A supervised fingerprint-based strategy to connect natural product mass spectrometry fragmentation data to their biosynthetic gene clusters

Authors: Tiago F. Leao[1], Mingxun Wang[1,2], Ricardo da Silva[3], Justin J.J. van der Hooft[4], Anelize Bauermeister[1], Asker Brejnrod[1], Evgenia Glukhov[5], Lena Gerwick[5], William H. Gerwick[1,5], Nuno Bandeira[1,2], Pieter C. Dorrestein[1,6,7,*].

Author Affiliations:

1 – Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA.
2 – Center for Computational Mass Spectrometry, University of California San Diego, La Jolla, California, USA.
3 – NPPNS, Physic and Chemistry Department, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil.
4 – Bioinformatics Group, Wageningen University, Wageningen, the Netherlands.
5 – Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA.
6 – Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA.
7 – Departments of Pharmacology and Pediatrics, University of California San Diego, La Jolla, California, USA.

* Corresponding author

Abstract

Microbial natural products, in particular secondary or specialized metabolites, are an important source and inspiration for many pharmaceutical and biotechnological products. However, bioactivity-guided methods widely employed in natural product discovery programs do not explore the full biosynthetic potential of microorganisms, and they usually miss metabolites that are produced at low titer. As a complementary method, the use of genome-based mining in natural products research has facilitated the charting of many novel natural products in the form of predicted biosynthetic gene clusters that encode for their production. Linking the biosynthetic potential inferred from genomics to the specialized metabolome measured by metabolomics would accelerate natural product discovery programs. Here, we applied a supervised machine learning approach, the *K*-Nearest Neighbor (KNN) classifier, for systematically connecting metabolite mass spectrometry data to their biosynthetic gene clusters. This pipeline offers a method for annotating the biosynthetic genes for known, analogous to known and cryptic metabolites that are detected via mass spectrometry. We demonstrate this approach by automated linking of six different natural product mass spectra, and their analogs, to their corresponding biosynthetic genes. Our approach can be applied to bacterial, fungal, algal and plant systems where genomes are paired with corresponding MS/MS spectra. Additionally, an approach that connects known metabolites to their biosynthetic genes

potentially allows for bulk production via heterologous expression and it is especially useful for cases where the metabolites are produced at low amounts in the original producer.

Significance

The pace of natural products discovery has remained relatively constant over the last two decades. At the same time, there is an urgent need to find new therapeutics to fight antibiotic resistant bacteria, cancer, tropical parasites, pathogenic viruses, and other severe diseases. To spark the enhanced discovery of structurally novel and bioactive natural products, we here introduce a supervised learning algorithm (*K*-Nearest Neighbor) that can connect known and analogous to known, as well as MS/MS spectra of yet unknowns to their corresponding biosynthetic gene clusters. Our Natural Products Mixed Omics tool provides access to genomic information for bioactivity prediction, class prediction, substrate predictions, and stereochemistry predictions to prioritize relevant metabolite products and facilitate their structural elucidation.

Introduction

Microbial natural products (NPs), also referred to as secondary or specialized metabolites, are often made by biosynthetic genes that are physically grouped into clusters (biosynthetic gene clusters or BGCs). Its been found that algae and plants can also contain BGCs, to some extent organized in a similar manner (1, 2). One of the challenges in the genome mining field is to connect microbial metabolites to their BGCs. Even the genome of *Streptomyces coelicolor* A3(2), one of the first sequenced microbial genomes, still contains a number of cryptic BGCs (BGCs without known metabolites)(3). In 2011, the bioinformatics tool antiSMASH (4) drastically improved the identification and annotation of BGCs based on automated genome mining. Similarly, since 2018, the program BiG-SCAPE (5) can reliably calculate the similarity between pairs of BGCs, grouping them into gene cluster families (GCFs). Recently, a number of approaches and tools have been created to connect NPs to their biosynthetic gene clusters, such as Pattern-based Genome Mining (6, 7), MetaMiner (8), CycloNovo (9), and NPLinker (10), recently reviewed by Van der Hooft *et al*., 2020 (11). However, most of these tools are not high-throughput or can only be used for a particular class of BGC (e.g., peptides or BGCs homologous to known BGCs). It has been challenging to create a systematic tool that can work at a repository scale to connect NP genotypes (BGCs) with their phenotypes (for example MS/MS spectra from untargeted mass spectrometry fragmentation profiles, LC-MS/MS). As a result, a large disparity exists between the number of known NPs versus the number of known BGCs. For example, the recently designated cyanobacterial genus *Moorena* has already yielded over 200 new metabolites, yet only a dozen of validated BGCs are currently deposited for this genus in the expert-annotated Minimum Information about a Biosynthetic Gene cluster (MIBiG) database (12). Connecting the molecules to the genes would facilitate research into the ecological role and functions of the specialized metabolome by studying the regulation of the expression of their biosynthetic gene clusters.

To begin to address this gene cluster annotation gap, we deployed a *K*-Nearest Neighbor (KNN) algorithm that uses a similarity/absence BGC fingerprints and analogous similarity/absence MS/MS fingerprints to classify gene cluster family (GCF, a group of similar BGCs) candidates for each MS/MS spectrum (Fig. 1). We recently sequenced draft metagenomic-assembled genomes (MAGs) for 60 cyanobacteria, mostly from tropical marine environments. The most complete drafts were reported in Leao *et al.*, 2021 (13), and for these we also obtained untargeted metabolomic data via LC-MS/MS (36 deposited in the PoDP platform and 24 not published due to the quality of their paired MAGs). Despite the bad quality of some of these MAGs, we could still annotated BGCs. As a first test for our NPOmix workflow, using this cyanobacterial dataset, we connected curacin A's MS/MS spectrum with its correct GCF/BGC. The performance of our KNN approach was superior to using a Mantel correlation method (the Jupyter notebook for this correlation is available at the GitHub repository: https://github.com/tiagolbiotech/NPOmix). The major limitation for evaluation of our method was the lack of available test data for structures that are linked to their MS/MS spectra and biosynthetic gene clusters.

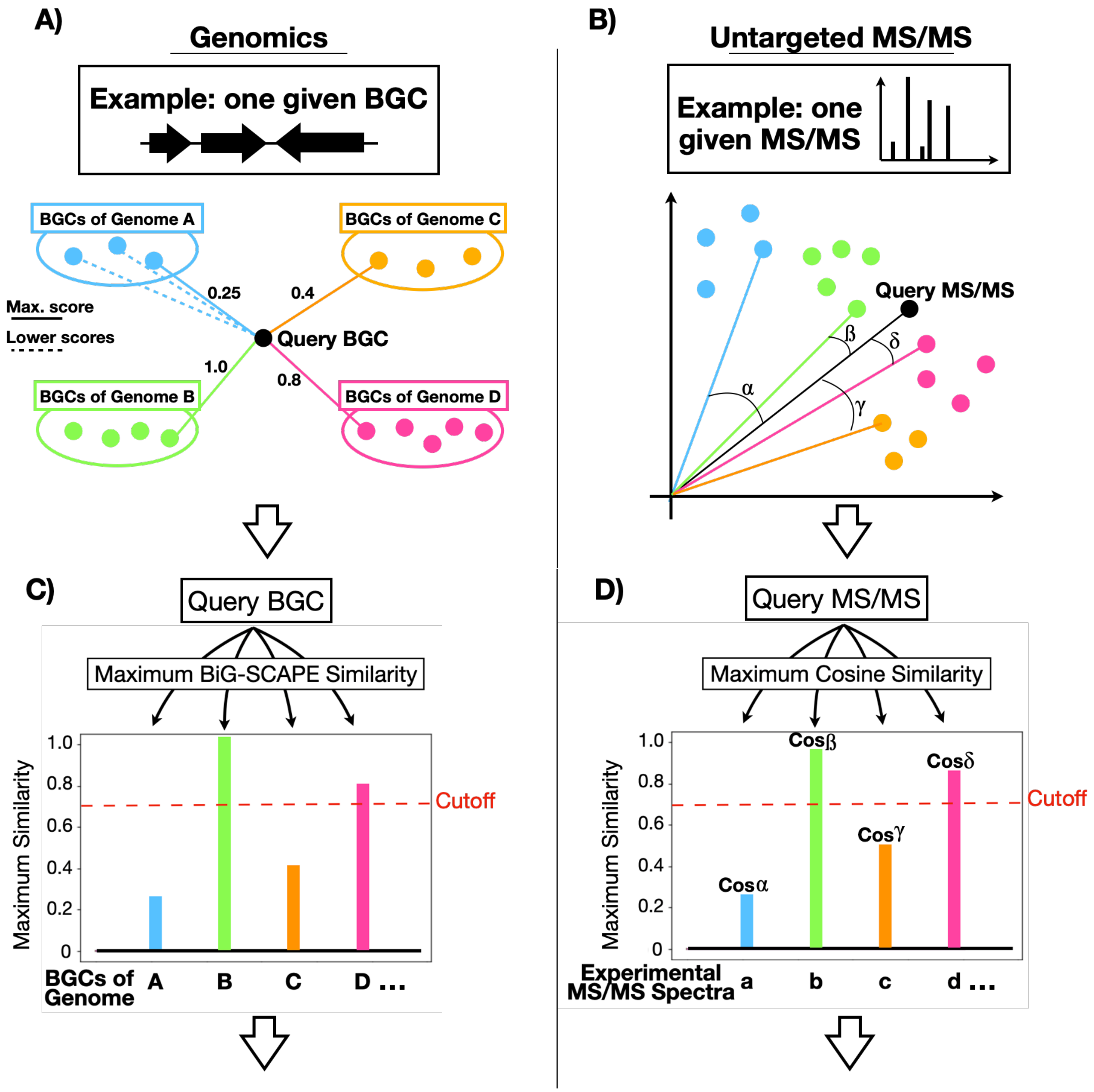
However, the training and testing set was expanded by the paired omics dataset from the recently built Paired Omics Data Platform (PoDP) (14), and enabled a further evaluation of our KNN tool (named NPOmix). The PoDP is the first community effort to make available validated links between BGCs, structures, and MS/MS spectra. In the present work, we used 36 out of the 71 paired metadatasets (listed in Dataset S1, sheet one). We selected genomic samples that contained a valid Genome ID or BioSample ID to aid in downloading them from the National Center for Biotechnology Information (NCBI) database, resulting in 732 genomes/MAGs obtained from these 36 PoDP metadatasets. Following the same procedure of the genomes, we also selected and assembled 1,034 metagenomes from part of these PoDP datasets. Additionally, using already linked MS/MS-BGC information from the PoDP and from a NPLinker dataset (10), we obtained validated data for eight metabolite families (major compounds and analogs). These compound families were orfamides, albicidins, bafilomycin, nevaltophin D, jamaicamide, hectochlorin, palmyramide and cryptomaldamide (totaling 15 reference MS/MS spectra due to the presence of analogs and sometimes more than one spectrum per metabolite). By training with the BGC fingerprints and testing these 15 validated links, we were able to correctly predict GCFs for 66.66% of the tested MS/MS fingerprints (10/15 reference MS/MS spectra were correctly classified using k = 3). Well-annotated links can be quickly prioritized by comparing substructures to mass differences in the fragmentation spectrum and/or predicted structures. A two-dimensional comparison of both types of fingerprints (BGC and MS/MS) can be a proxy for distinguishing some true positives from false positives. Critically, we filtered for BGC-MS/MS links wherein the query MS/MS spectra were mainly present in the same strains that the query BGCs were found (cutoff of 90% concordance between both BGC and MS/MS fingerprints). Once the PoDP data was filtered, our approach could connect BGCs with three types of mass spectra: known molecules (e.g., links that are validated experimentally), analogs of known molecules (e.g., links not validated but similar to validated reference spectra from the MS/MS database) or cryptic molecules (e.g., links without any library match, absent from the MS/MS database). We exemplify how it is possible to connect known BGCs to cryptic MS/MS spectra, new spectra that can be added to the current MS/MS databases. The same approach can be used for connecting new BGCs to cryptic MS/MS spectra that can be validated experimentally. While our approach uses unique fingerprints and a machine learning approach for connecting metabolites to BGCs, it can be considered a type of

