Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Nat. Prod. Rep.*, 2021, **38**, 2100

Genome mining methods to discover bioactive natural products

Katherine D. Bauman, [a] Keelie S. Butler, [b] Bradley S. Moore *[ac] and Jonathan R. Chekan *[b]

Covering: 2016 to 2021

With genetic information available for hundreds of thousands of organisms in publicly accessible databases, scientists have an unprecedented opportunity to meticulously survey the diversity and inner workings of life. The natural product research community has harnessed this breadth of sequence information to mine microbes, plants, and animals for biosynthetic enzymes capable of producing bioactive compounds. Several orthogonal genome mining strategies have been developed in recent years to target specific chemical features or biological properties of bioactive molecules using biosynthetic, resistance, or transporter proteins. These "biosynthetic hooks" allow researchers to query for biosynthetic gene clusters with a high probability of encoding previously undiscovered, bioactive compounds. This review highlights recent case studies that feature orthogonal approaches that exploit genomic information to specifically discover bioactive natural products and their gene clusters.

Received 22nd May 2021

DOI: 10.1039/d1np00032b

rsc.li/npr

1. Introduction

In the natural world, communication is largely chemical. Organisms produce small chemical compounds, which we refer to as natural products, to interact with the biological and physical world around them.[1] While much remains to be discovered about the ecological roles of these compounds, humankind has harnessed the incredible chemical intricacies produced by nature for a variety of purposes.[2] These molecules have evolved over millennia to interact with a specialized biological target, resulting in complex chemical structures and potent bioactivities that are difficult to replicate in the laboratory. Most famously, natural products play a critical role in

[a]*Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, 92093, USA. E-mail: bsmoore@ucsd.edu*

[b]*Department of Chemistry and Biochemistry, University of North Carolina Greensboro, Greensboro, NC, 27402, USA. E-mail: jrchekan@uncg.edu*

[c]*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, 92093, USA*

modern medicine, serving as the source of and inspiration for many lifesaving drugs.[3]

For decades, the discovery of natural products was largely accomplished by a bioactivity-guided isolation approach. Since bioactivity was used as the guiding force during isolation, this method naturally had a very high success rate of identifying bioactive natural products, but this approach also came with its own set of challenges. Dereplication of an active compound from a complex mixture can be incredibly time consuming, and this untargeted approach often resulted in the re-isolation of the same compounds.[4] However, the field of natural products discovery underwent a fundamental shift in the early 2000s, when the first *Streptomyces* bacterial genomes were sequenced.[5,6] These genomes revealed that the vast majority of small molecules produced by microbes had yet to be discovered,[7] thus opening the door for future discovery efforts by complementary genome-focused approaches.

Researchers have increasingly relied on genome mining strategies to discover new natural products, which classically has referred to using genomic sequence data to identify and predict genes that encode the production of novel compounds.[8] While chemical structures can be remarkably diverse, nature often converges on a few mechanisms to generate the same chemical building blocks, thereby allowing researchers to exploit genetic signatures of enzymes to identify new biosynthetic pathways. Once the genetic basis behind production of a molecule is identified, it opens the door to further discovery. Scientists are now incredibly competent at surveying genomic sequence, predicting whether or not it encodes a biosynthetic gene cluster (BGC), and targeting that BGC for isolation of the encoded product.[9,10] However, unlike the bioactivity-guided isolation approach, genome-based methods to discovery do not obviously reveal whether the encoded natural product is likely to be bioactive. Bioactivity is in fact often the last feature known about a molecule; only once it is isolated and characterized does the compound get tested in a bioassay.

Katherine D. Bauman is an NIH NRSA F31 predoctoral fellow in the lab of Bradley S. Moore at the Scripps Institution of Oceanography, UC San Diego. She is pursuing a PhD in Marine Chemical Biology where her research focuses on the biosynthesis of natural products. Prior to her graduate studies, Katherine studied chemistry at Middlebury College where she studied natural products with Professor Lesley-Ann Giddings.

Bradley Moore is a Distinguished Professor of Marine Chemical Biology and Pharmaceutical Chemistry at UC San Diego. His research focuses on the molecular and genomic basis of natural product biosynthesis and the application of new genetic tools and biocatalysts to produce bioactive small molecules. He is the former chair of the editorial board of Natural Product Reports *and the recipient of the Natural Product Chemistry Award from the Royal Society of Chemistry in 2018 and the Ernest Guenther Award in the Chemistry of Natural Products from the American Chemical Society in 2021.*

Keelie S. Butler is a NIH NCCIH T32 trainee in the lab of Jonathan R. Chekan at the University of North Carolina at Greensboro. She is pursuing a PhD in Chemistry and Biochemistry, with a research focus on the biosynthesis of natural products. Prior to her graduate studies, Keelie studied chemistry and biochemistry at Campbell University, where she studied biosurfactant production in marine microbes with Dr Michelle S. Thomas.

Jonathan R. Chekan is an Assistant Professor of Chemistry and Biochemistry at the University of North Carolina at Greensboro. He received his PhD in 2016 from the University of Illinois Urbana-Champaign under the supervision of Satish K. Nair. As the Simons Foundation Post-Doctoral Fellow of the Life Science Research Foundation in the laboratory of Bradley S. Moore, he investigated the biosynthesis of neurotoxic marine natural products. In 2020, he joined University of North Carolina at Greensboro where his research group focuses on the bioinformatic guided discovery of new natural products and characterization of their biosynthetic pathways.

While many recent reviews have focused on genome mining strategies,[11–15] the purpose of this review is to highlight recent examples in the literature that utilize genome mining strategies that target molecules suspected to have bioactivity. We have organized this review into four sections, each encompassing a distinct strategy commonly used for genome mining with the goal of bioactive natural product discovery. Moreover, while there are many genome mining reports that successfully mine a single bacterial genome or a handful of genomes for the presence of a BGC encoding a likely bioactive compound, this review instead focuses on larger-scale genome mining efforts. We highlight examples where researchers used targeted, hypothesis-driven approaches to mine large datasets with the express purpose of identifying new bioactive natural products. While not exhaustive, the examples described here aim to cover a variety of bioactive chemical features, compound families, and pharmacological properties.

2. Bioactive feature targeting

While natural products can be large and complex, they often contain smaller chemical features that directly lead to bioactivity, which we refer to as the "bioactive feature" of a molecule. For the purposes of this article, we have categorized these bioactive features into two distinct groups. The first group encompasses reactive features, which includes functional groups with electrophilic, radical, or nucleophilic reactivity that often result in covalent binding of the ligand to the protein target (Section 2.1). The second group includes structural features important for the natural product's ability to bind non-covalently to a biological or chemical target, which can vary from a macromolecular protein to small metal ions (Section 2.2). For both types of bioactive features, reactive and ligand binding, detailed studies have shown how a diversity of these chemical features are biosynthesized and installed into natural products. These insights can be utilized for genome mining efforts to identify orphan BGCs predicted to produce natural products with the queried bioactivity-associated chemical moiety (Fig. 1). The examples described here all successfully utilize an enzyme responsible for the installation of a bioactive chemical feature as a genome mining hook to discover new natural products with that target moiety. Importantly, the resulting molecule may exist in an entirely different compound family (peptide *versus* polyketide, for example) but still contain the cognate ligand feature (β-lactone, for example). We refer to this strategy as "bioactive feature targeting".

Because the diversity of bioactive functional groups is particularly broad, we have tabulated a selection of reactive chemical features commonly found in natural products and their associated biosynthetic routes (Table 1). Some of these chemical features (*i.e.*, enediynes, β-lactones, epoxyketones, and disulfides) have been the target of large-scale genome mining efforts, resulting in the discovery of new bioactive molecules. Those are elaborated on in Section 2.1. However, for

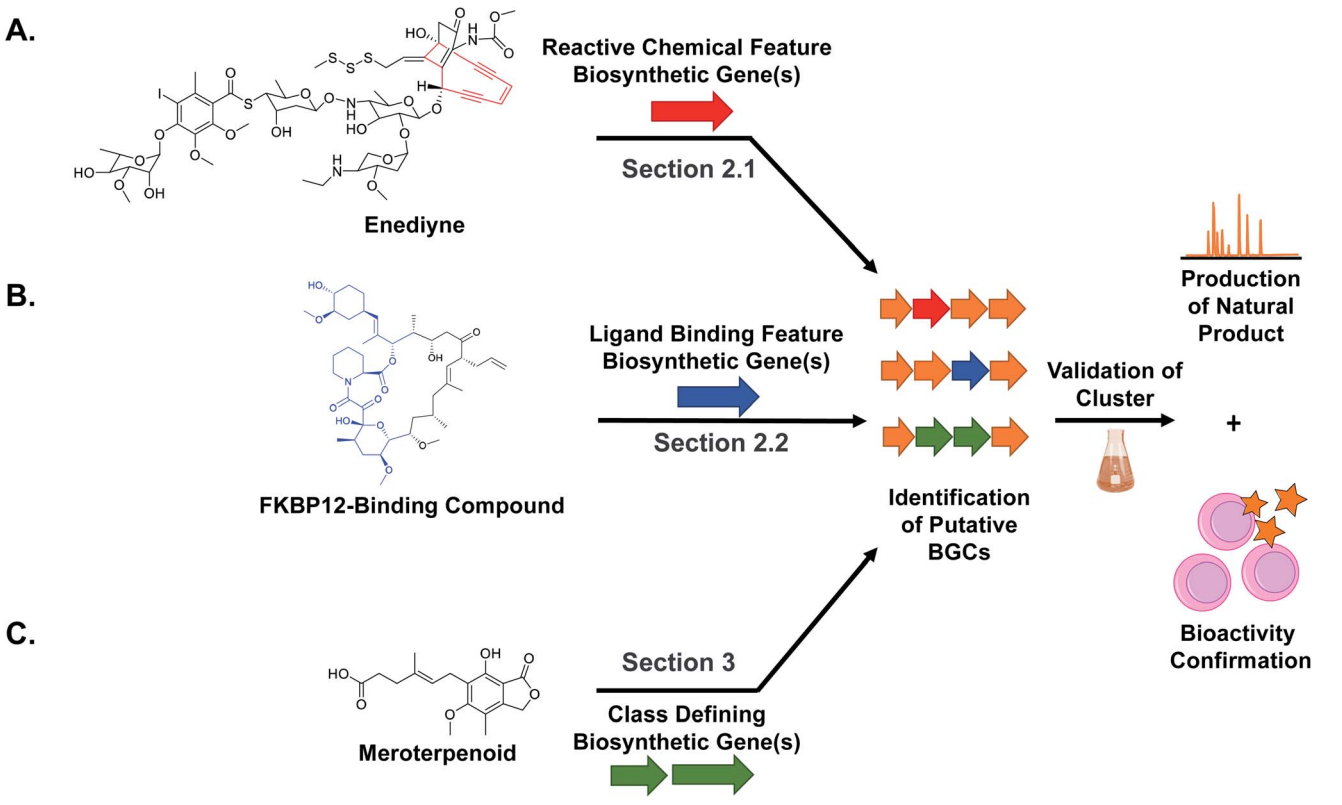
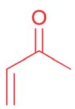
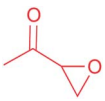
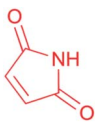
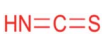
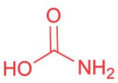


Fig. 1 Overall scheme for genome mining new natural products with different target chemical features including (A) reactive chemical features, (B) ligand binding features, and (C) compound family defining features. In each case, the diagnostic biosynthetic genes are used to bioinformatically identify candidate gene clusters that putatively encode the production of a target natural product. Subsequent production, isolation, characterization, and bioactivity assay tests validate the bioinformatic prediction.

Table 1 Reactive chemical features found in natural products

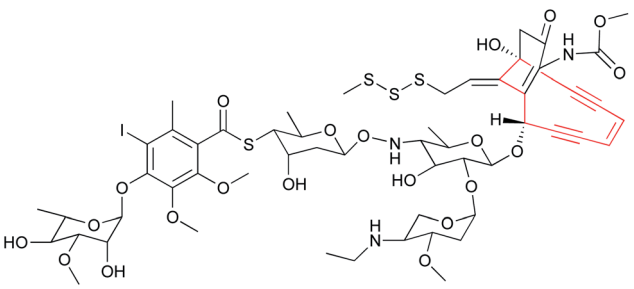
Reactive chemical feature	Structure	Enzymatic route(s) to installation	Genome mining efforts using reactive chemical feature biosynthetic enzymes
Enediyne		PKS[16,17]	Large scale genome mining[18,19]
β-Lactam		(1) β-Lactam synthetase (β-LS)[20] (2) Carbapenam synthetase (CPS)[21] (3) IPNS[22] (4) Condensation domain[23]	Single genome mining[24]
β-Lactone		(1) β-Lactone synthetase[25,26] (2) Thioesterase (TE)[27] (3) Hydrolase[28]	Large scale genome mining[26] Single genome mining[29]
Michael acceptor: α,β-unsaturated carbonyl		(1) Terpenes synthase[30] (2) PKSs[31] (3) Hybrid NRPS-PKS[32]	
Epoxyketone		(1) Flavin-dependent decarboxylase–dehydrogenase–monooxygenase[33,34]	Large scale genome mining[35]
Epoxide		(1) P450 (epothilone)[36] (2) Flavin-dependent epoxidases[37] (3) Dioxygenases (epoxyquinones)[38] (4) Non-heme iron-dependent epoxidases (fosfomycin)[39,40]	Single genome mining[41] Single genome mining[42]
Aziridine		Unknown	
Maleimide		(1) Flavin-dependent oxidase[43] (2) PKS-NRPS[44]	
Sulfonamide/sulfone		(1) Radical-forming, SO_2 incorporating flavoprotein[45]	
Furan		Terpene oxidation[46]	Large scale genome mining[47]
Disulfide		(1) FAD-dependent dithiol oxidase (holomycin, gliotoxin, FK228)[48–50] (2) DUF-SH didomain[51]	Single genome mining[52] Large scale genome mining[53]
Isothiocyanate		Putative isonitrile synthase[54]	Large scale genome mining[54]
Carbamate		Carbamoyltransferase[55]	

some well-known reactive chemical features, such as β-lactams and α,β-unsaturated carbonyl Michael acceptors, dedicated large-scale genome mining efforts have yet to be reported.

2.1 Reactive chemical features

2.1.1 Enediynes. Enediynes are a highly reactive class of polyketide synthase (PKS)-derived natural products characterized by an alkene flanked by a pair of alkynes within a 9- or 10-membered ring.[56] This reactive chemical feature undergoes a Bergman cycloaromatization reaction to form an aromatic diradical species that can interact with the minor groove of DNA thereby facilitating interstrand cross links or double strand breaks. Both actions are mutagenic and lead to cytotoxicity. As of 2016, 11 enediyne-contain natural products, such as calicheamicin γ1 (**1**), had been characterized with four either approved as drugs or as drug candidates in clinical trials as antibody–drug conjugates.[19] This high translational success rate motivated the Shen lab to search for new enediyne natural products using a genome mining approach.[19] Previous sequence alignments of enediyne biosynthetic gene clusters revealed a conserved set of PKS genes responsible for installation of the enediyne warhead.[17,57,58] An initial survey of the NCBI and JGI

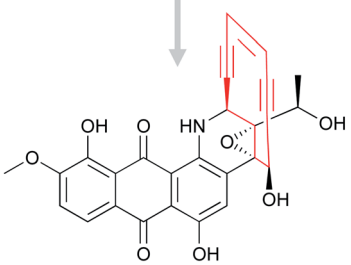
genomic databases using these biosynthetic genes suggested that enediyne containing natural products are not especially rare as they identified 87 putative enediyne BGCs across 78 bacterial strains.[59] Using these insights, the Shen lab surveyed the 3400 strain actinomycete collection housed at the Scripps Research Institute (TSRI) to prioritize strains for isolation of enediynes.[19] Because genomic information was not available for the entire collection, a real-time PCR method was employed to target two different enediyne biosynthetic genes. This analysis revealed 81 producing strains, and phylogenetic analysis of a 1 kb gene fragment suggested that many of the clusters were distinct from known ones. To confirm this observation, whole genome sequencing was completed for 31 representative strains, and a Genome Neighborhood Network (GNN) composed of the newly sequenced clusters revealed the potential for new chemistry. Of particular interest was a gene cluster from *Streptomyces* sp. strain CB03234 that appeared to encode a 10-membered enediyene related to uncialamycin.[60] Subsequent isolation work identified tiancimycin A (**2**) as a new uncialamycin analog. As with other characterized enediynes, tiancimycin A was highly cytotoxic with IC_{50}s in the subnanomolar range across different cancer lines and with potential to be developed as an antibody–drug conjugate. The potent activity of enediynes and presence of many promising uncharacterized BGCs make this reactive feature a promising target for future discovery efforts.



