Although genome mining tools like SMURF[22] and antiSMASH[17] greatly facilitate BGC analysis, the core NRPs (let alone mature NRPs) for the vast majority of sequenced NRP-producing BGCs (>99%) remain unknown[23,24]. Identification of NRP-producing BGCs, without revealing the final molecular products they encode, does not capture its full potential for finding bioactive compounds[25]. Thus, integrating (meta)genome mining with metabolomics is necessary for realizing the true promise of large-scale NRP discovery[4]. However, the existing genome mining strategies fail to reveal the chemical diversity of NRPs. For example, these methods fall short in correctly identifying PAMs, which are a unique feature of NRPs that make them the most diverse class of natural products[26] and play a crucial role in their mode of action[27,28]. As a result, the promise of large-scale NRP discovery has not yet been realized[29].

Discovery of NRPs involves a multitude of challenges such as PAM identification (with exception of methylation and epimerization[17], genome mining tools fail to identify PAMs) and accounting for substrate promiscuity of A-domains. The substrate promiscuity in NRP biosynthesis refers to the ability of an A-domain in an NRPS to incorporate several different amino acids into the NRP. The existing genome mining tools often predict a set of incorporated amino acids and output a ranked list of multiple amino acids for each A-domain. Allowing for all amino acid possibilities for each A-domain in an NRPS module results in a large number of putative NRPs predicted from each BGC. Without additional complementary data (such as mass spectra of NRPs), the genome mining approaches cannot identify the correct NRP among the multitude of putative NRPs[29,30].

Another challenge in discovering NRPs is due to the non-canonical assembly lines. While in many NRPSs each A-domain incorporates exactly one designated amino acid and the sequence of amino acids in NRP matches the order of the A-domains in the BGC[13,31,32] (see Supplementary Fig. 1a), there are many NRP families that violate this pattern[7,11,32–39]. Since an NRPS system may have multiple assembly lines[40], one needs to consider different combinations of NRPS units encoded by each open reading frames (ORFs) for finding the core NRPs[27,40]. In some non-canonical assembly lines, A-domains encoded by at least one ORF may be incorporated multiple times (in tandem) in the NRPS[7,34–36] (Supplementary Fig. 1b). For example, during biosynthesis of rhabdopeptides[34,38] and lugdunins[7], a single ORF encodes for one Val-specific NRPS module that loads multiple Val in the final NRPs. Moreover, in some NRPS assembly lines, the A-domains in some ORFs do not contribute to the core NRP[32,37,41] (see Supplementary Fig. 1c). For example, surugamide BGC[30,32,42,43] from *Streptomyces albus* produces two completely distinct NRPs through different non-canonical assembly lines (Supplementary Fig. 2). The non-canonical biosynthesis of surugamide makes its discovery particularly difficult as one need to account for these non-canonical assembly lines by generating different combinations of ORFs in the process of building a database of putative NRPs for each BGC.

Other hurdles include lack of sufficient training data for many A-domains, which can lead to specificity mispredictions[18] and complications in the genome mining due to fragmented assemblies (e.g. failure to capture a BGC in a single contig[44]). These challenges, in combination with those mentioned above, make it nearly impossible to accurately predict NRPs based solely on genome mining. The problem gets even more severe for NRP discovery from microbial communities.

To address these challenges, multiple peptidogenomics approaches have been developed for discovering peptidic natural products by combining genome mining and mass spectrometry (MS) information[30,45]. These approaches often use antiSMASH[16] to find all NRPS BGCs in the input genome, use NRPSPredictor2 (ref. [15]) to generate putative core NRPs encoded by each BGC, and attempt to match mass spectra against these putative NRPs. Kersten et al.[44] described a peptidogenomics approach based on manually inferring amino acid sequence tags (that represent a partial sequence of an NRP) from mass spectra and matching these tags against information about the substrate specificity generated by NRPSpredictor2 (ref. [15]). Nguyen et al.[46,47] and Tobias et al.[31] presented a manual approach for combining genome mining with molecular networking. In this approach, which is limited to the identification of previously unreported variants of known NRPs, molecules present in spectral families with known compounds are compared to BGCs.

Medema et al.[40] complemented the manual approach from Kersten et al.[44] by the NRP2Path[40] tool for searching the sequence tags against a collection of BGCs. NRP2Path starts with a set of sequence tags manually generated for each spectrum, considers multiple assembly lines for each identified BGC, and forms a database of all possible core NRPs for this BGC. Then, NRP2Path[40] computes a match score between each tag and each core NRP (using the specificity scores provided by NRPSpredictor2 (ref. [15])) and reports high-scoring matches as putative core NRPs. The success of this approach relies on inferring long sequence tags of 4–5 amino acids, which are usually absent in spectra of non-linear peptides. Such long sequence tags are often missing in NRPs with macrocyclic backbones and complex modifications, limiting the applicability of NRP2Path[44,48]. Moreover, NRP2Path is not able to identify enzymatic modifications (e.g. methylation) and PAMs in the final NRPs and is unable to predict the backbone structure of the mature NRPs (e.g. linear/cyclic/branch-cyclic).

Mohimani et al.[30] developed an automated NRPquest approach that takes paired MS and genomic datasets as input and searches each mass spectrum against all structures generated from putative core NRPs to identify high-scoring peptide-spectrum matches (PSMs). NRPquest leverages the entire mass spectrum (instead of just the sequence tags) to provide further insights into the final structure of the identified NRPs. They proposed using modification-tolerant search of spectral datasets against the core NRPs structures, for identifying PAMs in a blind mode (that is without knowing which PAMs exist in the sample). This is similar to identifying post-translational modifications in traditional proteomics[49]. The presence of covalent modifications in peptides affects the molecular weight of the modified amino acids; therefore, the mass increment or deficit can be detected using MS data[43,49]. However, as NRPquest uses a naïve pairwise scoring of all NRP structures against all mass spectra for PAM identification, it is prohibitively slow when searching for PAMs[30]. Furthermore, NRPquest does not handle non-canonical NRPS assembly lines and it does not provide statistical significance of identified NRPs, a crucial step for large-scale analysis.

On the other hand, development of high-throughput MS-based experimental and computational natural products discovery pipelines[29] such as the Global Natural Products Social (GNPS) molecular networking[50], PRISM[51], GNP[52], RODEO[53], Dereplicator+[54], CSI:FingerID[55], NAP[56], and CycloNovo[48] have permanently changed the field of peptide natural product discovery. The GNPS project has already generated nearly half a billion of information-rich tandem mass spectra (MS), an untapped resource for discovering bioactive molecules. However, the utility of the GNPS network is mainly limited to the identification of previously discovered molecules and their analogs. Currently, only about 5% of the GNPS spectra are annotated[50], emphasizing the need for algorithms that can annotate such large spectral datasets.

In this work, we present NRPminer a scalable modification-tolerant tool for analyzing paired MS and (meta)genomic datasets (Fig. 1). NRPminer uses the specificity scores of the amino acids appearing in core NRPs to perform an efficient search of all spectra against all core NRPs. In addition to predicting the amino acid sequence of an NRP generated by a BGC, NRPminer also analyzes various non-canonical assembly lines and efficiently predicts potential PAMs and backbone structures. We show NRPminer identifies 180 unique NRPs representing 18 distinct NRP families, including four previously unreported ones, by analyzing only four MS datasets in GNPS against their corresponding reference genomes.

Results

Outline of the NRPminer algorithm. Figure 1 illustrates the NRPminer algorithm. All NRPminer's steps are described in detail in the "Methods" section. Briefly, NRPminer starts by (a) identifying the NRPS BGCs in each genome (using antiSMASH[16]) and (b) predicting the putative amino acids for each identified A-domain (using NRPSpredictor2 (ref. [15])). Then, it accounts for (c) different NRPS assembly lines by considering various combinations of ORFs in the BGCs. NRPminer (d) filters the set of all core NRPs based on the specificity scores of their amino acids and selects those with high scores. It, next, (e) searches each BGC to find known modification enzymes (e.g. methylation) and incorporates them in the corresponding core NRPs. Then, (f) NRPminer constructs a database of putative NRP structures by considering linear, cyclic, and branch-cyclic backbone structures for each core NRP. Afterwards, (g) it performs a modification-tolerant search of the input spectra against the constructed database of putative NRPs and computes the statistical significance of PSMs. Finally, (h) NRPminer reports the statistically significant PSMs. These identifications are then (i) expanded using spectral networks[57] approach.

Datasets. We analyzed four microbial isolate datasets from *Xenorhabdus* and *Photorhabdus* families (XPF), *Staphylococcus* (SkinStaph), soil-dwelling Actinobacteria (SoilActi), and a collection of soil-associated bacteria within *Bacillus*, *Pseudomonas*, *Buttiauxella*, and *Rahnella* genera generated under the Tiny Earth antibiotic discovery project[58,59] (TinyEarth); all available from GNPS/MassIVE repository. The process of growth of the isolates and MS experiments are described in the "Methods" section (under "Sample preparation and MS experiments). The spectra collected on each of these datasets are referred to as $\text{spectra}_{\text{XPF}}$, $\text{spectra}_{\text{SkinStaph}}$, $\text{spectra}_{\text{SoilActi}}$, $\text{spectra}_{\text{TinyEarth}}$, and the genomes are referred as $\text{genome}_{\text{XPF}}$, $\text{genome}_{\text{SkinStaph}}$, $\text{genome}_{\text{SoilActi}}$, and $\text{genome}_{\text{TinyEarth}}$, respectively.

Summary of NRPminer results. Table 1 summarizes the NRPminer results for each dataset. NRPminer classifies a PSM as statistically significant if its p value is below the default conservative threshold 10^{-15}. The number of distinct NRPs and NRP families was estimated using MS-Cluster[60] and SpecNets[50] using the threshold cos < 0.7 (see "Methods" section). Two peptides are considered to be variants/modifications of each other if they differ in a single modified residue due to changes by tailoring enzymes, enzyme promiscuity, or through changes in the amino acid specificity at the genetic level[47]. Known NRPs (NRP families) are identified either by Dereplicator[42] search against the database of all known peptidic natural products[43] (referred to as PNPdatabase) using the p value threshold 10^{-15}, and/or by SpecNet[57] search against the library of all annotated spectra available on GNPS[50]. NRPminer ignores any BGCs with less than three A-domains and spectra that include less than 20 peaks.

Generating putative core NRPs. Table 1 presents the number of NRP-producing BGCs and the number of putative core NRPs generated by NRPminer for each analyzed genome in XPF (before and after filtering). For example, NRPminer identified eight NRP-producing BGCs and generated 253,027,076,774 putative core NRPs for *X. szentirmaii* DSM genome. After filtering putative core NRPs based on the sum of the specificity scores reported by NRPSpredictor2 (ref. [15]), only 29,957 putative core NRPs were retained (see "Methods" section for the details of filtering). Therefore, filtering putative core NRPs is an essential step for making the search feasible.

Analysis of the paired genomic and spectral datasets. NRPminer has a one-vs-one mode (each MS dataset is searched against a single genomic dataset) and a one-vs-all mode (each MS dataset

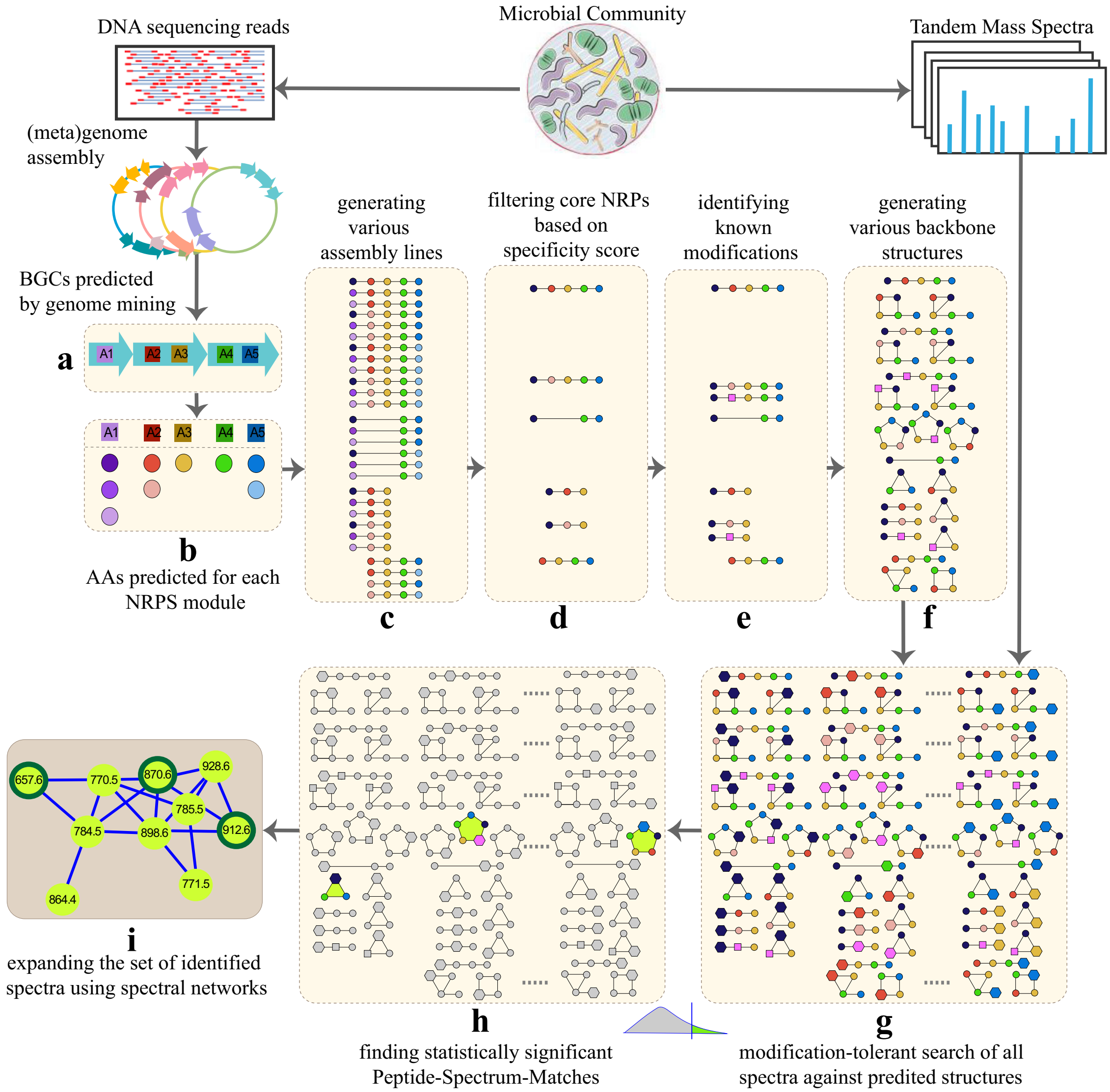

DNA sequencing reads
Microbial Community
Tandem Mass Spectra
(meta)genome assembly
BGCs predicted by genome mining
a
A1
A2
A3
A4
A5
b
AAs predicted for each NRPS module
generating various assembly lines
c
filtering core NRPs based on specificity score
d
identifying known modifications
e
generating various backbone structures
f
657.6
770.5
870.6
928.6
785.5
784.5
898.6
912.6
864.4
771.5
i
expanding the set of identified spectra using spectral networks
h
finding statistically significant Peptide-Spectrum-Matches
g
modification-tolerant search of all spectra against predited structures

Fig. 1 NRPminer pipeline. a Predicting NRPS BGCs using antiSMASH[16]. Each ORF is represented by an arrow, and each A-domain is represented by a square, **b** predicting putative amino acids for each NRP residue using NRPSpredictor2 (ref. [15]), colored circles represents different amino acids (AAs), **c** generating multiple assembly lines by considering various combinations of ORFs and generating all putative core NRPs for each assembly line in the identified BGC (for brevity only assembly lines generated by deleting a single NRPS unit are shown; in practice, NRPminer considers loss of up to two NRPS units, as well as single and double duplication of each NRPS unit), **d** filtering the core NRPs based on their specificity scores, **e** identifying domains corresponding to known modifications and incorporating them in the selected core NRPs (modified amino acids are represented by purple squares), **f** generating linear, cyclic and branch-cyclic backbone structures for each core NRP, **g** generating a set of high-scoring PSMs using modification-tolerant VarQuest[43] search of spectra against the database of the constructed putative NRP structures. NRPminer considers all possible mature NRPs with up to one PAM (shown as hexagons) in each NRP structure. For brevity some of the structures are not shown. **h** Computing statistical significance of PSMs and reporting the significant PSMs, and **i** expanding the set of identified spectra using spectral networks[57]. Nodes in the spectral network represent spectra and edges connect "similar" spectra (see "Methods").

Table 1 Summary of NRPminer search results on the XPF, SkinStaph, SoilActi, and TinyEarth datasets.

Dataset	#strains	#identified PSMs/ #spectra	#distinct NRPs (families)	#known NRPs (families)	#preiviously unreported variants of known NRPs	#previuosly unreported NRPs (families)
XPF	27	3023/263,768	122 (12)	21 (9)	79	22 (3)
SkinStaph	171	23/2,657,398	3 (1)	2 (1)	1	0
SoilActi	20	206/362,421	24 (2)	7 (1)	14	3 (1)
TinyEarth	28	498/380,414	31 (3)	29 (3)	2	0

Column "#strains" shows the number of microbial strains. Column "#identified PSMs/#spectra" shows the number of PSMs identified by NRPminer and the total number of spectra. The column "#distinct NRPs (families)" shows the number of unique NRPs (unique families). The number of unique NRPs is estimated using MS-Cluster60, and the number of unique families is estimated using SpecNets50. The column "#known NRPs (families)" shows the number of known NRPs (families) among all identified NRPs (families). Column "#previously unreported variants of known NRPs" shows the number of NRPs in the known families that were not reported before. Column "#previously unreported NRPs (families)" shows the number of previously unreported NRPs (families) that are not variants of any known NRPs.

is searched against a collection of genomic datasets within a taxonomic clade). While the one-vs-all mode is slower than the one-vs-one mode, it is usually more sensitive. For example, a BGC may be fragmented (or misassembled) in the draft assembly of one strain, but a related BGC may be correctly assembled and captured within a single contig in a related well-assembled strain. If these two BGCs synthesize the same (or even similar) NRP, NRPminer may be able to match the spectra from a poorly assembled strain to a BGC from a related well-assembled strain.

For example, NRPminer search of spectra$_{XPF}$ against genome$_{XPF}$ generated 3023 PSMs that represent 122 NRPs from 12 NRP families. Figure 2 shows the spectral network representing 12 NRP families identified by NRPminer in the XPF dataset.

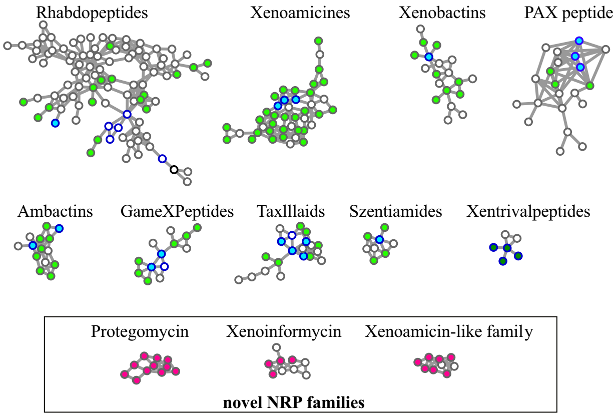


Fig. 2 Spectral networks for nine known and three previously unreported NRP families identified by NRPminer in the XPF dataset. Each node represents a spectrum. The spectra of known NRPs (as identified by spectral library search against the library of all known compounds in GNPS) are shown with a dark blue border. A node is colored if the corresponding spectrum forms a statistically significant PSM and not colored otherwise. We distinguish between identified spectra of known NRPs with known BGCs[31] (colored by light blue) and identified spectra of known NRPs (from xentrivalpeptide family) with previously unknown BGC (colored by dark green). Identified spectra of previously unreported NRPs from known NRP families (previously unreported NRP variants) are colored in light green. Identified spectra of NRPs from previously unreported NRP families are colored in magenta. Proteogomycins and xenoinformycin subnetworks represent previously unreported NRP families. The Xenoamicin-like subnetwork revealed a BGC family distantly related to xenoamicins (6 out 13 amino acids are identical). For simplicity only spectra at charge state +1 are used for the analysis.

SpecNet analysis against the annotated spectra in GNPS[50] showed that 9 out of 12 identified NRP families is known (reported by Tobias et al.[31]). NRPminer identified PAX-peptides family and their corresponding BGC in *X. nematophila* ATCC 19061 in the XPF dataset even though these NRPs include lipid side chains that are not predictable via genome mining. NRPminer failed to identify only one additional known family which was reported by Tobias et al.[31] (xefoampeptides) that has an ester bond between a hydroxy-fatty acid and the terminal amino acid with total mass exceeding the default NRPminer threshold (150 Da). Xefoampeptides are depsipeptides composed of a 3-hydroxy-fatty acid (total mass over 200 Da) and only three amino acids, resulting in a poorly fragmented spectrum that did not generate statistically significant PSMs against the putative structures generated from their corresponding core NRPs. Table 2 provides information about NRPminer-generated PSMs representing known NRP families. Among the nine known NRP families (in the XPF dataset) listed in Table 2, eight families have been connected to their BGCs in the previous studies, and for these families, the corresponding BGCs discovered by NRPminer are consistent with the literature[31] (see Supplementary Table 2 for the list of identified BGCs). Supplementary Figure 3 presents an example of an identified NRP family, szentiamide, and its corresponding BGC in *X. szentirmaii*. For one family (xentrivalpeptides) with no known BGC, we were able to predict the putative BGC (Supplementary Fig. 4). Furthermore, NRPminer identified 79 previously unreported NRP variants across these nine known NRP families. In addition to the known NRP families, NRPminer also discovered three NRP families (protegomycins, xenoinformycins, and xenoamicin-like family) in XPF dataset that includes no previously reported NRPs.

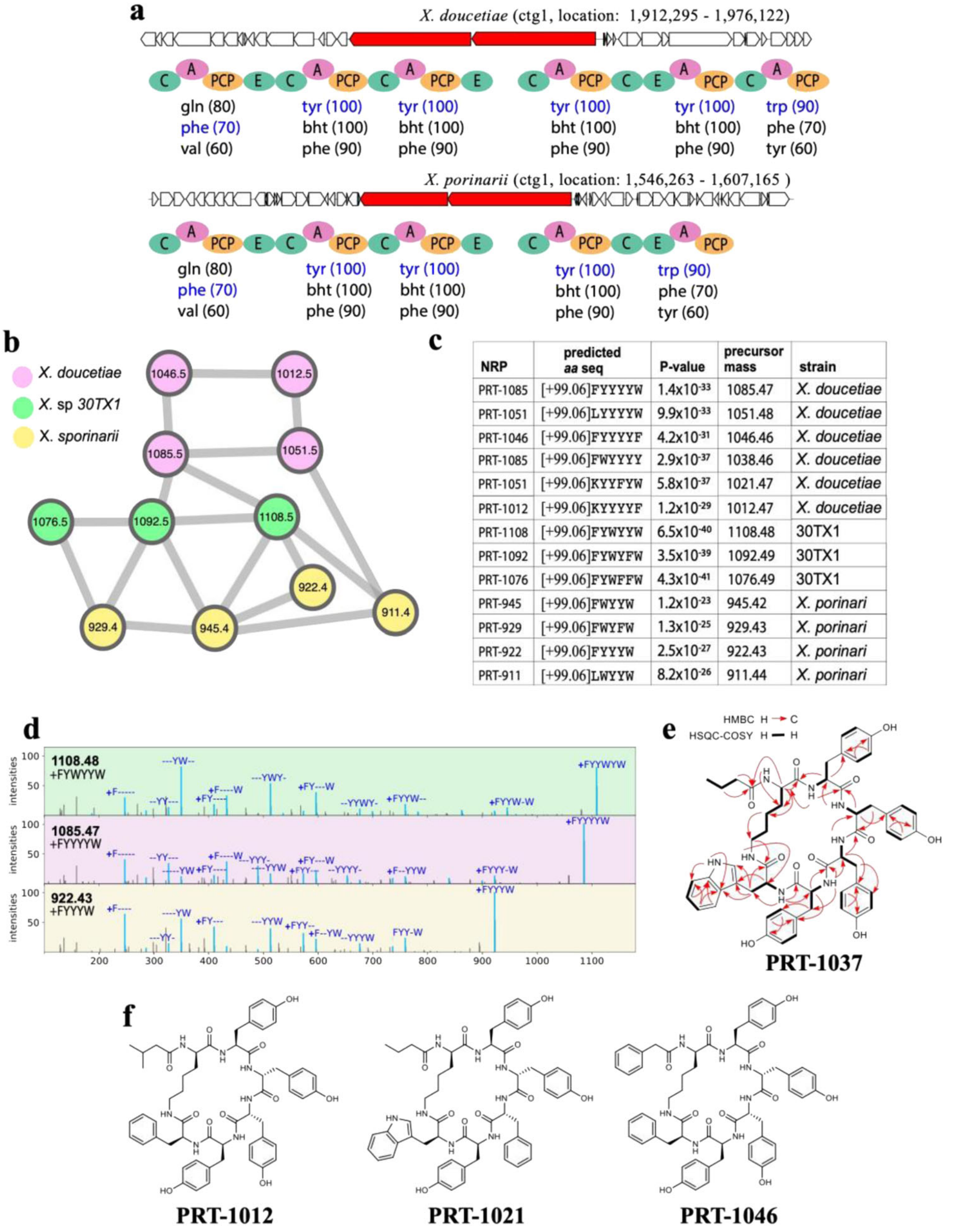
We named each identified NRP in a previously unreported family by combining the name of that family with the nominal precursor mass of the spectrum representing that NRP (with the lowest *p* value among all spectra originating from the same NRP). In what follows, we describe the four previously unreported NRP families identified by NRPminer (protegomycin, xenoinformycin, and xenoamicin-like family in the XPF dataset and aminformatide in SoilActi), as well as the previously unreported variants in two additional NRP families (lugdunin in SkinStaph and surugamide in SoilActi).

Discovery of protegomycin (PRT) NRP family in the XPF dataset. NRPminer matched 28 spectra representing 11 previously unreported cyclic NRPs to two BGCs. These spectra are from species *X. doucetiae*, *Xenorhabdus* sp. 30TX1, and *X. poinarii*. The BGCs were from *X. doucetiae* and *X. poinarii* with six and five A-domains, respectively, with one PAM (Fig. 3).

Table 2 Predicted amino acids for the eight A-domains appearing on cyclic surugamides A–D assembly line SurugamideAL.

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
Val (100)	**Phe (100)**	Tyr (100)	Val (100)	**Ala (100)**	Val (100)	Val (100)	Met (100)
Ile (80)	Tyr (90)	Phe (100)	**Ile (100)**	Ser (87)	**Ile (100)**	**Ile (100)**	Apa (100)
Abu (70)	Bht (90)	**Leu (100)**	Abu (70)	Pro (75)	Abu (70)	Abu (70)	Glu (86)
				Val (75)			Arg (86)
				Cys (75)			Gln (86)
				Phe (75)			**Lys (86)**
				Gly (75)			Asp (86)
							Val (86)
							Orn (86)

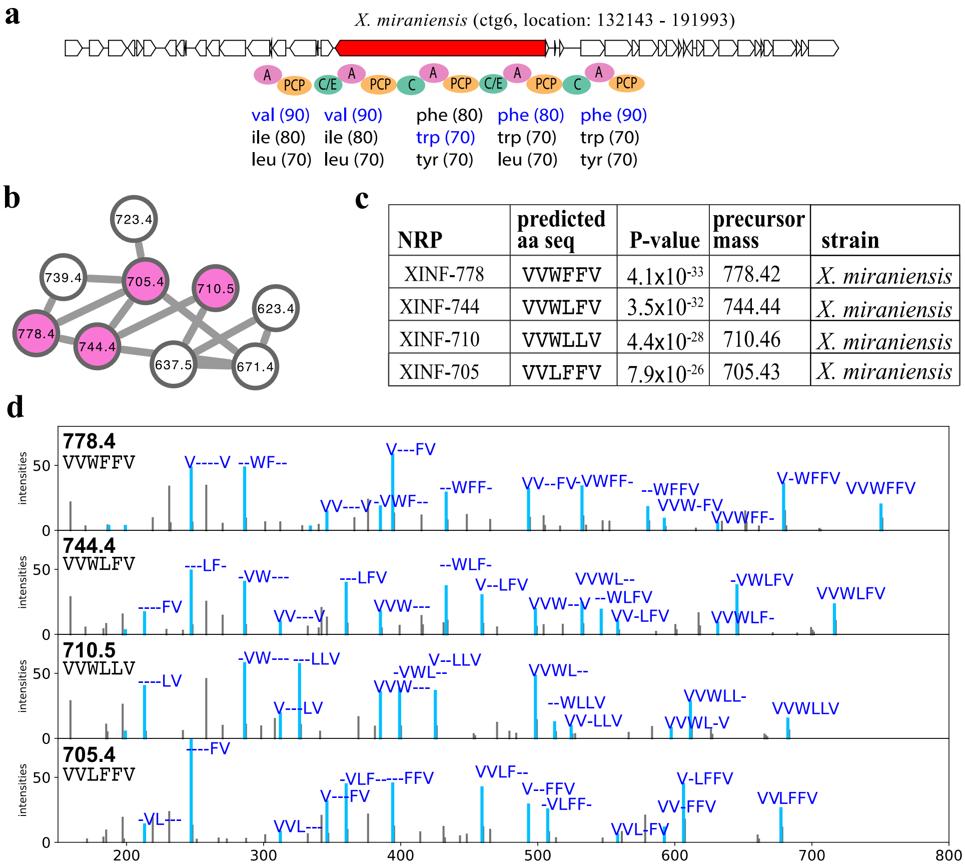
A_i represents the set of amino acids predicted for the *i*th A-domain in SurugamideAL. For each A_i at least three amino acids with the highest normalized specificity scores (listed in parentheses) are presented. Amino acids appearing in surugamide A (IFLIAIIK) are shown in bold. NRPminer considers all amino acids with the same normalized specificity score, as illustrated in the case of the fifth and the eighth A-domains.


a
X. doucetiae (ctg1, location: 1,912,295 - 1,976,122)
C A PCP E C A PCP C A PCP E C A PCP C E A PCP C A PCP
gln (80) phe (70) val (60)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
trp (90) phe (70) tyr (60)
X. porinarii (ctg1, location: 1,546,263 - 1,607,165)
C A PCP E C A PCP C A PCP E C A PCP C E A PCP
gln (80) phe (70) val (60)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
trp (90) phe (70) tyr (60)
b
X. doucetiae
X. sp 30TX1
X. sporinarii
1046.5
1012.5
1085.5
1051.5
1076.5
1092.5
1108.5
922.4
911.4
929.4
945.4
c
d
intensities
100
50
1108.48 +FYWYYW
+FYYWYW
---YW--
---YWY-
+F-----
--YY---
+F----W
+FY----
+FY---W
--YYWY-
+FYYW--
+FYYW-W
1085.47 +FYYYYW
+FYYYYW
+F-----
--YY---
+F----W
--YW
+FY----
-YYY-
--YYW
+FY---W
+FY---
--YYYY-
+F--YYW
+FYYY-W
922.43 +FYYYW
+FYYYW
+F----
----YW
---YY-
+FY---
---YYW
+FYY-
+F--YW
--YYYW
FYY-W
200 300 400 500 600 700 800 900 1000 1100
e
HMBC H → C
HSQC-COSY H — H
PRT-1037
f
PRT-1012
PRT-1021
PRT-1046

NRP	predicted *aa* seq	P-value	precursor mass	strain
PRT-1085	[+99.06]FYYYYW	1.4×10^{-33}	1085.47	*X. doucetiae*
PRT-1051	[+99.06]LYYYYW	9.9×10^{-33}	1051.48	*X. doucetiae*
PRT-1046	[+99.06]FYYYYF	4.2×10^{-31}	1046.46	*X. doucetiae*
PRT-1085	[+99.06]FWYYYY	2.9×10^{-37}	1038.46	*X. doucetiae*
PRT-1051	[+99.06]KYYFYW	5.8×10^{-37}	1021.47	*X. doucetiae*
PRT-1012	[+99.06]KYYYYF	1.2×10^{-29}	1012.47	*X. doucetiae*
PRT-1108	[+99.06]FYWYYW	6.5×10^{-40}	1108.48	30TX1
PRT-1092	[+99.06]FYWYFW	3.5×10^{-39}	1092.49	30TX1
PRT-1076	[+99.06]FYWFFW	4.3×10^{-41}	1076.49	30TX1
PRT-945	[+99.06]FWYYW	1.2×10^{-23}	945.42	*X. porinari*
PRT-929	[+99.06]FWYFW	1.3×10^{-25}	929.43	*X. porinari*
PRT-922	[+99.06]FYYYW	2.5×10^{-27}	922.43	*X. porinari*
PRT-911	[+99.06]LWYYW	8.2×10^{-26}	911.44	*X. porinari*

Fig. 3 Identifying protegomycin (PRT) NRP family. a The BGCs generating the NRP in *X. doucetiae* (top) and *X. porinarii* (bottom) along with NRPS genes (shown in red) and A-, C-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs is shown in white. No BGC was found in *Xenorhabdus* sp. 30TX1. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2 (ref. [15]) predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRPs [+99.06]FYYYYW and [+99.06]FYYYW identified by NRPminer (with the lowest *p* value) are shown in blue. **b** Spectral network formed by the spectra that originate from NRPs in the protegomycin family. **c** Sequences of the identified NRPs in the protegomycin family (with the lowest *p* value among all spectra originating from the same NRP). PRT represents protegomycin. For MS details see Supplementary Table 3. The *p* values are computed based on MCMC approach using MS-DPR[89] with 10,000 simulations. **d** For each strain, an annotated spectrum representing the lowest *p* value is shown. The spectra were annotated based on predicted NRPs [+99.06]FYYWYW, [+99.06] FYYYYW, and [+99.06]FYYYW from top to bottom. The "+" sign represents the addition of [+99.06 Da]. Colors in parts **b** and **d** are coordinated. Supplementary Figures 6–8 show the annotated spectra for all NRPs shown in part (**c**). **e** Key HMBC and HSQC-COSY correlations in PRT-1037. **f** Structures for selected PRT derivatives produced by *X. doucetiae* including amino acid configuration as concluded from the presence of epimerization domains in the corresponding NRPSs and acyl residues as concluded from feeding experiments (Supplementary Fig. 9). Predicted structures for all identified PRT derivatives from *X. doucetiae*, *X. poinarii*, and 30TX1 are shown in Supplementary Figs. 10 and S11.

Additional derivatives were found in large-scale cultivation of wild type and Δ*hfq* mutants of *X. doucetiae* (Supplementary Fig. 5 and "Methods" section under "Additional Analyses for Protegomycin Family"). No BGC was found in *Xenorhabdus* sp. 30TX1 due to highly fragmented assembly. The spectra representing the three protegomycins produced by *Xenorhabdus* sp. 30TX1 did not match any core NRP generated from its genome because the corresponding BGC was not assembled in a single contig in this genome. However, they were identified with statistically significant *p* values using the one-vs-all search when these spectra were searched against core NRPs from *X. doucetiae* genome (Fig. 3) that included an orthologous BGC in a single contig.



NRP	predicted aa seq	P-value	precursor mass	strain
XINF-778	VVWFFV	4.1×10^{-33}	778.42	*X. miraniensis*
XINF-744	VVWLFV	3.5×10^{-32}	744.44	*X. miraniensis*
XINF-710	VVWLLV	4.4×10^{-28}	710.46	*X. miraniensis*
XINF-705	VVLFFV	7.9×10^{-26}	705.43	*X. miraniensis*

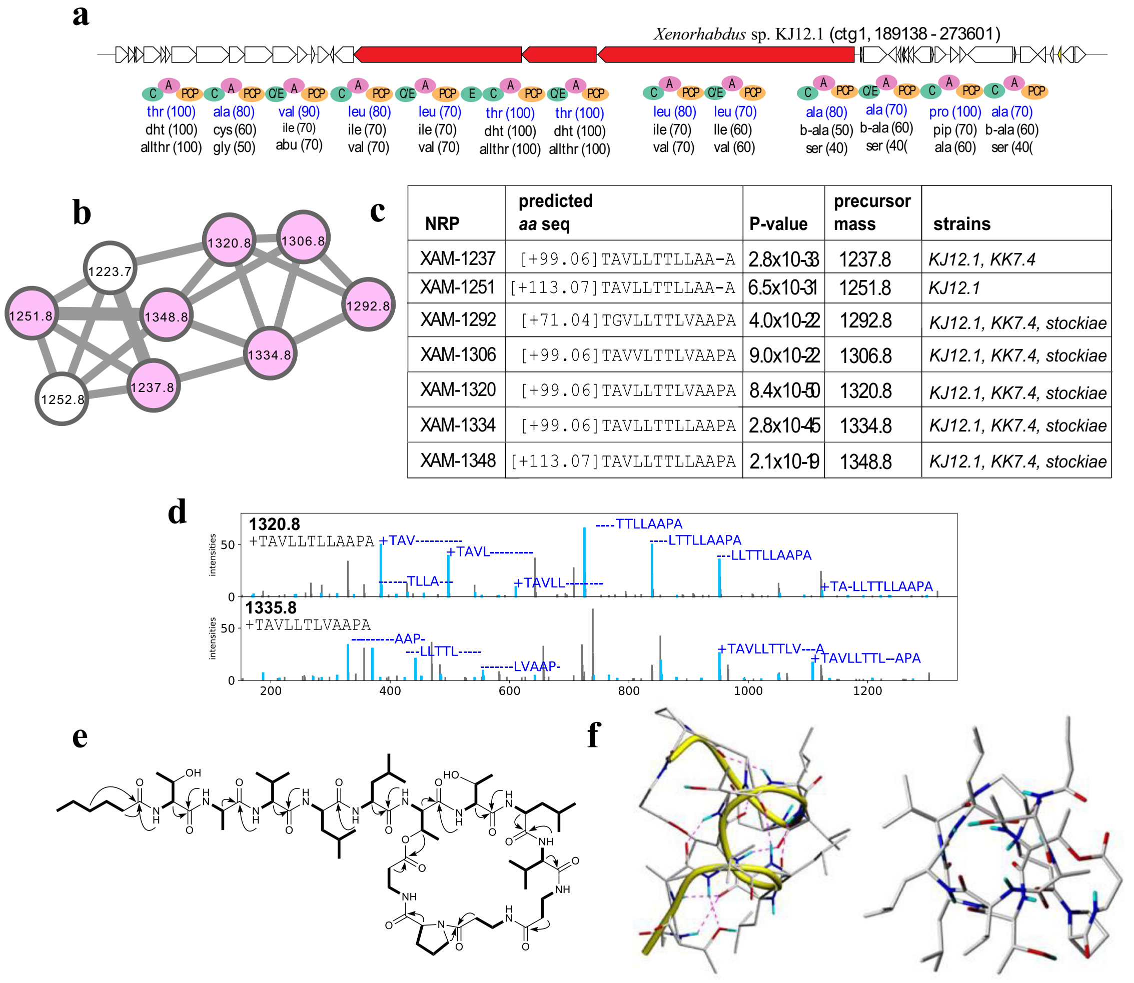
Fig. 4 Identifying xenoinformycin (XINF) NRP family. a The BGC generating the NRP in *X. miraniensis* along with NRPS genes (shown in red) and the A-, C-, PCP-, and C/E-domains appearing on the corresponding NRPS. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in this BGC (according to NRPSpredictor2 (ref. [15]) predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP VVWFF identified by NRPminer (with the lowest *p* value) are shown in blue. **b** Spectral network formed by the spectra that originate from NRPs in the xenoinformycin family. A node is colored if the corresponding spectrum forms a statistically significant PSM (with *p* value threshold 10^{-15}) and not colored otherwise. **c** Sequences of the identified NRPs in the xenoinformycin family (with the lowest *p* value among all spectra originating from the same NRP). XINF represents xenoinformycin. The *p* values are computed based on MCMC approach using MS-DPR[89] with 10,000 simulations. **d** For each identified NRP, an annotated spectrum forming a PSM with the lowest *p* value is shown.

Figure 3, Supplementary Figs. 6–11, and Supplementary Table 3 present information about protegomycin BGC and NRPs.

We further conducted nuclear magnetic resonance (NMR) spectroscopy on one of the major derivatives (Fig. 3e, f and Supplementary Figs. 12–18 and Supplementary Table 4). Our NMR results confirmed the MS results, with the distinction that NMR revealed a short chain fatty acid like phenylacetic acid (PAA) as a starting unit (incorporated by the C-starter domain), followed by a Lys that is cyclized to the terminal thioester by the C-terminal TE domain. NRPminer predicted Phe instead of the correct amino acid Lys, since NRPSpredictor2 made an error in identifying the amino acid for the corresponding A-domain (see Fig. 3a for the list of predicted amino acids). It has been shown that NRPSpredictor2 (ref. [15]) often fails to predict Lys residues, due to lack of training data for this amino acid[15]. Furthermore, as with any other MS-based method, NRPminer was not able to distinguish between residues with the same molar mass in the structure of final NRP, such as the pair Ala and β-Ala. All other NRPminer predictions of individual amino acids were consistent with NMR.

Besides PAA, other starter acyl units are isovaleric acid (in PRT-1012; NRPminer prediction 99.06+Leu; see Fig. 3f) and butyric acid (in PRT-1037; see Fig. 3e). Supplementary Figure 9 describes labeling data and mass spectra for the identified protegomycins in *X. doucetiae*. The isolated derivatives PRT-1037 and PRT-1021 (Fig. 3e, f) were tested against various protozoa and showed a weak activity against *Trypanosoma brucei rhodesiense* (IC_{50} [mg/L] 79 and 53) and *Plasmodium falciparum* (IC_{50} [mg/L] > 50 and 33) with no toxicity against L6 rat myoblast cells (IC_{50} [mg/L] both >100).

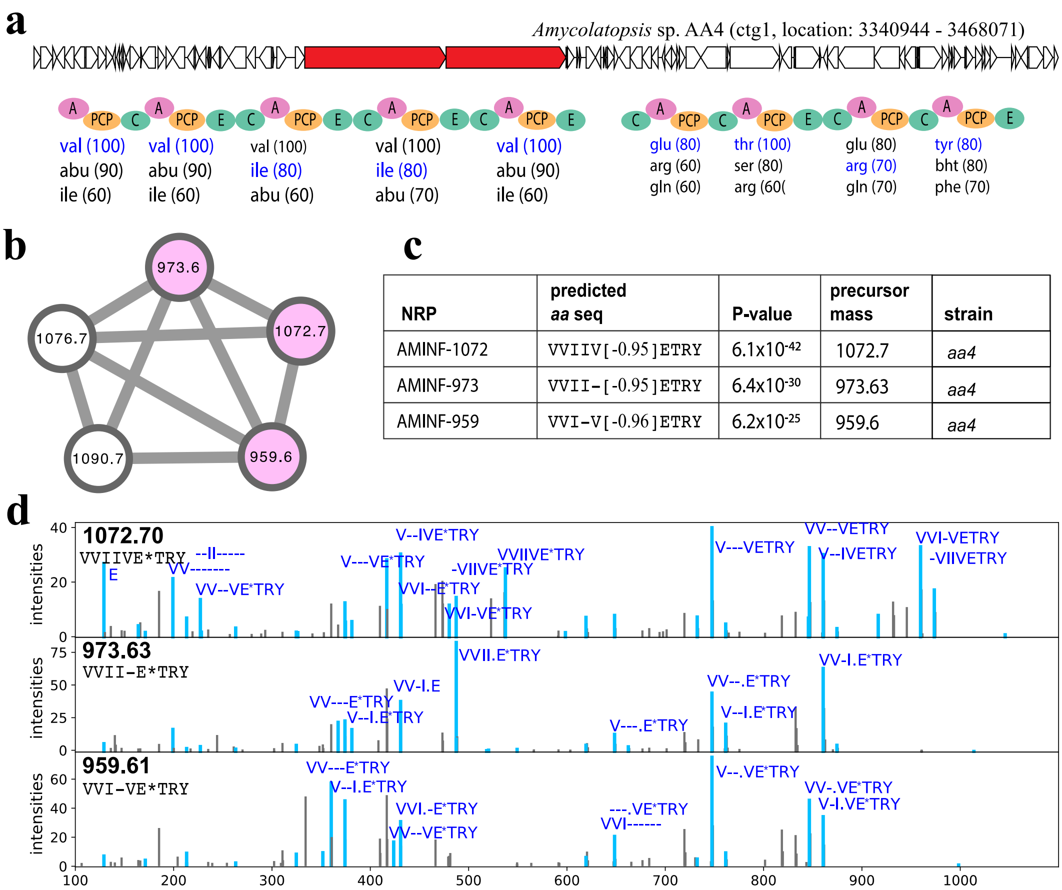
Discovery of xenoinformycin (XINF) NRP family in the XPF dataset. NRPminer matched four spectra representing four cyclic NRPs *X. miraniensis* dataset to a previously uncharacterized BGC in its genome (Fig. 4). NRPminer reported a modification with a total mass of 99.068 for all the four identified NRPs, which matches the valine mass. We hypothesize that one of the valine-specific adenylation domains is responsible for the activation of two consecutive valine units, suggesting an iterative use of the Val-incorporating module (similar to stuttering observed in polyketide synthases[61,62]) but this is yet to be experimentally verified. Interestingly, the predicted xenoinformycin producing NRPS XinfS is highly similar to the widespread NRPS GxpS found in *Xenorhabdus* and *Photorhabdus*, responsible for the GameXPeptide production[31,63]. While both XinfS and GxpS have five modules, XinfS has a C-domain instead of the usual C/E-domain in the last module, suggesting a different configuration of the amino acid Phe or Leu (corresponding to the second last A-domain on their NRPSs), respectively.


a
Xenorhabdus sp. KJ12.1 (ctg1, 189138 - 273601)
C A PCP C A PCP C/E A PCP C A PCP C/E A PCP E C A PCP C/E A PCP C A PCP C/E A PCP C A PCP C/E A PCP C A PCP C A PCP
thr (100) dht (100) allthr (100)
ala (80) cys (60) gly (50)
val (90) ile (70) abu (70)
leu (80) ile (70) val (70)
leu (70) ile (70) val (70)
thr (100) dht (100) allthr (100)
thr (100) dht (100) allthr (100)
leu (80) ile (70) val (70)
leu (70) lle (60) val (60)
ala (80) b-ala (50) ser (40)
ala (70) b-ala (60) ser (40(
pro (100) pip (70) ala (60)
ala (70) b-ala (60) ser (40(
b
1320.8
1306.8
1223.7
1292.8
1251.8
1348.8
1334.8
1252.8
1237.8
c
d
1320.8
+TAVLLTLLAAPA
+TAV---------
+TAVL---------
-----TLLA--
+TAVLL--------
----TTLLAAPA
----LTTLLAAPA
---LLTTLLAAPA
+TA-LLTTLLAAPA
1335.8
+TAVLLTLVAAPA
---------AAP-
---LLTTL-----
-------LVAAP-
+TAVLLTTLV---A
+TAVLLTTL--APA
intensities
50
0
200
400
600
800
1000
1200
e
OH
HO
NH
N H
O
f

NRP	predicted *aa* seq	P-value	precursor mass	strains
XAM-1237	[+99.06]TAVLLTTLLAA-A	2.8x10-33	1237.8	*KJ12.1, KK7.4*
XAM-1251	[+113.07]TAVLLTTLLAA-A	6.5x10-31	1251.8	*KJ12.1*
XAM-1292	[+71.04]TGVLLTTLVAAPA	4.0x10-22	1292.8	*KJ12.1, KK7.4, stockiae*
XAM-1306	[+99.06]TAVVLTTLVAAPA	9.0x10-22	1306.8	*KJ12.1, KK7.4, stockiae*
XAM-1320	[+99.06]TAVLLTTLVAAPA	8.4x10-50	1320.8	*KJ12.1, KK7.4, stockiae*
XAM-1334	[+99.06]TAVLLTTLLAAPA	2.8x10-45	1334.8	*KJ12.1, KK7.4, stockiae*
XAM-1348	[+113.07]TAVLLTTLLAAPA	2.1x10-19	1348.8	*KJ12.1, KK7.4, stockiae*

Fig. 5 Identifying xenoamicin-like (XAM) NRP family. a The BGCs generating the NRP in *Xenorhabdus* sp. KJ12 along with NRPS genes (shown in red) and A-, C-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2 (ref. [15]) predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP [+99.06]TAVLLTTLLAAPA identified by NRPminer (with the lowest *p* value) are shown in blue. **b** Spectral network formed by the spectra that originate from NRPs in the XAM family. **c** Sequences of the identified NRPs in this family (with the lowest *p* value among all spectra originating from the same NRP). The *p* values are computed based on MCMC approach using MS-DPR[89] with 10,000 simulations. **d** For each strain, an annotated spectrum representing the lowest *p* value is shown. The spectra were annotated based on predicted NRPs [+99.06]TAVLLTTLLAAPA and [+99.06]TAVLLTTLVAAPA from top to bottom. The "+" sign represents the addition of [+99.06]. Supplementary Figures 23 and S24 show the annotated spectra for the other NRPs shown in part (**c**). **e** NMR-based correlations of XAM-1320 (*m/z* 1320.8 $[M+H]^+$) produced by *Xenorhabdus* KJ12.1 (Supplementary Table 5 and Supplementary Figs. 25–29). HSQC-TOCSY (bold lines) and key ROESY correlations (arrows) are shown. **f** 3D structure of XAM-1320 derived from 121 ROE-derived distance constraints (Supplementary Table 6), molecular dynamics, and energy minimization. Peptide backbone is visualized with a yellow bar (left). Predicted hydrogen bonds stabilizing the β-helix are shown as dashed lines. View from above at the pore formed by XAM-1320 (right). NRPminer identified this NRP with *p* value 8.4×10^{-50}.

Discovery of xenoamicin-like (XAM) NRP family in the XPF dataset. NRPminer discovered an NRP family that includes eight distinct NRPs, along with their BGC (Fig. 5). While the matched BGC for this family is evolutionary related to the xenoamicin BGC[64] and both BGCs include 13 A-domains, 7 out of 13 amino acids in XAM differ from the corresponding amino acids in xenoamicin A (Supplementary Fig. 19). We named this previously unreported class of xenoamicins class III. Interestingly, the occurrence of XAM-1237 and XAM-1251 suggest a loss of Pro in their structure indicating another possibility of NRP diversification, namely module skipping as previously observed in other NRPSs[61,65,66]. We confirmed the sum formula of XAM-1320 (*m/z* 1320.793 $[M+H^+]$; $C_{63}H_{109}N_{13}O_{17}$) and XAM-1334 (*m/z* 1334.810 $[M+H^+]$; $C_{64}H_{111}N_{13}O_{17}$) by feeding (Supplementary Figs. 20 and 21) and MS–MS experiments (Supplementary Fig. 22 and Methods section under "Additional analysis for xenoamicin-like family") and were also able to isolate the major derivative XAM-1320 from *Xenorhabdus* sp. KJ12.1 and to elucidate its structure by NMR including its 3D solution structure (Supplementary Tables 5 and 6 and Supplementary Figs. 25–S29) that confirms its β-helical structure from the alternating D/L configurations (confirmed by the advanced Marfey's analysis; Supplementary Fig. 23 and "Methods" section) throughout the peptide chain from the presence of C/E domains, except for the C-terminal part shown in Fig. 5. XAM-1320 was also tested against protozoa and showed a good activity against *T. brucei rhodesiense* (IC_{50} [mg/L] 3.9) but much lower activity against *Trypanosoma cruzi*, *Plasmodium falciparum* and rat L6 cells (IC_{50} [mg/L] 25.5, 56.2, and 46.0, respectively). Supplementary Figure 24 provides information about the isolation and structure



NRP	predicted aa seq	P-value	precursor mass	strain
AMINF-1072	VVIIV[-0.95]ETRY	$6.1x10^{-42}$	1072.7	*aa4*
AMINF-973	VVII-[-0.95]ETRY	$6.4x10^{-30}$	973.63	*aa4*
AMINF-959	VVI-V[-0.96]ETRY	$6.2x10^{-25}$	959.6	*aa4*

Fig. 6 Identifying aminformatide (AMINF) NRP family discovered by NRPminer in the *SoilActi* dataset. a The BGC generating the core NRP in *Amycolatopsis* sp. AA4 along with NRPS genes (shown in red) and the A-, C-, PCP, and E-domains appearing in the corresponding NRPS. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in this BGC (according to NRPSpredictor2 (ref. [15]) predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP VVIVETRY identified by NRPminer (with the lowest *p* value) are shown in blue. **b** Spectral network formed by spectra that originate from the AMINF NRPs. A node is colored if the corresponding spectrum forms a statistically significant PSM and not colored otherwise. The *p* values are computed based on MCMC approach using MS-DPR[89] with 10,000 simulations. **c** Sequences of the NRPs identified by NRPminer in the aminformatide family (with the lowest *p* value among all PSMs originating from the same NRP). NRPminer predicted a PAM with loss of ~0.96 Da on E, represented by E*. AMINF represents aminformatide. **d** For each identified NRP, an annotated spectrum representing the lowest *p* value is shown.

elucidation of XAM-1320, XAM-1278, XAM-1292, and XAM-1348 that differed in the starter acyl unit and the following amino acid (Ala or Gly).

Discovery of aminformatide NRP family produced by *Amycolatopsis* sp. aa4 in the SoilActi dataset. Supplementary Table 7 presents the number of NRP-producing BGCs and the number of putative core NRPs generated by NRPminer for each analyzed genome in XPF (before and after filtering). NRPminer identified 11 PSMs (representing three NRPs) when searching the SoilActi spectral dataset against *Amycolatopsis* sp. aa4 genome (Fig. 6). Previously, another NRP family, siderophore amychelin, and its corresponding BGC was reported from this organism[67]. Using the NRPSpreidctor2 (ref. [15])-predicted amino acids NRPminer predicted a modification of ~0.95 Da on the Glu in aminoformatide-1072 VVII[E-1.0]TRY. Since NRPSpredictor2 is the least sensitive in recognizing Lys (as compared to other amino acids)[15], we hypothesize that this amino acid is in fact a Lys as we have seen in the case of protegomycins (with Lys), but this is yet to be determined.

Identifying lugdunin NRP family in the SkinStaph dataset. Antibiotics lugdunins[7] represent the only NRP family reported in the human commensal microbiota. NRPminer matched nine spectra representing three NRPs from a single family in the $\text{spectra}_{\text{SkinStaph}}$ dataset against *Staphylococcus lugdunensin* genome. In addition to the two known cyclic variants of lugdunin, NRPminer also discovered a previously unreported lugdunin variant with precursor mass 801.52 (Supplementary Fig. 30). Due to a +18.01 Da mass difference, NRPminer predicted a linear structure for this variant that represents the linear version of the known one. Since NRPminer predicts sequence VWLVVVt for the linear lugdunin, with the breakage between valine and Cys-derived thiazolidine, we hypothesize that this is a naturally occurring linear derivative in the lugdunin family. Lugdunins, synthesized by a non-canonical assembly line, were predicted using the non-canonical assembly line feature of NRPminer (Fig. 7).

Identifying lipopeptides in the TinyEarth dataset. Our NRPminer analysis of the TinyEarth dataset generated 498 PSMs representing 31 NRPs from three families, using the 200 Da threshold for PAM identification. Supplementary Table 9 provides information about the NRPminer-generated PSMs representing these three NRP families. *Bacillus* derived surfactins[68] and plipastatin[69] are bioactive lipopeptide with wide variety

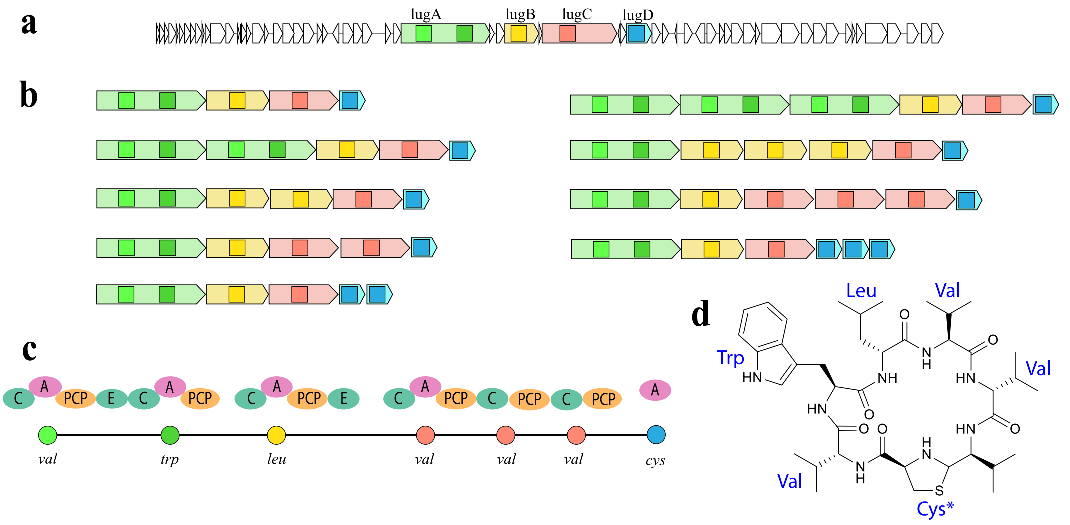


Fig. 7 Lugdunin BGC and the assembly lines formed by NRPminer using the OrfDup option. a Lugdunin BGC with the four ORFs shown in different colors. The squares represent the A-domains. **b** Assembly lines formed by duplication of a single NRPS subunit (corresponding to each ORF) zero, one, and two times are pictured. NRPminer explores all assembly lines generated by duplicating each ORF up to two times when the "OrfDup" option is selected. **c** The NRPS assembly lined (with A-, C-, PCP-, and E-domains pictured) appearing in the NRPS that synthesizes lugdunin, where one Val-specific A-domain loads three amino acids (*valines*) to the growing peptide. Amino acids corresponding to lugdunin structure are shown below each A-domain. Circles represent amino acids (different amino acids are shown by different colors). **d** Cyclic structure of lugdunin with the amino acids highlighted in blue. The "Cys*" represent Cys-derived thiazolidine in lugdunin structure.

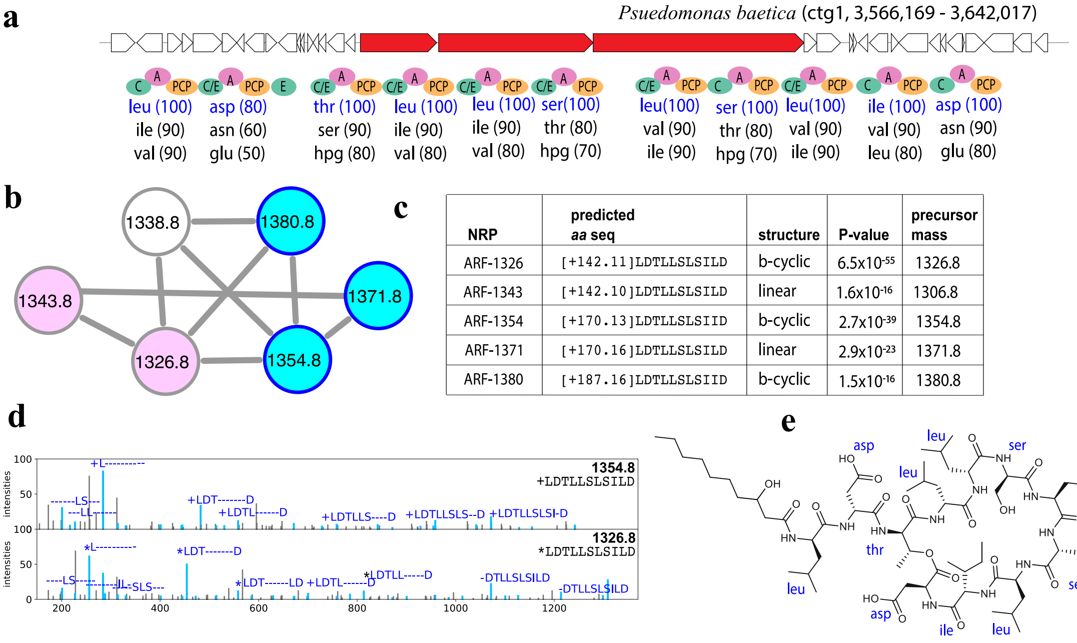
of activities. Surfactins are reported to have anti-viral[70,71], anti-tumor[72], anti-fungal[73], and anti-microbial[74] functions[75–78] and plipastatins have known anti-fungal activities[79]. In the analysis of *Bacillus amyloliquefaciens* sp. GZYCT-4-2, NRPminer correctly reported all known surfactins (17 NRPs) and plipastatins (9 NRPs) identified in this dataset (PSMs listed in Supplementary Table 10). Moreover, NRPminer search of spectra$_{\text{TinyEarth}}$ against putative NRP structures generated from *Pseudomonas baetica* sp. 04-6(1) genome identified 63 PSMs representing the arthrofactins (ARF) NRP family (Fig. 8). NRPminer identified the known branch-cyclic arthrofactins[80] that only differ in the fatty acid tail (namely ARF-1354 and ARF-1380) and a known linear arthrofactin ARF-1372 (the linear version of ARF-1354). Furthermore, it identified two previously unreported arthrofactin variants: ARF-1326 (predicted to only differ in its side chain from the known branch-cyclic ARF-1354 shown in Fig. 8e) and ARF-1343 (predicted to be the linear version of the putative ARF-1326). NRPminer missed one known NRP family identified in spectra$_{\text{TinyEarth}}$ (xantholysins[81]) since the xantholysin BGC was split among multiple contigs in the *Pseudomonas plecoglossicida* sp. YNA158 genome assembly.

Identifying surugamides in the SoilActi dataset. NRPminer identified 183 spectra representing 25 NRPs when searching spectra$_{\text{SoilActi}}$ against *S. albus* J10174 genome, hence extending the set of known surugamide variants from 8 to 21 (Supplementary Table 8 and Supplementary Fig. 2). Spectral network analysis revealed that these spectra originated from two NRP families. VarQuest search of this spectral dataset against PNPdatabase[43] identified only 14 of these 21 NRPs. The remarkable diversity of surugamide NRPs, which range in length from 5 to 10 amino acids, is explained by the non-canonical assembly lines[13,43]. Using the "orfDel" option when analyzing surugamide BGC, with four ORFs (see Fig. S31), NRPminer generated 11 assembly lines. Supplementary Table 12 presents the number of core NRPs generated from the assembly line formed by SurA and SurD genes, based on their scores; 1104 core NRPs are retained out of 45,927 possible core NRPs generated from this assembly line. In total, 14,345 core NRPs from the original 3,927,949,830 core NRPs of the 11 assembly lines of surugamide BGC are retained. In addition to the surugamides synthesized by the SurA-SurD pair, NRPminer also discovered Surugamide G synthesized by the SurB-SurC pair (Supplementary Fig. 2d). In comparison with surugamide F from *Streptomyces albus*[32], this NPR lacks the N-terminal tryptophan. Surugamide F was not identified in the spectral dataset from *Streptomyces albus*.

Discussion

We developed the scalable and modification-tolerant NRPminer tool for automated NRP discovery by integrating genomics and metabolomics data. We used NRPminer to match multiple publicly available spectral datasets against 241 genomes from RefSeq[82] and genome online database (GOLD)[83]. NRPminer identified 55 known NRPs (13 families) whose BGCs have been identified previously, without having any prior knowledge of them (Figs. 2 and 7, Supplementary Fig. 2, S3, and S25, and Supplementary Table 2 and S8). Furthermore, NRPminer identified the BGC for an orphan NRP family (xentrivalpeptides) with previously unknown BGC. In addition to the known NRPs, NRPminer reported 121 previously unreported NRPs from a diverse set of microbial organisms. Remarkably, NRPminer identified four NRP families, representing 25 previously unreported NRPs with no known variants, three families in the XPF dataset (Figs. 3–5) and one in the SoilActi dataset (Fig. 6), illustrating that it can match large spectral datasets against multiple bacterial genomes for discovering NRPs that evaded identification using previous methods. We further validated two of the previously unreported families predicted by NRPminer using NMR and demonstrated their anti-parasite activities.

Existing peptidogenomics approaches are too slow (and often memory-intensive) to conduct searches of large MS datasets against many genomes. Moreover, these approaches are limited to NRPs synthesized by canonical assembly lines and without PAMs, which limits the power of these methods for discovering NRPs. NRPminer is the first peptidogenomics tool that efficiently filters core NRPs based on their specificity scores without losing



NRP	predicted *aa* seq	structure	P-value	precursor mass
ARF-1326	[+142.11]LDTLLSLSILD	b-cyclic	$6.5x10^{-55}$	1326.8
ARF-1343	[+142.10]LDTLLSLSILD	linear	$1.6x10^{-16}$	1306.8
ARF-1354	[+170.13]LDTLLSLSIID	b-cyclic	$2.7x10^{-39}$	1354.8
ARF-1371	[+170.16]LDTLLSLSILD	linear	$2.9x10^{-23}$	1371.8
ARF-1380	[+187.16]LDTLLSLSIID	b-cyclic	$1.5x10^{-16}$	1380.8

Fig. 8 Arthrofactin (ARF) NRP family. a The BGCs generating the NRP in *Pseudomonas baetica* sp. 04-6(1) along with the NRPS genes (shown in red) and A-, C-, C/E-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2 (ref. [15]) predictions) are shown below the corresponding A-domains. Amino acids appearing in the known NRP ARF-1354 with amino acid sequence [+170.13]LDTLLSLSILD are shown in blue. **b** Spectral network formed by the spectra that originate from NRPs in the ARF family. The known arthrofactins are shown in blue, while the purples nodes represent the previously unreported variants identified by NRPminer. All identified athrofactins share the same core NRP LDTLLSLSILD. **c** Sequences of the identified NRPs in this family (with the lowest *p* value among all spectra originating from the same NRP). Column "structure" shows if the predicted structure for the identified NRPs is linear or branch-cyclic (shown by b-cyclic). The *p* values are computed based on MCMC approach using MS-DPR[89] with 10,000 simulations. **d** Two annotated spectra representing the PSMs (with the lowest *p* values among spectra originating from the same NRPs) corresponding to ARF-1354 and 1326. The two spectra were annotated based on predicted NRPs [+170.13]LDTLLSLSILD (PSM *p* value 2.7×10^{-39}) and [+142.11]LDTLLSLSILD (PSM *p* value 6.5×10^{-55}), from top to bottom. The "+" and "*" signs represent the addition of [+170.13] and [+142.11], respectively. **e** The 2D structure of known arthrofactin ARF-1354 (ref. [80]). NRPminer identified this NRP with *p* value 2.7×10^{-39}.

sensitivity and enables searching millions of spectra against thousands of microbial genomes. Furthermore, NRPminer can identify NRPs with non-canonical assembly lines of different types (e.g., surugamides, xenoinformycin, and lugdunin) and PAMs (e.g. surfactins, arthrofactins, plipastatins, protegomycins, and PAX peptides).

Majority of the spectral datasets in GNPS are currently not accompanied by genomics/metagenomics data. To address this limitation, NRPminer can search a spectral dataset against all genomes from RefSeq[82] or GOLD databases[83] within a user-defined taxonomic clade. This one-vs-all mode enables analysis of spectral datasets that are not paired with genomic/metagenomic data by searching them against multiple genomes. This mode, which relies on the scalability of NRPminer, enabled NRPminer to identify the lugdunin family (by searching the SkinStaph spectral dataset) even though the paired genome sequence from the same strain was not available.

In contrast to the previous peptidogenomics approaches, NRPminer is robust against errors in specificity prediction in genome mining tools and can efficiently identify mature NRPs with PAMs. This feature was crucial for discovering protegomycins that include a PAM (lipid chain) and a mis-prediction (Phe instead of Lys), as well as for identifying the lipopeptide biosurfactant in the TinyEarth dataset. While NRPminer is a powerful tool for discovering NRPs it can only succeed if the genome mining algorithms successfully identify an NRP-encoding BGC and predict the correct amino acids for nearly all A-domains. One of the bottlenecks of genome mining methods for NRP discovery is the lack of training data for many non-standard amino acids from under-explored taxonomic clades. We anticipate that more NRPs will be discovered using automated methods, and these discoveries will increase the number of A-domain with known specificity, which in turn will pave the path toward the development of more accurate machine learning techniques for A-domains specificity prediction.

In case of metagenomic datasets, NRPminer's one-vs-all function allows for searching the spectral dataset against all the metagenomic assemblies generated from the same sample. However, the success of genome mining crucially depends on capturing the entire BGCs in a single contig during genome assembly. NRPS BGCs are long (average length ~60 kb[45]) and repetitive (made up of multiple highly similar domains), making it difficult to assemble them into a single contig. Meleshko et al.[45], recently developed the biosyntheticSPAdes tool for BGC reconstruction in short-read isolate assemblies, but at the same time acknowledged that short-reads metagenome assemblies are not adequate for full-length BGC identification. Even with biosyntheticSPAdes[45], it remains difficult to capture long and repetitive BGCs within a single contig. With recent advances in long-read sequencing technologies, more contiguous microbial genome assemblies are becoming available[84,85], increasing the power of NRPminer.

Another challenge in applications of NRPminer to complex microbiome data is that, with the current state of MS technology, many spectra originate from host molecules (in the case of host-associated microbiomes) or environmental contaminations. For example, the majority of spectra collected on human skin microbiome correspond to deodorants, shampoos, and other beauty products, rather than microbial products[86]. The advent of sensitive MS data acquisition techniques could enable capturing low abundant microbial products from complex environmental and host-oriented samples.

NRPminer only considers methylation and epimerization tailoring enzymes in the BGCs and does not recognize any other modification enzymes that modify NRPs, such as glycosylation and acylation[87]. These modifications can only be predicted as blind modifications using the modification-tolerant search of their corresponding spectral datasets against the input genomes.

Currently, NRPminer identifies ~1% of spectra of isolated microbes as NRPs. However, ~99% of spectra in these datasets remain unidentified, representing the dark matter of metabolomics. These spectra could represent primary metabolites (e.g. amino acids), other classes of secondary metabolites (e.g. RiPPs, polyketides, lipids, terpenes), media contaminations, and lower intensity/quality spectra that are difficult to identify. Thus, further advances in experimental and computational MS are needed toward a comprehensive illumination of the dark matter of metabolomics.

Methods

Outline of the NRPminer algorithm. NRPminer expands on the existing tools for automated NRP discovery[30,40] by utilizing algorithms that enable high-throughput analysis and handle non-canonical assembly lines and PAMs. Below we describe various steps of the NRPminer pipeline:

(a) *Predicting NRPS BGCs in (meta)genome sequences by genome mining.* NRPminer uses Biopython[88] and antiSMASH[17] to identify the NRP-producing BGCs in the assembled genome. Given a genome (or a set of contigs), antiSMASH uses HMMs to find NRP-producing BGCs. The NRPminer software package also includes biosyntheticSPAdes[45], a specialized short-read BGC assembler.

(b) *Predicting putative amino acids for each A-domain in the identified BGCs.* NRPminer uses NRPSpredictor2 (ref. [15]) to predict putative amino acids for each position in an NRP. Given an A-domain, NRPSpredictor2 uses support vector machines (trained on a set of A-domains with known specificities) to predict the amino acids that are likely to be recruited by this A-domain. NRPSpredictor2 provides a specificity for each predicted amino acid that is based on the similarity between the analyzed A-domain and the previously characterized A-domains[16,18]. NRPminer uses NRPSpredictor2 (ref. [15]) predictions to calculate the specificity scores for each predicted amino acid (see "Methods" section under "Specificity Scores of Putative Amino Acids)".

(c) *Generating multiple NRPS assembly lines.* NRPminer generates multiple NRPS assembly lines by allowing for the option to either delete an entire ORFs, referred to as "orfDel" (Fig. 1c) or duplicate A-domains encoded by an ORF, referred to as "*orfDup*" (Fig. 1b). In the default "orfDel" setting, NRPminer considers all assembly lines formed by deleting up to two ORFs. With "orfDup" option, NRPminer generates non-canonical assembly lines that tandemly duplicate all A-domains appearing in a single ORF.

We represent an NRPS assembly line as a sequence of sets of amino acids, $\mathscr{A}_1$, …,$\mathscr{A}_k$ where each $\mathscr{A}_i$ represents the set of amino acids predicted for the *i*th A-domain of this assembly line along with their specificity scores. Given an NRPS assembly line with *k* A-domains and the corresponding sets $\mathscr{A}_1,\ldots,\mathscr{A}_k$, the set of all possible core NRPs for this assembly line is given by the cartesian product $\mathscr{A}_1 \times \ldots \times \mathscr{A}_k$. See "Methods" section under "Generating Assembly-lines Using NRPminer" for more information.

(d) *Filtering the core NRPs based on their specificity scores.* Supplementary Table 1 and Supplementary Table 7 illustrate that some BGC-rich genomes give rise to trillions of putative core NRPs. NRPminer uses the specificity scores of amino acids in each core NRP to select a smaller set of core NRPs for downstream analyses. Given an assembly line $\mathscr{A}_1,\ldots,\mathscr{A}_k$, for each amino acid $a \in \mathscr{A}_i$ ($i = 1,\ldots, k$), NRPminer first divides the specificity score of *a* by the maximum specificity score observed across all amino acids in $\mathscr{A}_i$ (see "Methods" section under "Filtering the Core NRPs Based on their Specificity Scores)"; we refer to the integer value of the percentage of this number as the "normalized specificity score" of *a*. We define the score of a core NRP to be the sum of the normalized scores of its amino acids.

NRPminer uses a dynamic programming algorithm to efficiently find *N* highest-scoring core NRPs for further analyses (the default value is $N = 1000$), which enables peptidogenomics analysis of BGCs with many A-domains. The "Methods" section provides more information.

(e) *Identifying domains corresponding to known modifications and incorporating them in the core NRPs.* NRPminer searches each BGC for methylation domains (PF08242) and accounts for the possible methylations on corresponding residues for all resulting core NRPs (corresponding to +14.01 Da mass shift). NRPminer also searches for epimerization domains in each BGC (as well as dual condensation-epimerization domains) that provide information about the structure of the final NRP (D- or L-amino acids).

(f) *Generating linear, cyclic, and branch-cyclic backbone structures for each core NRP.* NRPminer generates linear and cyclic structures for all core NRPs. Similar to NRPquest[30], whenever NRPminer finds a cytochrome P450 domain, it also generates branched-cyclic NRPs by considering a side-chain bond between any pair of residues in the peptide.

(g) *Modification-tolerant search of spectra against the constructed backbone structures.* Similar to PSMs in proteomics, a PSM in peptidogenomics is scored based on similarities between the theoretical spectrum of the peptide and the mass spectrum[43] (see "Methods" section under "Forming Peptide-Spectrum-Matches (PSMs) and Calculating PSM score)". The standard search of a spectrum against a peptide database refers to finding a peptide in the database that forms a highest-scoring PSM with this spectrum. Similarly, the modification-tolerant search of a spectrum against the peptide database refers to finding a variant of a peptide in the database that forms a highest-scoring PSM with this spectrum. In the case of NRPs, it is crucial to conduct modification-tolerant search in a blind mode in order to account for unanticipated PAMs in the mature NRP.

Existing peptidogenomics methods utilize a brute-force approach for modification-tolerant search, by creating a database of all possible unanticipated modifications[30]. For example, given a spectrum and a core NRP structure with *n* amino acids, these methods consider a modification of mass δ on all possible amino acids in the NRP, where δ is the mass difference between the spectrum and the NRP. Gurevich et al.[43] developed the VarQuest tool for modification-tolerant search of large spectral datasets against databases of peptidic natural products that is two orders of magnitude faster than the brute-force approach. NRPminer utilizes VarQuest for identification of PAMs with masses up to MaxMass with the default value MaxMass=150 Da (see "Methods" section for more informatoin). This approach also allows NRPminer to identify loss or addition of an amino acid (for amino acids with molecular mass up to MaxMass Da). Note that, similar to identification of PAMs in linear proteomics[30], MS-based methods for NRP discovery are limited to finding modification masses and cannot provide information about the exact chemistry of the identified modifications.

NRPminer has the one-vs-one mode for searching a spectral dataset against the genome corresponding to its producer. Additionally, NRPminer features the one-vs-all mode that a spectral dataset is searched against all genomes in the corresponding taxonomic clade (or any given set of genomes). One-vs-all is useful in cases when an entire BGC is not assembled in a single contig in the producer's genome, but well-assembled in a related genome.

In scoring PSMs, NRPminer has a user-adjustable threshold for the accuracy of precursor and products ions, thus improving the accuracy of PSM scoring in the case of modification-tolerant search of high-resolution spectral datasets. This feature improves on NRPquest whose applications are largely limited to low-resolution spectra.

(h) *Computing statistical significance of PSMs.* NRPminer uses MS-DPR[89] to compute *p* values of the identified PSMs. Given a PSM, MS-DPR computes the probability (*p* value) that a random peptide has a score greater than or equal to the PSM score (see "Methods" section under "Computing P-values and Peptide-Spectrum-Matches"). The default *p* value threshold (10^{-15}) is chosen based on the previous studies where the *p* value cut-off 10^{-15} was necessary for reaching a false discovery rate (FDR) below 1% against NRPs[42,43]. Furthermore, NRPminer filters the PSMs based on the FDR values reported by VarQuest (default threshold 1%). The user can change the *p* value and FDR thresholds (using "—*p* value" and "—fdr" handles) depending on their study. *E*-values are also calculated by multiplying *p* values with the number of spectra and NRPs computed.

(i) *Expanding the set of identified NRPs using spectral networks*. Spectral datasets often contain multiple spectra originating from the same compound. NRPminer clusters similar spectra using MS-Cluster[60] and estimates the number of distinct NRPs as the number of clusters. It further constructs the spectral network[50,57] of all identified spectra and estimates the number of distinct NRP families as the number of connected components in this network.

Spectral networks reveal the spectra of related peptides without knowing their amino acid sequences[57]. 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. Ideally, each connected component of a spectral network corresponds to a single NRP family[57] representing a set of similar NRPs. In this study, we only report an identified NRP family if at least one NRP in the family is identified with a PSM *p* value at least 10^{-20}. NRPminer utilizes spectral networks for expanding the set of identified NRPs.

Sample preparation and MS experiments. *General experimental procedures.* ^{1}H, ^{13}C, HSQC, HMBC, HSQC-COSY, HSQC-TOCSY, and ROESY spectra were measured on Bruker AV500, AV600, and AV900 spectrometers, using DMSO-d_6 and $CDCl_3$ as solvent. Coupling constants are expressed in Hz and chemical shifts are given on a ppm scale. HRESIMS was performed on an UltiMate 3000 system (Thermo Fisher) coupled to an Impact II qTof mass spectrometer (Bruker Daltonik GmbH). Preparative HPLC was performed on an Agilent 1260 HPLC/MS system with a ZORBAX StableBond 300 C18 (21.2 mm × 250 mm, 7.0 μm, Agilent). Semi-preparative HPLC was performed on an Agilent 1260 HPLC/MS system with a ZORBAX StableBond 300 C18 (9.4 mm × 250 mm, 5.0 μm, Agilent).

Below we describe sample preparation and mass spectra generation for all analyzed datasets in more details.

XPF: A total of 27 strains from soil nematode symbiont *Xenorhabdus* and *Photorhabdus* families were grown in lysogeny broth and agar and were extracted with methanol as described previously (Supplementary Table 1). Briefly, the crude extracts were diluted 1:25 (vol/vol) with methanol and analyzed by UPLC-ESI coupled with Impact II qTof mass spectrometer. MS dataset spectra$_{XPF}$[31] contains 27 spectral sub-datasets representing each sample for a total of 263,768 spectra across all strains (GNPS-accession #: MSV000081063). The genome$_{XPF}$ dataset contains 27 draft genomes generated by DNA sequencing from the same samples as reported by Tobias et al.[31] (available from RefSeq[82]). See the sections below for detailed information about experiments regarding protegomycin and xenoamicin-like families, respectively.

SkinStaph: A total of 171 *Staphylococcus* strains isolated from skin of healthy individuals were grown in 500 mL Tryptic Soy Broth (TSB) liquid medium in Nunc 2.0 mL DeepWell plates (Thermo Catalog# 278743) by Zhou et al.[90]. An aliquot of each culture was used to measure optical density. Cultures that effectively grew were transferred to a new deep well plate. Cultures were placed in a −80 °C freezer for 10 min and then allowed to thaw at room temperature three times, to lyse bacterial cells. Two hundred microliters of the supernatant collected from cell cultures were filtered using a Phree Phospholipid Removal kit (Phenomenex). Sample clean up was performed following the manufacturer's protocol described here (https://phenomenex.blob.core.windows.net/documents/c1ac3a84-e363-416e-9f26-f809c67cf020.pdf). Briefly, the Phree kit plate was conditioned using 50% MeOH; bacterial supernatant were then added to the conditioned wells followed by sample clean up using 100% MeOH (a 4:1 v/v ratio of MeOH:bacterial supernatant). The plate was centrifuged 5 min at 500*g* and the clean up extracts were lyophilized using a FreeZone 4.5 L Benchtop Freeze Dryer with Centrivap Concentrator (Labconco). Wells were resuspended in 200 μL of resuspension solvent (80% MeOH spiked with 1.0 μM Amitriptyline), vortexed for 1 min, and shaken at 2000 r.p.m. for 15 min at 4 °C. One hundred and fifty microliters of the supernatant was transferred into a 96-well plate and maintained at 20 °C prior to liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. Bacterial extracts were analyzed using a ThermoScientific UltiMate 3000 UPLC system for liquid chromatography and a Maxis Q-TOF (Quadrupole-Time-of-Flight) mass spectrometer (Bruker Daltonics), controlled by the Otof Control and Hystar software packages (Bruker Daltonics) and equipped with ESI source. Untargeted metabolomics data were collected using a previously validated UPLC-MS/MS method[91,92]. The spectra$_{SkinStaph}$ dataset contains 2,657,398 spectra from bacterial extracts of 171 *Staphylococcus* strains (GNPS- accession #: MSV000083956). The genome$_{SkinStaph}$ dataset contains draft genomes of these species (available from RefSeq).

SoilActi: A total of 20 strains of soil-dwelling *Actinobacteria* were grown on A1, MS, and R5 agar, extracted sequentially with ethyl acetate, butanol, methanol, and analyzed on Agilent 6530 Accurate-Mass QTOF spectrometer coupled with Agilent 1260 LC System. The spectra$_{SoilAct}$ dataset contains 362,421 spectra generated from extracts of these 20 Actinobacteria strains (GNPS-accession #: MSV000078604 (ref. [93])) includes 20 sub-datasets representing each strain. The genome$_{SoilActi}$ dataset contains draft genomes of these strains (available via RefSeq).

TinyEarth: A total of 23 bacterial strains extracted from the soil in Wisconsin were grown in microscale liquid cultures and extracted using solid phase extraction with in methanol. These samples were and analyzed by LC-MS/MS on a Thermo Fisher Q-Exactive mass spectrometer coupled with a Thermo Fisher Vanquish UPLC system. The spectra$_{TinyEarth}$ dataset contains 380,414 spectra generated from extracts of these 23 strains (GNPS-accession #: MSV000084951) includes 23 sub-datasets representing each strain (4 *Bacillus*, 16 *Pseudomonas*, 1 *Buttiauxella*, and 1 *Citrobacter*). The genome$_{TinyEarth}$ dataset contains draft genomes of these strains (available via Gold OnLine Database[83] under study ID Gs0135839).

Additional analyses for protegomycin family. *X. doucetiae*-Δ*hfq* was constructed as described before[94]. Exchange of the natural promoter against the inducible P$_{BAD}$ was performed as described[95]. Briefly, the first 598 base pairs of *prtA* were amplified with primer pEB_317-fw TTTGGGCTAACAGGAGGCTAGCATAT GAGAATACCTGAAGGTTCG and PEB_318-rv TCTGCAGAGCTCGAGCA TGCACATCGTAATGAAACGAGTTCAGG (Supplementary Table 11). The resulting fragment was cloned via hot fusion cloning into pCEP-km. The resulting construct pCEP *prtA*-km was transformed into *E. coli* S17-1 λpir resulting in *E. coli* pCEP_*prtA*. Conjugation of this strain with *X. doucetiae* wt or *X. doucetiae*-Δ*hfq* was followed by integration of pCEP_*prtA*-km into the acceptors genome via homologous recombination[94,95]. In *X. doucetiae*-Δ*hfq*-P$_{BAD}$-*prtA* the production of protegomycin was induced by adding 0.2% L-arabinose into the fresh inoculated medium[94].

For large-scale production of protegomycin, 6 × 1 L LB medium was inoculated with *X. doucetiae*-Δ*hfq*_P$_{BAD}$-*prtA* preculture 0.02%. Two percent Amberlite® XAD-16 adsorber resin was added and the production was induced with 0.2% L-arabinose. The cultures were constantly shaked at 130 r.p.m. at 30 °C. After 72 h the XAD beads were harvested and protegomycins extracted using 3 L of methanol. The solvent was evaporated, and the crude extract was used for isolation and analysis of protegomycin derivatives. Part of the crude extraction was purified by preparative HPLC with a gradient mobile from 5 to 95% ACN in H_2O (v/v) in 30 min followed by semi-preparative HPLC (ACN–H_2O, 35–45% in 30 min, v/v) to yield PRT-1037 (24.4 mg).

For structure elucidation and determination of incorporated C- and N-atoms and amino acids into protegomycins, cultivation of *X. doucetiae*-Δ*hfq*_P$_{BAD}$-*prtA* and *X. doucetiae*_ P$_{BAD}$-*prtA*, induced with 0.2% L-arabinose was performed in 5 mL LB (^{12}C), ^{13}C-, and ^{15}N-isogro® medium (Sigma-Aldrich). The cultures were supplemented with 2% Amberlite® XAD-16 adsorber resin. To analyze the incorporated amino acids, induced mutants were grown in LB medium supplemented with selected ^{13}C-labeled amino acids with a concentration of 2 mM. After 48 h cultivation at 30 °C, constantly shaking at 200 r.p.m., Amberlite® XAD-16 beads were harvested and extracted with 5 mL MeOH for 45 min. Samples were taken from the filtered extracts and centrifuged for 15 min at 17,000*g* for further HPLC-MS analysis (Dionex Ultimate 3000 coupled to a Bruker AmaZon X ion trap). Generated HPLC-MS data were interpreted as described previously[94,96].

Additional analyses for Xenoamicin-like family. *Cultivation of strains*: *Xenorhabdus* KJ12.1 was routinely cultivated in Luria-Bertani (LB) medium (pH 7.0) at 30 °C and 200 r.p.m. on a rotary shaker and on LB agar plates at 30 °C. Inverse feeding experiments were applied in either ISOGRO® ^{13}C medium, ISOGRO® ^{15}N medium. Fifty microliters ISOGRO® medium was prepared with ISOGRO® powder (0.5 g), K_2HPO_4 (1.8 g/L), KH_2PO_4 (1.4 g/L), $MgSO_4$ (1 g/L), and $CaCl_2$ (0.01 g/L) solved in water. Feeding experiments in ISOGRO® ^{13}C medium supplemented with ^{12}C amino acids was inoculated with ISOGRO® washed overnight cultures.

Production cultures were grown in LB media containing 2% Amberlite® XAD-16 resin inoculated with 1% overnight culture. Promotor exchange mutants were induced with 0.2% arabinose at the beginning of the cultivation. Resin beads and bacterial cells were harvested by centrifugation after 72 h cultivation time, washed twice with one culture volume methanol. The crude extracts were analyzed by means of MALDI-MS and HPLC-MS (Bruker AmaZon).

HPLC-based purification: XAM-1320 was isolated by a two-step chromatography. Strain KJ12.1 was cultivated in a BIOSTAT A plus fermenter (Sartorius) equipped with a 2-L vessel in 1.5 L of LB broth at 30 °C for 12 h. For the inoculation, 1% overnight preculture was used and 2% XAD-16 were added. Additionally, 10 g of glucose and 5 mL Antifoam 204 (Sigma-Aldrich) were added. The fermentation was performed with an aeration of 2.25 vvm, constant stirring at 300 rpm and at pH 7, stabilized by the addition of 0.1 N phosphoric acid or 0.1 N sodium hydroxide. The XAD resin was washed with methanol to get the extract after evaporation. Xenoamicin III A was isolated by a two-step chromatography. In the first step the extract was fractionated with a 5–95% water/acetonitrile gradient over 15 min on a Luna C_{18} 10 μm 50 × 50 mm column (Phenomenex). In the second step XAM-1320 was isolated with a 40–60% water–acetonitrile gradient over 19 min on Luna C_{18} 5 μm 30 × 75 mm column (Phenomenex).

MS analysis: MS analysis was carried out by using an Ultimate 3000 LC system (Dionex) coupled to an AmaZon X electronspray ionization mass spectrometer (Bruker Daltonics). Separation was done on a C18 column (ACQITY UPLC BEH, 1.7 mm, 2.1 × 50 mm, flow rate 0.4 mL/min, Waters). Acetonitrile/water containing 0.1% formic acid was used as a mobile phase. The gradient started with 5% acetonitrile continuous over 2 min. Over 0.5 min under a linear gradient acetonitrile reaches 40%. Following an equilibration phase over 1.5 min with 40% acetonitrile takes place. For separation a linear gradient from 40–95% acetonitrile over 10.5 min was used. The gradient ends up with 95% acetonitrile continuous over 1.5 min. Collision-induced dissociation (CID) was performed on ion trap in the AmaZon X in positive mode. HR-ESI-HPLC-MS data were obtained on a LC-coupled Impact II ESI-TOF spectrometer (Bruker Daltonics).

Advanced Marfey's method: The advanced Marfey's method to determine the configurations of the amino acid residues was performed as described previously[64].

Calculating specificity scores of putative amino acids. During NRP synthetase, the A-domains recognize and activate the specific amino acid that will be appended to the growing peptide chain by other NRPS enzymes. Conti et al.[97] showed that some residues at certain positions on each A-domain are critical for substrate activation and bonding; they reported 10 such positions. Stachelhaus et al.[98] showed that for each A-domain AD, the residues at these decisive 10 positions can be extracted to form a specificity-conferring code called non-ribosomal code of AD. They demonstrated that the specificity of an uncharacterized A-domain can be inferred based on the sequence similarity of its non-ribosomal code to those of the A-domains with known specificities[98].

Given an input A-domain AD, NRPSpredictor2 (ref. [15]) first compares the sequence of the non-ribosomal code of AD to those of the already characterized A-domains in the NRPSpredictor2 (ref. [15]) database. Afterwards, for each amino acid *a*, NRPSpredictor2 (ref. [15]) reports the Stachelhaus score of (specificity of) *a* for A-domain AD, that is (the integer value of) the percentage of sequence identity between the non-ribosomal code of AD and that of the most similar A-domain within NRPSpredictor2 (ref. [15]) search space that encodes for *a*.

Furthermore, Rausch et al.[99] expanded the set of specificity-conferring positions on A-domains to 34 residue positions and proposed a predictive model trained on residues at these 34 positions (instead of just the 10 included in Stachelhaus code) to provide further specificity predictions[15]. Given an A-domain, they used a Support Vector Machine (SVM) method trained on previously annotated A-domains. For each input A-domain, this approach[99] predicts three sets of amino acids in three different hierarchical levels based on the physio-chemical properties of the predicted amino acids: large clusters[99] (each large cluster is at most eight

Table 3 Number of core NRPs of SurugamideAL (assembly line corresponding to cyclic surugamides A-D) according to their adenylation scores.

s	800	790	788	786	780	778	776	774	772	Total
$\text{numCoreNRPs}_{\text{SurugamideAL}}(8, s)$	24	48	24	192	24	48	384	192	168	1104

Only values of s with non-zero number of cores and corresponding to the top 1000 high-scoring core NRPs are shown[103].

amino acids), small clusters[99] (each small cluster is at most three amino acids), and single amino acid prediction (the single amino acid most likely to be activated by the given A-domain), as described by Rausch et al.[99] For a given A-domain AD, we use the terms large cluster, small cluster, and single prediction of AD to describe the sets of amino acids predicted at each of these hierarchical levels. While Rausch et al.[99] demonstrated that their approach reports better specificity predictions for less commonly observed A-domains, they also showed that integrating their score with the sequence similarity approach described by Stachelhaus et al.[98] results in the highest accuracy[99].

Similar to the approach used by NRP2Path[40], NRPminer combines the two predictions provided by NRPSpredictor2 (ref. [15]). Given an A-domain AD and an amino acid a, NRPminer defines the SVM score of a for AD to be 100 if a matches the single amino acid prediction, 90 if a appears in the small cluster predictions, and 80 if a appears in the large cluster. If a does not appear in any of these sets, NRPminer defines the SVM score of a for AD to be 0. The total number of amino acids per A-domain with SVM score above 0 is at most 12 (considering all three sets of amino acids). For a given A-domain AD, NRPminer only considers amino acids with a predicted Stachelhaus score>50 and a predicted SVM score>0 for AD. Finally, NRPminer defines the specificity (or NRPSpredictor2) score of a for AD as the mean of Stachelhaus and SVM scores of a for AD.

Generating NRPS assembly lines using NRPminer. Given a BGC, an assembly line refers to a sequence of NRPS modules in this BGC that together assemble the core NRP. NRPminer represents an assembly line as the sequence of A-domains appearing in its NRP modules and allows a user to explore various assembly lines using OrfDel and OrfDup options. Each portion of an NRPS that is encoded by a single ORF is an *NRPS subunit*. With OrfDel option, NRPminer considers skipping up to two entire NRPS subunits. Figure S31b illustrates the assembly lines generated from surugamide BGC by deleting A-domains appearing on zero, one, and two NRPS subunits, out of the four NRPS subunits encoded by the four ORFs appearing in this BGC. For example, for surugamide BGC with four ORFs (shown in yellow in Fig. S31a), with “orfDel” option, NRPminer generates six NRP assembly lines formed by two ORFs (Fig. S31b), four assembly lines formed by three ORFs, and one canonical assembly line formed by all four ORFs. Figure S31c illustrates that for surugamide NRPS assembly line formed by SurA and SurD genes, $\mathscr{A}_1 = \{\text{val, ile, abu}\}$, $\mathscr{A}_2 = \{\text{phe, tyr, bht}\}$, etc.

Using the OrfDup option, NRPminer also considers assembly lines that are generated by multiple incorporation of A-domains appearing on a single NRPS subunit. For example, Supplementary Fig. 7 shows the lugdunin BGC with four ORFs encoding for five A-domains. This figure illustrates that using OrfDup option, NRPminer forms nine assembly lines: one representing the canonical assembly line (each NRPS subunit appears once), four assembly lines that are generated by duplicating the A-domains appearing in one NRPS subunit once (one subunit appearing two times in tandem), and four non-canonical assembly lines by duplicating them twice (one subunit appearing three times in tandem). NRPminer considers all assembly lines made up of at least three and at most 20 NRPS modules.

Filtering the core NRPs based on their specificity scores. Given an NRPS assembly line $A = A_1,\ldots,A_n$, where A_i is the set of amino acids predicted for the ith A-domain of A, for every $a \in A_i$ ($i = 1,\ldots,n$), let SpecificityScore(Ai) (a) be the specificity score of a for the ith A-domain of A as described in Supplementary Note 3. Then, for each integer $1 \le i \le n$ and $a \in A_i$, we define normalized specificity score of a for ith A-domain of A, denoted by S_A (i,a), to be the nearest integer to the following value:

$$\frac{\text{Specificity Score}_{A_i}(a)}{\max\limits_{b \in A_i} \text{Specificity Score}_{A_i}(b)} \times 100$$

We use this scoring function (instead of SpecificityScore) to reduce the bias towards the more frequently observed A-domains that usually result in higher specificity scores compared to the less commonly observed ones, which do not have closely related A-domains in NRPSpredictor2 training datasets[15]. Consider the assembly line of cyclic surugamides A–D shown in Fig. S31c (corresponding to *SurA-SurD* gene pairs in surugamide BGC) which is made up of eight A-domains, we refer to this assembly line by SurugamideAL. Table 2 presents the values of $S_{\text{SurugamideAL}}$ for integers $1 \le i \le 8$ and (at least) the three amino acids with the highest normalized specificity scores for each A-domain in this assembly line.

Given $A = A_1,\ldots,A_n$ we call the set of all core NRPs generated by the cartesian product $A_1 \times \ldots \times A_n$ as the core NRPs of A. For each core NRP of A, $a_1a_2\ldots a_n$, we define the adenylation score of $a_1a_2\ldots a_n$, denoted by $\text{Score}_A(a_1a_2\ldots a_n)$, to be the sum of the normalized specificity scores of all of its amino acids:

$$\text{Score}_A(a_1 a_2 \ldots a_n) = \sum_{i=1}^{n} S_A(i, a_i)$$

Therefore, given assembly line SurugamideAL and core NRP, P=IAIIKIFL (the core NRP corresponding to surugamide A), $\text{Score}_{\text{SurugamideAL}}(P) = 80 + 100 + 100 + 100 + 100 + 100 + 100 + 86 = 766$. Note that, for any assembly line A, the maximum value of Score_A denoted by $\text{maxScore}_A = \sum_{i=1}^{n} \max_{a_i \in A_i} S_A(i, a_i) = 100n$.

For many organisms, the total number of possible core NRPs is prohibitively large, making it infeasible to conduct search against massive spectral repositories. Currently, even the fastest state-of-the-art spectral search methods are slow for searching millions of input spectra against databases with over 10^5 peptides in a modification-tolerant manner as the runtime grows exceedingly large when the database size grows[43]. Supplementary Tables S2 and S7 shows that for 24 (22) out for 27 organisms in XPF dataset and 9 (7) out of 20 organisms in SoilActi dataset, the total number of core NRPs exceed 10^5 (10^6). Therefore, to enable scalable peptidogenomics for NRP discovery, for each constructed assembly line NRPminer, selects a set of candidate core NRPs. To do so, NRPminer starts by finding the number of core NRPs of A according to their adenylation scores (Problem 1) and then it uses these numbers for generating all core NRPs of A with adenylation scores higher than a threshold (Problem 2).

Problem 1. Given $A = A_1,\ldots,A_n$ and a positive integer s, find the number of all core NRPs of A with adenylation score equal to s.

Let $k = \max\limits_{i \in \{1,\ldots,\text{n}\}} (|A_i|)$ where $|A_i|$ shows the number of amino acids in A_i. For any positive integers i and s satisfying,$1 \le i \le n$ and $s \le \text{maxScore}_A$, let $\text{numCoreNRPs}_{\text{A}}$ $(i,\ s)$ denote the number of core NRPs, of assembly line $A_1,\ldots,\text{A}_i$ with $\text{Score}_{\text{A}_1,\ldots,\text{A}_\text{i}}$ equal to s. Let $\text{numCoreNRPs}_{\text{A}}$ $(0,s) = 0$ for any positive integer s, and $\text{numCoreNRPs}_{\text{A}}$ $(i,\ s) = 0$ for any integer $s < 0$, across all possible values of i. Then, for any positive integers i and s satisfying $1 \le i \le n$ and $0 < s \le \text{maxScore}_A$, we have

$$\text{numCoreNRPs}_A(i, s) = \sum_{a_i \in A_i} \text{numCoreNRPs}_A(i - 1, s - S_A(i, a_i)) \tag{1}$$

Using recursive formula (1), NRPminer calculates $\text{numCoreNRPs}_{\text{A}}$ using parametric dynamic programming in a bottom-up manner: NRPminer first, computes numCoreNRPs$A(1,s)$, for all positive integers $s \le \text{maxScore}_A$. then proceeds to numCoreNRPsA(2,s) for all such s, and so on, computing $\text{numCoreNRPs}_A(n,s)$ for all such $0 < s$. Using this approach, for each value of i and s, NRPminer computes $\text{numCoreNRPs}_{\text{A}}$ (i,s) by summing over at most k values. Therefore, NRPminer calculates all values of numCoreNRPs_A with time complexity $O(k \times n \times \text{maxScore}_{\text{A}})$.

Given a positive integer $N < 10^5$, let $scor_{(\text{A,N})}$ be the greatest integer $s' \le \text{maxScore}_A$ such that, $N \le \Sigma_{s' \le s \le \text{maxScore}} \text{numCoreNRPs}_A$ (n,s).

Then, we define

$$\text{thresholdScore}_A(N) = \begin{cases} \text{score}_\text{N} & \text{if score}_N < \text{score}_{10^5} \\ \text{score}_N - 1 & \text{if score}_N = \text{score}_{10^5} \end{cases} \tag{2}$$

NRPminer selects, $\text{candidateCoreNRPs}_{\text{A}}(N)$, defined as the set of all core NRPs of A, with adenylation score at least thresholdScore_A (N). NRPminer selects core NRPs $\text{candidateCoreNRPs}_A(N)$ for downstream spectral analyses. Using this approach, NRPminer is guaranteed to be scalable as at most 10^5 candidate core NRPS are explored per assembly line.

Table 3 presents the values of $\text{numCoreNRPs}_{\text{SurugamideAL}}(8,s)$ for various values of s. Note that, this table presents the number of core NRP only for a single assembly line, SurugamideAL, corresponding to cyclic surugamides (surugamide A–D). In total, 14,345 core NRPs were retained from the original 3,927,949,830 core NRPs of the 11 assembly lines of surugamide’s BGC.

Problem 2. Given an assembly line A and a positive integer N, generate $\text{candidateCoreNRPs}_A(N)$, defined as all core NRPs of A with adenylation scores at least $\text{thresholdScore}_A(N)$.

NRPminer follows a graph-theoretic approach to quickly generate $\text{candidateCoreNRPs}_A(N)$ by using the computed values of numCoreNRPs. Let G (A) be the acyclic directed graph with nodes corresponding to pairs of positive integers $i \le n$ and $s \le \text{maxScore}_A$, such that $\text{numCoreNRPs}_A(i,s) > 0$, denoted by $v_{i,s}$. For every node $v_{i,s}$ $(i = 1,\ldots,n)$ and every $a \in A_i$ such that $\text{numCoreNRPs}_A(i{-}1,s{-}S_A(i,a)) > 0$, there exists a directed edge from $v_{i-1,s-S_A(i,a)}$ to

$v_{i,s}$. Let Source be $v_{0,0}$ and let Sink be the set of all nodes $v_{n,s}$ such that thresholdScore$_A$(*N*) $\leq s$. We call each directed path in *G*(*A*) from Source to the nodes in Sink as a candidate path of *G(A)*.

Each candidate path of *G*(*A*) corresponds to a distinct core NRP of *A* with adenylation score at least thresholdScore$_A$(*N*) and vice versa. Therefore, the problem of finding all core NRPs of *A* with adenylation score at least thresholdScore$_A$(*N*) corresponds to the problem of finding all candidate paths of *G* (*A*). While enumerating all paths with *n* nodes in a directed acyclic graph can grow exponentially large (as there can be exponentially many such paths), but due to our choice of thresholdScore$_A$(*N*), the number of candidate paths of *G*(*A*) is bound by 10^5 (or *N* if $\text{score}_N = \text{score}_{10^5}$). NRPminer uses the default value $N = 1000$. Moreover, $n \leq 20$ (only assembly lines made up of up to 20 A-domains are considered) and $k \leq 12$.

Forming PSMs and calculating PSM scores. PSMs and their PSM scores are described by Gurevich et al.[43]. Given a peptide *P* (with any backbone structure), we define Mass(*P*) as the sum of masses of all amino acids present in P. Furthermore, we define the graph of *P* as a graph with nodes corresponding to amino acids in *P* and edges corresponding to *generalized peptide bonds* as described in Mohimani et al.[100]. Then, we define theoretical spectrum of P (as opposed to the experimental spectrum) is the set of masses of all fragments generated by removing pairs of bonds corresponding to two-cuts in graph of *P* or by removing single bonds corresponding to the bridges in the graph of *P* as described by Mohimani et al.[100]. Each mass in this set is called a theoretical peak. Then, given the spectrum *S*, if precursor mass of *S* and Mass(*P*) are within a threshold Δ Da (where the default value of Δ is 0.02), we define the score of *P* against *S*, shown by SPCScore(*P*,*S*), as the number of peaks in theoretical spectrum of *P* that are within *ε* Da of a peak in *S* (where the default value of *ε* is 0.02). NRPminer only considers high-resolution data.

If (A_1, …, A_n) is the list of amino acid masses in a peptide *P*, we define Variant (*P*,*i*,*δ*) as (A_1,…, $A_i + \delta$, …, A_n), where *P* and Variant(*P*,*i*,*δ*) have the same topology and $A_i + \delta \geq 0$. VariableScore(*P*,*S*) is defined as

$$(\text{SPCScore}(\text{Variant}(P, i, \omega), S))$$

where *ω* is Mass(*P*) − Mass(*S*) and *i* varies from 1 to *n* (*n* stands for the number of amino acids in the peptide *P*)[43]. We define a variant of peptide *P* derived from a spectrum *S* as Variant(*P*,*i*,*ω*) of peptide *P*, which maximizes SPCScore(Variant(*P*,*i*, *ω*),*S*) across all positions *i* in *P*. For simplicity, we refer to this variant as Variant(*P*, *S*). Given *P* and *S*, VarQuest[43] uses a heuristic approach to efficiently find Variant (*P*,*S*).

NRPminer uses VarQuest[43] to perform modification-tolerant search of the input spectral datasets against the constructed peptide structures generated from selected core NRPs (see the NRPminer step "generating linear, cyclic, and branch-cyclic backbone structures for each core NRP" in Fig. 2 and "Method section"). Given a positive number MaxMass representing the maximum allowed modification mass (default value of MaxMass = 150), for each constructed structure *P* and input spectrum *S*, if |Mass*(P)*−Mass*(S)*| ≤ MaxMass, NRPminer uses VarQuest[43] to find the Variant*(P, S)*. In this context, Variant*(P,S)* represents the mature NRP with a single PAM on *P* that resulted in the mass difference |Mass *(S)*−Mass*(P)*|. Similar idea has been applied to identification of post-translational modifications in traditional proteomics[49,101].

Computing *P* values of PSMs. NRPminer uses the MS-DPR[89] to compute the statistical significance (*p* value) of each identified PSM. Given PSM(*P*, *S*) where *P* is a peptide with length *n* and *S* is a spectrum, MS-DPR estimates the probability that a random peptide, say *P*′, with length *n*, has SPCScore(P', *S*) ≥ SPCScore(*P*, *S*). We refer to this probability as *p* value of PSM(*P*, *S*). Monte Carlo approach can estimate the *p* value by generating a population of random peptides with length *n*, and scoring them against the spectrum *S*.

In case of MS-based experiments for identifying NRPs[102], we are often interested in PSMs with *p* value < 10^{-12} (the *p* values corresponding to high-scoring PSMs)[102]. But naive Monte Carlo approach is infeasible for evaluating such rare events as the number of trials necessary for exploring such low *p* value is too large to practically explore. To resolve this issue, MS-DPR[89] uses multilevel splitting technique for estimating the probability of rare event (i.e. high-scoring PSMs). MS-DPR[89] constructs a Markov Chain over the scores of all peptides with length *n* and then uses multilevel splitting to steer toward peptides that are more likely to form high PSM scores against *S*. Using this approach, MS-DPR[89] can efficiently estimate an extreme tail of the scores of all possible peptides against *S* which is then used to compute the *p* value of the PSM(*P*, *S*).

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

Data availability

All described datasets are available through the corresponding public repositories. XPF, SkinStaph, SoilActi, and TinyEarth datasets are available via MSV000081063, MSV000083956, MSV000078604, and MSV00084951 GNPS-accessions, respectively.

Code availability

NRPminer is available as both a stand-alone tool (https://github.com/mohimanilab/NRPminer) and as a web application via GNPS in silico toolbox. We used NPDtools, antiSMASH 3.5.0 and Biopython 1.78.

Received: 23 September 2020; Accepted: 4 May 2021;
Published online: 28 May 2021

References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).
2. Li, J. W. H. & Vederas, J. C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161–165 (2009).
3. Ling, L. L. et al. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
4. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14**, 111–129 (2015).
5. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc. Natl Acad. Sci. USA* **111**, 9259–9264 (2014).
6. Donia, M. S. et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
7. Zipperer, A. et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535**, 511–516 (2016).
8. Wilson, M. R. et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
9. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nat. Chem.* **7**, 411–417 (2015).
10. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651–2674 (1997).
11. Süssmuth, R. D. & Mainz, A. Nonribosomal peptide synthesis—principles and prospects. *Angew. Chem.—Int. Ed.* **56**, 3770–3821 (2017).
12. Renier, A. et al. Substrate specificity-conferring regions of the nonribosomal peptide synthetase adenylation domains involved in albicidin pathotoxin biosynthesis are highly conserved within the species Xanthomonas albilineans. *Appl. Environ. Microbiol.* **73**, 5523–5530 (2007).
13. Juguet, M. et al. An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in Streptomyces ambofaciens. *Chem. Biol.* **16**, 421–431 (2009).
14. Yu, D., Xu, F., Zhang, S. & Zhan, J. Decoding and reprogramming fungal iterative nonribosomal peptide synthetases. *Nat. Commun.* **8**, 15349 (2017).
15. Röttig, M. et al. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, 362–367 (2011).
16. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, 339–346 (2011).
17. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, 81–87 (2019).
18. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
19. Mori, T. et al. Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated Entotheonella sponge symbionts. *Proc. Natl Acad. Sci. USA* **33**, 3202–3210 (2018).
20. Hover, B. M. et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
21. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).
22. Khaldi, N. et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47**, 736–741 (2010).
23. Palaniappan, K. et al. IMG-ABC v. 5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* **48**, D422–D430 (2020).
24. Kautsar, S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
25. Medema, M. H. Computational genomics of specialized metabolism: from natural product discovery to microbiome ecology. *mSystems* **3**, e000182 (2018).

26. Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.* **12**, 233–239 (2016).
27. Weissman, K. J. The structural biology of biosynthetic megaenzymes. *Nat. Chem. Biol.* **11**, 660 (2015).
28. Caboche, S., Leclère, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
29. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
30. Mohimani, H. et al. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *J. Nat. Products* **77**, 1902–1909 (2014).
31. Tobias, N. J. et al. Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. *Nat. Microbiol.* **2**, 1676–1685 (2017).
32. Ninomiya, A. et al. Biosynthetic gene cluster for Surugamide A encompasses an unrelated decapeptide, Surugamide F. *ChemBioChem* **17**, 1709–1712 (2016).
33. Goyal, R. K. & Mattoo, A. K. Multitasking antimicrobial peptides in plant development and host defense against biotic/abiotic stress. *Plant Sci.* **228**, 135–149 (2014).
34. Reimer, D. et al. Rhabdopeptides as insect-specific virulence factors from entomopathogenic bacteria. *ChemBioChem* **14**, 1991–1997 (2013).
35. Hacker, C. et al. Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. *Nat. Commun.* **9**, 1–11 (2018).
36. Hoyer, K. M., Mahlert, C. & Marahiel, M. A. The Iterative Gramicidin S thioesterase catalyzes peptide ligation and cyclization. *Chem. Biol.* **14**, 13–22 (2007).
37. Li, S., Wu, X., Zhang, L., Shen, Y. & Du, L. Activation of a cryptic gene cluster in lysobacter enzymogenes reveals a module/domain portable mechanism of nonribosomal peptide synthetases in the biosynthesis of pyrrolopyrazines. *Org. Lett.* **19**, 5010–5013 (2017).
38. Cai, X. et al. Entomopathogenic bacteria use multiple mechanisms for bioactive peptide library design. *Nat. Chem.* **9**, 379 (2017).
39. Crosa, J. H. & Walsh, C. T. Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **66**, 223–249 (2002).
40. Medema, M. H. et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* **10**, e1003822 (2014).
41. Moss, N. A. et al. Nature's combinatorial biosynthesis produces Vatiamides A–F. *Angew. Chem.* **58**, 9027–9031 (2019).
42. Mohimani, H. et al. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
43. 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).
44. Meleshko, D. et al. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res.* **29**, 1352–1362 (2019).
45. Kersten, R. D. et al. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).
46. Nguyen, D. D. et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl Acad. Sci. USA* **110**, E2611–E2620 (2013).
47. Nguyen, D. D. et al. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. Nature. *Microbiology* **2**, 1–10 (2016).
48. Behsaz, B. et al. De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments. *Cell Syst.* **10**, 99–108 (2020).
49. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567 (2005).
50. 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).
51. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
52. Johnston, C. W. et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6**, 1–11 (2015).
53. Tietz, J. I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470 (2017).
54. Mohimani, H. et al. Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).
55. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
56. da Silva, R. R. et al. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, e1006089 (2018).
57. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proc. Natl Acad. Sci. USA* **104**, 6140–6145 (2007).
58. Handelsman, J. *Tiny Earth—Studentsourcing Antibiotic Discovery*. In Tiny Earth. https://tinyearth.wisc.edu (2018).
59. Hurley, A. et al. Tiny earth: a big idea for stem education and antibiotic discovery. *mBio* **12**, 1 (2021).
60. Frank, A. M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**, 587–591 (2011).
61. Moss, S. J., Martin, C. J. & Wilkinson, B. Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. *Nat. Prod. Rep.* **21**, 575–593 (2004).
62. He, J. & Hertweck, C. Iteration as programmed event during polyketide assembly; molecular analysis of the aureothin biosynthesis gene cluster. *Chem. Biol.* **10**, 1225–1232 (2003).
63. Nollmann, F. I. et al. Insect-specific production of new GameXPeptides in Photorhabdus luminescens TTO1, widespread natural products in entomopathogenic bacteria. *ChemBioChem* **16**, 205–208 (2015).
64. Zhou, Q. et al. Structure and biosynthesis of xenoamicins from entomopathogenic xenorhabdus. *Chemistry* **19**, 16772–16779 (2013).
65. Wenzel, S. C., Meiser, P., Binz, T. M., Mahmud, T. & Müller, R. Nonribosomal peptide biosynthesis: point mutations and module skipping lead to chemical diversity. *Angew. Chem. Int. Ed.* **45**, 2296–22301 (2006).
66. Wenzel, S. C. et al. Structure and biosynthesis of myxochromides S1-3 in Stigmatella aurantiaca: Evidence for an iterative bacterial type I polyketide synthase and for module skipping in nonribosomal peptide biosynthesis. *ChemBioChem* **6**, 375–385 (2005).
67. Seyedsayamdost, M. R., Traxler, M. F., Zheng, S. L., Kolter, R. & Clardy, J. Structure and biosynthesis of amychelin, an unusual mixed-ligand siderophore from amycolatopsis sp. AA4. *J. Am. Chem. Soc.* **133**, 11434–11437 (2011).
68. Arima, K., Kakinuma, A. & Tamura, G. Surfactin, a crystalline peptidelipid surfactant produced by Bacillus subtilis: isolation, characterization and its inhibition of fibrin clot formation. *Biochem. Biophys. Res. Commun.* **31**, 488–494 (1968).
69. Nishikiori, T., Naganawa, H., Muraoka, Y., Aoyagi, T. & Umezawa, H. Plipastatins: new inhibitors of phospholipase A2, produced by bacillus cereus BMG302-fF67: II. structure of fatty acid residue and amino acid sequence. *J. Antibiotics* **39**, 745–754 (1986).
70. Vollenbroich, D., Özel, M., Vater, J., Kamp, R. M. & Pauli, G. Mechanism of inactivation of enveloped viruses by the biosurfactant surfactin from Bacillus subtilis. *Biologicals* **25**, 289–297 (1997).
71. Huang, X. et al. Antiviral activity of antimicrobial lipopeptide from Bacillus subtilis fmbj against Pseudorabies virus, Porcine Parvovirus, Newcastle Disease virus and Infectious Bursal Disease virus in vitro. *Int. J. Pept. Res. Therapeutics* **12**, 373–377 (2006).
72. Wu, Y. S. et al. Anticancer activities of surfactin potential application of nanotechnology assisted surfactin delivery. *Front. Pharmacol.* **8**, 761 (2017).
73. Sandrin, C., Peypoux, F. & Michel, G. Coproduction of surfactin and iturin A, lipopeptides with surfactant and antifungal properties, by Bacillus subtilis. *Biotechnol. Appl. Biochem.* **12**, 370–375 (1990).
74. Cochrane, S. A. & Vederas, J. C. Lipopeptides from Bacillus and Paenibacillus spp.: a gold mine of antibiotic candidates. *Med. Res. Rev.* **36**, 4–31 (2016).
75. Rodrigues, L., Banat, I. M., Teixeira, J. & Oliveira, R. Biosurfactants: potential applications in medicine. *J. Antimicrob. Chemother.* **57**, 609–618 (2006).
76. Wang, C. L., Ng, T. B., Yuan, F., Liu, Z. K. & Liu, F. Induction of apoptosis in human leukemia K562 cells by cyclic lipopeptide from Bacillus subtilis natto T-2. *Peptides* 28, 1344–1350 (2007).
77. Agrawal, S., Acharya, D., Adholeya, A., Barrow, C. J. & Deshmukh, S. K. Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential. *Front. Pharmacol.* **21**, 828 (2017).
78. Zhao, H. et al. Effect of cell culture models on the evaluation of anticancer activity and mechanism analysis of the potential bioactive compound, iturin A, produced by: Bacillus subtilis. *Food Funct.* **10**, 1478–1489 (2019).
79. Gong, A. D. et al. Antagonistic mechanism of iturin a and plipastatin a from Bacillus amyloliquefaciens S76-3 from wheat spikes against Fusarium graminearum. *PLoS ONE* **10**, e0116871 (2015).
80. Lange, A., Sun, H., Pilger, J., Reinscheid, U. M. & Gross, H. Predicting the structure of cyclic lipopeptides by bioinformatics: structure revision of arthrofactin. *ChemBioChem* **13**, 2671–2675 (2012).
81. Li, W. et al. The Antimicrobial compound xantholysin defines a new hroup of Pseudomonas cyclic lipopeptides. *PLoS ONE* **8**, e62946 (2013).
82. Kim, D. P., Tatiana, T. & Donna, R. M. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2006).
83. Mukherjee, S. et al. Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Res.* **47**, D649–D659 (2019).

84. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
85. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1–8 (2020).
86. Bouslimani, A. et al. Lifestyle chemistries from phones for individual profiling. *Proc. Natl Acad. Sci. USA* **113**, E7645–E7654 (2016).
87. Hur, G. H., Vickery, C. R. & Burkart, M. D. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat. Prod. Rep.* **29**, 1074–1098 (2012).
88. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
89. Mohimani, H., Kim, S. & Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *J. Proteome Res.* **12**, 1560–1568 (2013).
90. Zhou, W. et al. Host-specific evolutionary and transmission dynamics shape the functional diversification of Staphylococcus epidermidis in human skin. *Cell* **180**, 454–470 (2020).
91. Bouslimani, A. et al. Molecular cartography of the human skin surface in 3D. *Proc. Natl Acad. Sci. USA* **112**, E2120–E2129 (2015).
92. Bouslimani, A. et al. The impact of skin care products on skin chemistry and microbiome dynamics. *BMC Biol.* **17**, 47 (2019).
93. Mohimani, H. et al. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **11**, 3642–3650 (2011).
94. Bode, E. et al. Promoter activation in Δhfq mutants as an efficient tool for specialized metabolite production enabling direct bioactivity testing. *Angew. Chem.* **131**, 19133–19139 (2019).
95. Bode, E. et al. Simple 'on-demand' production of bioactive natural products. *ChemBioChem* **16**, 1115–1119 (2015).
96. Bode, H. B. et al. Determination of the absolute configuration of peptide natural products by using stable isotope labeling and mass spectrometry. *Chemistry* **18**, 2342–2348 (2012).
97. Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174–4183 (1997).
98. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
99. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799–5808 (2005).
100. 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).
101. Tanner, S. et al. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626–4639 (2005).
102. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
103. Mohimani, H. et al. Multiplex de novo sequencing of peptide antibiotics. *J. Comput. Biol.* **18**, 1371–1381 (2011).

Acknowledgements

The work of B.B. and P.A.P. was supported by the 2018 Grand Challenge Award from the UC San Diego Center for Microbiome Innovation. The work of H.M. was supported by a research fellowship from the Alfred P. Sloan Foundation, a National Institutes of Health New Innovator Award DP2GM137413. The work of B.B. and H.M. were also supported by a U.S. Department of Energy award DE-SC0021340. The work in the Bode lab was supported by the LOEWE-Center Translational Biodiversity Genomics (TBG) and the LOEWE Schwerpunkt MegaSyn, both funded by the State of Hesse and an ERC Advanced Grant (SYNPEP, grant agreement number 835108). The work of A.G. was supported by the Russian Science Foundation (grant 20-74-00032). J.O. was supported by the NIH/NIGMS (1 DP2 GM126893-01) and the Department of Defense (W81XWH1810229). A.M.C.-R. and P.C.D. were supported by the National Sciences Foundation grant IOS-1656481 and National Institutes of Health (NIH) grant 1DP2GM137413-01.

Author contributions

B.B. and H.M. designed the NRPminer algorithm and B.B. developed NRPminer. B.B. performed the benchmarking and spectral network and VarQuest analysis for all datasets included in this study. A.G. created the NRPminer web application and tutorials on GNPS. E.B., Y.-N.S., F.G., A.L., H.B.B. generated XPF dataset and executed all the experimental analyses for Protegomycin, Xenoinformycin, and Xenoamicin-like NRP families. D.A. generated the *TinyEarth* dataset. A.M.C.-R. and P.C.D analyzed *SoilActi* dataset. A.B., M.P., C.G., and J.O. generated the *SkinActi* dataset. H.B.B., P.A.P., and H.M. directed the work. B.B., P.A.P., and H.M. wrote the manuscript with contributions from all the co-authors.

Competing interests

P.A.P. 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. B.B. and H.M. are co-founders and have equity interests from Chemia.ai, LLC. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-23502-4.

Correspondence and requests for materials should be addressed to H.B.B., P.A.P. or H.M.

Peer review information *Nature Communications* thanks Rafael Cuadrat and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.