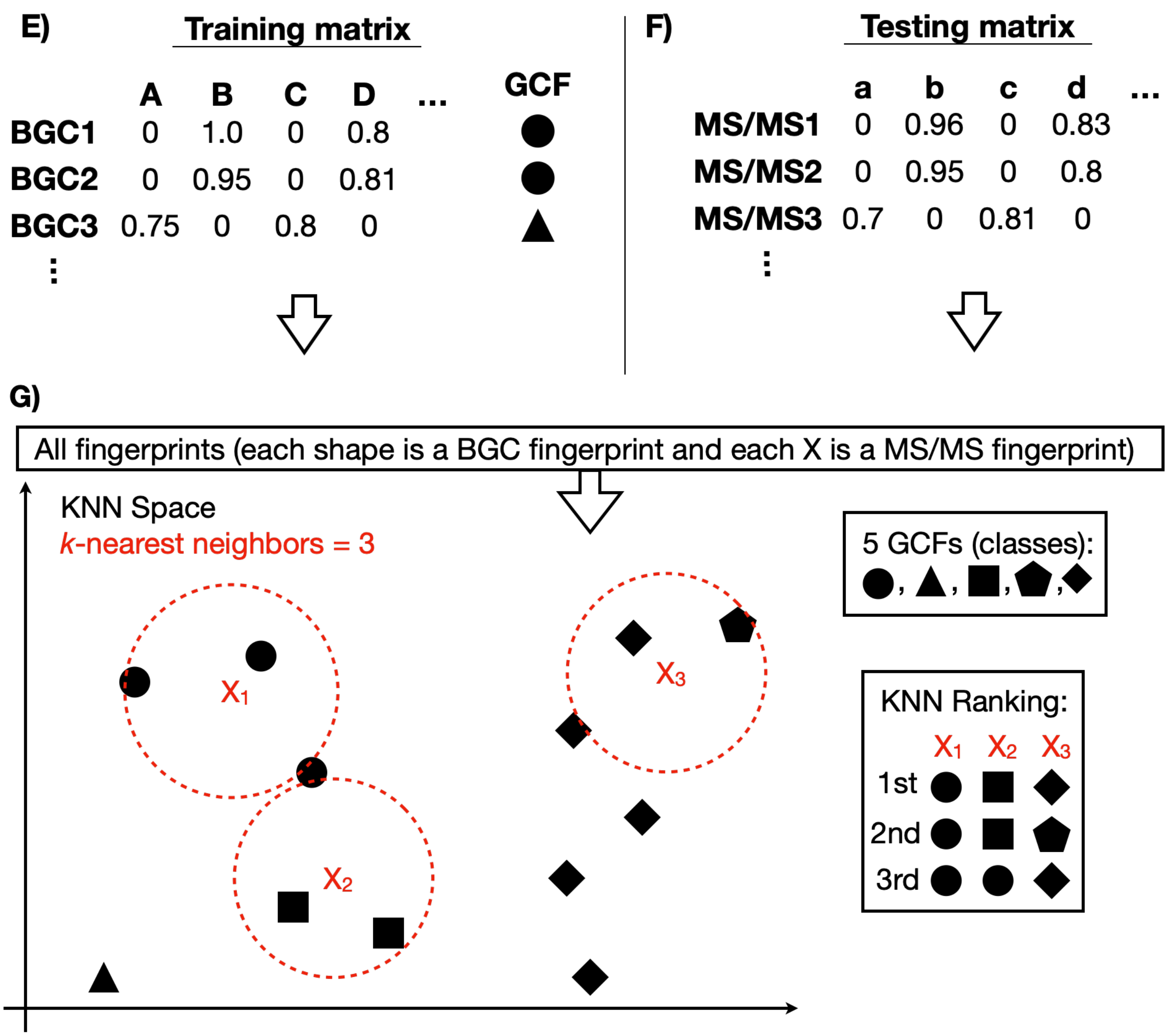
Pattern-based Genome Mining (PBGM) which was previously reported by Doroghazi *et al.* in 2014 and Duncan *et al.* in 2015 (6, 7). PBGM is based on the idea that the distribution of a given secondary metabolite should be comparable to the distribution of the BGCs responsible for their production.

Generally, finding novel metabolites for cryptic BGCs or even known BGCs (e.g., novel analogs) is very useful to accelerate natural products discovery, however, connection of known metabolites to their biosynthetic gene clusters is also important. Newly linked BGCs for known metabolites can lead to the discovery of new enzymatic processes. For example, in the strain *Anabaena variabilis* ATCC 29413, a NRPS gene is responsible for the attachment of a serine residue to generate the final mycosporine-like amino acids (MAA) product. However, in the strain *Nostoc punctiforme* ATCC 29133, this same step is performed by an ATP-grasp ligase (15). This highlights that different microbes can generate the same specialized metabolites through different biosynthetic routes, and therefore, we believe that our NPOmix tool will assist with the discovery of both novel metabolites as well as known metabolites with new biosynthesis.

Results and Discussion

The Natural Products Mixed Omics (NPOmix) Approach: Description of the Genomic and Metabolomic Pipelines. To use the NPOmix approach (Fig. 1 shows a conceptual example using only four samples), it is required to have a dataset of paired genomic and MS/MS information. The genomic information can be either that of a genome or metagenome, and the MS/MS spectra should be obtained via untargeted LC-MS/MS. Paired datasets have become available at the Paired omics Data Platform (PoDP)(14), one of the first initiatives to gather paired genomic and MS/MS information. Using BiG-SCAPE (5), each biosynthetic gene cluster (BGC) in the genome to be queried undergoes a pairwise similarity comparison (Fig. 1A) to every other BGC in the query set (e.g., the set of genomes used for the training, for example, the genomes downloaded from the PoDP), and similarity scores are computed as “1 minus BiG-SCAPE raw distance” to assign BGCs to Gene Cluster Families (GCFs), if possible. In order to create a BGC fingerprint (Fig. 1C), we identify the similarity between the query BGC and each of the BGCs in each genome in the training dataset. The BGC fingerprint that emerges is a series of columns for each compared genome, the column value of which represents the similarity score between the query BGC and the BGC to which it is maximally similar in a given genome (column). Similarity scores range from 0.0 to 1.0; identical BGCs have perfect similarity and are scored as 1.0 whereas a score of 0.8 would indicate that a homologous BGC is present in the genome. A score below the similarity cutoff of 0.7 indicates that the queried BGC is likely absent in the genome. A similar process is used to create MS/MS fingerprints (Fig. 1B); a query MS/MS spectrum is compared to all of the MS/MS spectra in the query set. This query spectrum could be either a reference spectrum from GNPS (16, 17) or a cryptic MS/MS spectrum from a new sample that contains a sequenced genome and experimental MS/MS spectra. In the case of MS/MS fingerprints (Fig. 1D), GNPS molecular networking was used to calculate the pairwise modified cosine score and then the maximum similarity was identified between the query MS/MS spectrum and the many MS/MS spectra in each experimental sample. This analysis only used the GNPS functions that are required to calculate a modified cosine similarity score between a pair of MS/MS spectra. The BGC fingerprints were used to create a training matrix (Fig. 1E) where rows are the maximum similarity scores for each BGC. Typically, this results in thousands of rows, and for our first release of NPOmix, we have captured this analysis for 5,421 BGCs that were present in 1,040 networked genomes/metagenomes (DNA samples can be downloaded using code from the GitHub repository, notebook 1), where each column is a genome and each value is the maximum similarity between the query BGC and the BGCs in this given genome. This BGC training matrix can be fed into the *K*-Nearest Neighbor (KNN) algorithm in order to train it with the genomic data. Additionally, one extra column is required in this BGC data matrix, a column that labels each BGC fingerprint with a GCF so the KNN algorithm will know the fingerprint patterns that belong together. The KNN algorithm plots the BGC fingerprints in the KNN feature space (in Fig. 1G). The KNN feature space is exemplified by only two dimensions as 1,040 dimensional space is not feasible to visualize (one dimension per sample). More details of how this multidimensional plotting occurs are described in the Fig. S1. where 3 BGCs are plotted in the three-dimensional space according to the scores from genomes A-C. The axis represent the genomes and the similarity values are coordinates in three-dimensional space. Next, the MS/MS fingerprints form a testing matrix (Fig. 1F), in this case, the matrix also contains 1,040 columns due to the 1,040 sets of paired experimental MS/MS spectra (samples can be downloaded using the ftp links from Dataset S1, sheet two). For example, for our first release, this testing matrix contained 15 MS/MS fingerprints (rows) for MS/MS reference spectra from the GNPS database (also present at the PoDP). Each query MS/MS fingerprint (a row in the testing metabolomic matrix and columns are the experimental MS/MS spectra per sample) are plotted into the same KNN feature space (Fig. 1G) so the algorithm can obtain the GCF labels for the nearest neighbors to the query MS/MS fingerprint (e.g., for three most similar BGC neighbors, k = 3). We note that GCF labels can be present more than once in the returned list if two or more BGC nearest neighbors belong to the same GCF. This repetition on the GCF classification is a common behavior of the KNN approach. Our approach is suitable for bacterial, fungal, algal and plant genomes and MS/MS spectra obtained from the same organism. Metagenomes and metagenome-assembled genomes (MAGs) can also be used instead of genomes, however, complete genomes are preferred. This KNN approach also supports LC-MS/MS from fractions or from different culture conditions; multiple LC-MS/MS files for the same genome were merged together into a single set of experimental MS/MS spectra.


A)
Genomics
Example: one given BGC
BGCs of Genome A
BGCs of Genome C
BGCs of Genome B
BGCs of Genome D
Max. score
Lower scores
0.25
0.4
1.0
0.8
Query BGC
B)
Untargeted MS/MS
Example: one given MS/MS
Query MS/MS
α
β
δ
γ
C)
Query BGC
Maximum BiG-SCAPE Similarity
Maximum Similarity
1.0
0.8
0.6
0.4
0.2
0
Cutoff
BGCs of Genome
A
B
C
D ...
D)
Query MS/MS
Maximum Cosine Similarity
Maximum Similarity
1.0
0.8
0.6
0.4
0.2
0
Cosα
Cosβ
Cosγ
Cosδ
Cutoff
Experimental MS/MS Spectra
a
b
c
d ...



E) Training matrix

	A	B	C	D	...	GCF
BGC1	0	1.0	0	0.8		●
BGC2	0	0.95	0	0.81		●
BGC3	0.75	0	0.8	0		▲
⋮						

F) Testing matrix

	a	b	c	d	...
MS/MS1	0	0.96	0	0.83	
MS/MS2	0	0.95	0	0.8	
MS/MS3	0.7	0	0.81	0	
⋮					

G)
All fingerprints (each shape is a BGC fingerprint and each X is a MS/MS fingerprint)
KNN Space
k-nearest neighbors = 3
X1
X2
X3
5 GCFs (classes):
●, ▲, ■, ⬟, ◆

KNN Ranking:

	X1	X2	X3
1st	●	■	◆
2nd	●	■	⬟
3rd	●	●	◆

Fig. 1. The genomics and metabolomics pipelines to use the proposed KNN approach for a hypothetical dataset with 4 paired genomes-MS/MS samples. Representation of how to calculate the similarity scores between BGCs (A) and between MS/MS spectra (B). Schematic of how to create BGCs (C) and MS/MS (D) fingerprints using a paired genomics-metabolomics dataset of four samples (genomes, metagenomes or MAGs)(samples A-D) and similarity scores from BiG-SCAPE and GNPS. The dashed red line represents the selected cutoff of 0.7. The query BGC is highly similar to a BGC in sample B (indicating as identical BGC), while it is probably absent in sample A and C. The BGC fingerprints are grouped together in a training matrix (E) and the MS/MS fingerprints compose the testing matrix (F). All fingerprints are plotted in the multi-dimensional KNN space (G, here represented in only 2D for simplification) where each shape represents a BGC fingerprint and each X represents an MS/MS fingerprint. BGCs are labeled according to one of the five GCFs (five different shapes). KNN ranking of neighbors is based in the proximity between the query MS/MS fingerprint and the neighboring BGC fingerprints. In this example, a KNN = 3 (three closest neighbors) is depicted. BGC = biosynthetic gene cluster; MS/MS = mass fragmentation spectrum; KNN = *K*-Nearest Neighbor; BiG-SCAPE =

software to calculate pairwise BGC-BGC similarity; Cosine score = modified cosine score from GNPS to calculate pairwise spectrum-spectrum similarity.

Cyanobacterial dataset: connecting a known metabolite (link validated experimentally) with a cyanobacterial BGC. Marine cyanobacteria living on coral reefs have resulted in the discovery of many novel NPs (13, 18). We collected, sequenced and binned 60 cyanobacterial MAGs, mainly from the NP rich genera of *Moorena*, *Okeania*, *Symploca*, *Leptolyngbya, Oscillatoria* and *Spirulina* (13). Strains with good quality MAGs and paired LC-MS/MS data were published at PoDP under the ID “864909ec-e716-4c5a-bfe3-ce3a169b8844.2”. We clustered 2,558 BGCs (not including the BGCs from MIBiG) and we obtained high resolution LC-MS/MS for the same set of marine cultures/environmental samples. Previous investigations (19–26) reported the discovery of 8 cyanobacterial metabolites (Fig. 2) and their BGCs from a subset of these 60 marine cyanobacteria. Hence, we used these 8 BGC-MS/MS links, with a total of 39 different MS/MS spectra, to validate our KNN algorithm for a small, uniformly built and not so sparse dataset. There are multiple spectra per compound due to different types of molecular ions (protonated, sodiated, halogenated, etc.). From this relatively small dataset, we were already able to connect one MS/MS spectrum to its correct BGC – curacin A (23), marked in red in Fig. 2 – thus providing a fairly low precision of 1/39 (2.56%). However, the BGC fingerprints had a very small number of similarity scores and it is expected that the fingerprints and the algorithm’s precision would improve with a larger dataset with more complete BGCs (many of the 60 MAGs contained several fragmented BGCs). Despite its low precision, this approach is already an improvement over an earlier attempt that used a presence/absence Mantel correlation, as that effort to connect genomes and metabolomes only yielded false positives for this same small cyanobacterial dataset (Mantel correlation generated 51 GCF-MF links, all false positives). Mantel correlation is an approach that combines two presence/absence matrices (one for genomics and one for experimental MS/MS spectra) into a single output, creating a pairwise association between a given row of the genomics matrix with a second row from the metabolomics matrix. The Mantel correlation code is available in a Jupyter notebook found at the GitHub repository: https://github.com/tiagolbiotech/NPOmix.

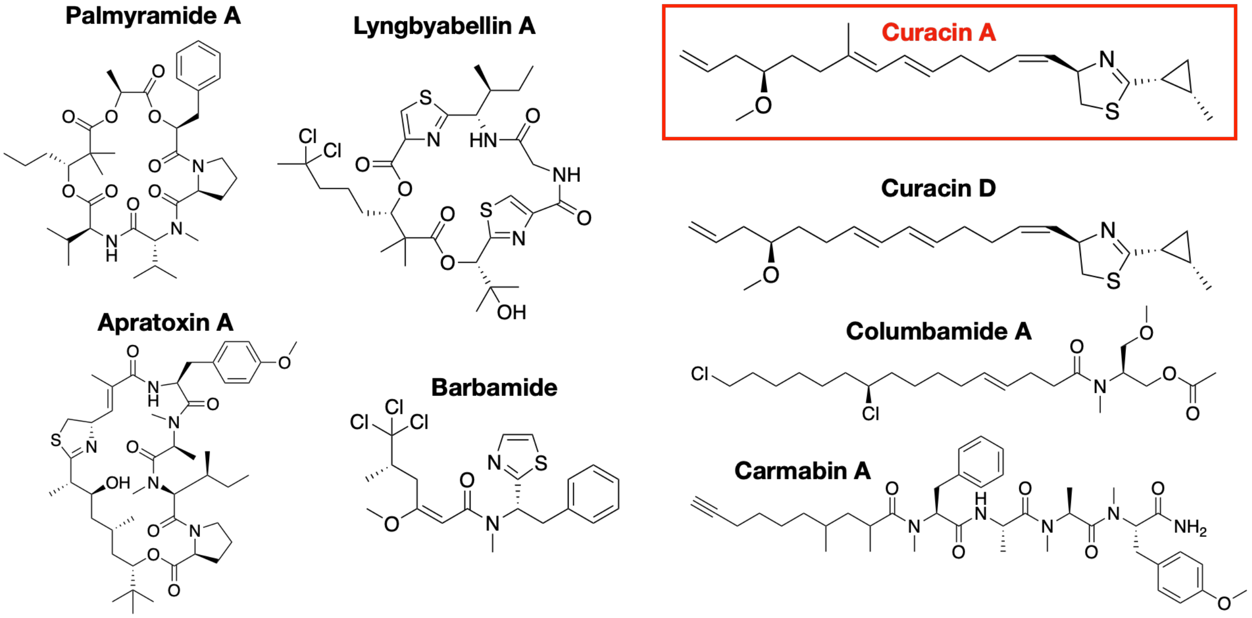


Fig. 2. Structures of compounds used for validating links between BGC and MS/MS spectra for the 60 cyanobacterial samples. Highlighted in red is curacin A, the one correct link that was predicted via this KNN approach.

PoDP dataset: connecting known metabolites (links validated experimentally) to PoDP BGCs. To further validate our NPOmix approach, we used 36 out of 71 datasets from the PoDP (from February 2021, listed at Dataset S1, sheet one). We selected genomic samples that contained a valid Genome ID or BioSample ID to aid their downloading from the NCBI database and totaling 732 genomes/MAGs obtained from these 36 metadatasets. We also selected and assembled 1,034 metagenomes from two major metagenomic datasets: 1) MSV000082969 and PoDP ID cd327ceb-f92b-4cd3-a545-39d29c602b6b.1 - 556 cheetah fecal samples and environmental samples; 2) MSV000080179 and PoDP ID 50f9540c-9c9c-44e6-956c-87eabc960d7b.3 - The American Gut Project (27) that contains fecal samples from 481 human subjects. These (meta)genomes were automatically downloaded with the code shared at the GitHub repository https://github.com/tiagolbiotech/NPOmix, notebook 1. The LC-MS/MS files can be downloaded using "ftp" from links found at Dataset 1, sheet two. We were able to cluster 1,040 (meta)genomes that contained 5,681 BGCs (including 260 BGCs from the MIBiG database) distributed into 997 GCFs. In the untargeted metabolomics data, we matched 3,248 LC-MS/MS files to 15 GNPS (16, 17) reference library spectra in order to create the MS/MS fingerprints for testing the KNN classification (one fingerprint per spectra). In the near future, we envision creating a balanced, diverse and less sparse training dataset. To maximize precision rates in the future, we plan to purchase cultures from collections that have well assembled genomes so we can obtain the paired LC-MS/MS. However, the current dataset produced highly supportive results by testing validated links from the PoDP, links generated by the Gerwick lab dataset, and validated links used in the NPLinker publication (10). We attempted to test all 242 metabolite-BGC links from NPLinker (totaling 2,069 unique MS/MS spectra, Dataset S1, sheet four), 109 manually added MS/MS spectra (connected to BGCs, annotated by experts at the PoDP, Dataset S1, sheet three) and 406 MS/MS spectra from metabolites isolated by the Gerwick lab. Although, most of these validated links were not present in the 1,040 paired (meta)genomes-MS/MS samples from the PoDP (as NPLinker used BGCs from MIBiG and not PoDP) or their BGC scores did not co-occur with their MS/MS scores because they were not present in the same sample. Hence, our validation dataset was limited to 8 validated links found in the paired (meta)genomes-MS/MS samples (orfamides, albicidins, bafilomycin, nevaltophin D, jamaicamide, hectochlorin, palmyramide and cryptomaldamide, totaling 15 reference MS/MS spectra that were present in the GNPS database). We stress that a larger training dataset with more complete genomes is likely to increase the size of the validation set by adding more valid BGCs into the analysis. We also combined the NPOmix program with *in silico* tools like Dereplicator+ (28) to make new links between MS/MS spectra, BGCs and molecular structures. This was accomplished by annotating cryptic MS/MS spectra (without a GNPS library hit and therefore not present in either the GNPS or the PoDP databases) to known BGCs. Such new links could be confirmed experimentally to improve the size of the validation set, as well as to expand MS/MS databases by adding these cryptic spectra to them.

A two-dimensional comparison of both types of fingerprints (BGC and MS/MS) can be a proxy for distinguishing some true positives from false positives. As observed in Fig. S2, we can visualize a mismatch between the BGC fingerprints (one GCF) and the MS/MS fingerprint in the "reduced" KNN-space (represented schematically in only two dimensions), indicative of a possible false positive link. This GCF is dereplicated as the known metabolite, pyocyanin, and it was incorrectly associated with the metabolite 2,4-diacetylphloroglucinol, confirming the false positive (at *k* = 3). In contrast, Fig. 3 illustrates that 5 metabolites, 2 albicidins and 3 albicidin analogs, could be correctly assigned to their corresponding GCF that contains 2 BGCs. In this case, the BGC fingerprints match the MS/MS fingerprints (Fig. 3C, 3D). Using this second larger dataset comprised of 1,040 samples instead of only 60 yielded a precision of 66.7% as 10 out of 15 reference MS/MS spectra were correctly labeled when top-*n* = 3 (*k* also equal to 3). Top-*n* represents how often the correct GCF label was found among the top *n* labels classified by the KNN approach (see Tables 1 and 2). The observed precision was much higher than with the cyanobacterial dataset because the PoDP dataset has a larger number of samples and it also contains a larger diversity of microbial entries thus providing fingerprint-based approaches more resolution. Lastly, we regard our NPOmix approach as multi-omics enabled dereplication because the 5 MS/MS albicidin labels were automatically assigned to a known GCF that confirmed their metabolite labels, thereby minimizing the necessity to purchase standards, to perform isolation and NMR characterization, gene knockout or heterologous expression.

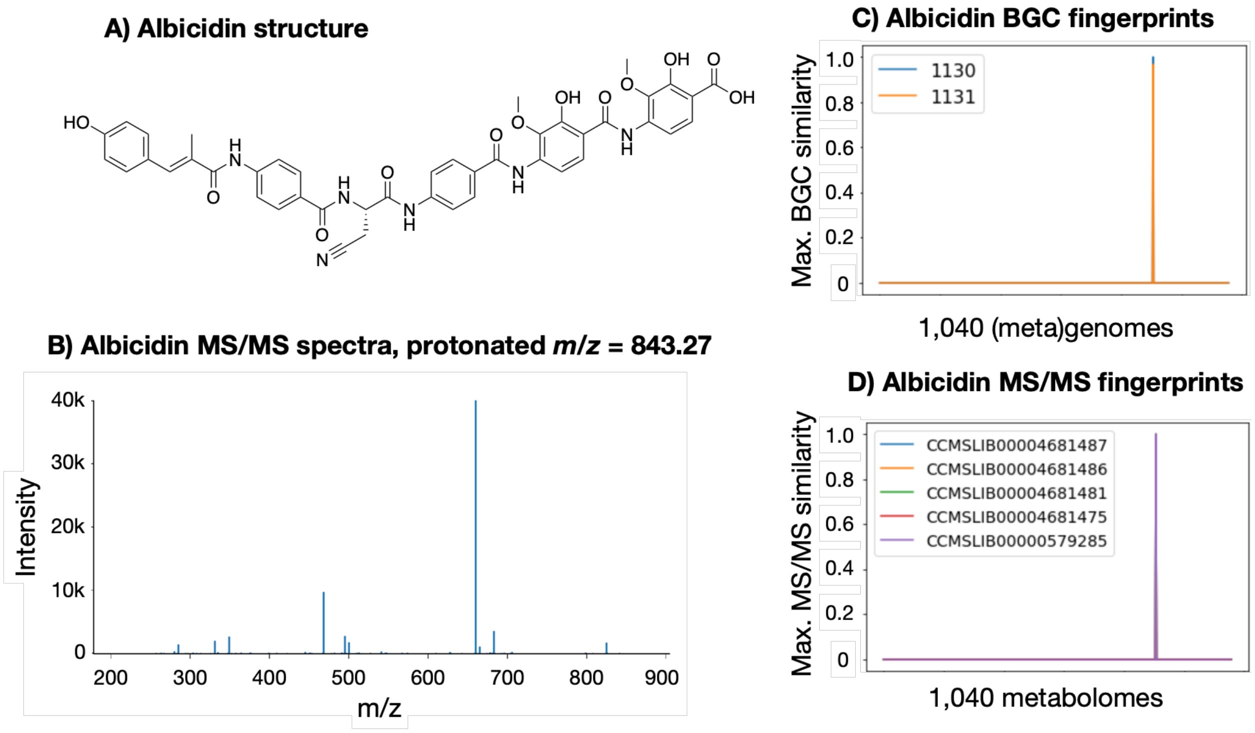


Fig. 3. Multi-omics enabled dereplication of albicidin by automatically predicting a true BGC-metabolite link. Structure of the dereplicated metabolite (A) and its corresponding representative MS/MS spectrum (B, spectrum example from GNPS ID CCMSLIB00000579285 and *m/z* of 843.27), obtained via Metabolite Spectrum Resolver (29). The two BGC fingerprints (1130 and 1131) are represented in a 2D plot (C) and they match the 2D plot for the 5 MS/MS fingerprints obtained from GNPS for albicidin and its analogs (D). BGC = biosynthetic gene cluster; MS/MS = mass fragmentation spectrum; *m/z* = mass over charge calculated via mass spectrometry.

Connecting analogs to BGCs: the example of orfamide C. An NPOmix link can be further confirmed by matching the AA predictions from the BGC with the structure prediction for the query metabolite based on library match or *in silico* annotations (Fig. 4). For example, the BGC (genes 1-6 in Fig. 4) for the metabolite orfamide C (MIBiG ID BGC0000399) was automatically connected by our KNN approach to a GNPS metabolite labeled "putative orfamide C" (CCMSLIB00004679300). This MS/MS spectrum was obtained from the same strain where the BGC was first identified (*Pseudomonas protegens* Pf-5, Genbank ID GCA_000012265)(30). The nine amino acid (AA) predictions for this BGC, based on the specificity of adenylation domains, match the structure for orfamide C in the correct order: leu, asp, thr, ile, leu, ser, leu, leu and ser. AntiSMASH was not able to predict the tenth and last in the biosynthetic series, namely valine. The matching between the predicted structures confirmed the multi-omics enabled dereplication of orfamide C (using k = 3, BGC predictions and predicted metabolite structure are represented in Fig. 4). The KNN GCF predictions do not use structures/substructures for linking MS/MS spectra to BGCs; hence, as demonstrated in Fig. 4, these substructure predictions can be an extra dimension for selecting links that are true positives over false positives.

We have determined that the use of three neighbors is the optimal performance, providing a good balance between precision and number of links to validate (top-3 = 66.7% and randomness equal to 0, as detailed in Table 1). Randomness is observed by shuffling the testing columns, experimental MS/MS names, and counting how many correct links are present between the top-*n* GCF candidates. This parameter (*n* and k = 3) enabled the dereplication of the albidicins, orfamides B-C, jamaicamides A and C and cryptomaldamide, totaling 4 different metabolite families (and analogs) that were correctly predicted by our KNN approach using the PoDP dataset. Noteworthy, the top-10 precision had a maximum score of 73.33% with randomness still equal to 0. However, 10 GCF candidates is practically too large for useful genome mining as all those candidates would need to be tested experimentally. We expect that our approach will improve with a larger training set and with further improvement of the features in the BGC and MS/MS fingerprints (e.g., based on substructure presence/absence). The 15 BGC-MS/MS validated links reported herein and their predictions using k = 3 are found in Table 2 that provides the GCF labels for the three closest BGCs to a given MS/MS fingerprint (the 10 correct GCF predictions are colored red and highlighted in bold). We confirm that all 10 correct GCF predictions reported here were found in the original producer of the identified metabolites and they matched the reported masses. With 49 known GCF-MS/MS links were present in the 1,040 samples with paired data, the annotation rate was reasonably high (around 30%, 15 out of 49 links were retained after the co-occurrence filter, a filter to keep only the metabolites that are found among the same samples that contain the candidate BGCs).

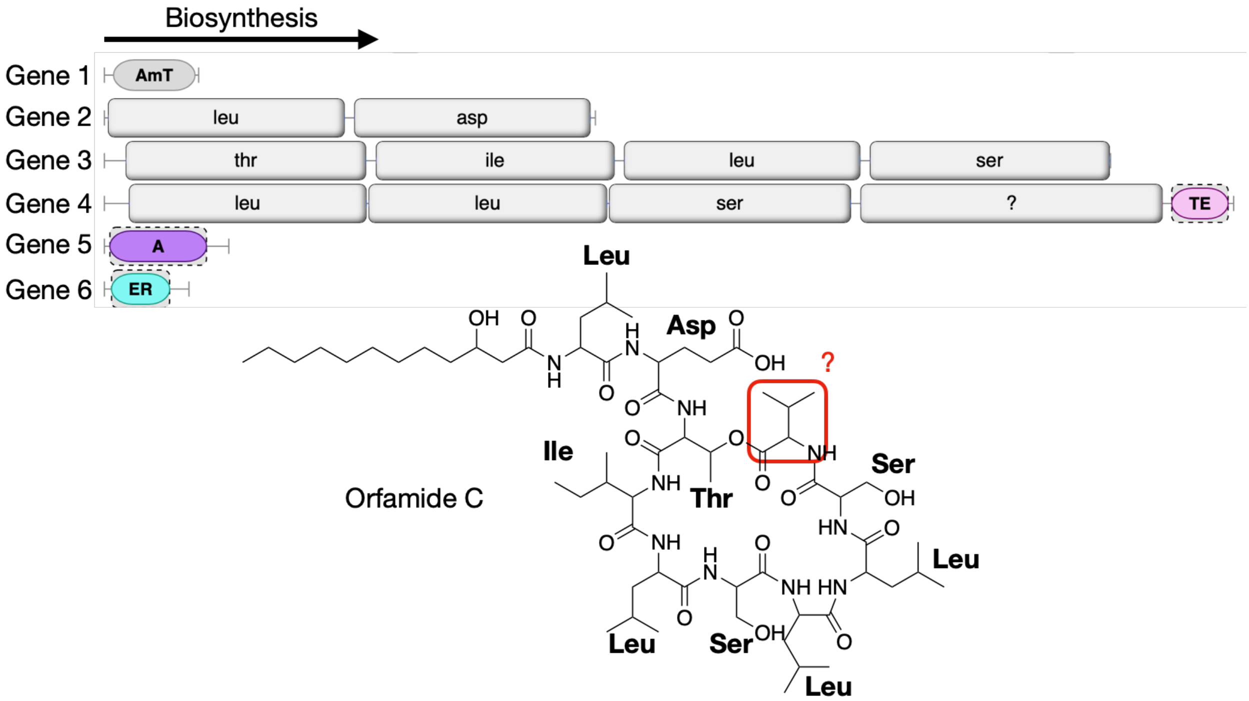


Fig. 4. NPOmix automatically connected an MS/MS spectrum annotated as "putative orfamide C" to the MIBiG BGC annotated as orfamide C. The figure illustrates the matches between the BGC's AA predictions (via antiSMASH) and the predicted metabolite structure (orfamide C, predicted via MS/MS spectral matching). Only one AA (valine, in red) out of 10 AA could not be predicted by the BGC annotation tool (antiSMASH), however, this valine residue was predicted by the MS/MS spectrum. BGC = biosynthetic gene cluster; AA = amino acid; AmT = aminotransferase; TE = thioesterase; A = adenylation domain; ER = enol reductase; "?" in the BGC represents that one AA could not be predicted by antiSMASH.

Table 1. Top-*n* precision scores (how often the correct GCF label was found among the top *n* labels classified by the KNN approach) for 15 reference GNPS MS/MS spectra connected to a BGC found in the paired 1,040 (meta)genomes-MS/MS downloaded from the PoDP. These links were obtained from the NPLinker dataset, GNPS and PoDP databases. Randomness is observed by shuffling the testing columns, experimental MS/MS names, and counting how many correct links are present between the top-*n* GCF candidates. Based on this, we believe the best performance is *n* = 3 for the examined dataset.

	Top-1	Top-3	Top-5	Top-10	Top-50	Top-100
Data	46.66%	66.66%	66.66%	73.33%	73.33%	73.33%
Random	0%	0%	0%	0%	0%	20%

Table 2. 15 links between GNPS MS/MS spectra (with CCMS metabolite ID) and networked gene cluster family (true GCF). The table also includes their KNN predictions (*k* = 3); the predicted GCFs are ordered according to the value for *k*, from 1 (nearest) to 3 (furthest), and the first correct family is marked in bold red font. GCF labels can be repeated because multiple BGCs from the same GCF can be predicted as the nearest neighbors. Classification is considered correct if the true GCF is among the top-3 candidates. Annotations are according to each MIBiG BGC(s) found in the true GCFs. The "orphan" label indicates that the BGC was not networked in the current dataset.

CCMS metabolite ID	True GCF	Predicted GCFs for *k* = 3	Annotation
CCMSLIB00000479759	GCF320	GCF122, GCF115, GCF112	Bafilomycin
CCMSLIB00000579285	GCF476	**GCF476**, GCF180, GCF476	Albicidin
CCMSLIB00000840594	GCF488	GCF740, GCF740, GCF739	Nevaltophin D
CCMSLIB00004679298	GCF450	GCF465, GCF445, GCF439	Orfamide A
CCMSLIB00004679299	GCF450	GCF465, GCF445, **GCF450**	Orfamide B
CCMSLIB00004679300	GCF450	GCF465, GCF445, **GCF450**	Orfamide C
CCMSLIB00004681475	GCF476	**GCF476**, GCF180, GCF476	Propionyl-albicidin
CCMSLIB00004681481	GCF476	**GCF476**, GCF180, GCF476	Beta-methoxy-albicidin
CCMSLIB00004681486	GCF476	**GCF476**, GCF180, GCF476	Carbamoyl-beta-methoxy-albicidin
CCMSLIB00004681487	GCF476	**GCF476**, GCF180, GCF476	Albicidin
CCMSLIB00000001706	GCF471	**GCF471**, GCF498, GCF471	Jamaicamide A
CCMSLIB00005724004	GCF498	GCF471, **GCF498**, GCF471	Cryptomaldamide
CCMSLIB00000001553	Orphan	GCF471, GCF498, GCF471	Hectochlorin
CCMSLIB00000001751	Orphan	GCF471, GCF498, GCF471	Palmyramide A
CCMSLIB00000001708	GCF471	**GCF471**, GCF498, GCF471	Jamaicamide C

Connecting cryptic metabolites (without GNPS library matches) to BGCs: the example of brasilicardin A. We used a combination of MS/MS fingerprints (notebook 2), BGC fingerprints (notebook 3), MZmine (31) and Dereplicator+ (28) in order to annotate brasilicardin A. This approach differs from the previous NPOmix analysis because it uses MZmine to select the MS/MS spectra instead of collecting spectra from the GNPS and PoDP databases. After selecting 300 MS/MS spectra from the 16 most diverse genomes in the dataset with 1,040 samples, Dereplicator+ had three *in silico* predictions and one of them was the unique tricyclic glycosylated terpene brasilicardin A. The observed *m/z* matches the value previously reported in the literature)(32), identifying an MS/MS spectrum that is currently absent from both the GNPS and the PoDP databases. NPOmix connected the MS/MS spectrum (predicted to be brasilicardin A by Dereplicator+, information not used in the NPOmix training) with the correct BGC (brasilicardin A MIBiG ID BGC0000632 from the strain *Nocardia terpenica* IFM 0406, GenBank ID GCA_001625105)(33), highlighting how NPOmix can connect cryptic molecules without library matches (absent from MS/MS databases) to their corresponding BGCs. Predicted fragmentation (Fig. S3 and table with deltas in Dataset S1, sheet seven) strongly suggests that the query MS/MS spectrum is indeed brasilicardin A (all differences between exact *m/z* and observed *m/z* were extremely low). This pipeline provided additional 70 links between cryptic MS/MS spectra and BGCs from the most diverse strains (links listed at Dataset S1, sheet six) and potentially new BGCs can be explored experimentally (e.g., BGC knock-out, heterologous expression or isolation and NMR structure elucidation), especially if coupled to NMR SMART analysis (34, 35) to confirm their novelty.

Improving the fingerprint for known metabolites using biosynthetic class. In order to increase the precision of our NPOmix algorithm, we added the biosynthetic classes (PKSs, NRPSs, terpenes, siderophores, RiPPs, phosphonates, oligosaccharides, phenolic metabolites, others/unknowns and other minor classes) to the BGC and MS/MS fingerprints as presence/absence in the training set (5,681 BGCs). For example, if a given BGC is a hybrid PKS-NRPS, it was annotated as 1 in the PKS and NRPS columns, and with a 0 in the remaining classes (additional columns). For the MS/MS fingerprints in the validation set (testing set), we manually annotated these same features (biosynthetic classes) because the structures for these testing MS/MS spectra were known. In cases where the structure is unknown, tools like CANOPUS (36) and MolNetEnhancer (37) can provide a similar biosynthetic class prediction, and these predictions can be further confirmed using substructures predicted with unsupervised tools like MS2LDA (38) or dedicated tools like MassQL (based on specific MS/MS fragments found in the spectra, manuscript in preparation) or CSI:FingerID via SIRIUS 4 (39). As observed in the precision curves from Fig. S4 for version 1.0 (fingerprints without biosynthetic classes) and version 2.0 (fingerprints with biosynthetic classes), the precision increased for top-3 and top-5 testing results, for top-3 it increased from 66.66% without the biosynthetic class (good score with a lower number of GCF candidates than top-10) to 73.33% with the biosynthetic class added, requiring less GCF candidates to obtain a similar precision as the top-10 without inclusion of the biosynthetic class. Consequently, we observed a better ranking of the predicted GCFs when the new class features were added.

Conclusion

We created a machine learning solution, a K-Nearest Neighbors algorithm named NPOmix, to connect specialized metabolites observed by untargeted mass spectrometry to their biosynthetic gene clusters (BGCs). We demonstrated that the tool performs reasonably well for a small dataset that was sequenced and collected in a uniform fashion; in this case, the dataset was constructed from 60 marine cyanobacterial samples with MAGs and high resolution untargeted LC-MS/MS spectra. These were mostly from tropical marine cyanobacteria, which are known to be rich producers of NPs. Nevertheless, performance was limited by the small size of the dataset of good cyanobacterial genomes. We showed that a larger dataset, deriving from heterogeneous sources such as the ones currently available in the Paired omics Data Platform (PoDP), can create better fingerprints and can thus more successfully connect known metabolites to their corresponding BGCs, such as albicidin and its analogs to a BGC in *Xanthomonas albilineans* GPE PC73 (GenBank ID GCA_000087965.1), orfamides A-C to a BGC in *Pseudomonas protegens* Pf-5 (GCA_000012265), and cryptomaldamide and jamaicamide A and C to BGCs in *Moorena producens* JHB (GCA_001854205). All three of these strains were the original producers of these metabolites. In Fig. 4, we illustrated how the BGC predictions (such as predicted moieties) can help to prioritize true links over false positives via matching of predicted structures between a given MS/MS spectrum and its BGC candidates.

In this work we demonstrated the use of machine learning and genome mining to process several thousand LC-MS/MS files and a thousand genomes to connect MS/MS spectra to GCFs. Our approach can systematically connect MS/MS spectra from known metabolites (links validated experimentally), spectra from metabolites analogous to known (links with GNPS library matches) and spectra from cryptic metabolites (links without GNPS library matches and therefore absent from the MS/MS database, as exemplified by brasilicardin A). The advantage of using paired data is that the genomic information represents the full metabolic potential of an organism, and hence, we can prioritize the discovery of the most diverse BGCs via genome mining. Additionally, the use of genetic information can help in the structure elucidation and prediction of bioactivity (40), highlighting the advantage of using the BGC information in the drug discovery process. Moreover, predicting linked MS/MS spectra for a promising BGC can facilitate their heterologous expression as expression can be difficult if the target molecule is not known. Furthermore, we show how cryptic MS/MS spectra (absent from MS/MS databases like GNPS) can be annotated using NPOmix, MZmine (31) and Dereplicator+ (28), allowing expansion of the current MS/MS databases. We also demonstrated how our methodology is suitable for linking cryptic MS/MS spectra with putative BGC candidates that can assist in the isolation of novel natural product scaffolds. Despite the relatively small size of the training dataset (in comparison to other machine learning approaches, 1,040 paired samples and 5,681 BGCs from the PoDP database), we observed good precision scores of top-3 = 66.66% and top-10 = 73.33% (both with randomness equal to 0). By including the biosynthetic class in the fingerprints, the best precision score was top-3 = 73.33%. In effect, this latter analysis required less GCF candidates to obtain a similar precision as the top-10 without inclusion of the biosynthetic class. We observed an annotation rate of around 30%, as 15 out of 49 GCF-MS/MS validated links were retained after the co-occurrence filter.

The use of complete genomes over MAGs and metagenomes is preferred to create a more “complete” training set; we predict that this would result in better precision than if the training set is populated with several fragmented BGCs. Our results highlight the importance of making genomics and metabolomics data publicly available with curated metadata, because more available paired data would enable better training of models, and therefore, better tools for the research community. Future plans include the testing of other similarity metrics for networking and fingerprinting such as BiG-SLICE (41) for genomics and Spec2Vec (42) and MS2DeepScore (43) for the metabolomics. We will also look for synergy with correlation scores from NPLinker to better annotate paired datasets. We intend to implement structure and substructure predictions from the MS/MS fragmentation spectra using tools like SIRIUS 4 (39), MS2LDA (44), MolNetEnhancer (37) or CANOPUS (36), prioritizing candidates that have several substructures or predicted chemical compound classes matching between BGCs and MS/MS spectra. The GNPS molecular family information could be used to select a consensus prediction among different MS/MS spectra from the same family. The BGCs assembled from the metagenomic samples could be improved using tools like metaBGC (45) and BiG-Mex (46). Enrichment of the current Paired Omics Data Platform dataset (we could now use 1,040 PoDP samples) with higher quality samples as well as more validated BGC-MS/MS links will further drive the development of tools such as NPOmix, and this will spark the discovery of more novel NPs. Furthermore, machine learning can be used to connect promising BGCs with their biological activities (anticancer, antimicrobial and antifungal)(40). Finally, we would like to stress that all true positive BGC-MS/MS validated links reported here were found in the original producer of the metabolites and they matched the reported masses. We expect that NPOmix is a promising tool to search for new natural products in paired omics data of natural extracts by using links between cryptic MS/MS and putative BGCs. This will, for example, facilitate the use of genome mining in drug discovery pipelines.

Code and Data Availability

The code (a collection of Jupyter notebooks) required to reproduce this work and to use the NPOmix tool for new samples can be found in the following GitHub repository page: https://github.com/tiagolbiotech/NPOmix. The repository also includes short video explanations on how the tool works and its importance for natural product discovery. The (meta)genomes used to create the NPOmix training dataset for validation were downloaded from the Paired omics Data Platform (PoDP)(14) using notebook 1 from the GitHub repository. The paired experimental MS/MS files were downloaded using the ftp links (also from the Paired omics Data Platform) found in Dataset S1, sheet two. The testing set included MS/MS spectra from PoDP, spectra from the Global Natural Products Social Molecular Networking database (GNPS)(16) and also spectra used in the NPLinker dataset (10). If the potential users find the tool challenging to run, we have our contact information at the GitHub web page (link above) to submit samples and we expect that promising results will lead to fruitful collaborations. In the near future, we will have a web-based interface for direct submission of samples.

Author Contributions

T.F.L. conceptualized the software; T.F.L., R.d.S. and A.B (Asker Brejnrod) programmed the software; M.W. assembled the metagenomic reads and annotated all biosynthetic gene clusters; E.G. cultured cyanobacterial samples and collected the cyanobacterial LCMS data; A.B. (Anelize Bauermeister) developed the predicted fragmentation for brasilicardin A; T.F.L, J.J.J.v.d.H. and M.W. curated the dataset; T.F.L, J.J.J.v.d.H. and A.B. (Anelize Bauermeister) wrote the manuscript; L.G., W.H.G, N.B. and P.C.D. funded and designed the research; L.G., W.H.G, N.B. and P.C.D. edited the manuscript; all authors read, reviewed and agreed to the published version of the manuscript.

Acknowledgements

This research was supported by National Institutes of Health (NIH) Grants GM107550 (to P.D., W.H.G. and L.G.) and GM118815 (to W.H.G. and L.G.). This work was supported in part by a seed grant from the Center for Microbiome Innovation at University of California San Diego. R.S. was supported by the São Paulo Research Foundation (Awards FAPESP 2017/18922– 2 and 2019/05026–4). J. J. J. v. d. H. was supported by the Netherlands eScience Center (ASDI eScience grant ASDI.2017.030).

Conflict of Interest

W.H.G. has an equity interest in NMRFinder and in SirenasMD Inc., companies that may potentially benefit from the research results and W.H.G. also serves on the companies' Scientific Advisory Boards. 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. P.C.D. is a scientific advisor to SirenasMD Inc., Galileo and Cybele, and cofounder and scientific advisor to Ometa and Enveda with approval by the University of California San Diego. M.W. is a cofounder of Ometa Labs, LLC.

Methods

Obtaining paired data. Sixty cyanobacterial samples were collected via SCUBA diving or snorkeling along coastal shores around the globe and subjected to processing as described by Leao *et al.*, 2021, (13). High quality genomes were published at NCBI database and LC-MS/MS data were collected for the same set of samples, also as described by Leao *et al.*, (2021)(13). The paired data is available at the PoDP (ID “864909ec-e716-4c5a-bfe3-ce3a169b8844.2”). We automatically downloaded the paired (meta)genomics-metabolomics data from the samples in the PoDP according to the code in the notebook 1 at the GitHub repository described below. The cyanobacterial high resolution LC-MS/MS data was obtained according to the methods in by Luzzatto-Knaan *et al.* (47).

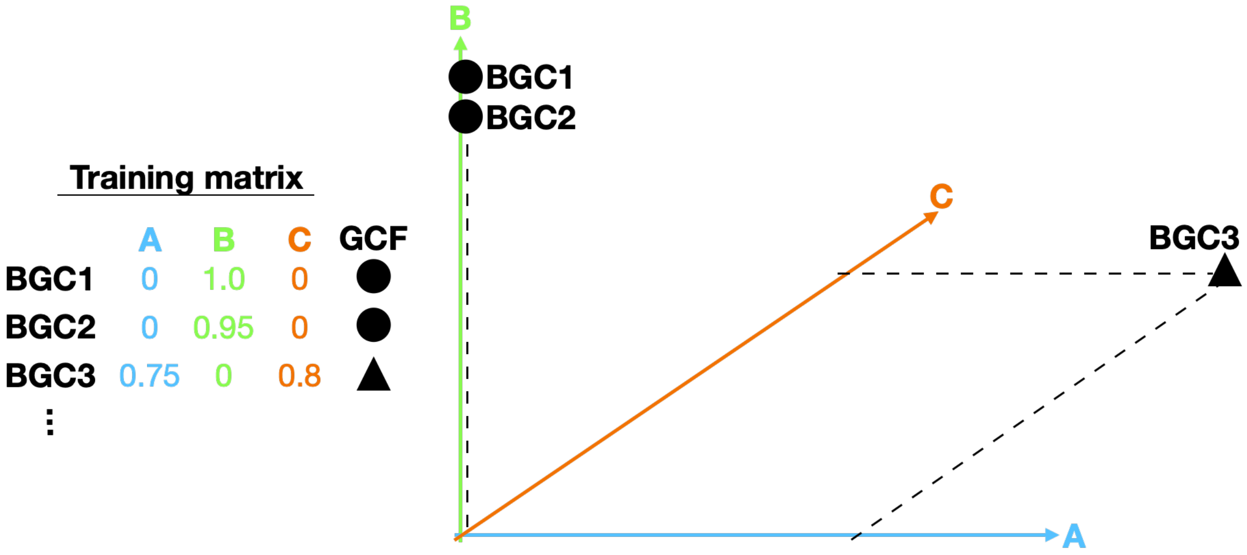
Genome assembly and annotation, BGC and MS/MS similarity calculation. Metagenomic reads were assembled with SPAdes 3.15.2. (48). For BGC annotation, we used antiSMASH 5.0 (49) and for gene cluster networking we used BiG-SCAPE 1.0 (similarity cutoff of 0.7) (5). BiG-SCAPE raw distance is measured via the domain sequence similarity (DSS) index, an index that calculates the Pfam domain copy number differences and sequence identity (5). For networking metabolites, we used GNPS classical molecular networking release 27 (similarity cutoff of 0.7). We did not use the full classical molecular networking capabilities in the NPOmix approach, as only the functions required to calculate a modified cosine score between a pair of MS/MS spectra were needed.

Creating fingerprints. We developed python scripts and we combined with scripts from sklearn (https://scikit-learn.org/stable/index.html) to create both BGC and MS/MS fingerprints and to run the KNN algorithm. A BGC fingerprint is created by pairwise BiG-SCAPE comparison between the queried BGC and all the BGCs found in the (meta)genomes in the training set, selecting the highest similarity scores for each (meta)genomes. An MS/MS fingerprint (part of the testing set) is created by pairwise modified cosine comparison between the queried MS/MS and all the MS/MS present in the LC-MS/MS files paired with the genomes from the training set, also selecting only the highest similarity scores per set of experimental MS/MS spectra.

Jupyter notebooks. All scripts used in this research can be found at this GitHub repository: https://github.com/tiagolbiotech/NPOmix. Notebook 1 can be used to download (meta)genomes and metagenome-assembled genomes (MAGs) that contain paired untargeted metabolomics (LC-MS/MS)(metabolomic files will also be downloaded by the notebook). We selected genomic samples that contained a valid Genome ID or BioSample ID, resulting in 732 genomes/MAGs. We also selected and assembled 1,034 metagenomes. Notebook 2 can be used to process downloaded metabolomics files and a selected set of “.mgf” reference MS/MS spectra, creating a matrix containing the MS/MS fingerprints for the selected set of reference spectra (reference MS/MS spectra for the validation but for using the tool these reference spectra will be replaced by cryptic MS/MS spectra). If there are more than one LC-MS/MS file per genome (for example different media conditions or different chemical fractions), these files were merged into a single file representing these experimental MS/MS spectra. Notebook 3 can be used to process the antiSMASH results to create BGC fingerprints and use those to train the KNN algorithm. The MS/MS fingerprints are used to predict a/multiple GCF(s) for each tested reference MS/MS spectra found in the paired genomes-MS/MS data. We filtered the GCF-MS/MS links for cases that the top GCF candidate had co-occurrence (GCF and MS/MS scores were present in the same set of samples, as illustrated in Fig. 3C and 3D). Notebook 3 also performs cross-validation (dividing the data into 5 parts) and the average precision score for the cross-validation was 56.9%. Notebook 4 can be used to generate metadata such as the type of GCF or the count of BGCs per each genus in the database. The code for making the Mantel correlation, an approach that combines two presence/absence matrices, can be found in notebook 5. Notebook 6 presents the code for genome mining that yielded the annotation of brasilicardin A (more details below). Notebook 7 expanded the similarity/absence fingerprints by including the biosynthetic class (NPOmix version 2.0).

Genome mining for new MS/MS spectra using Dereplicator+ and NPOmix. In order to use the NPOmix approach to find new NPs without any GNPS library matches (absent from the MS/MS database), we developed a pipeline combining NPOmix, MZmine (31) and Dereplicator+ (28). First, a number of strains were selected using MZmine, here exemplified with 16 strains, based on their BGC beta-diversity scores. The Jaccard beta-diversity score metric of the similarity between a pair of strains was calculated as the intersection over the union of the detected gene cluster families. Using MZmine, we select peaks that were above a certain intensity threshold (we used base peak relative abundance of 1E6) in order to prioritize the chromatographic peaks that could reasonably be isolated for structure elucidation. In this example, we detected approximately 3,800 peaks with MS/MS spectra found in the analysis of the 16 most diverse strains. This MZmine list of peaks that have associated MS/MS data was filtered for minimum precursor mass of *m/z* 500 to promote the presence of multiple moieties (substructures) in the predicted structures, generating 300 ".mgf" files. These mgf files were used by NPOmix to predict the GCFs/BGCs for each of the 300 MS/MS spectra. We filtered for BGC-MS/MS links that the query MS/MS spectra existed in the same strains that the query BGCs were found (e.g., Fig 3C-D) and not across different strains (e.g., Fig. S2), using the Jaccard index in the presence/absence of fingerprints, essentially a pairwise analysis between the BGC fingerprint and the MS/MS fingerprint. This second filter narrowed down the number of mgf files to 72, as listed in Dataset S1, sheet six. These 72 mgf files were processed by Dereplicator+ for predicting structures for each MS/MS spectrum, leading to the annotation of brasilicardin A. Two other Dereplicator+ hits did not match the predicted GCFs. MZmine parameters were as follows: noise level of 1E6 for MS1 and 1E3 for MS/MS, minimum group size in number of scans of 4, group intensity threshold of 1E6, minimum highest intensity of 3E6, *m/z* tolerance of 10 ppm, retention time tolerance of 0.2, weight for *m/z* of 75%, and weight for retention time of 25%.

Expanding BGC and MS/MS fingerprints using biosynthetic classes. In notebook 7, the BGC classes were annotated and included in the BGC fingerprints. To accomplish this, all of the antiSMASH annotations for a given BGC were added to the presence of all predicted classes. Each class represented a new column in the fingerprints and the columns were filled with 1 (if the class was present) and 0 (if the class was absent). We observed the following classes in our dataset: PKSs, NRPSs, terpenes, siderophores, RiPPs, phosphonates, oligosaccharides, phenolic metabolites, others/unknowns and other minor classes. In the MS/MS fingerprint, for each one of the 15 validated MS/MS spectra, we annotated the presence/absence of the biosynthetic classes based on the known structures. These new fingerprints were used in the machine learning process, analogously to the notebook 3.



Training matrix

	A	B	C	GCF
BGC1	0	1.0	0	●
BGC2	0	0.95	0	●
BGC3	0.75	0	0.8	▲
⋮				

Fig. S1. Representation of how BGCs can be plotted in the KNN space by using the values in the training matrix, each column represents a genome in the training set and it also represents a dimension in the KNN space (1,040 genomes distributed in 1,040 columns). This example has three dimensions because it uses only three genomes; the actual training matrix used in this study had 1,040 genomes and therefore 1,040 dimensions.

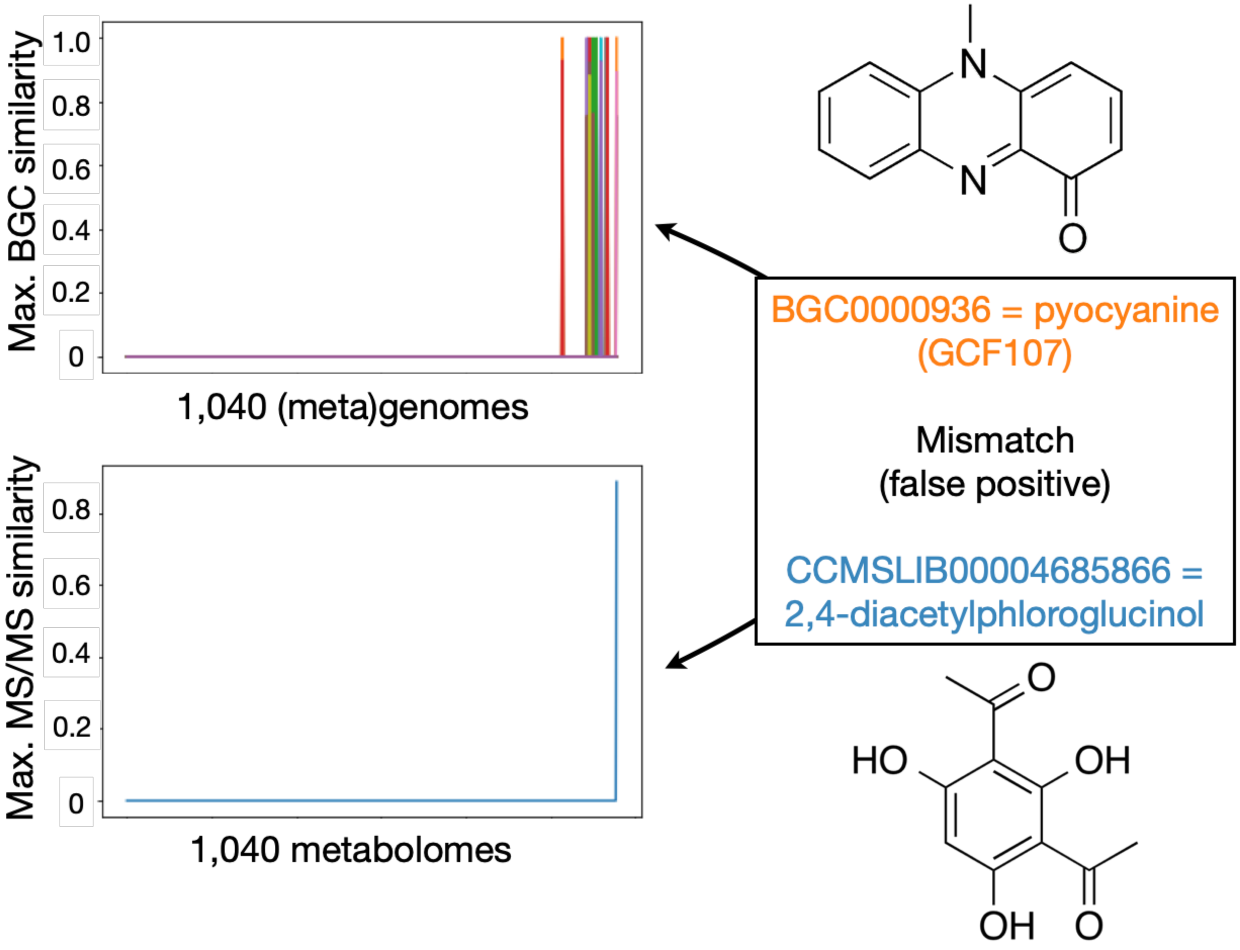
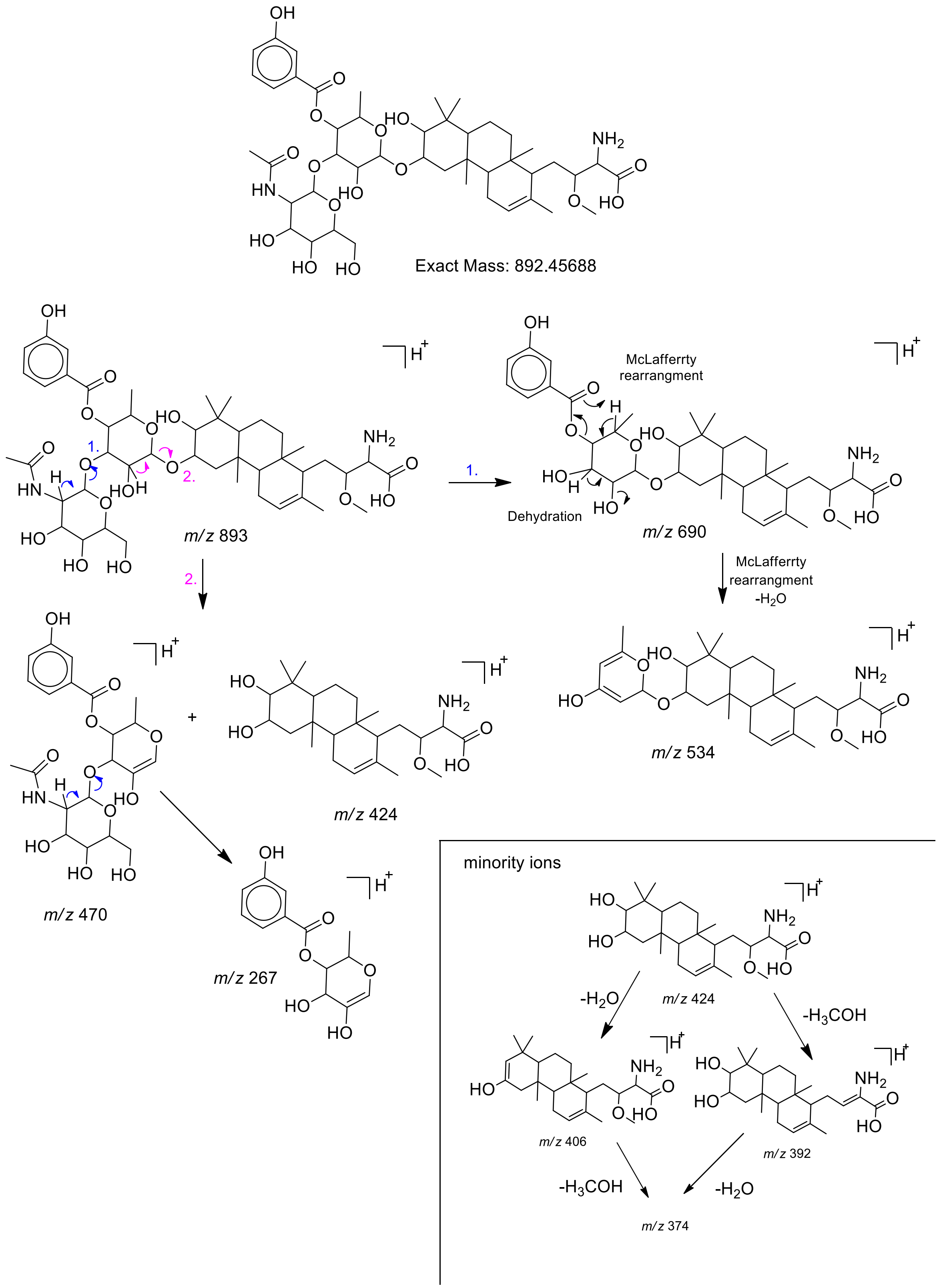


Fig. S2. Representation of a mismatch linked by the KNN algorithm using *k* = 3. It is visually clear that the closest neighboring BGC fingerprints for pyocyanine does not properly match the MS/MS fingerprint from the metabolite 2,4- diacetylphloroglucinol, indicating that NPOmix suggested the wrong GCF for the 2,4- diacetylphloroglucinol MS/MS spectrum.


Exact Mass: 892.45688
H+
1.
2.
m/z 893
McLafferrty rearrangment
Dehydration
m/z 690
McLafferrty rearrangment -H2O
m/z 534
+
m/z 470
m/z 424
m/z 267
minority ions
m/z 424
-H2O
-H3COH
m/z 406
m/z 392
-H3COH
-H2O
m/z 374
OH
O
HO
NH2
HN
H

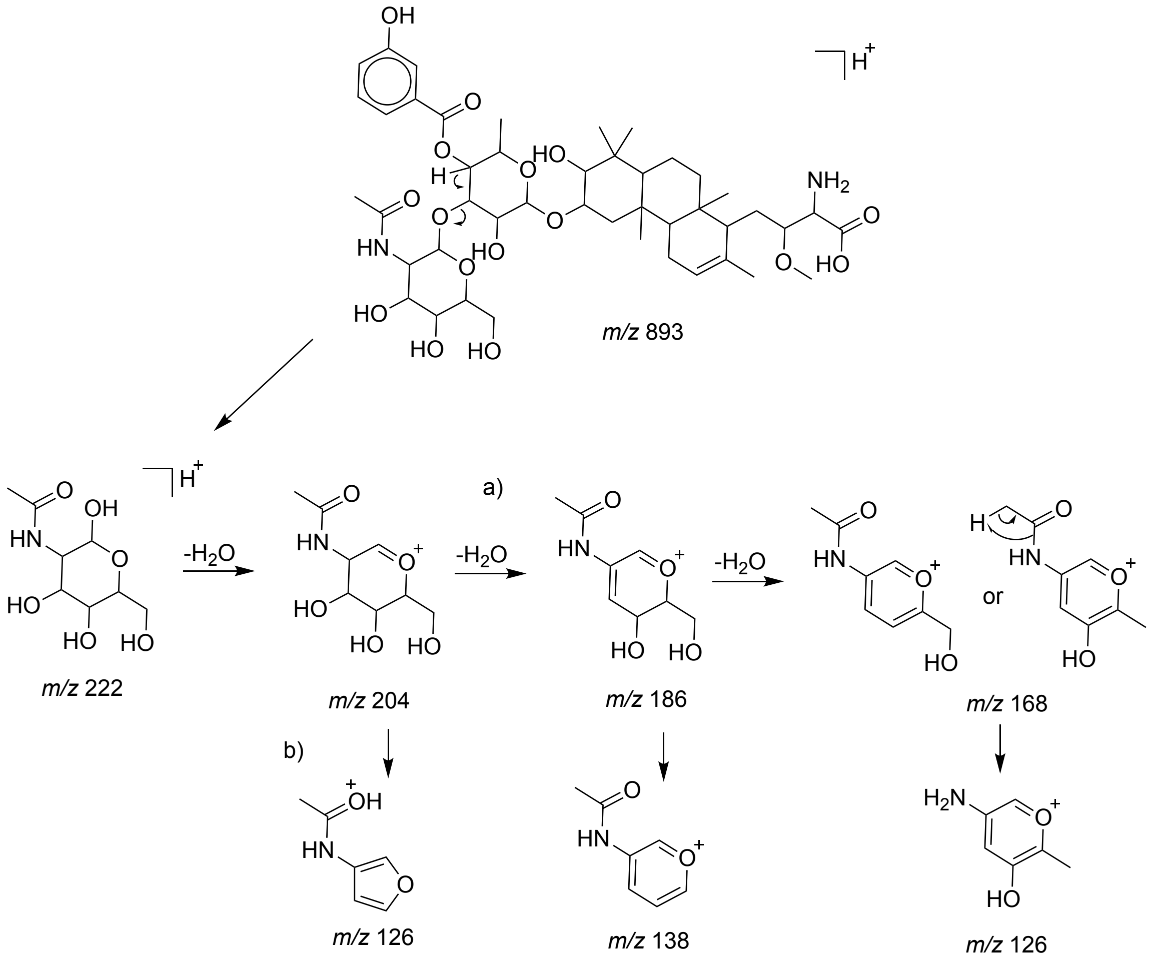


Fig. S3. Proposed mechanism for the fragmentation of brasilicardin A by ESI mass spectrometry. The structure was proposed by NPOmix as a possible match for the MS/MS spectrum with protonated *m/z* 893.4624. Dataset S1, sheet seven, shows the SMILES strings and delta *m/z* values for the predicted structural fragments and the observed fragments in the MS/MS spectrum. All delta *m/z* values in the table were extremely small, strongly indicating that brasilicardin A is the correct structure for this MS/MS spectrum and it matches well with the BCG identified in genome of *Nocardia terpenica* IFM 0406 (BGC known to produce brasilicardin A, ID BGC0000632).

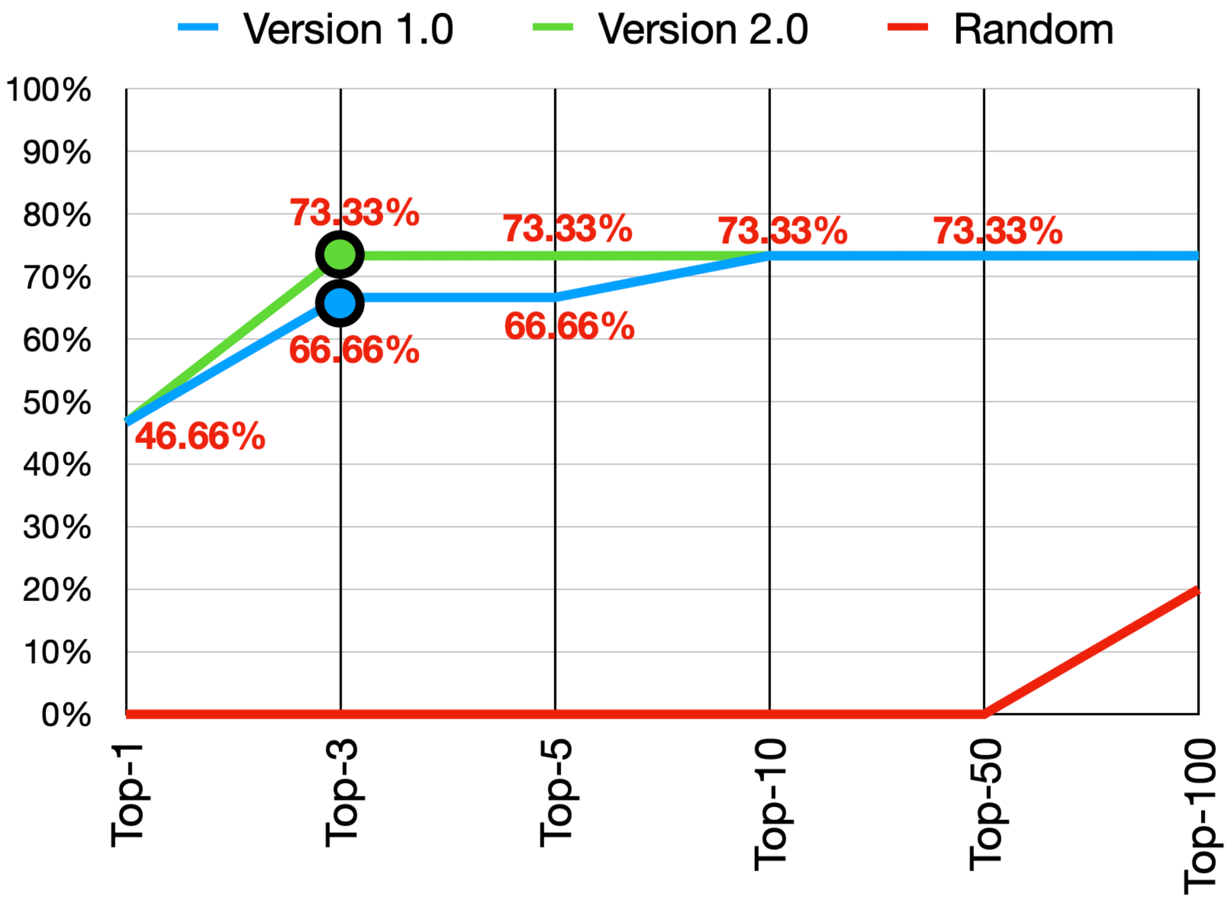


Fig. S4. Comparison of precision curves before (blue line, version 1.0) and after addition of the biosynthetic class (green line, version 2.0). Best precisions are marked by dots (version 1.0 is top-3 = 66.66% and version 2.0 is top-3 = 73.33%). Randomness is represented by the red line.

References

1. E. O'Neill, Mining natural product biosynthesis in eukaryotic algae. *Mar. Drugs* **18**, 90 (2020).
2. S. A. Kautsar, H. G. Suarez Duran, K. Blin, A. Osbourn, M. H. Medema, PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
3. S. D. Bentley, *et al.*, Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147 (2002).
4. M. H. Medema, *et al.*, antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339-46 (2011).
5. J. C. Navarro-muñoz, *et al.*, A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2019).
6. J. R. Doroghazi, *et al.*, A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
7. K. R. Duncan, *et al.*, Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. *Chem. Biol.*, **22**, 60-68 (2015).
8. L. Cao, *et al.*, MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities. *Cell Syst.* **9**, 600-608.e4 (2019).
9. B. Behsaz, *et al.*, De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Syst.* **10**, 99-108.e5 (2020).
10. G. Hjörleifsson Eldjárn, *et al.*, Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLOS Comput. Biol.* **17**, e1008920 (2021).
11. J. J. J. Van Der Hooft, *et al.*, Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297-3314 (2020).
12. S. A. Kautsar, *et al.*, MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
13. T. Leão, *et al.*, A Multi-Omics Characterization of the Natural Product Potential of Tropical Filamentous Marine Cyanobacteria. *Mar. Drugs* **19**, 20 (2021).
14. M. A. Schorn, *et al.*, A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).
15. E. P. Balskus, C. T. Walsh, The Genetic and Molecular Basis for Sunscreen Biosynthesis in Cyanobacteria. *Science (80-.).* **329**, 1653–1656 (2010).
16. M. Wang, *et al.*, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
17. A. T. Aron, *et al.*, Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
18. T. Leao, *et al.*, Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. *Proc. Natl. Acad. Sci.* **114**, 3198–3203

(2017).

19. M. Taniguchi, *et al.*, Palmyramide a, a cyclic depsipeptide from a palmyra atoll collection of the marine cyanobacterium lyngbya majuscula. *J. Nat. Prod.* **73**, 393–398 (2010).

20. H. Luesch, W. Y. Yoshida, R. E. Moore, V. J. Paul, S. L. Mooberry, Isolation, structure determination, and biological activity of Lyngbyabellin A from the marine cyanobacterium lyngbya majuscula. *J. Nat. Prod.* **63**, 611–615 (2000).

21. R. V Grindberg, *et al.*, Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* **6**, e18565 (2011).

22. J. Orjala, W. H. Gerwick, Barbamide, a chlorinated metabolite with molluscicidal activity from the Caribbean cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **59**, 427–430 (1996).

23. Z. Chang, *et al.*, Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **67**, 1356–1367 (2004).

24. B. Márquez, P. Verdier-Pinard, E. Hamel, W. H. Gerwick, Curacin D, an antimitotic agent from the marine cyanobacterium Lyngbya majuscula. *Phytochemistry* **49**, 2387–2389 (1998).

25. K. Kleigrewe, *et al.*, Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. *J. Nat. Prod.* **78**, 1671–1682 (2015).

26. G. J. Hooper, J. Orjala, R. C. Schatzman, W. H. Gerwick, Carmabins A and B , New Lipopeptides from the Caribbean Cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **61**, 529–533 (1998).

27. D. McDonald, *et al.*, American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, e00031-1 (2018).

28. H. Mohimani, *et al.*, Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).

29. W. Bittremieux, *et al.*, Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. *bioRxiv* (2020).

30. H. Gross, *et al.*, The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. *Chem. Biol.* **14**, 53–63 (2007).

31. T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).

32. H. Komaki, *et al.*, Brasilicardin A, a new terpenoid antibiotic from pathogenic Nocardia brasiliensis: Fermentation, isolation and biological activity. *J. Antibiot. (Tokyo).* **52**, 13-19 (1999).

33. Y. Hayashi, *et al.*, Cloning of the gene cluster responsible for the biosynthesis of brasilicardin a, a unique diterpenoid. *J. Antibiot. (Tokyo).* **61**, 164–174 (2008).

34. C. Zhang, *et al.*, Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep.*, **7**, 14243 (2017).

35. R. Reher, *et al.*, A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J. Am. Chem. Soc.* **142**, 4114–4120

(2020).

36. K. Dührkop, *et al.*, Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2020).

37. M. Ernst, *et al.*, Molnetenhancer: Enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **9**, 144 (2019).

38. J. J. J. Van Der Hooft, *et al.*, Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Anal. Chem.* **89**, 7569–7577 (2017).

39. K. Dührkop, *et al.*, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods.* **39**, 462–471 (2019).

40. A. S. Walker, J. Clardy, A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* **61**, 2560-2571 (2021).

41. S. A. Kautsar, J. J. J. Van Der Hooft, D. De Ridder, M. H. Medema, BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* **10**, 1-17 (2021).

42. F. Huber, L. Ridder, S. Rogers, J. J. J. van der Hooft, Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **17**, e1008724 (2020).

43. F. Huber, S. van der Burg, J. J. J. van der Hooft, L. Ridder, MS2DeepScore - a novel deep learning similarity measure for mass fragmentation spectrum comparisons. *bioRxiv* (2021).

44. J. J. J. Van Der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, S. Rogers, Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738-13743 (2016).

45. Y. Sugimoto, *et al.*, A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science.* **366**, 1309 (2019).

46. E. Pereira-Flores, *et al.*, Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques. *bioRxiv* (2021).

47. T. Luzzatto-Knaan, *et al.*, Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. *Elife* **6**, 1686–1699 (2017).

48. A. Bankevich, *et al.*, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

49. K. Blin, *et al.*, AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).