1, calicheamicin γ1

enediyne forms highly reactive diradical anticancer activity

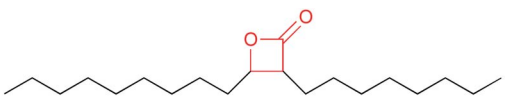
Genome mining target: conserved PKS



2, tiancimycin A

enediyne anticancer activity

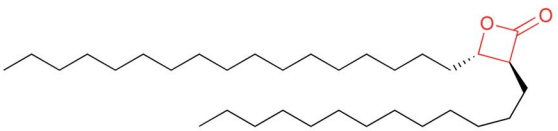
2.1.2 β-Lactones. Electrophilic moieties in natural products are often responsible for the bioactivity of a molecule due to their ability to form covalent interactions with the nucleophilic centers present in a wide variety of biological targets.[61] Well known examples include Michael acceptor systems such as α,β-unsaturated ketones, β-lactams, epoxyketones, and β-lactones (Table 1). β-Lactones are often referred to as "privileged" structures due to their highly strained and reactive nature.[62] These electrophilic warheads are particularly intriguing as they react with nucleophiles *via* ring opening reactions and covalently bind their biological target. Therefore, the presence of a β-lactone often endows a molecule with bioactivity. However, this inherent reactivity has made their isolation and characterization difficult, which has hampered efforts towards understanding their biosynthesis. Three examples of β-lactone biosynthesis have been biochemically validated, and all proceed *via* a different enzymatic mechanism (Table 1). The first enzymatic route to β-lactone biosynthesis was discovered only recently in the pathway for olefinic hydrocarbons.[25] The enzyme OleC is a member of the Acyl-CoA ligases, Nonribosomal peptide synthetases (NRPSs), and Luciferases (ANL) superfamily. In the olefin pathway OleC instead acts as an ATP-dependent β-lactone synthetase that converts *syn*- and *anti*-β-hydroxy acids into *cis*- and *trans* β-lactones. Following this discovery, the Wackett group generated a web-based predictive tool to take a genome mining approach to discover novel β-lactone natural products that are biosynthesized *via* similar β-lactone synthetase chemistry.[26] Their tool, AdenylPred, utilizes machine learning to predict ANL superfamily enzyme function and substrate specificity. Using AdenylPred, they mined their collection of over 50 000 BGCs for β-lactone synthetases and identified over 90 candidates. They proceeded to experimentally validate one of these predicted enzymes found in an



OleC product

β-lactone strained ring is highly reactive

Genome mining target: β-lactone synthetase (OleC)



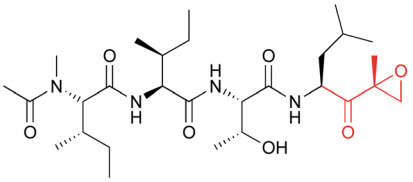
3, nocardiolactone

β-lactone antibiotic

uncharacterized BGC in pathogenic *Nocardia* bacteria. *In vitro* reconstitution of the four gene pathway led to the production of nocardiolactone (**3**), a previously isolated natural product[63] but one for which no BGC had yet been characterized. By focusing on the enzyme responsible for the synthesis of the β-lactone warhead, the Wackett group was able to use genome mining to not only link a previously isolated natural product to its orphan gene cluster, but also identify a treasure trove of uncharacterized pathways predicted to encode novel β-lactone natural products that await discovery.

2.1.3 Epoxyketones. Epoxyketones are well known reactive chemical features that perform covalent modifications of their protein target. Natural products bearing epoxyketones may be rare, but their potent bioactivity has been well recognized.[64] Eponemycin and epoxomicin (**4**) were the first epoxyketone-bearing proteasome inhibitors to be linked to their biosynthetic genes,[65,66] and while neither of these compounds are used clinically, derivatives such as carfilzomib are FDA-approved chemotherapeutics.[67] With the discovery of a biosynthetic route to epoxyketone construction established in 2014,[34] genome mining for bioactive natural products containing epoxyketones became possible.

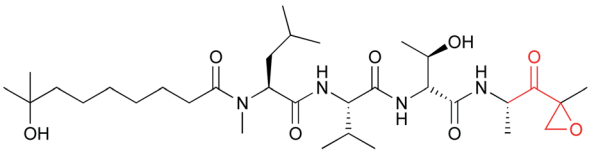
To find new epoxyketone-bearing molecules, the Brady lab utilized their previously developed informatics platform eSNaPD (environmental Surveyor of Natural Product Diversity).[35,68] This platform compares PCR generated sequences from metagenomes that target conserved biosynthetic motifs, called natural product sequence tags (NPSTs), to a reference set of gene clusters. This comparison reveals the biosynthetic potential of the original metagenomic sample. To apply this approach specifically to mine for epoxyketones, they utilized the lone ketosynthase (KS) domain essential for the biosynthesis of the conserved epoxyketone in eponemycin and epoxomicin as their NPST target.[34] By using that sequence as a genome mining hook, they used the eSNaPD platform to mine 185 soil metagenomes, which likely include tens of thousands of unique bacterial genomes. This search returned 99 hits indicative of a sequence encoding an enzyme capable of installing an epoxyketone moiety. Four of the metagenomes with epoxyketone hits had cosmid libraries that allowed the authors to identify 11 putative BGCs that contained the hit sequence. Out of these 11 BGCs, 9 were suggested by *in silico* analysis to produce an epoxyketone-containing natural product, which was further verified by the proteasome inhibitory activity of culture extracts of these BGCs. The Brady lab was ultimately able to identify seven novel compounds from these strains, the clarepoxcins (clarepoxcin A, **5**) and the landepoxcins (landepoxcin A, **6**), all nanomolar inhibitors of the human 20S proteasome. Both series of compounds contain an epoxyketone moiety attached to a hydrophobic peptidic backbone, but the structures differ from any previously described molecule. This example demonstrates that enzymes responsible for installation of a specific bioactive chemical feature can prioritize BGCs of interest from even massive metagenomic datasets, ultimately leading to the discovery of new bioactive natural products.



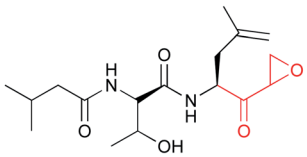
4, epoxomicin

epoxyketone electrophilic warhead
proteasome inhibitory activity

Genome mining target:
KS required for epoxyketone



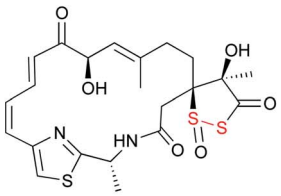
5, clarepoxcin A



6, landepoxcin A

epoxyketones
proteasome inhibitory activity

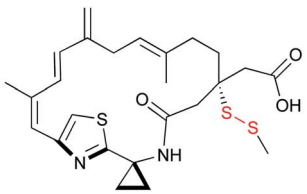
2.1.4 Disulfide bonds. Disulfide bonds are well recognized structural components in proteins, but they can also serve as highly reactive features in small molecules leading to pronounced bioactivity.[69] Disulfide bonds act as prodrugs – the inactive disulfide bond can be readily transformed to an active dithiol *via* cleavage in the cellular environment.[70] Leinamycin (**7**) is one such disulfide-containing small molecule with potent bioactivity.[71] Isolated from *Streptomyces atroolivaceus*, leinamycin contains a unique 1,3-dioxo-1,2-dithiolane moiety spirofused to a macrolactam ring and displays potent anticancer bioactivity *via* an unusual mechanism of action.[71,72] Because leinamycin's bioactivity is directly linked to the reactive 1,3-dioxo-1,2-dithiolane moiety, we will refer to this as the reactive chemical feature. Cellular thiols reduce the 1,3-dioxo-1,2-dithiolane moiety thereby forming an episulfonium ion capable of alkylating DNA, ultimately resulting in DNA cleavage and death.[73,74]



7, leinamycin A

reactive disulfide cleaved to dithiol
episulfonium ion alkylates DNA
anticancer

Genome mining target:
DUF-SH domain



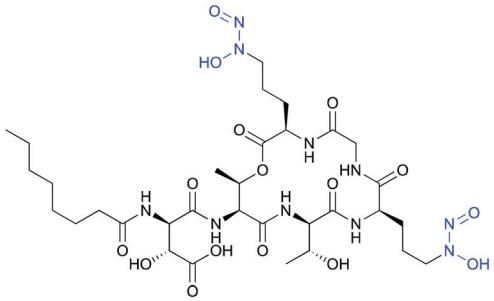
8, guangnamycin A

disulfide
episulfonium ion formation
anticancer

Despite its unusual chemical features and potent bioactivity, leinamycin remained the only member of its family until 2017, when the Shen lab aimed to add to the leinamycin family of natural products *via* a genome mining approach.[53] Because they knew that leinamycin's bioactivity was linked to the 1,3-dioxo-1,2-dithiolane moiety, they used a bioactive feature mining-based approach to identify new compounds containing this reactive chemical feature. Previous biochemical work on the leinamycin pathway (*lnm*) had revealed a domain of unknown function (DUF) and a cysteine lyase domain (SH), referred to as the DUF-SH didomain, was responsible for sulfur incorporation.[51,75] Therefore the authors mined all public databases using the DUF-SH didomain and identified 19 potential *lnm* BGCs. An additional 30 BGCs were selected by mining TSRI's strain collection. Phylogenetic analysis of this DUF-SH protein revealed 18 distinct clades containing 28 unique BGCs. A series of bioinformatic analyses were used to make structural prediction and aid in strain prioritization. Ultimately, two strains were chosen for future analysis. Using gene knockouts, Shen and coworkers identified two series of leinamycin analogs, the guangnanmycins from the *gnm* BGC and the weishanmycins from the *wsm* BGC. Guangnanmycin A (**8**) had the same reactivity towards oxidants as leinamycin, suggesting it could possess cytotoxic activity using a similar episulfonium ion mediated DNA alkylating mechanism. Weishanmycin A lacks the disulfide bond found in leinamycin. Instead, the free thiol may be activated by reactive oxygen species similar to leinamycin E1.[73] After 30 years without the discovery of any leinamycin analog, a bioactive feature-based genome mining approach ultimately resulted in the expansion of the leinamycin family.

2.2 Ligand binding features

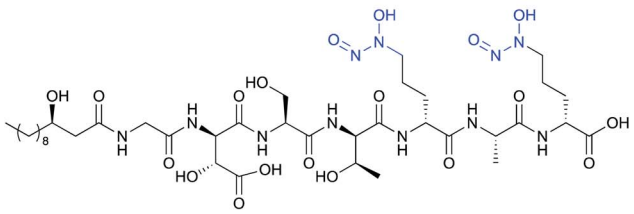
2.2.1 Diazeniumdiolates. Metallophores are small molecules with electron-rich ligand features that bind metal ions, such as iron and cobalt, that are critical for an organism's survival. While this ecological role is well recognized, these compounds are also important from a biomedical perspective. Their metal mobilization abilities have been used clinically to treat metal toxicity,[76] and are now utilized to create siderophore–antibiotic conjugates,



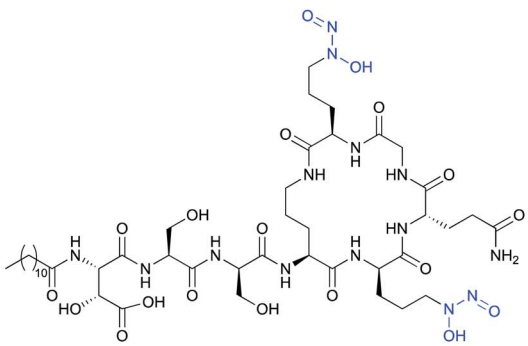
9, gramibactin A

diazeniumdiolate
binds iron

Genome mining target:
diazeniumdiolate
(GrbD, GrbE)



10, megapolibactin B



11, plantaribactin

diazeniumdiolate
binds iron

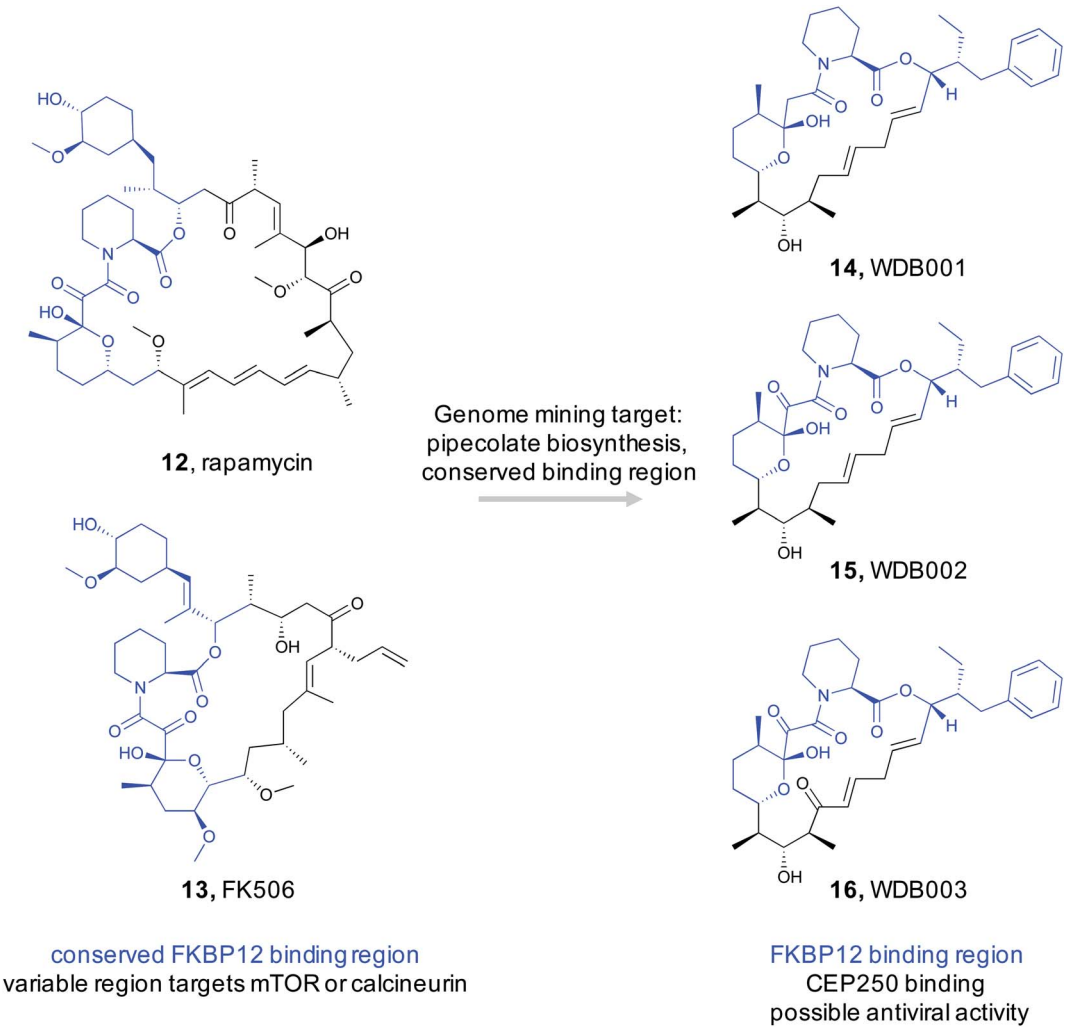
capable of stealthily entering bacterial cells *via* the endogenous metal transport system where they can then deploy an antibiotic payload.[77,78] Despite the fact that over 500 different siderophores have been isolated, showcasing tremendous structural diversity, these small molecule metallophores employ just a small set of functional groups for binding trace metals,[79] which endow the molecule with its bioactivity.

In 2018, Hertweck and coworkers discovered a new bacterial siderophore with a novel iron-binding functional group produced by plant-associated symbiotic bacteria.[80] This metallophone, named gramibactin A (**9**), is unusual among siderophores because it harbors two diazeniumdiolate (*N*-nitrosohydroxylamine) moieties. The diazeniumdiolate moiety, while never previously recognized as an iron-binding feature, was shown to be directly involved in iron complex formation, and thus represents the bioactive feature critical for the ligand's metal-binding ability. Because of this important biological activity and unusual chemical structure, the Hertweck group took a genome mining approach to identify novel diazeniumdiolate-containing siderophores.[81]

Using a targeted knockout approach, two enzymes, GrbD and GrbE, were identified as essential for graminine biosynthesis. In the hopes of identifying novel natural products that contain the diazeniumdiolate ligand, they used BLAST to query GrbD against the non-redundant UniProt protein sequence database. Genome neighborhood analysis identified 37 probable gene clusters. Notably, all clusters were found in organisms from the Burkholderiaceae family, but the strains harboring these BGCs were divided into two distinct groups. One group of BGCs was found in plant symbiotic bacterial strains, while the other group was found in plant-pathogenic strains. The authors cultured a selection of these strains and screened *via* MS/MS for the characteristic fragment resulting from cleavage of the diazeniumdiolate moieties. From the symbiotic bacteria, they ultimately identified a gramibactin congener, gramibactin B, and a new family of six peptide compounds bearing the diazeniumdiolate moiety that they named the megapolibactins A–F. From the culture extract of a strain from the group of plant-pathogenic strains they isolated a new compound they named plantaribactin. Bioactivity testing of megapolibactin B (**10**) and plantaribactin (**11**) revealed that beyond just iron chelation, these diazeniumdiolate-containing compounds are also capable of serving as nitric oxide (NO) donors. The discovery of new diazeniumdiolate siderophores from bacteria further suggests that ligand binding-based genome mining is an attractive approach for the targeted discovery of natural products with specific bioactivities, and that this approach transcends the nature of the ligand's target. As this example showed, bioactive feature targeting is not limited to features important for macromolecular binding but can be utilized to target bioactive features important for metal ion binding as well.

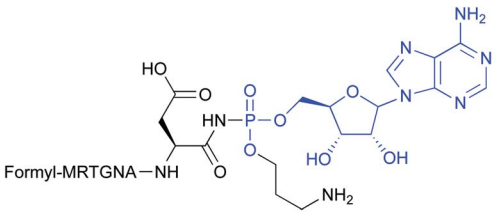
2.2.2 FKBP12-binding compounds. Protein–protein interactions have been seen as a promising target for new drugs. Unfortunately, it has proven difficult to design or discover small molecules that can alter how proteins interact with each other.[82] This issue is so problematic that some promising targets are considered "undruggable".[83] A particularly interesting strategy to overcome this challenge is observed in the FK506/rapamycin family of natural products.[84] These NRPS/PKS hybrid natural products possess immunosuppressive bioactivity by modulating the kinase mTOR or phosphatase calcineurin. The bioactivity of rapamycin (**12**)/FK506 (**13**) can be attributed to two structural elements. The first is a conserved region that contains a key pipecolate moiety. This portion of the molecule tightly binds to the common cellular protein peptidyl prolyl isomerase FK506-binding protein (FKBP12).[83] The rapamycin-FKBP12 binary complex then displays the second structural feature of rapamycin, a PKS derived variable region that directly interacts with mTOR and modulates its activity. Changes to this variable region allow rapamycin-like molecules to target different proteins, making this class of molecules a promising source of new bioactive compounds. Critically, rapamycin-like natural products cannot bind to their protein targets, such as mTOR, alone. Instead, the initial binding to FKBP12 *via* the pipecolate-containing conserved region is necessary. Thus the conserved ligand binding motif is critically important for the class defining bioactivity.

By exploiting the two structural features of rapamycin, a recent study conducted by scientists at Warp Drive Bio sought to use a genome mining approach to discover new FK506/rapamycin family members with new bioactivity.[85] To find new FK506/rapamycin analogs, the authors focused on the pipecolate moiety found in the ligand binding feature that binds in the FKBP12 active site. The pipecolate residue is biosynthesized by lysine cyclodeaminase and is subsequently incorporated as the final step of the NRPS assembly line.[86] The authors therefore reasoned that the lysine cyclodeaminase would be an ideal search query for a genome mining based discovery effort. With an inhouse strain database of ~135 000 *Actinomycetes*, it was not feasible to assemble and search each genome for lysine cyclodeaminases in an efficient way. Instead, the authors pooled between 40 and 480 strains at a time and conducted initial high throughput sequencing and assembly. The resulting contigs were then searched for the presence of a lysine cyclodeaminase analog. Strains in the positive pools were subsequently screened by PCR and followed up with complete genome assemblies. This process rediscovered the four known rapamycin family members along with seven additional clusters that appeared to produce new analogs with similar conserved regions and diverse variable regions. Five of the clusters were over expressed and their products were structurally characterized. A cluster that was designated as X1 from *Streptomyces malaysiensis* DSM 41697 proved promising. X1 produced three new analogs (WDB001-003) with the predicted conserved and variable regions (**14–16**). Detailed studies demonstrated that this series of natural products behaved



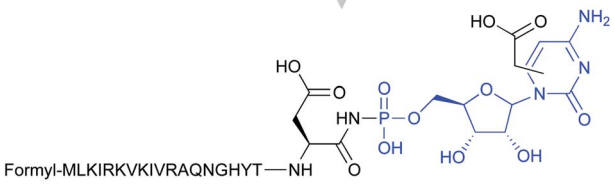
similar to FK506/rapamycin in that they tightly bind to FKBP12 *via* the conserved region. The variable region is then exposed to interact with the protein target. Rigorous studies revealed the target to be a coiled-coil region of a centrosome-associated protein called CEP250. Cell-based assays demonstrated that WDB002 recruits FKBP12 to CEP250 in the centromere and decreases centromere separation. A recent screening assay indicated that CEP250 may interact with Nef13 from SARS-CoV-2,[87] the virus responsible for COVID-19. This suggests that WDB002 may warrant further study as antiviral.

2.2.3 Nucleotidyl phosphoramidates. Enzymatic reactions often make use of chemically activated, high energy intermediates to facilitate catalysis. This is most often observed with ATP-dependent enzymes that utilize phosphorylated or adenylated intermediates. Bioactive natural products that mimic these activated intermediates with non-hydrolyzable analogs of the labile phosphoester bond can therefore tightly bind to and inhibit a target enzyme.[88] For example, the ribosomally synthesized and post-translationally modified peptide (RiPP) natural product microcin C7 (McC, **17**) utilizes this approach to inhibit aspartyl-tRNA synthetase. McC is a seven amino acid long peptide that contains a C-terminal aspartyl-adenylate formed with a stable phosphoramidate bond, the critical ligand binding feature.[89] Upon import of McC into the cell[90] and its subsequent hydrolysis into constitutive amino acids, the modified aspartyl residue is released and specifically targets aspartyl-tRNA synthetase because it mimics the aspartyl adenylate mechanistic intermediate.[91] Mining for amino acid nucleotidyl phosphoramidates should therefore reveal new natural products that similarly target and inhibit tRNA-synthetases.



17, microcin C7
nucleotidyl phosphoramidate,
antibiotic (aspartyl tRNA synthetase inhibitor)

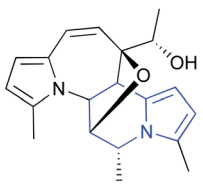
Genome mining target:
nucleotidyl
phosphoramidate formation
(MccB)



18, *B. amyloliquefaciens* microcin C7
nucleotidyl phosphoramidate
antibiotic

Detailed study of the McC biosynthetic gene cluster revealed that MccB, a ThiF-like adenyltransferase, is responsible for installation of critical phosphoramidate bond from an essential C-terminal asparagine residue of the precursor.[92–94] Genome mining for MccB, along with a peptide exporter, identified homologs from diverse bacterial species.[95,96] Sequence analysis suggested that many of these clusters contained precursor peptides with the obligate C-terminal asparagine, indicating that these strains could be producing McC-like antibiotics that still target aspartyl-tRNA synthetase.[95] Some of these new clusters appeared to contain additional enzymes that were not present in the original *Escherichia coli mcC* operon, indicating that further modifications to the precursor peptide could be found in the resulting natural products. In particular, the Severinov and Dubiley groups focused on a cluster found in *Bacillus amyloliquefaciens*.[97] Because metabolite analysis of *B. amyloliquefaciens* DSM 7 did not indicate the presence of McC analogs, the gene cluster was heterologously expressed in *Bacillus subtilis* 168. Mass-spectrometric analysis revealed the presence of a modified precursor peptide with an unknown nucleobase. *In vitro* reconstitution of the biosynthetic pathway revealed that *B. amyloliquefaciens* MccB favors CTP instead of ATP to form an aspartyl-cytidylate mimic that retains the non-hydrolyzable phosphoramidate bond (**18**). Moreover, this nucleobase is further modified with a carboxymethyl group. Activity assays demonstrated that the new McC analog was still a potent aspartyl-tRNA synthetase inhibitor despite the change in nucleobase due to the conserved ligand binding feature. Notably, *B. amyloliquefaciens* McC was active in the presence of the acylating McC resistance gene *mccE*, allowing it to overcome one of the two modes of McC resistance found in *E. coli*.[97] This combination of genome mining and biochemical study could lead to the discovery of other novel McC antibiotics that use modified peptidyl-nucleobases to inhibit tRNA synthetases.

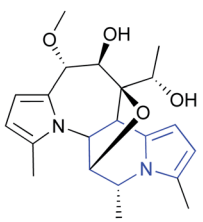
2.2.4 Indolizidines. The examples discussed in this section all showcase genome mining studies that targeted specific features important for a ligand's ability to bind a known target, whether that target be a cellular protein or extracellular iron. However, in some cases a specific feature is identified as important for ligand binding to a biological target, but that target remains unknown. Such is the case for the indolizidine alkaloids, a diverse class of compounds characterized by an indolizidine ring system: fused 5- and 6-membered rings sharing a nitrogen atom at the bridgehead position.[98] This ring structure is a privileged feature associated with bioactive molecules,[99] and structure–activity relationship studies have shown the bioactivity of indolizidine alkaloids is associated with the indolizidine moiety.[100] This potent bioactivity has created tremendous interest in this family of natural products and inspired the search for new members of this class of compounds, resulting in the discovery of curvulamine (**19**) and curindolizine, two compounds isolated from a fish-associated fungi with antibacterial and anti-inflammatory activities, respectively.[101,102]



19, curvulamine

indolizidine ring critical for bioactivity
antibiotic

Genome mining target:
indolizidine ring
(CuaB)



20, bipolamine G

indolizidine ring
antibiotic

Dai *et al.*, identified the BGC responsible for curvulamine production using a combination of antiSMASH, RT-PCR experiments, and targeted gene deletions.[103] Ultimately, heterologous expression of a PLP-dependent aminotransferase, CuaB, revealed its role as a bifunctional aminotransferase capable of both a Claisen condensation to form new C–C bonds and the α-hydroxylation of an amino acid. The unusual biosynthetic ability of this enzyme to install a potent pharmacophore prompted the authors to use it as a genome mining hook to discover new antibacterial indolizidine alkaloids. By searching the NCBI and JGI databases, they uncovered five distinct BGCs that encode CuaB homologs as well as other putative biosynthetic enzymes. These included a BGC from *Bipolaris maydis* ATCC 48331 that they referred to as the *bip* cluster. Upregulation of the putative transcription factor *bipF* using a strong promoter resulted in the production of nine new polyketide indolizidine alkaloids they named bipolamines A–I. Antibacterial bioactivity assays revealed that many of the bipolamines were bioactive, with bipolamine G (**20**) exhibiting even more potency than curvulamine, the small molecule that initially originated this search. This work thus demonstrates that even when a ligand's binding partner is unknown, genome mining using a bioactive feature targeting approach can still be effective for the discovery of new bioactive natural products.

3. Compound family mining

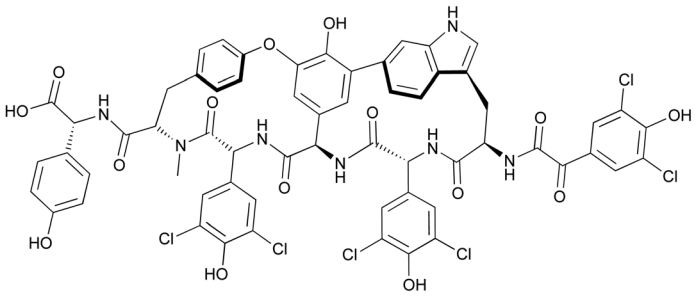
The bioactivity-guided isolation approach that dominated natural product discovery for decades, while untargeted and time consuming, resulted in the discovery of a tremendous number of bioactive small molecules, many of which are still used clinically today.[104] This discovery work revealed that specific compound families have proven time and time again to likely possess bioactivity. We define 'compound families' as a structurally related group of compounds that share a biosynthetic origin. With the advent of genome mining, researchers can now utilize a targeted, hypothesis-driven approach to search for a new member of a bioactive compound family, with the assumption that related members of this compound family will also be bioactive and likely have the same molecular target.

In a compound family genome mining approach, biosynthetic genes responsible for the overall chemical structure of a bioactive molecule are used as genome mining hooks to identify related analogs (Fig. 1C). This differs from the bioactive feature mining approach (Section 2) because the goal is to extend a structural series of compounds to better understand and optimize structure–activity relationships around a lead molecule rather than to mine for the presence of a specific functional group. To search for new analogs of a known bioactive compound family, the genome mining hook might be a single gene, a small biosynthetic cassette, or even an entire BGC. In the case of single gene hooks, sequence similarity networks such as a those generated by the Enzyme Function Initiative Enzyme-Similarity Tool (EFI-EST) can rapidly identify protein homologs for consideration.[105,106] While not inclusive, the diversity of examples included in this section is intended to highlight the utility of this genome mining approach to discover new bioactive molecules across compound classes.

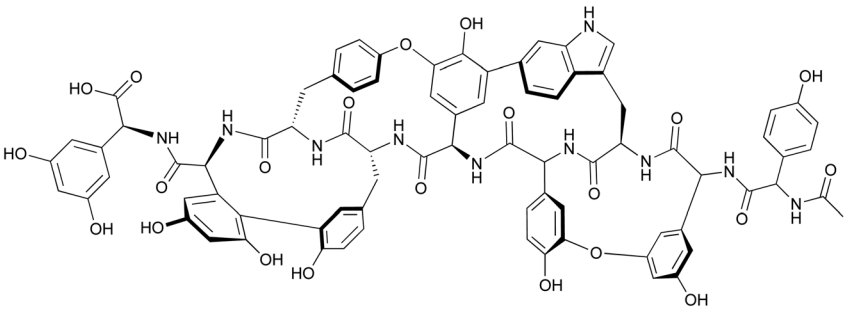
3.1 Glycopeptide antibiotics

Glycopeptide antibiotics are an important family of natural products and include the clinically valuable drugs vancomycin, teicoplanin, and their semi-synthetic derivatives.[107] These NRPS derived products are composed of seven amino acids that are modified into typically tri- or tetra-cyclic structures. Additional tailoring reactions, such as glycosylations and halogenations, are also common.[108] Glycopeptide antibiotics often have a conserved mode of action wherein they bind to lipid II of the bacterial cell wall, inhibiting the transpeptidation or transglycosylation reactions required for bacterial cell wall synthesis. Because of this potent bioactivity, the Wright lab sought to identify new glycopeptide antibiotics that function *via* a novel mode of action instead of the well-known lipid II binding activity.[109]

To accomplish this, Wright and coworkers focused on 71 glycopeptide antibiotic BGCs. Hypothesizing that phylogenetic diversity leads to molecular diversity, they generated phylogenetic trees that were based on condensation domains of the C-2 module rather than the phylogeny of the organism itself.[109,110] The presence of known antibiotic resistance genes in the cluster was then mapped onto the phylogenetic tree, revealing a large conserved clade of traditional glycopeptide antibiotics with the anticipated lipid II binding mode of action. Further examination of the tree revealed two divergent clades that lacked an identifiable resistance gene. One of these produced the known molecule complestatin (**21**),[111] while the other produced a novel compound that was named corbomycin (**22**). Even though these compounds belonged to a known family of glycopeptide alkaloids, the lack of a recognizable resistance gene suggested a novel mode of action. Indeed, corbomycin and complestatin do not bind to lipid II and inhibit peptidoglycan biosynthesis. Instead, they bind to peptidoglycan itself and prevent cell wall remodeling by peptidoglycan hydrolases called autolysins.[112] These enzymes are essential for cell growth and division, and their inhibition represents a novel mode of action. Overall, this study demonstrates that a combination of genome mining and knowledge of the resistance mechanism can facilitate discovery of new molecules from a known compound family with unprecedented modes of action.



21, complestatin
glycopeptide
antibiotic (binds peptidoglycan)



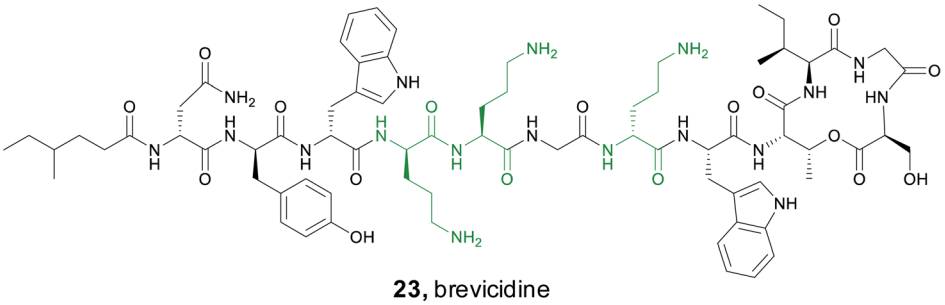
22, corbomycin
glycopeptide
antibiotic (binds peptidoglycan)

3.2 Cationic nonribosomal peptides

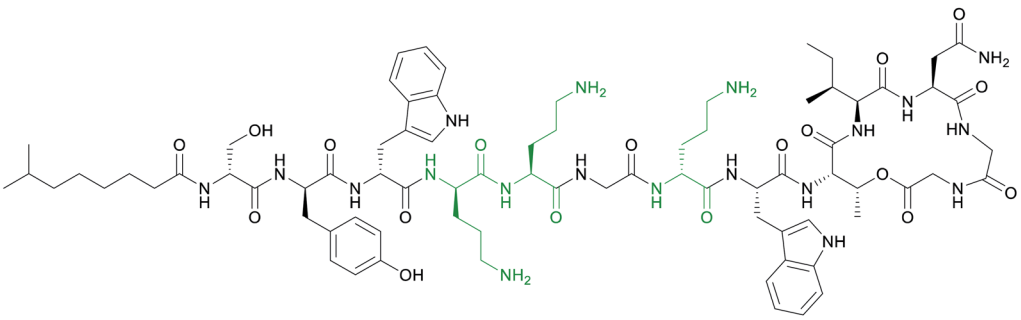
Recently, a major public health crisis has revolved around Gram-negative bacteria and their growing resistance to many antibiotics. Cationic non-ribosomal peptides (CNRPs) have become a prominent contender for treatment of these pathogens.[113] Members of this natural product family are rich in positively charged amino acids, allowing them to cross the outer membrane and interact with anionic targets. To discover new NRPS products rich in cationic residues, the Qian group utilized an antiSMASH[9] based genome mining approach.[114] They analyzed roughly 7400 complete or draft bacterial genomes from public databases using antiSMASH to search for putative NRPS biosynthetic gene clusters. Based on the predicted substrate specificity of the adenylation domains found in the gene clusters, they anticipated that 807 of the non-ribosomal peptides could be classified as CNRPs. They were all at least 6 amino acids in size and contained at least two positively charged amino acids, such as arginine, lysine, and histidine.

Using the results from the antiSMASH data, the Qian group generated a protein sequence similarity network (SSN) of the predicted peptide sequences and concluded that 91% of these cationic peptides were not closely related to a known CNRP.[114] Amongst all CNRPs, some *N*-acylated CNRPs (cationic lipopeptides), such as the tridecaptins, hold particular promise as a treatment for Gram-negative pathogens.[115] A specific search for cationic lipopeptides revealed that they represent 32% of all cationic non-ribosomal peptides and 85% of these did not cluster with a known CNRP. Following this genome mining, the Qian group focused their isolation efforts on clusters from two bacilli, *Brevibacillus laterosporus* and *Paenibacillus alvei*. They successfully isolated and characterized two cationic lipopeptides, later named brevicidine (**23**) and laterocidine (**24**).[114] Both consist of a linear, positively charged section with three ornithine residues and a small hydrophobic peptide ring, representing a novel class of cyclic cationic peptides.

To fully characterize these newly isolated compounds, bioactivity assays were completed against ESKAPE pathogens, which are often the cause of nosocomial infections.[116] Both brevicidine and laterocidine showed significant *in vitro* efficacy against these Gram-negative pathogens with no resistance development. Further *in vivo* studies in a mouse model demonstrated these compounds were effective in treating an *E. coli* infection. While the exact mechanism by which they operate is still being researched, brevicidine and laterocidine appear to be promising leads due to their *in vivo* efficacy, selectivity, and a low rate of bacterial resistance.



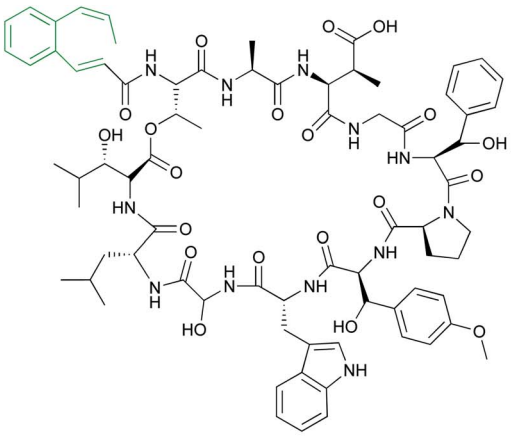
23, brevicidine
cationic nonribosomal peptide
antibiotic



24, laterocidine
cationic nonribosomal peptide
antibiotic

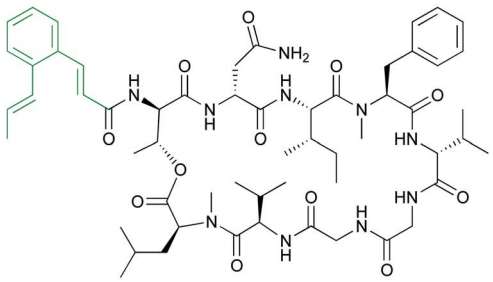
3.3 Cinnamoyl-containing nonribosomal peptides

In some cases, an unusual chemical moiety is not the pharmacophore itself, but is often associated with bioactive natural products, and therefore becomes the defining feature of a compound family. An interesting example of this correlation is the 2-[1-(*Z*)-propenyl]-cinnamoyl moiety. This substructure is found in only a small number of PKS/NRPS natural products, but each compound in this family has a different bioactivity. For instance, skyllamycin (**25**) is a platelet-derived growth factor inhibitor,[117] WS9326 is a tachykinin receptor agonist,[118] mohangamides inhibit isocitrate lyase,[119] and coprisamides enhance quinone reductase activity.[119] The Ge Lab capitalized on the bioactivity of this compound family and conducted a study to search for new cinnamoyl-containing nonribosomal peptides.[120] The cinnamoyl moiety itself is constructed from a type II PKS as a linear polyene that is isomerized and then cyclized.[117,121,122] To mine for new examples of this family of natural products, a series of SSNs were created that included different PKS subunits from polyene producing type II PKSs, type I PKSs, aromatic type II PKSs, and type II fatty acid synthases (FASs). The Ge Lab found that not all the polyene PKS subunits that generate the cinnamoyl structure are consistently clustered together. Instead, only the KSα subunit was a reliable predictor. In addition to the PKS architecture, they found that both the isomerase responsible for introducing the *cis* bond into the polyene chain and ketoreductase domain was diagnostic of the cinnamoyl motif compared to simple linear polyenes. Using these three domains, the Ge Lab then conducted a bioinformatic search for bacteria in the NCBI and JGI databases that contained both the KSα subunit and isomerase. These three genes were present in 192 bacteria and further filtering with antiSMASH identified 51 gene clusters that contained an intact type II PKS adjacent to an NRPS. A GNN was then generated and revealed that only two of the 51 gene clusters appeared similar to known natural products and suggested there were opportunities for discovery of novel bioactive molecules. The Ge Lab chose to focus on a glycosyltransferase containing cluster from *Kitasatospora* sp. CGMCC 16924. Gene knockout and isolation studies lead to the discovery of six compounds named kitacinnamycin A–F. The isolated kitacinnamycins A–F along with intermediates obtained from the knockout and biosynthetic studies were tested for stimulator of interferon genes (STING) activation. One of the intermediates, kitacinnamycin H (**26**) was an effective activator of the STING signaling pathway. These results further solidify the value of targeting the cinnamoyl containing NRPS family for the discovery of diverse bioactive molecules.

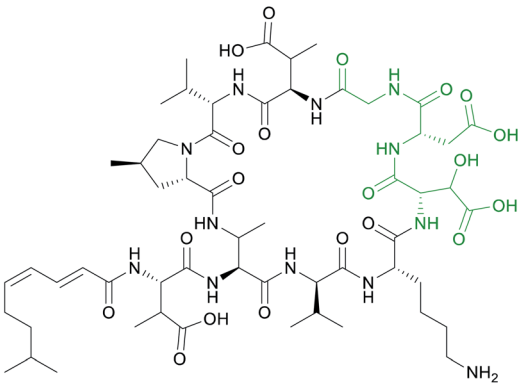


25, skyllamycin A
cinnamoyl moiety
growth factor inhibitor

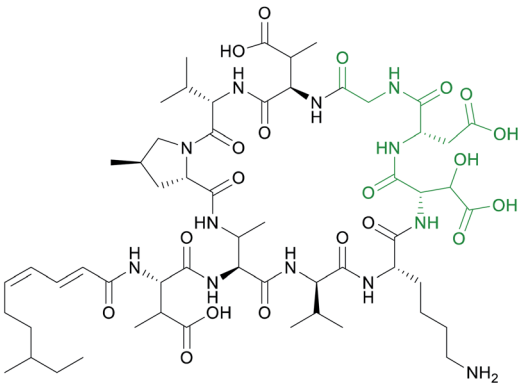
Genome mining target:
KSα subunit



26, kitacinnamycin H
cinnamoyl moiety
STING activator



27, malacidin A
calcium dependent antibiotic
antibiotic (binds lipid II)



28, malacidin B
calcium dependent antibiotic
antibiotic (binds lipid II)

3.4 Calcium dependent antibiotics

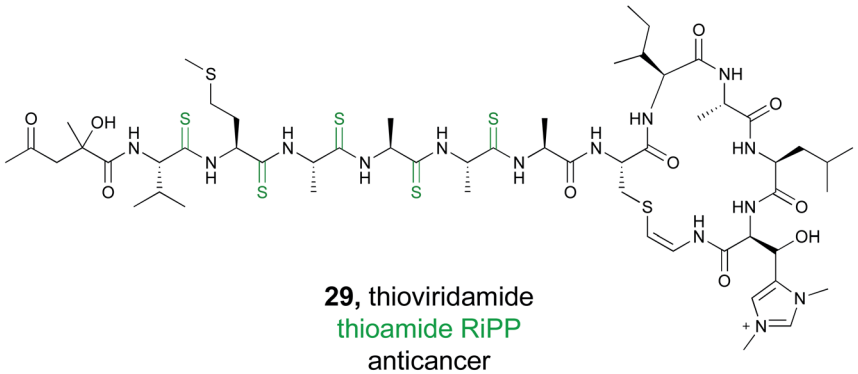
Calcium dependent antibiotics are NRPS derived natural products that require calcium for activity.[123] They typically contain an aspartate rich motif that is responsible for calcium binding, which is required for penetration of the bacterial cell membrane.[124] Because calcium binding defines this class of peptidic natural products, members have diverse modes of action including targeting of cell wall biosynthesis or cell membrane integrity.[123] They also hold great promise as new antibiotics. The most well-known calcium dependent antibiotic, daptomycin, has been clinically proven to be an effective treatment of Gram-positive bacterial infections[125] and has been classified as a critically important antimicrobial for human medicine by the World Health.

A recent study by the Brady Lab sought to discover new calcium dependent antibiotics by mining soil metagenomes for BGCs that should produce non-ribosomal peptides with the characteristic Asp–X–Asp–Gly calcium binding motif.[126] However, simply shotgun genome sequencing each soil sample would not provide enough sequencing depth to reliably detect low abundance BGCs. Instead, the Brady lab used a PCR based approach to enhance the sensitivity of detection of target biosynthetic gene clusters within a particular soil sample.[127,128] Specifically, they targeted conserved regions in the adenylation domains of non-ribosomal peptide synthetases to generate next-generation sequencing reads.[126] Using eSNaPD,[68] the Brady lab analyzed the sequencing data for a conserved adenylation domain responsible for the first Asp in the Asp–X–Asp–Gly calcium binding sequence. Phylogenetic analysis indicated numerous sequencing tags fell into new clades, suggesting novel products. For detailed study, they decided to pursue a particularly abundant clade found in 19% of the metagenomes and named it the malacidins (metagenomic acidic lipopeptide antibiotic-cidins). Traditional cloning followed by cosmid screening led to the recovery of the malacidin gene

cluster. Heterologous expression permitted isolation and structure elucidation of malacidin A and B (**27** and **28**), cyclic lipopeptides that typically differ by a methylene on the lipid tail terminus and four non-coded amino acids at the core. Although the malacidins were found to be calcium dependent, they do not contain this conserved Asp–X–Asp–Gly motif, but rather have a different motif, 3-hydroxylAsp–Asp–Gly. Further experiments demonstrated that the malacidins target lipid II, have a low rate of developed resistance, and could be effective against multidrug resistant Gram-positive pathogens.[126]

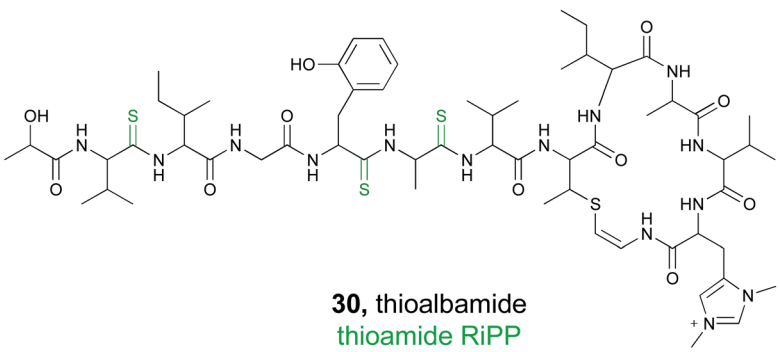
3.5 Thioamide RiPPs

Typically used by synthetic chemists to form stable isosteres, thioamides are a rare functional group in natural products, but

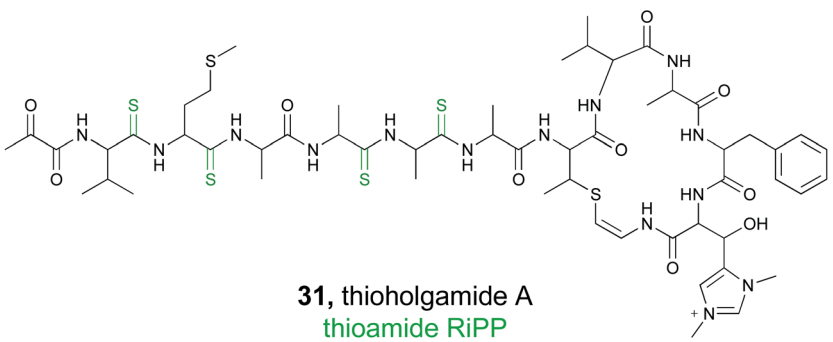


29, thioviridamide
thioamide RiPP
anticancer

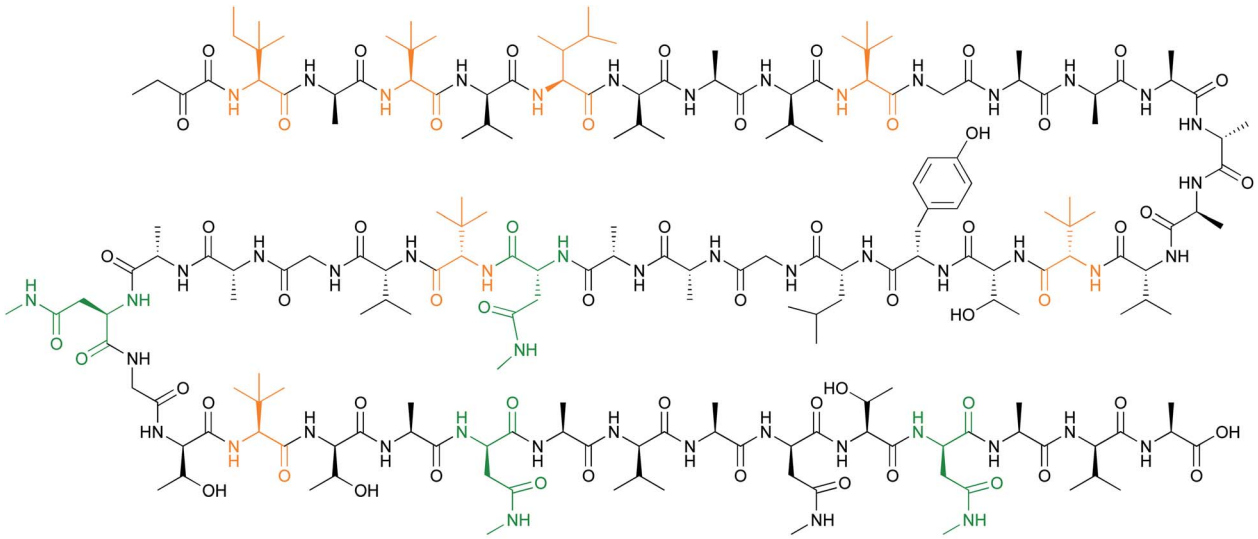
Genome mining:
Tva biosynthetic genes



30, thioalbamide
thioamide RiPP
anticancer



31, thioholgamide A
thioamide RiPP
anticancer



32, aeronamide A
pore forming proteusin, cytotoxic
Residues in orange are methylated, but the exact location is unknown.
A possible location of the methylation is shown.

one with recognized bioactivity.[129,130] Famously thioamide-containing drugs have been used to treat thyroid disorders.[131] Because of this, the biological activity tied to thioamide moieties found in natural products is currently under investigation. Thioviridamide (**29**) is one such thioamide-containing RiPP natural product with potent anticancer activity.[132] Due to the unusual chemical structure and intriguing bioactivity of thioviridamide, two research groups in 2017 almost simultaneously undertook very similar genome mining approaches to expand the thioviridamide family of natural products.[133,134]

The Truman group chose to mine bacterial pathways using a YcaO domain protein, TvaH, found in the recently identified thioviridamide BGC (*tva*) as a genome mining hook.[133,135] While TvaH was not fully functionally characterized at the time, YcaO proteins are typically involved in the biosynthesis of thiazoline and (methyl)oxazoline moieties *via* an ATP-dependent cyclodehydration reaction, and so TvaH was predicted to play a role in thioamide biosynthesis.[136] A BLAST search for TvaH homologs followed by MultiGeneBlast ultimately identified 14 thioviridamide-like BGCs from bacterial genomes. Five publicly available strains were selected for fermentation, and three thioviridamide-like molecules were isolated and structurally characterized. These compounds were designated thioalbamide (**30**), isolated from *Amycolatopsis alba* DSM 44262, and thiostreptamides S4 and S87, isolated from *Streptomyces* sp. NRRL S-4 and NRRL S-87, respectively. All three compounds contained characteristic thioamide moieties present in the peptide backbone. Importantly, thioalbamide was found to be highly bioactive, with an IC_{50} in the nanomolar range, but with high levels of selectivity for tumor cells as opposed to healthy human cells.

Concurrent with this study, the Müller group utilized a similar approach to identify novel members of the thioviridamide compound class.[134] Rather than genome mining using the TvaH enzyme putatively involved in thioamide synthesis, they utilized the thioviridamide precursor peptide TvaA to probe sequence databases. This mining effort yielded 13 hits that resembled the thioviridamide BGC, and they obtained four of these strains for further investigation. Notably, three of these four strains were also identified by the Truman group. However, fermentation of the fourth strain, *Streptomyces malaysiense* MUSC 13657, resulted in the production, isolation, and structural characterization of two novel compounds, thioholgamide A and B. Thioholgamide A (**31**) showed anticancer activity more pronounced than that of thioviridamide A. Both groups were thus independently able to use a genome mining-based approach, utilizing different biosynthetic enzymes, to expand a family of bioactive compounds.

3.6 Proteusins

Some bioactive natural products do not have a single important chemical feature, but instead the entire structure acts in concert to produce the observed bioactivity. An example of natural products that fall into this category are pore forming molecules, such as the polytheonamides. These highly modified peptides were first isolated from sponges and shown to be highly cytotoxic.[137–139]

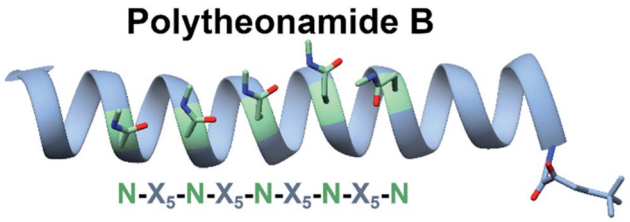
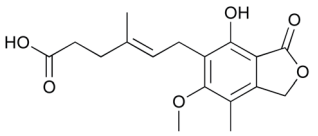


Fig. 2 NRM solution structure of polytheonamide B (PDB: 2RQO) reveals N-methyl Asns are found on one face of the β-helix. This repeating N–X_5–N motif was targeted for genome mining efforts.

While initially hypothesized to be biosynthesized by NRPS machinery, they were later demonstrated to actually be members of the proteusin class of RiPP natural products.[140,141] These long peptides are nearly 50 amino acids in size with extensive epimerizations and methylations that promote formation of a β-helix.[142] This lipophilic structure can act as a cation-channel that inserts into membranes.[143] However, efforts to study and engineer the entire biosynthetic pathway were hindered by the fact that the producing organism, *Candidatus* Entotheonella factor, is a bacterial sponge symbiont and unculturable. To solve this problem, the Piel lab sought to find alternative sources of proteusins from culturable organisms that still retained the potent cytotoxicity observed in polytheonamide.[144] Previous NMR[142] and molecular dynamics[145] studies indicated that the bioactive β-helix structure is stabilized by evenly spaced *N*-methylated Asn residues defined by an N–X_5–N motif (Fig. 2). Therefore, the Piel lab conducted a bioinformatic search for clusters that contained an N–X_5–N–proteusin precursor along with an epimerase and methyltransferase.[144] Three clusters from diverse bacteria were identified, one of which came from the culturable bacteria *Microvirgula aerodenitrificans*. After establishing a conjugation-based plasmid system, production of the expected aeronamide A (**32**) product was confirmed using a hybrid *in vivo*/*in vitro* production system. Because aeronamide A contained the critical repeating N–X_5–N motif that promotes β-helix formation, it possessed high cytotoxic activity (IC_{50} of 1.48 nM in HeLa cells). Moreover, the ability to create a genetic system for *M. aerodenitrificans* enabled processing of alternative precursor peptides to make non-native products.

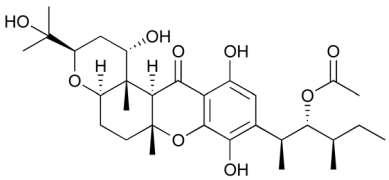
3.7 Meroterpenoids

Meroterpenoids are hybrid natural products that arise from a combination of terpenoid biosynthesis coupled with structural features originating from alternative biosynthetic origin.[146–148] These compounds are ubiquitous in nature, found across all domains of life, and comprise a tremendous variety of structural diversity.[149] Importantly, many of these compounds are well recognized for their bioactivity. Fungal meroterpenoids, such as the clinically used immunosuppressant mycophenolic acid (**33**), demonstrate the potential utility of this compound class.[148] Because of the tremendous structural diversity and high rate of bioactivity in this compound class, Zhang *et al.*, decided to capitalize on the known biosynthetic origins of this family to discover new, likely bioactive, fungal meroterpenoids.[150]



33, mycophenolic acid
meroterpenoid
immunosuppressant

Genome mining:
PKS, PT, TC



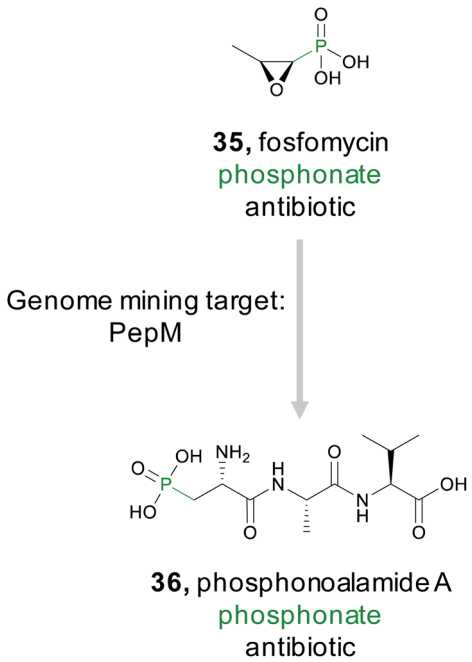
34, arthripenoid C
meroterpenoid
immunosuppressant

The fungal biosynthesis of meroterpenoids includes the co-clustering of genes encoding polyketide synthases, prenyltransferases (PTs), and terpenoid cyclases (TCs).[151] By exploiting this knowledge, the authors were able to mine fungal genomes from their extensive strain collection for the co-occurrence of these genes, which led to the discovery of two homologous BGCs containing all three required biosynthetic elements.[150] Fermentation of these strains resulted in the production of ten related meroterpenoids, eight of which represent novel compounds. These molecules were then definitively linked back to their gene clusters *via* gene inactivation experiments. Six of these meroterpenoids, isolated from *Arthrinium* sp. NF2194, were named arthripenoid A–F and were used as a model system to explore the biosynthesis of this compound family. A series of gene inactivation experiments coupled with *in vitro* protein expression and enzymatic activity assays were used to parse out the biosynthetic route to these compounds. Finally, assays demonstrated that arthripenoid C (**34**) is capable of immunosuppressive activity on CD4+ T cells. By targeting proximal genes for meroterpenoid biosynthesis, the authors were able to identify two similar BGCs in phylogenetically distinct fungi and isolate bioactive natural products. Thus, genome mining using the key enzymes that biosynthesize a compound family previously recognized for its bioactivity as a hook has proven successful in prioritizing BGCs of interest in the hunt for novel therapeutics.

3.8 Phosphonates

The phosphonate class of natural products are characterized by the presence of a stable C–P bond that mimics the labile phosphoester bond found in numerous phosphorylated primary metabolites.[152] These natural products have diverse bioactivities

including antifungal, antimalarial, antibacterial, and herbicidal. Several commercial products, such as the clinical antibiotic fosfomycin (**35**), are also members of this class. While the presence of a phosphonate bond defines this class, this molecular feature is not necessarily a pharmacophore unlike the microcin C7 compounds where the stable phosphoramidate serves as the critical ligand binding feature for bioactivity (Section 2.2.3). For example, the epoxide ring in fosfomycin serves as an electrophile to form a covalent adduct with UDP-GlcNAc enolpyruvyl transferase (MurA),[153] thereby granting the molecule its potent bioactivity. In contrast, FR900098 and fosfomidimycin are tight binding competitive inhibitors of D-1-deoxyxylulose-5-phosphate reductoisomerase, a key enzyme in the non-mevalonate pathway for isoprenoid biosynthesis.[154]



Because phosphonates have been a prolific source of potent natural products that utilize diverse modes of action, there have been concerted efforts to identify novel members of this natural product family. Genome mining for phosphonates relies on the first committed step of the biosynthesis which is nearly always catalyzed by phosphoenolpyruvate mutase (PepM).[155] A detailed investigation and discovery by the Metcalf group effort revealed that the phylogeny and sequence identity of this single gene can be diagnostic for identifying new classes of phosphonates. Specifically, *pepM* genes that share at least 80% sequence identity are likely to be found in gene clusters that produce related compounds.[156,157] This observation led to strain prioritization and discovery of several new bioactive phosphonates.[157] A recent report from the Ju lab applied this observation using a genome first approach. They first mined *pepM* sequences from all public databases to identify 448 putative phosphonate producing gene clusters.[158] A SSN was then generated using the 80% sequence identity threshold to generate 113 gene cluster families predicted to produce novel compounds. The PepM reaction is thermodynamically unfavorable and requires a coupled secondary reaction, such as decarboxylation, to push the reaction forward.[152,155] Therefore, to identify a novel phosphonate scaffold, the Ju lab examined each of the 113 gene cluster families for enzymes that catalyze the thermodynamically favored biosynthetic step. In 6% of gene cluster families, no candidate enzyme was identified. Examination of the largest of these, an 11-membered gene cluster family, lead to the discovery of the phosphonoalamides (**36**), a new class of phosphonates that contains a phosphoalanine moiety and demonstrates antibacterial activity.

4. Target directed genome mining

Both the bioactive feature targeting approach (Section 2) and the compound family strategy (Section 3) take a "chemistry first" approach to genome mining, where a bioactive chemical feature or family is known prior to any genome mining attempts. However, researchers can also take a "gene first" approach to genome mining to prioritize specific BGCs over others for further investigation. In this strategy, there is no known compound guiding the genome mining. Instead BGCs are prioritized from sequence data alone that portend function or bioactivity. Recent prioritization methods have relied on the fact that bioactive molecule producing organisms must be resistant to their own products to avoid inadvertent suicide.[159,160] Microorganisms typically utilize three main strategies to avoid self-toxicity: (1) product detoxification, (2) binding and removal of the product *via* transports and (3) target duplication or modification, where an organism encodes a duplicated but resistant copy of the target gene. In each case, genes encoding these resistance strategies are often colocalized in the cognate BGC. The third mechanism described above has led to a new strategy for genome mining that enables prioritization of BGCs with potential bioactivity. By searching for duplicated housekeeping genes located near BGCs, it is possible to identify BGCs with predicted bioactivity towards that target (Fig. 3). This approach is therefore chemistry independent, as the BGC is not identified for any specific chemical feature, but rather by the

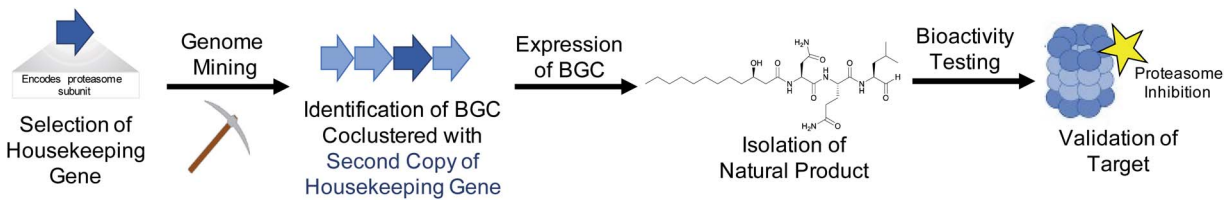
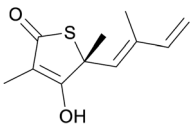


Fig. 3 By mining for the presence of duplicate housekeeping genes, it is possible to discover natural products and their biological target simultaneously.

presence of a colocalized duplicated housekeeping gene. This method to prioritize BGCs with potential bioactivity has been termed "target-directed genome mining" as well as "resistance-gene genome mining". In this section we will discuss four examples of target-directed genome mining that led to the directed discovery of natural products with anticipated bioactivity.

4.1 Thiotetronic acid

In 2015, the Moore lab successfully used a target-directed genome mining approach with the express purpose of natural product discovery for the first time.[161] They proposed that by identifying gene clusters that also contained a duplicated copy of a housekeeping gene that serves as a mechanism of self-resistance, they could prioritize bioactive BGCs and discover natural products with bioactivity towards that particular target.[162,163] By querying the genomes of 86 *Salinispora* strains, the Moore group was able to identify 912 orthologous groups (OGs) that were both duplicated copies of core housekeeping genes *and* were associated with previously identified BGCs. They were particularly interested in BGCs that contained a duplicated copy of proteins in the fatty acid synthase (FASII) complex, as this is an attractive drug target due to its lack of similarity to fatty acid synthesis in mammals, which would result in ideal selectivity for a potential antibiotic.[164–166] One gene (*tlmE*) encoding a FabB/F homolog was promising due to its high sequence similarity to the previously characterized self-resistance proteins PtmP3 and PtnP3 found in other BGCs that produce FASII inhibitors.[167] That suggested that this gene was indeed a resistance gene, and not involved in biosynthesis.

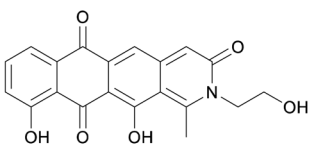


37, thiolactomycin
Resistance gene: FabB/F
Target: Fatty acid biosynthesis

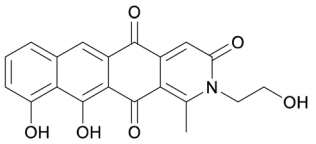
The putative FASII resistance gene was found co-localized within a hybrid ~22 kb PKS/NRPS BGC that was captured by transformation-associated recombination (TAR) cloning in yeast and expressed in a *Streptomyces* heterologous host. This BGC produced four natural products that share a rare thiotetronic acid moiety, including the previously characterized antibiotic thiolactomycin (TLM, **37**).[168,169] This discovery ultimately allowed the authors to use this resistance gene (*tlmE*) as a genome mining hook to discover other, related, natural products from *Streptomyces*. A second gene cluster (*ttm*) contained two copies of a *tlmE* homolog, and subsequent TAR cloning and heterologous expression resulted in the production of a new series of TLM analogs. This work utilized a systematic approach to identify BGCs that might produce bioactive natural products based on the presence of a duplicated housekeeping gene found within the cluster that confers resistance to the BGC product. This target-directed genome mining strategy allows for the discovery of BGCs without prior knowledge of the molecules produced, and it has since been used extensively to prioritize BGCs that potentially produce bioactive compounds. Finally, this method is now the basis of the ARTs (Antibiotic Resistant Target Seeker) program, which allows for efficient *in silico* genome mining based on putative bioactivity.[170]

4.2 Pyxidicyclines

Myxobacteria are ubiquitous but incredibly interesting soil dwelling bacteria with large genomes and complex life cycles, including the tendency to form multicellular fruiting bodies.[171] These organisms have long been overlooked as sources of natural products, but genomic analysis suggests they are a prolific source for new molecules.[172] The Müller group utilized a target-directed approach to mine the genomes of 93 myxobacterial strains to find novel topoisomerase inhibitors.[173] Topoisomerase inhibitors are used as antibiotics as well as chemotherapeutics, depending on the specificity of the inhibitor. The Müller group utilized CysO, a topoisomerase-targeting pentapeptide repeat protein responsible for self-resistance against the gyrase inhibitor cystobactamid, as their genetic search hook.[174] This pentapeptide repeat protein mimics the structure of DNA, thereby protecting DNA gyrase from inhibitors like cystobactamid. Their search revealed 31 candidate sequences, eight of which were located near uncharacterized BGCs. They became particularly interested in a candidate gene, which they named *pcyT*, located next to a BGC that putatively encoded a type II PKS. Type II PKS BGCs are rare and uncharacterized in myxobacteria, making this cluster an interesting target for further investigation. Because the gene cluster was silent in the wild type organism, the authors pursued an activation *via* promoter exchange strategy. By replacing two native promoters in the BGC, the authors were ultimately able to isolate two novel type II PKS-derived nitrogen-containing tetracene quinones which they named pyxidicycline A and B (**38** and **39**). The pyxidicyclines exhibited strong cytotoxicity with activity in the nanomolar range, which prompted the authors to pursue a heterologous expression strategy to facilitate investigation of this BGC. The Müller lab was thus able to use a target directed approach to identify and prioritize a BGC predicted to encode a topoisomerase inhibitor, isolate two chemically novel products, and confirm they had the expected topoisomerase inhibitory activity.



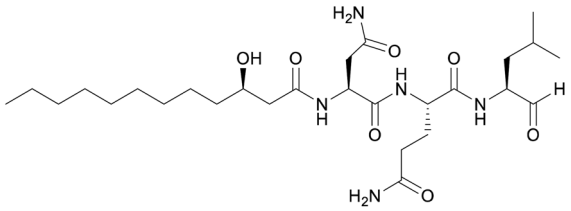
38, pyxidicycline A



39, pyxidicycline B
Resistance gene: pentapeptide repeat protein
Target: topoisomerase

4.3 Fellutamide B

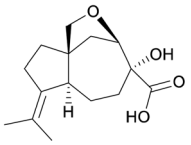
While the resistance gene genome mining strategy was initially utilized in bacteria, it has also been successfully applied to fungi as well. A 2016 study from the Wang and Oakley groups applied a target-directed genome mining approach to fungi by examining uncharacterized BGCs in *Aspergillus nidulans*.[175] They identified one BGC of interest containing a gene (*inpE*) that encodes a putative proteasome subunit. This suggested that the gene cluster might encode a proteasome inhibitor, which is of particular interest due to the clinical success of proteasome inhibitors such as the marine natural product salinosporamide A.[176] However, this cluster was silent in the native organism, and no small molecule products had been identified. The authors utilized a serial promoter replacement strategy to activate this BGC in the native producing organism, replacing every native promoter in the cluster with the regulatable *alcA* promoter. Ultimately, this led to the production of the tripeptide fellutamide B (**40**), a potent proteasome inhibitor that had been previously isolated from fungi,[177] although no gene cluster had been described. Interestingly, the authors also found that the resistance gene, *inpE*, was constitutively expressed and not coregulated with the other genes of the *inp* cluster, which indicates that the fellutamide-resistant proteasome subunit is always expressed. This work shows that target-directed genome mining is useful for the discovery, or rediscovery, of natural products from eukaryotic organisms as well.



40, fellutamide B
Resistance gene: proteasome subunit
Target: proteasome

4.4 Aspterric acid

Resistance gene-based genome mining is valuable not only because it can efficiently link a natural product to a target, but it also provides a mechanism of action. The utility of this approach was recently demonstrated by the Tang group.[178] Their study specifically targeted dihydroxyacid dehydratase (DHAD), an enzyme in the branched chain amino acid (BCAA) biosynthetic pathway. No natural product inhibitors of DHAD were known at that time, but DHAD is a particularly compelling potential target for an herbicide as the BCAA pathway is essential for plants but not found in animals. DHAD, despite being highly conserved, has never been targeted by an herbicide.[179] The authors proposed that a fungal gene cluster which encodes a DHAD inhibitor may contain an additional copy of DHAD that provides a mechanism of self-resistance, as fungi do synthesize BCAAs *de novo*. *Via* genome mining, the authors identified three genes responsible for the biosynthesis of the tricyclic terpenoid aspterric acid (**41**) found co-clustered with a duplicated copy of DHAD, named *astD*. Although aspterric acid was a previously isolated molecule, until this work the molecular target of aspterric acid remained unknown, as did its biosynthesis.[180]



41, aspterric acid
Resistance gene: dihydroxyacid dehydratase (DHAD)
Target: branched chain amino acid biosynthesis

Additional structural biology work of the wild type fungal DHAD as well as the duplicated copy *astD* allowed the researchers to understand the mechanism of *astD* resistance to aspterric acid on a molecular level.[178] This understanding enabled their development of a trans gene system, where *astD* was deployed in *Arabidopsis thaliana* to provide resistance to aspterric acid, which could then be used as an effective and selective herbicide, in the vein of the RoundUp/Roundup Ready glyphosate system.[181] Although aspterric acid is a previously known compound, this paper is an excellent example of the ability to repurpose known compounds by uncovering their mechanism of action through a target-directed genome mining approach.

5. (Bio)synthetic production of genome mined natural products

With the abundance of available bioinformatic tools, it has become relatively straightforward to identify a new putative natural product gene cluster *in silico*. The challenge then becomes obtaining the producing organism, growing it under conditions that permit production of the compound, and isolating enough of the natural product for characterization and bioactivity testing. If interesting bioactivity is observed, the compound must then be produced on scale using chemical synthesis, fermentation, or a mixture of the two approaches. Recent studies that we discuss here have demonstrated that this workflow can be avoided all together. Instead of growing the producing organism to isolate the target compound, the natural product is bioinformatically predicted and then directly produced, either by chemical synthesis or with promiscuous biosynthetic enzymes (Fig. 4). This approach avoids the issues of production, isolation, and scalability. The caveat to this methodology is that there needs to be sufficient understanding of the biosynthetic enzymes to predict a product. Moreover, the new compound may not directly match the authentic natural product but should at least closely resemble it. Here we describe examples that demonstrate the utility of this (bio)synthetic production approach to discover potent antibacterial agents from mined gene clusters.

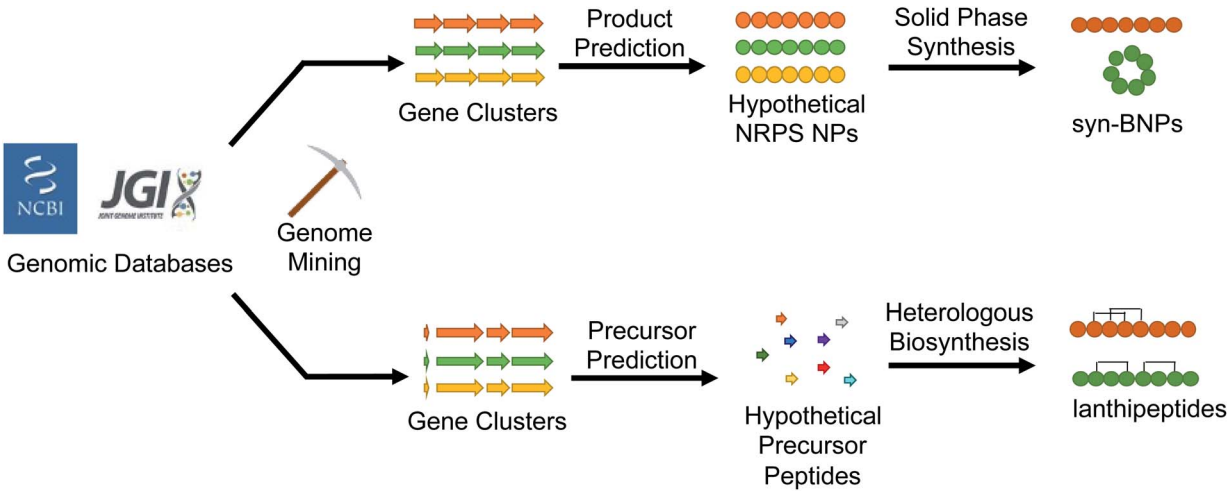
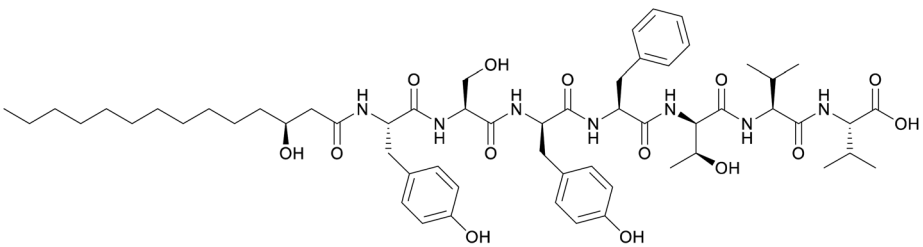


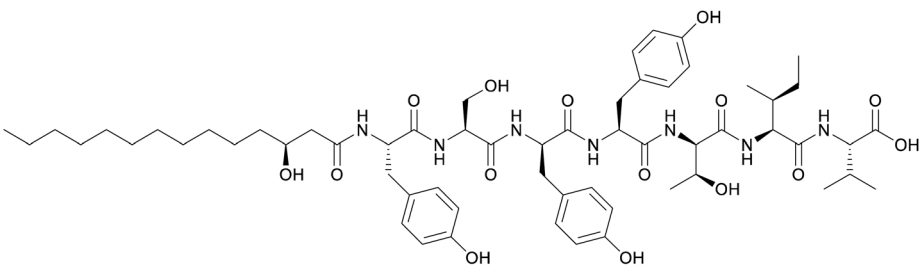
Fig. 4 Bioinformatically identified natural products can be produced either synthetically or biosynthetically to generate new molecules that mimic the authentic compound.

5.1 Antibacterial syn-BNPs from the human microbiome

The human gut microbiome is home to a vast array of diverse bacterial species that produce an extensive collection of small molecules. Some of these compounds appear to directly interact with our bodies and have important consequences for our health. Others may be antibiotics that help provide a growth advantage against the competitive human gut environment. Studying these small molecules can be challenging because many of the bacteria are obligate anaerobes or difficult to cultivate. However, the extensive analysis of sequencing data from the human microbiome demonstrates that numerous biosynthetic gene clusters do exist and could provide a wealth of useful compounds.[182] In 2016, the Brady lab described their development of the synthetic-bioinformatic natural products (syn-BNPs) approach to address this problem.[183] They decided to focus on the NRPS class of natural products for several reasons. First, the bioinformatic prediction of NRPS clusters and adenylation module specificity is relatively mature and accurate.[184,185] This allows for successful *in silico* prediction of at least a close approximation of the authentic natural product. Second, the predicted molecules can be rapidly produced by



42, humimycin A



43, humimycin B

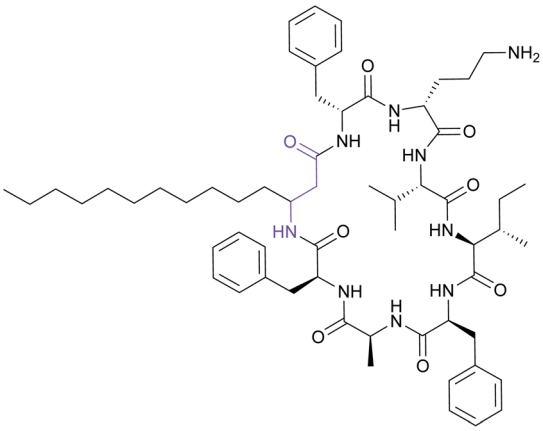
solid phase peptide synthesis or purchased at a low cost. Therefore, the Brady lab mined genomes from both the Human Microbiome Project and Human Oral Microbiome Database for the presence of NRPSs.[183] They culled the mined NRPSs to only include those that produced products at least five amino acids in size. These larger products were thought to be less heavily modified and therefore more amenable to solid phase synthesis. In total, 25 of the 30 targeted peptides were successfully synthesized. After screening against a panel of bacterial strains, two syn-BNPs possessed promising antibacterial activity and were named humimycin A and B (**42** and **43**). Critically, these two related peptides were not detected in the host organisms, demonstrating the utility of this approach is accessing natural products without relying on the host organism.

5.2 Design of cyclic syn-BNPs

Initial studies on syn-BNPs focused on synthesizing and testing only linear NRPS derived peptides. However, the majority of NRPS natural products are cyclized in some way. Therefore, the Brady lab expanded upon their initial work with two studies that examine cyclic syn-BNPs. In the first study, the genomes of human-associated bacteria were again mined for the presence of large NRPS gene clusters to generate a list of 25 candidate peptides.[186] Bioinformatic analysis of NRPS domains can predict the peptide sequence of the resulting natural product with a relatively high degree of certainty. However, current software programs are unable to accurately predict the mode of cyclization. Multiple different moieties can serve as the nucleophile in the offloading process, such as the N-terminal amine or a Ser/Thr sidechain.[187] To account for this ambiguity, each peptide was synthesized with up to three cyclic scaffolds to generate 72 syn-BNPs. These compounds were used in assays targeting both Gram-positive and Gram-negative bacteria, fungi, and HeLa cells. While no compounds were found to be active in the antibacterial or antifungal assays, five of the cyclic syn-BNPs elicited a response in the HeLa cell MTT metabolic activity assay.

In a second study, the Brady lab completed a larger scale analysis of cyclic syn-BNPs.[188] Analysis of 3000 bacterial genomes led to 96 NRPS natural product predictions. Instead of synthesizing every possible cyclic scaffold for the 96 peptide targets, a series of rules was created based on examining authentic NRPS natural products. Of 171 designed peptides, 157 were successfully synthesized and tested in a semi-pure form for antibiotic activity against a panel of bacteria that included the ESKAPE pathogens. After initial testing and subsequent validation, nine of the cyclic syn-BNPs demonstrated promising activity. Further analysis indicated that only two of the nine compounds possessed cytotoxic activity, indicating that the majority are specifically bacterial antibiotics. Detailed mode of action studies revealed diverse mechanisms of bioactivity including cell lysis, inhibition of cell wall biosynthesis, cell membrane depolarization, and dysregulation of the ClpP protease. Amongst the compounds, mucilasyn (**44**) was particularly promising due to an unidentified mode of action, low rate of resistance, and efficacy against *Acinetobacter baumannii*.

A recent study by the Parkinson lab used a related strategy to synthesize a library of predicted natural products.[189] By focusing on NRPs that are cyclized by penicillin binding protein-like cyclases, the authors could better hypothesize the authentic cyclic scaffold and were able to ultimately identify 14 cyclic peptides with antibiotic activity out of a library of 51.

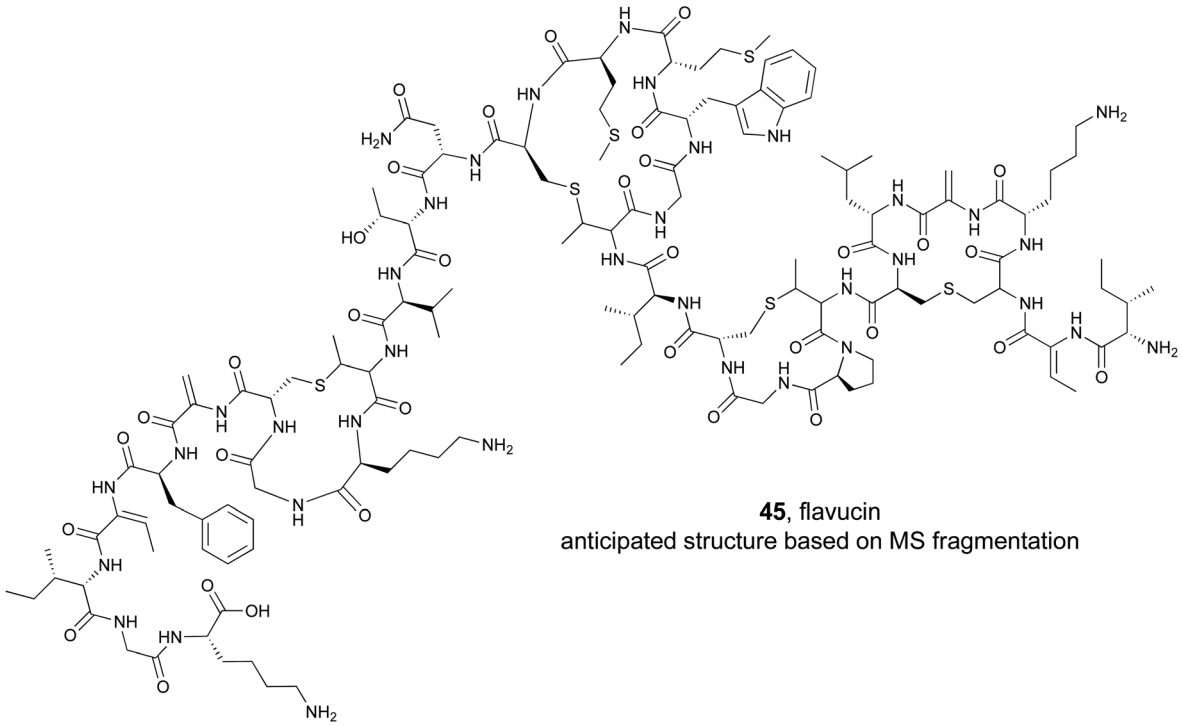


44, mucilasyn
cyclization site

5.3 Lanthipeptides

RiPP natural products are particularly intriguing from a genome mining perspective because the substrates for the biosynthetic pathways (precursors peptides) are directly encoded for in the gene cluster. Several published software programs have taken advantage of this and can rapidly discover thousands of putative precursor peptides *in silico*.[190] However, unlike the NRPS derived syn-BNPs, total synthesis of most RiPPs is incredibly challenging and cannot be completed in an automated, high-throughput manner. Instead, researchers have focused on using the inherent promiscuity of many RiPP biosynthetic enzymes to produce new RiPP compounds. Several studies have demonstrated this approach can work both in a random high throughput manner[191,192] or targeted approach using genome mined precursor peptides[193,194] to generate new bioactive lanthipeptide analogs.

Lanthipeptide natural products are defined by the presence of a (methyl)lanthionine bridge between a cysteine residue and a dehydrated serine or threonine.[195] They also typically contain additional dehydroalanine or dehydrobutyrine residues derived from serine or threonine respectively. To explore the activity of bioinformatically identified lanthipeptides, the Kuipers lab used a heterologous production method to evaluate 54 putative precursor peptides.[193] They employed an *E. coli* strain that was first transformed with a plasmid containing the core class I lanthipeptide biosynthetic genes (dehydratase, cyclase, and transporter). They added a second plasmid that contained a gene for one of the mined precursor peptides. The 54 target precursor peptides were chosen based on several features, such as the ability to predict the location of the leader cleavage site, and represented diverse examples. As with the syn-BNPs, these heterologously produced lanthipeptides will not necessarily be



45, flavucin
anticipated structure based on MS fragmentation

structurally identical to the authentic natural product but may be similar enough to still be bioactive. Of the 54 precursor peptides, 31 could be detected analytically. Of those, 18 appeared to be good substrates for the biosynthetic enzymes and were heavily modified. Bioactivity testing revealed that five were active against a panel of Gram-positive pathogens. This represents a nearly 10% success rate of the tested precursor peptide to bioactive molecule pipeline. MS/MS based structural analysis of one of the newly produced lanthipeptides, flavucin (**45**), demonstrated that (methyl)lanthionine bridges were formed and matched the predicted pattern observed in known lanthipeptides. Efforts to isolate the authentic flavucin from its native host, *Corynebacterium lipophiloflavum* DSM 44219, were unsuccessful as no production was observed. This again highlights the value of a (bio)synthetic production approach to study genome mined natural products because the host organism often does not produce the desired compound under laboratory conditions.

6. Conclusions and outlook

The 2020s have ushered in an era of biological revolutions, largely heralded by advancements in so-called 'omic' technologies. The convergence of these biological and technological advances has resulted in a sharp drop in the cost of sequencing DNA; in less than 20 years the cost of sequencing the human genome has plummeted, from millions of dollars in 2003 to just around $1000 today.[196] Sequencing a single bacterial genome costs just a fraction of that. Because of this unprecedented access to genomic data, genome mining has become integral to the field of natural products[12] and birthed a renaissance in natural product discovery.[197]

This review focused on four approaches taken by researchers for genome mining large genomic datasets with a particular focus on the discovery of bioactive compounds. Two of these strategies, bioactive feature mining and compound family mining, we categorize as "chemistry first" approaches. Both strategies require previously isolated chemistry to have exhibited interesting bioactivity, which allows known enzymatic reactions to be exploited for mining large databases for similar BGCs and their products, under the assumption that related compounds are likely to be bioactive as well. The other two strategies, target-directed genome mining and the (bio) synthetic production approach to genome mining natural products, we classify as "gene first" methods. These strategies rely on sequence data alone to either help prioritize a BGC based on the presence of a putative biological target, or to directly predict a chemical structure ultimately made *via* chemical or enzymatic synthesis.

By prioritizing bioactivity, these four strategies serve as the 21st century genome-oriented answer to traditional bioactivity-guided isolation strategies. These examples also make clear the continued importance of collaboration between the field of natural product discovery and other research disciplines. Target-directed genome mining will only improve as new protein targets for drug discovery are identified by cell

biologists. Bioactive feature targeting relies on the identification of these specific chemical moieties/pharmacophores, which speaks to the continued importance of mechanism of action and structure–activity relationship studies of natural products to explicitly determine what chemical feature endows a molecule with a specific bioactivity. As Table 1 indicates, there is room for bioinformaticians to introduce pipelines to mine for specific pharmacophores, much like AdenylPred did for β-lactone mining.[26] Similar programs could be developed for other bioactive features, such as epoxyketones or β-lactams. While programs like antiSMASH can indicate gene clusters that encode enzymes responsible for the installation of these features, the program is not oriented towards this purpose like AdenylPred is for β-lactones. Table 1 is designed to help the reader understand where these genome mining gaps may exist, and where feature-specific pipelines could be developed. To complement this, dedicated genome mining algorithms could be specifically designed to rapidly search through large genomic data sets and expand known bioactive natural product (sub) families. This approach has gained significant attention in the RiPP community with several family specific genome mining programs available.[189] Together, these types of specific tools could aid discovery pipelines and help identify new bioactive natural products.

Additional approaches could further exploit the fact that genes responsible for the installation of specific chemical features can sometimes be found in so-called "sub-clusters" within a larger BGC (Fig. 5).[198] These smaller biosynthetic units are interesting from an evolutionary perspective as they are thought to consist of co-evolving genes required for construction of a specific biosynthetic feature and they function almost like an independent unit. These sub-clusters may be found in

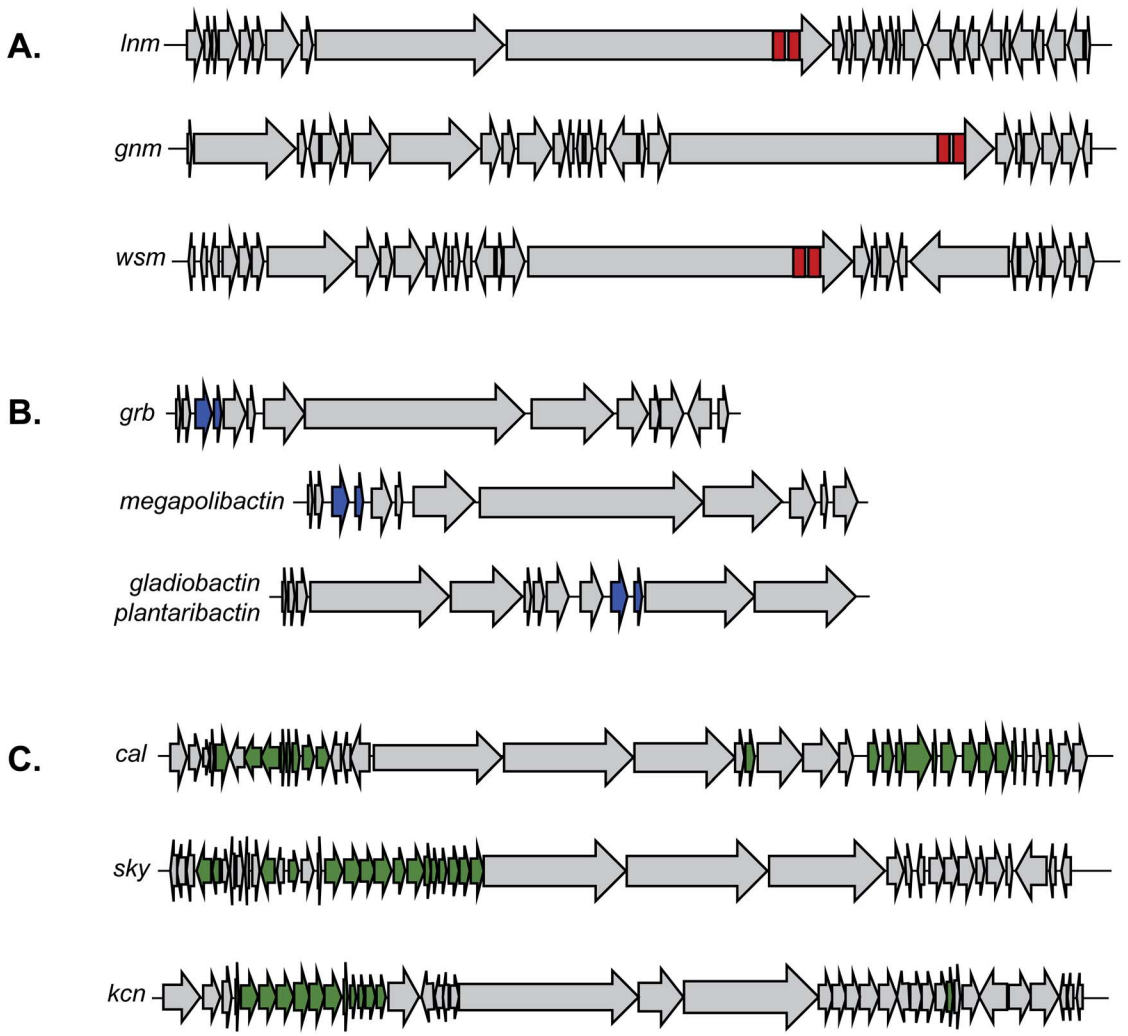


Fig. 5 Sub-clustering of genes is responsible for the installation of bioactive chemical features and class-defining chemical features. Gene clusters not shown to scale. (A) Sub-clustering of the DUF and cysteine lyase domains (red) is responsible for the installation of reactive 1,3-dioxo-1,2-dithiolane moiety and is seen in leinamycin (*lnm*) BGC, guangnanmycin (*gnm*) BGC, and weishanmycin (*wsm*) BGC (Section 2.1.4). (B) Sub-clustering of two genes *grbD* and *grbE* (blue) that are responsible for the installation of the diazeniumdiolate moiety is evident in the gramibactin BGC (*grb*) as well as the BGCs responsible for production of megapolibactin from a symbiotic bacterial strain and gladiobactin/plantaribactin from a pathogenic bacterial strain (Section 2.2.1). C. Genes required for installation of the cinnamoyl-moiety (green) are found sub-clustered in the BGCs for WS9326 (*cal*), skyllamycin (*sky*), and kitacinnamycin (*kcn*) (Section 3.3).

otherwise unrelated BGCs, resulting in the same chemical feature appearing across compound classes. There is a functional aspect to this sub-clustering phenomenon as well; from a discovery perspective sub-clusters allow MultiGeneBlast[199] or GNNs[106] to identify co-localization of genes and guide identification in new genomic contexts. Beyond discovery, sub-clustering of genes often results in coordinated expression which can provide gene cassettes for experimentation and isolation of new molecules in the lab.

Finally, genome mining rests on detailed mechanistic enzymology that identifies and characterizes the enzymes responsible for biosynthetic installation of specific chemical features. The leinamycin story perfectly exemplifies this continued need for collaboration; it took 30 years of cell biology experiments, mechanism of action studies, bioinformatics advances, and mechanistic enzymology to develop the knowledge and tools for researchers to identify related leinamycin analogs *via* a genome mining approach.[53] The interdisciplinary nature of natural products research is a real strength of the field, and it is precisely this position at the interface of chemistry, biology, informatics, and medicine that will help expand the role of genome mining in natural product discovery in the years to come.

7. Conflicts of interest

There are no conflicts to declare.

8. Acknowledgements

We thank Ajalaye Sykes for her work on the table of contents image. This work was partially supported by the US National Institutes of Health (F31-HD101307 to K. D. B and R01-GM085770 to B. S. M.), University of North Carolina at Greensboro (research start-up funds to J. R. C.), and the National Center for Complementary and Integrative Health, a component of the US National Institutes of Health (T32AT008938 to K. S. B.).

9. References

1 A. L. Demain and A. Fang, *Adv. Biochem. Eng./Biotechnol.*, 2000, **132**, 1–39.
2 J. Berdi, *J. Antibiot.*, 2005, **58**, 1–26.
3 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770–803.
4 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, I. E. Orhan, M. Banach, J. M. Rollinger, D. Barreca, W. Weckwerth, R. Bauer, E. A. Bayer, M. Majeed, A. Bishayee, V. Bochkov, G. K. Bonn, N. Braidy, F. Bucar, A. Cifuentes, G. D'Onofrio, M. Bodkin, M. Diederich, A. T. Dinkova-Kostova, T. Efferth, K. El Bairi, N. Arkells, T. P. Fan, B. L. Fiebich, M. Freissmuth, M. I. Georgiev, S. Gibbons, K. M. Godfrey, C. W. Gruber, J. Heer, L. A. Huber, E. Ibanez, A. Kijjoa, A. K. Kiss, A. Lu, F. A. Macias, M. J. S. Miller, A. Mocan, R. Müller, F. Nicoletti, G. Perry, V. Pittalà, L. Rastrelli, M. Ristow, G. L. Russo, A. S. Silva, D. Schuster, H. Sheridan, K. Skalicka-Woźniak, L. Skaltsounis, E. Sobarzo-Sánchez, D. S. Bredt, H. Stuppner, A. Sureda, N. T. Tzvetkov, R. A. Vacca, B. B. Aggarwal, M. Battino, F. Giampieri, M. Wink, J. L. Wolfender, J. Xiao, A. W. K. Yeung, G. Lizard, M. A. Popp, M. Heinrich, I. Berindan-Neagoe, M. Stadler, M. Daglia, R. Verpoorte and C. T. Supuran, *Nat. Rev. Drug Discovery*, 2021, **20**, 200–216.
5 H. Ikeda, J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori and S. Omura, *Nat. Biotechnol.*, 2003, **21**, 526–531.
6 S. D. Bentley, K. F. Chater, A. M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill and D. A. Hopwood, *Nature*, 2002, **417**, 141–147.
7 M. Nett, H. Ikeda and B. S. Moore, *Nat. Prod. Rep.*, 2009, **26**, 1362–1384.
8 N. Ziemert, M. Alanjary and T. Weber, *Nat. Prod. Rep.*, 2016, **33**, 988–1005.
9 K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Wezel, M. H. Medema and T. Weber, *Nucleic Acids Res.*, 2021, **49**, W29–W35.
10 M. A. Skinnider, C. W. Johnston, M. Gunabalasingam, N. J. Merwin, A. M. Kieliszek, R. J. MacLellan, H. Li, M. R. M. Ranieri, A. L. H. Webster, M. P. T. Cao, A. Pfeifle, N. Spencer, Q. H. To, D. P. Wallace, C. A. Dejong and N. A. Magarvey, *Nat. Commun.*, 2020, **11**, 6058.
11 K. Scherlach and C. Hertweck, *Nat. Commun.*, 2021, **12**, 3864.
12 M. H. Medema, T. de Rond and B. S. Moore, *Nat. Rev. Genet.*, 2021, **22**, 553–571.
13 L. Albarano, R. Esposito, N. Ruocco and M. Costantini, *Mar. Drugs*, 2020, **18**, 199.
14 E. Kenshole, M. Herisse, M. Michael and S. J. Pidot, *Curr. Opin. Chem. Biol.*, 2021, **60**, 47–54.
15 R. H. Baltz, *J. Ind. Microbiol. Biotechnol.*, 2021, kuab044.
16 J. Ahlert, E. Shepard, N. Lomovskaya, E. Zazopoulos, A. Staffa, B. O. Bachmann, K. Huang, L. Fonstein, A. Czisny, R. E. Whitwam, C. M. Farnet and J. S. Thorson, *Science*, 2002, **297**, 1173–1176.
17 W. Liu, S. D. Christenson, S. Standage and B. Shen, *Science*, 2002, **297**, 1170–1173.
18 X. Yan, J. J. Chen, A. Adhikari, D. Yang, I. Crnovcic, N. Wang, C. Y. Chang, C. Rader and B. Shen, *Org. Lett.*, 2017, **19**, 6192–6195.
19 X. Yan, H. Ge, T. Huang, Hindra, D. Yang, Q. Teng, I. Crnovčić, X. Li, J. D. Rudolf, J. R. Lohman, Y. Gansemans, X. Zhu, Y. Huang, L. X. Zhao, Y. Jiang,

F. van Nieuwerburgh, C. Rader, Y. Duan and B. Shen, *mBio*, 2016, **7**, e02104–e02116.
20 B. O. Bachmann, R. Li and C. A. Townsend, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 9082–9086.
21 B. Gerratana, A. Stapon and C. A. Townsend, *Biochemistry*, 2003, **42**, 7836–7847.
22 B. J. Weigel, S. G. Burgett, V. J. Chen, P. L. Skatrud, C. A. Frolik, S. W. Queener and T. D. Ingolia, *J. Bacteriol.*, 1988, **170**, 3817–3826.
23 N. M. Gaudelli, D. H. Long and C. A. Townsend, *Nature*, 2015, **520**, 383–387.
24 G. Blanco, *Arch. Microbiol.*, 2012, **194**, 549–555.
25 J. K. Christenson, J. E. Richman, M. R. Jensen, J. Y. Neufeld, C. M. Wilmot and L. P. Wackett, *Biochemistry*, 2017, **56**, 348–351.
26 S. L. Robinson, B. R. Terlouw, M. D. Smith, S. J. Pidot, T. P. Stinear, M. H. Medema and L. P. Wackett, *J. Biol. Chem.*, 2020, **295**, 14826–14839.
27 J. E. Schaffer, M. R. Reck, N. K. Prasad and T. A. Wencewicz, *Nat. Chem. Biol.*, 2017, **13**, 737–744.
28 K. N. Feng, Y. L. Yang, Y. X. Xu, Y. Zhang, T. Feng, S. X. Huang, J. K. Liu and Y. Zeng, *Angew. Chem., Int. Ed.*, 2020, **59**, 7209–7213.
29 A. S. Eustáquio, S. J. Nam, K. Penn, A. Lechner, M. C. Wilson, W. Fenical, P. R. Jensen and B. S. Moore, *ChemBioChem*, 2011, **12**, 61–64.
30 N. D. Chaurasiya, N. S. Sangwan, F. Sabir, L. Misra and R. S. Sangwan, *Plant Cell Rep.*, 2012, **31**, 1889–1897.
31 J. Wu, T. J. Zaleski, C. Valenzano, C. Khosla and D. E. Cane, *J. Am. Chem. Soc.*, 2005, **127**, 17393–17404.
32 J. J. Zhang, X. Tang, T. Huan, A. C. Ross and B. S. Moore, *Nat. Chem. Biol.*, 2020, **16**, 42–49.
33 D. Zabala, J. W. Cartwright, D. M. Roberts, B. J. C. Law, L. Song, M. Samborskyy, P. F. Leadlay, J. Micklefield and G. L. Challis, *J. Am. Chem. Soc.*, 2016, **138**, 4342–4345.
34 M. Schorn, J. Zettler, J. P. Noel, P. C. Dorrestein, B. S. Moore and L. Kaysser, *ACS Chem. Biol.*, 2014, **9**, 301–309.
35 J. G. Owen, Z. Charlop-Powers, A. G. Smith, M. A. Ternei, P. Y. Calle, B. V. B. Reddy, D. Montiel and S. F. Brady, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 4221–4226.
36 L. Tang, S. Shah, L. Chung, J. Carney, L. Katz, C. Khosla and B. Julien, *Science*, 2000, **287**, 640–642.
37 C. J. Thibodeaux, W. C. Chang and H. W. Liu, *Chem. Rev.*, 2012, **112**, 1681–1709.
38 S. J. Gould, M. J. Kirchmeier and R. E. LaFever, *J. Am. Chem. Soc.*, 1996, **118**, 7663–7666.
39 P. Liu, K. Murakami, T. Seki, X. He, S. M. Yeung, T. Kuzuyama, H. Seto and H. Liu, *J. Am. Chem. Soc.*, 2001, **123**, 4619–4620.
40 T. Kuzuyama, T. Seki, S. Kobayashi, T. Hidaka and H. Seto, *Biosci., Biotechnol., Biochem.*, 1999, **63**, 2222–2224.
41 R. Ueoka, M. Bortfeld-Miller, B. I. Morinaka, J. A. Vorholt and J. Piel, *Angew. Chem., Int. Ed.*, 2018, **57**, 977–981.
42 X. Liu, S. Biswas, M. G. Berg, C. M. Antapli, F. Xie, Q. Wang, M. C. Tang, G. L. Tang, L. Zhang, G. Dreyfuss and Y. Q. Cheng, *J. Nat. Prod.*, 2013, **76**, 685–693.
43 D. Ren, M. Kim, S. A. Wang and H. Liu, *Angew. Chem., Int. Ed.*, 2021, **60**, 17148–17154.
44 M. Sato, J. E. Dander, C. Sato, Y. S. Hung, S. S. Gao, M. C. Tang, L. Hang, J. M. Winter, N. K. Garg, K. Watanabe and Y. Tang, *J. Am. Chem. Soc.*, 2017, **139**, 5317–5320.
45 M. Baunach, L. Ding, K. Willing and C. Hertweck, *Angew. Chem., Int. Ed.*, 2015, **54**, 13279–13283.
46 G.-Q. Wang, G.-D. Chen, S.-Y. Qin, D. Hu, T. Awakawa, S.-Y. Li, J.-M. Lv, C.-X. Wang, X.-S. Yao, I. Abe and H. Gao, *Nat. Commun.*, 2018, **9**, 1838.
47 K. C. Belknap, C. J. Park, B. M. Barth and C. P. Andam, *Sci. Rep.*, 2020, **10**, 2003.
48 B. Li and C. T. Walsh, *Biochemistry*, 2011, **50**, 4615–4622.
49 D. H. Scharf, N. Remme, T. Heinekamp, P. Hortschansky, A. A. Brakhage and C. Hertweck, *J. Am. Chem. Soc.*, 2010, **132**, 10136–10141.
50 C. Wang, S. R. Wesener, H. Zhang and Y. Q. Cheng, *Chem. Biol.*, 2009, **16**, 585–593.
51 M. Ma, J. R. Lohman, T. Liu and B. Shen, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 10359–10364.
52 H. Liu, J. Fan, P. Zhang, Y. Hu, X. Liu, S. M. Li and W. B. Yin, *Chem. Sci.*, 2021, **12**, 4132–4138.
53 G. Pan, Z. Xu, Z. Guo, Hindra, M. Ma, D. Yang, H. Zhou, Y. Gansemans, X. Zhu, Y. Huang, L.-X. X. Zhao, Y. Jiang, J. Cheng, F. Van Nieuwerburgh, J.-W. W. Suh, Y. Duan and B. Shen, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E11131–E11140.
54 B. Dose, S. P. Niehs, K. Scherlach, S. Shahda, L. V. Flórez, M. Kaltenpoth and C. Hertweck, *ChemBioChem*, 2021, **22**, 1920–1924.
55 J. Liu, T. Ng, Z. Rui, O. Ad and W. Zhang, *Angew. Chem., Int. Ed.*, 2014, **53**, 136–139.
56 Z. Liang, *Nat. Prod. Rep.*, 2010, **27**, 499.
57 E. Zazopoulos, K. Huang, A. Staffa, W. Liu, B. O. Bachmann, K. Nonaka, J. Ahlert, J. S. Thorson, B. Shen and C. M. Farnet, *Nat. Biotechnol.*, 2003, **21**, 187–190.
58 B. Shen, X. Yan, T. Huang, H. Ge, D. Yang, Q. Teng, J. D. Rudolf and J. R. Lohman, *Bioorg. Med. Chem. Lett.*, 2015, **25**, 9–15.
59 J. D. Rudolf, X. Yan and B. Shen, *J. Ind. Microbiol. Biotechnol.*, 2016, **43**, 261–276.
60 J. Davies, H. Wang, T. Taylor, K. Warabi, X. H. Huang and R. J. Andersen, *Org. Lett.*, 2005, **7**, 5233–5236.
61 M. Gersch, J. Kreuzer and S. A. Sieber, *Nat. Prod. Rep.*, 2012, **29**, 659–682.
62 S. L. Robinson, J. K. Christenson and L. P. Wackett, *Nat. Prod. Rep.*, 2019, **36**, 458–475.
63 Y. Mikami, Y. Yazawa, Y. Tanaka, M. Ritzau and U. Gräfe, *Nat. Prod. Lett.*, 1999, **13**, 277–284.
64 A. Zeeck, K. Schröder, K. Frobel, R. Grote and R. Thiericke, *J. Antibiot.*, 1987, **40**, 1530–1540.
65 M. Hanada, K. Sugawara, K. Kaneta, S. Toda, Y. Nishiyama, K. Tomita, H. Yamamoto, M. Konishi and T. Oki, *J. Antibiot.*, 1992, **45**, 1746–1752.
66 K. Sugawara, M. Hatori, Y. Nishiyama, K. Tomita, H. Kamei, M. Konishi and T. Oki, *J. Antibiot.*, 1990, **43**, 8–18.

67 D. J. Kuhn, Q. Chen, P. M. Voorhees, J. S. Strader, K. D. Shenk, C. M. Sun, S. D. Demo, M. K. Bennett, F. W. B. Van Leeuwen, A. A. Chanan-Khan and R. Z. Orlowski, *Blood*, 2007, **110**, 3281–3290.
68 B. V. B. Reddy, A. Milshteyn, Z. Charlop-Powers and S. F. Brady, *Chem. Biol.*, 2014, **21**, 1023–1033.
69 S. H. Lee, *Arch. Pharmacal Res.*, 2009, **32**, 299–315.
70 M. C. Yi and C. Khosla, *Annu. Rev. Chem. Biomol. Eng.*, 2016, **7**, 197–222.
71 M. Kara, K. Asano, I. Kawamoto, H. Nakano, T. Takiouchi, S. Katsumata and K. Takahashi, *J. Antibiot.*, 1989, **42**, 1768–1774.
72 M. Hara, I. Takahashi, M. Yoshida, K. Asano, I. Kawamoto, H. Nakano and M. Morimoto, *J. Antibiot.*, 1989, **42**, 333–335.
73 S. X. Huang, B. S. Yun, M. Ma, H. S. Basu, D. R. Church, G. Ingenhorst, Y. Huang, D. Yang, J. R. Lohman, G. L. Tang, J. Ju, T. Liu, G. Wilding and B. Shen, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 8278–8283.
74 V. Viswesh, K. Gates and D. Sun, *Chem. Res. Toxicol.*, 2010, **23**, 99–107.
75 G.-L. Tang, Y.-Q. Cheng and B. Shen, *Chem. Biol.*, 2004, **11**, 33–45.
76 L. P. H. Yang, S. J. Keam and G. M. Keating, *Drugs*, 2007, **67**, 2211–2230.
77 G. G. Zhanel, A. R. Golden, S. Zelenitsky, K. Wiebe, C. K. Lawrence, H. J. Adam, T. Idowu, R. Domalaon, F. Schweizer, M. A. Zhanel, P. R. S. Lagacé-Wiens, A. J. Walkty, A. Noreddin, J. P. Lynch and J. A. Karlowsky, *Drugs*, 2019, **79**, 271–289.
78 K. H. Negash, J. K. S. Norris and J. T. Hodgkinson, *Molecules*, 2019, **24**, 3314.
79 R. C. Hider and X. Kong, *Nat. Prod. Rep.*, 2010, **27**, 637–657.
80 R. Hermenau, K. Ishida, S. Gama, B. Hoffmann, M. Pfeifer-Leeg, W. Plass, J. F. Mohr, T. Wichard, H. P. Saluz and C. Hertweck, *Nat. Chem. Biol.*, 2018, **14**, 841–843.
81 R. Hermenau, J. L. Mehl, K. Ishida, B. Dose, S. J. Pidot, T. P. Stinear and C. Hertweck, *Angew. Chem., Int. Ed.*, 2019, **58**, 13024–13029.
82 D. E. Scott, A. R. Bayly, C. Abell and J. Skidmore, *Nat. Rev. Drug Discovery*, 2016, **15**, 533–550.
83 L. Mabonga and A. P. Kappo, *Biophys. Rev.*, 2019, **11**, 559–581.
84 Y. J. Yoo, H. Kim, S. R. Park and Y. J. Yoon, *J. Ind. Microbiol. Biotechnol.*, 2017, **44**, 537–553.
85 U. K. Shigdel, S. J. Lee, M. E. Sowa, B. R. Bowman, K. Robison, M. Zhou, K. H. Pua, D. T. Stiles, J. A. V. Blodgett, D. W. Udwary, A. T. Rajczewski, A. S. Mann, S. Mostafavi, T. Hardy, S. Arya, Z. Weng, M. Stewart, K. Kenyon, J. P. Morgenstern, E. Pan, D. C. Gray, R. M. Pollock, A. M. Fry, R. D. Klausner, S. A. Townson and G. L. Verdine, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 17195–17203.
86 G. J. Gatto, M. T. Boyne, N. L. Kelleher and C. T. Walsh, *J. Am. Chem. Soc.*, 2006, **128**, 3838–3847.
87 D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Hüttenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. McGregor, Q. Li, B. Meyer, F. Roesch, T. Vallet, A. Mac Kain, L. Miorin, E. Moreno, Z. Z. C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lyu, C. J. P. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Rakesh, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, X. P. Huang, Y. F. Liu, S. A. Wankowicz, M. Bohn, M. Safari, F. S. Ugur, C. Koh, N. S. Savar, Q. D. Tran, D. Shengjuler, S. J. Fletcher, M. C. O'Neal, Y. Cai, J. C. J. Chang, D. J. Broadhurst, S. Klippsten, P. P. Sharp, N. A. Wenzell, D. Kuzuoglu-Ozturk, H. Y. Wang, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R. M. Stroud, A. D. Frankel, O. S. Rosenberg, K. A. Verba, D. A. Agard, M. Ott, M. Emerman, N. Jura, M. von Zastrow, E. Verdin, A. Ashworth, O. Schwartz, C. d'Enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. N. Floor, J. S. Fraser, J. D. Gross, A. Sali, B. L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. García-Sastre, K. M. Shokat, B. K. Shoichet and N. J. Krogan, *Nature*, 2020, **583**, 459–468.
88 D. Y. Travin, K. Severinov and S. Dubiley, *RSC Chem. Biol.*, 2021, **2**, 468–485.
89 J. I. Guijarro, J. E. Gonzalez-Pastor, F. Baleux, J. L. San Millan, M. A. Castilla, M. Rico, F. Moreno and M. Delepierre, *J. Biol. Chem.*, 1995, **270**, 23520–23532.
90 M. Novikova, A. Metlitskaya, K. Datsenko, T. Kazakov, A. Kazakov, B. Wanner and K. Severinov, *J. Bacteriol.*, 2007, **189**, 8361–8365.
91 A. Metlitskaya, T. Kazakov, A. Kommer, O. Pavlova, M. Praetorius-Ibba, M. Ibba, I. Krasheninnikov, V. Kolb, I. Khmel and K. Severinov, *J. Biol. Chem.*, 2006, **281**, 18033–18042.
92 R. F. Roush, E. M. Nolan, F. Löhr and C. T. Walsh, *J. Am. Chem. Soc.*, 2008, **130**, 3603–3609.
93 S. H. Dong, A. Kulikovsky, I. Zukher, P. Estrada, S. Dubiley, K. Severinov and S. K. Nair, *Chem. Sci.*, 2019, **10**, 2391–2395.
94 C. A. Regni, R. F. Roush, D. J. Miller, A. Nourse, C. T. Walsh and B. A. Schulman, *EMBO J.*, 2009, **28**, 1953–1964.
95 K. Severinov, E. Semenova, A. Kazakov, T. Kazakov and M. S. Gelfand, *Mol. Microbiol.*, 2007, **65**, 1380–1394.
96 O. Bantysh, M. Serebryakova, K. S. Makarova, S. Dubiley, K. A. Datsenko and K. Severinov, *mBio*, 2014, **5**, e01059.
97 M. Serebryakova, D. Tsibulskaya, O. Mokina, A. Kulikovsky, M. Nautiyal, A. Van Aerschot, K. Severinov and S. Dubiley, *J. Am. Chem. Soc.*, 2016, **138**, 15690–15698.
98 J. P. Michael, *Nat. Prod. Rep.*, 2003, **20**, 476–493.
99 J. Zhang, S. L. Morris-Natschke, D. Ma, X. F. Shang, C. J. Yang, Y. Q. Liu and K. H. Lee, *Med. Res. Rev.*, 2021, **41**, 928–960.
100 S. Chemler, *Curr. Bioact. Compd.*, 2009, **5**, 2–19.

101 W. B. Han, A. H. Zhang, X. Z. Deng, X. Lei and R. X. Tan, *Org. Lett.*, 2016, **18**, 1816–1819.
102 W. B. Han, Y. H. Lu, A. H. Zhang, G. F. Zhang, Y. N. Mei, N. Jiang, X. Lei, Y. C. Song, S. W. Ng and R. X. Tan, *Org. Lett.*, 2014, **16**, 5366–5369.
103 G. Z. Dai, W. B. Han, Y. N. Mei, K. Xu, R. H. Jiao, H. M. Ge and R. X. Tan, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1174–1180.
104 D. A. Dias, S. Urban and U. Roessner, *Metabolites*, 2012, **2**, 303–336.
105 J. A. Gerlt, J. T. Bouvier, D. B. Davidson, H. J. Imker, B. Sadkhin, D. R. Slater and K. L. Whalen, *Biochim. Biophys. Acta*, 2015, **1854**, 1019–1037.
106 R. Zallot, N. O. Oberg and J. A. Gerlt, *Curr. Opin. Chem. Biol.*, 2018, **47**, 77–85.
107 M. S. Butler, K. A. Hansford, M. A. T. Blaskovich, R. Halai and M. A. Cooper, *J. Antibiot.*, 2014, **67**, 631–644.
108 G. Yim, M. N. Thaker, K. Koteva and G. Wright, *J. Antibiot.*, 2014, **67**, 31–41.
109 E. J. Culp, N. Waglechner, W. Wang, A. A. Fiebig-Comyn, Y.-P. Hsu, K. Koteva, D. Sychantha, B. K. Coombes, M. S. Van Nieuwenhze, Y. V. Brun and G. D. Wright, *Nature*, 2020, **578**, 582–587.
110 N. Waglechner, A. G. McArthur and G. D. Wright, *Nat. Microbiol.*, 2019, **4**, 1862–1871.
111 H. T. Chiu, B. K. Hubbard, A. N. Shah, J. Eide, R. A. Fredenburg, C. T. Walsh and C. Khosla, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 8548–8553.
112 T. J. Smith, S. A. Blackman and S. J. Foster, *Microbiology*, 2000, **146**, 249–262.
113 R. E. W. Hancock and R. Lehrer, *Trends Biotechnol.*, 1998, **16**, 82–88.
114 Y. X. Li, Z. Zhong, W. P. Zhang and P. