Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery

Bahar Behsaz1,2,3,12, Edna Bode4,12, Alexey Gurevich5, Yan-Ni Shi4, Florian Grundmann4, Deepa Acharya6, Andrés Mauricio Caraballo-Rodríguez7, Amina Bouslimani7, Morgan Panitchpakdi7, Annabell Linck4, Changhui Guan8, Julia Oh8, Pieter C. Dorrestein2,7, Helge B. Bode4,9,10✉ & Hosein Mohimani3✉

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.
Microbial natural products represent a major source of bioactive compounds for drug discovery. 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. NRPs have been reported in various habitats, from marine environments to soil and even human microbiomes. 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). 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 A-domain (A-domain) that is responsible for recognition and activation of the specific amino acid, which can be incorporated by that module through the non-ribosomal code. At minimum, each NRPS module also includes an Thiolation domain (T-domain) and a Condensation domain (C-domain) that are responsible for loading and elongation of the core 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.

In the past decade, genome mining methods have been developed for predicting the NRP sequences from their BGC sequences. Genome mining tools, such as antiSMASH, start developing NRPS modules and predict the amino acids incorporated by the A-domain in each module using the substrate prediction algorithm developed for predicting the NRP sequences from their BGC. Each NRPS module contains an A-domain that is responsible for recognition and activation of the specific amino acid, which can be incorporated by that module through the non-ribosomal code. At minimum, each NRPS module also includes an Thiolation domain (T-domain) and a Condensation domain (C-domain) that are responsible for loading and elongation of the core 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.

Discovery of NRPs involves a multitude of challenges such as PAM identification (with exception of methylation and epimerization), 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 NRP 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.

Another challenge in discovering NRPs is due to the non-canonical assembly lines. While in many NRPs 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 (see Supplementary Fig. 1a), there are many NRPs that violate this pattern (see Supplementary Fig. 1b). For example, during biosynthesis of rhabdopeptides and lugdunins, 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 (see Supplementary Fig. 1c). For example, surugamide BGC produces two completely distinct NRPs through different non-canonical assembly lines (see 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.

To address these challenges, multiple peptidogenomics approaches have been developed for discovering peptidic natural products by combining genome mining and mass spectrometry (MS) information. These approaches often use antiSMASH to find all NRPS BGCs in the input genome, use NRPSpredictor to generate putative core NRPs encoded by each BGC, and attempt to match mass spectra against these putative NRPs. Kersten et al. 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 NRPSpredictor. These approaches are limited to the identification of previously unreported variants of known NRPs, molecules present in spectral families with known compounds compared to BGCs.

Medema et al. complemented the manual approach from Kersten et al. by the NRP2Path 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 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. 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. 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. 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. However, as NRPquest uses a naive pairwise scoring of all NRP structures against all mass spectra for PAM identification, it is prohibitively slow when searching for PAMs. 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 such as the Global Natural Products Social (GNPS) molecular networking, PRISM, GNPS, ROEDO, Dereplicator, CSI:FingerIT, NAP, and CycloNovo 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, 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 NRfamilies, 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) 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 approach.

#### Datasets

We analyzed four microbial isolate datasets from *Xenorhabdus* and *Photobacterium* 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 (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 spectraXPF, spectraSkinStaph, spectraSoilActi, spectraTinyEarth and the genomes are referred as genomeXPF, genomeSkinStaph, genomeSoilActi, and genomeTinyEarth, 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 and SpecNets 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. Known NRPs (NRfamilies) are identified either by Dereplicator search against the database of all known peptidic natural products (referred to as PNDbase) using the p value threshold 10−15, and/or by SpecNet search against the library of all annotated spectra available on GNPS. 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. sertormai 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
**Fig. 1 NRPminer pipeline.** a Predicting NRPS BGCs using antiSMASH\textsuperscript{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. \textsuperscript{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\textsuperscript{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 statistically significant PSMs, and i expanding the set of identified spectra using spectral networks (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) | #previously unreported variants of known NRPs | #previously 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-Cluster\textsuperscript{60}, and the number of unique families is estimated using SpecNets\textsuperscript{50}. 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 the BGC from a related well-assembled strain.

For example, NRPminer search of spectraXPF against genomic 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.

SpecNet analysis against the annotated spectra in GNPS50 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 literature31 (see Supplementary Table 2 for the list of identified BGCs). Supplementary Figure 3 presents an example of an identified NRP family, s magnetide, and its corresponding BGC in *X. s entirrnaii*. For one family (xentivialpeptides) 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 xoenoamicin-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 xenoinformycins-like family in the XPF dataset and aminiformamide in SoilActi), as well as the previously unreported variants in two additional NRP families (ludgudin in SkinStaph and surugamide in SoilActi). Discovery of protegomycins (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.**

| A1       | A2       | A3       | A4       | A5       | A6       | A7       | A8       |
|----------|----------|----------|----------|----------|----------|----------|----------|
| 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)  |

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.
Additional derivatives were found in large-scale cultivation of wild type and Δhfq 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.
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 often fails to predict Lys residues, due to lack of training data for this amino acid. 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 proteomycins 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* (IC50 [mg/L] 79 and 53) and *Plasmodium falciparum* (IC50 [mg/L] > 50 and 33) with no toxicity against L6 rat myoblast cells (IC50 [mg/L] both >100).

**Discovery of xenoinformycin (XINF) NRP family in the XPF dataset.** NRPminer matched four spectra representing four cyclic NRPs in *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) 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. 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.
**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 BGCs 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 pre-NRP xenoamicin A (Supplementary Fig. 19). We named this pre-NRP xenoamicin A (Supplementary Fig. 19). We named this pre-NRP xenoamicin A (Supplementary Fig. 19). We named this pre-NRP xenoamicin A (Supplementary Fig. 19).
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[\(\text{E}-1.0\)]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 spectraSkinStaph dataset against *Staphylococcus lugdunensis* 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 VWLVVV\(t\) 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.

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. 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 the 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 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.](https://doi.org/10.1038/s41467-021-23502-4)
of activities. Surfactins are reported to have anti-viral\textsuperscript{70,71}, anti-tumor\textsuperscript{72}, anti-fungal\textsuperscript{73}, and anti-microbial\textsuperscript{74} functions\textsuperscript{5–78} and plipastatins have known anti-fungal activities\textsuperscript{79}. In the analysis of \textit{Bacillus amyloliquefaciens} sp. GZ7CCT-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\textsuperscript{tinyEarth} against putative NRP structures generated from \textit{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\textsuperscript{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\textsuperscript{tinyEarth} (xantholysins\textsuperscript{81}) since the xantholysin BGC was split among multiple contigs in the \textit{Pseudomonas plecoglossicida} sp. YNA158 genome assembly.

### Identifying surugamides in the SoilActi dataset

NRPminer identified 183 spectra representing 25 NRPs when searching spectra\textsubscript{soilActi} against \textit{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\textsuperscript{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\textsuperscript{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 \textit{Streptomyces albus}\textsuperscript{82}, this NRP lacks the N-terminal tryptophan. Surugamide F was not identified in the spectral dataset from \textit{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\textsuperscript{82} and genome online database (GOLD)\textsuperscript{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...
NRPminer identifies NRPs in the ARF family. The known arthrofactins are shown in blue, while the purple nodes represent the previously unreported variants.

The b-galactosidase 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 or GOLD datasets 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 with PAMs. This feature was crucial for discovering protegomycins that include a PAM (lipid chain) and a mis-prediction with PAMs.

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. NRP BGCs are long (average length ~60 kb) and repetitive BGCs within a single contig. With recent advances in long-read sequencing technologies, more contiguous microbial microbiome correspond to deodorants, shampoos, and other

sensitive levels 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 way toward the development of more accurate machine learning techniques for A-domains specificity prediction.
beauty products, rather than microbial products. 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. 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, microbes as NRPs). However, ~99% of spectra in these datasets metabolomics. therefore 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 discovery49,50 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 Biopython49 and antiSMASH17 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 biosyntheticSpAdes45, a specialized short-read BGC assembler. (b) Predicting putative amino acids for each A-domain in the identified BGCs. NRPminer uses NRPspredictor2 (ref. 13) to predict putative amino acids for each position in an A-domain. Given a 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-domains16,18. NRPminer uses NRPspredictor2 (ref. 13) 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 ORF, 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. (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 millions of putative core NRPs. NRPminer uses the specificity scores of amino acids in each set of core NRPs for filtering domains in spectral analyses. Given an assembly line $\ldots\alpha_i\ldots\alpha_k$, for each amino acid $a \in \alpha_i (i = 1,\ldots,k)$, NRPminer first divides the specificity score of $a$ by the maximum specificity score observed across all amino acids in $\alpha_i$ (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 contains more information about the structure of the final NRP (0- or 1-amino acids). (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 condensation-epimerization domains) that provide information about the structure of the final NRP (0- or 1-amino acids).
XFP: 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 previously (Supplementary Table 1). 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 spectra81 contains 27 spectral sub-datasets representing each sample for a total of 263,768 spectra across all strains (GNPS-accession #: MSV0000863). The genomeShasta dataset contains 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.93. 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.biblo.core.windows.net/documents/c1a3c84a-e363-416e-9d26-b089c67c0202.pdf). The Phree kit was conditioned using 50 mL 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 500g and the clean up extracts were lyophilized using a FreezeFree Dryer Freeze Dryer with Centrivap Concentrator (Labconco). Wells were resuspended in 200 µL of resuspension solvent with 1.0 µM (S)-1-Acetyltropolone, vortexed, and shaken at 2000 rpm 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. Extracts were analyzed using a Thermo Fisher UltraMate III System for liquid chromatography and a Maxis Q-TOF (Quadrupole-Time-of-Flight) mass spectrometer (Bruker Daltonics), controlled by the Otof Control and Hystar liquid chromatography and a Maxis Q-TOF (Quadrupole-Time-of-Flight) mass spectrometer extracts were diluted 1:25 (vol/vol) with methanol and analyzed by UPLC-ESI mass spectrometer with 0.2% formic acid was performed in 5 mL LB (1/2, 1/3, and 1/4 isoiso species) 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 1/2C-labelled amino acids with a concentration of 2 mM. After 48 h cultivation at 30 °C, constantly shaking at 200 rpm, Amberlite® XAD-16 beads were harvested and extracted with 5 mL MeOH for 45 min. Samples were then taken from the filtered extracts and concentrated for 15 min at 17,000 g for further HPLC-MS analysis (Dionex Ultimate 3000 coupled to a Bruker AmaZon ion trap). Generated HPLC-MS data were interpreted as described previously94,95.

Additional analyses for Xanthomonas-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™ 1/4C medium, ISOGRO™ 1/3 medium. Fifty microliters ISOGRO™ medium was prepared with ISOGRO™ powder (0.5 g), K2HPO4 (1.4 g/l), KH2PO4 (1.4 g/l), MgSO4 7H2O (1 g/l), and CaCl2 (0.01 g/l) for 1/3C medium supplemented with 1/2C 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 culture medium. The crude extracts were analyzed by means of MALDI-MS and HPLC-MS (Brunner AmaZon). HPLC-based purification: XAM-1320 was isolated by a two-step chromatography. Strain KJ12.1 was cultivated in a BIOSTAT A plus fermentor (Sartorius) equipped with a 2-L vessel in 1.5 L of LB broth at 37 °C. 0.5 mL of 1/2C medium was used to inoculate 10 mL of 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% phosphoric acid or 0.1 N NaOH. Ammonium chloride was added to the culture medium. The XAM-1320 was harvested with an extraction after evaporation. Xanthomonas III A was isolated by a two-step chromatography. In the first step the fraction was extracted with a 5–95% water/acetonirole gradient over 15 min on a Luna C8 10 µm 50 × 50 mm column (Phenomenex). In the second step XAM-1320 was isolated with a 40–60% water–acetonirole gradient over 19 min on Luna C18 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 (Brunner Daltonics). Separation was done on a C18 column (ACQUITY UPLC BEH, 1.7 mm, 2.1 × 50 mm, flow rate 0.4 mL/min, Waters). Acetonirole/water containing 0.1% formic acid was used as a mobile phase. The gradient started with 5% acetonirole continuous over 2 min. Over 0.5 min under a linear gradient acetonirole reaches 40%. Following an equilibration phase over 1.5 min with 40% acetonirole takes place. For separation a linear gradient from 40–95% acetonirole over 10.5 min was used. The gradient ends up with 95% acetonirole continuous over 1.5 min. Collision-induced dissociation (CID) was performed on ion trap in the 23 Hz ESI-HPLC/MS. After the obtained on a LC-coupled Impact II ESI-TOF spectrometer (Brunner Daltons).

Advanced Marfrey’s method: The advanced Marfrey’s method to determine the configurations of the amino acid residues was performed as described previously96.

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. Enzlin 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 inferred based on the sequence similarity of its non-ribosomal code to those of the A-domains with known speciﬁcity. Given an input A-domain AD, NRPSpredictor2 (ref.15) first compares the sequence of the non-ribosomal code of AD to those of already characterized A-domains stored in the NRSp predictor database. After that each amino acid a, NRPSpredictor2 (ref.15) reports the Stachelhaus score of (speciﬁcity of an A-domain AD, that is (the integral value of the) 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. Given an A-domain, they used a Support Vector Machine (SVM) method trained on previously annotated 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. Given an A-domain, they used a Support Vector Machine (SVM) method trained on previously annotated 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.
Given $A = A_1, \ldots, A_n$, we call the set of all core NRPs generated by the cartesian product $A_1, \ldots, A_n$, as the core NRPs of $A$. For each core NRP of $A$, $a_{i_{1}} \ldots a_{i_{k}}$, we define the adenylation score $s_{A_{i_{1}} \ldots a_{i_{k}}}$, denoted by $Score_{A}(a_{i_{1}} \ldots a_{i_{k}})$, to be the sum of the normalized specificity scores of all of its amino acids:

$Score_{A}(a_{i_{1}} \ldots a_{i_{k}}) = \sum_{i=1}^{k} \frac{s_{i}}{\sum_{j} s_{j}}$

Therefore, given assembly line SurugamideAL and core NRP, $P=\text{AIKIKIFEL}$ (the core NRP corresponding to surugamide A), $Score_{\text{SurugamideAL}}(P) = 80 + 100 + 100 + 100 + 100 + 100 + 86 = 766$. Note that, for any assembly line $A$, the maximum value of $Score_{A}$, denoted by $maxScore_{A}$, is $100n$. 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 $10^7$ peaks in a modification-tolerant manner as the runtime grows exceedingly large when the database size grows. 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 $10^4$ ($10^5$). 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 $k = \max \left\lceil \left\{ [\frac{1}{n}] \right\} \right\rceil$ where $[\cdot]$ shows the number of amino acids in $A_i$. For any positive integers $i$ and $s$ satisfying, $1 \leq i \leq n$ and $s \leq maxScore_{A}$, let $numCoreNRPs_{A}(i, s)$ denote the number of core NRPs, of assembly line $A_1 \ldots A_n$, with $Score_{A} \geq s$ equal to $s$. Let $numCoreNRPs_{A}(i, 0) = 0$ for any positive integer $s$, and $numCoreNRPs_{A}(i, s) = 0$ for any integer $s < 0$, across all possible values of $i$. Then, for any positive integers $i$ and $s$ satisfying $1 \leq i \leq n$ and $0 \leq s \leq maxScore_{A}$, we have

$$numCoreNRPs_{A}(i, s) = \sum_{s_{i} \leq s} numCoreNRPs_{A}(i-1, s_{i})$$

(1)

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

Given a positive integer $N < 10^5$, let $k_{\text{not}}(N)$ be the greatest integer $s \leq maxScore_{A}$ such that, $N \leq \sum_{s \leq s_{i}} numCoreNRPs_{A}(n, s)$. Then, we define

$$\text{thresholdScore}(N) = \begin{cases} \text{score}_{N} & \text{if score}_{N} < \text{score}_{\text{not}}(N) \\ \text{score}_{\text{not}}(N) - 1 & \text{if score}_{N} = \text{score}_{\text{not}}(N) \end{cases}$$

(2)

NRPminer selects, candidateCoreNRPs_{A}(N), defined as the set of all core NRPs of $A$, with adenylation score at least $\text{thresholdScore}(N)$. NRPminer selects core NRPs in candidateCoreNRPs_{A}(N) for downstream spectral analyses. Using this approach, NRPminer is guaranteed to be scalable as at most $10^3$ candidate core NRPs are explored per assembly line.

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

| s     | numCoreNRPs_{SurugamideAL}(8, s) |
|-------|----------------------------------|
| 800   | 790                              |
| 788   | 786                              |
| 780   | 778                              |
| 776   | 774                              |
| 772   | 770                              |
| Total |                                  |

Table 3 presents the values of $numCoreNRPs_{\text{SurugamideAL}}(8, s)$ 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 candidateCoreNRPs_{A}(N), defined as all core NRPs of A with adenylation scores at least thresholdScore_{A}(N).

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

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 NRPminer2 training datasets. 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 $\Sigma_{\text{SurugamideAL}}$ for integers 1 \leq i \leq 8 and (at least) the three amino acids with the highest normalized specificity scores for each A-domain in this assembly line.
Forming PSMs and calculating PSM scores. PSMs and their PSM scores are described by Gueriche et al.33. 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 the 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, we define VariableScore(P, S) as the sum of masses of all amino acids present in P. Furthermore, the number of peaks in theoretical spectrum of P (where the default value of i 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 i is 0.02). NRPminer only considers high-resolution data.

Given (εi, ..., εn) is the list of amino acid masses in a peptide P, we define Variant(P, εi) as (A1, ..., An, δ1, ..., δn), where P and Variant(P, εi) have the same topology and A1 + δ1 > 0. VariableScore(P, S) is defined as

\[
\text{SPCScore}(\text{Variant}(P, \epsilon_i), S)
\]

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). We define a variant of peptide P derived from a spectrum S as Variant(P, εi) of peptide P, which maximizes SPCScore(Variant(P, εi), S) across all positions i in P. For simplicity, we refer to this variant as Variant(P, S). Given P and S, VarQuest43 uses a heuristic approach to efficiently find Variant(P, S).

NRPminer uses VarQuest43 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 VarQuest43 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.98,101.

Computing P values of PSMs. NRPminer uses the MS-DPR98 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 NRPs99, we are often interested in PSMs with p value < 10−12 (the p values corresponding to high-scoring PSMs). 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-DPR98 uses multilevel splitting technique for estimating the probability of rare event (i.e. high-scoring PSMs). MS-DPR98 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 SPS scores against S. Using this approach, MS-DPR98 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, MSV000076804, and MSV000084951 GNPS-accessions, respectively.

Code availability
NRPminer is available as both a stand-alone tool (https://github.com/mohimaniab/NRPminer) and as a web application via GNPS in silico toolbox. We used NRDtools, 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. 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., Rouhainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of cyclic nonribosomal 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. Mazaehi, 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-confering regions of the nonribosomal peptide synthetase adenylation domains involved in albicidin pathotoxin biosynthesis are highly conserved within the species Xanthomonas albovirens. Appl. Environ. Microbiol. 73, 5523–5530 (2007).
13. Fuguet, 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 synthetase pathways. 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., Kohnbacher, 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 genomes validates rich and varied specialized metabolism of uncultivated Entotheonella sponge symbionts. Proc. Natl Acad. Sci. USA 133, 3202–3210 (2018).
20. Hower, 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. J. et al. Discovery of the tyrobertaine 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. MBIG 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, e00182 (2018).
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 Center for Microbiome Innovation. 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 the National Institutes of Health (NIH) grant AR0737413-01.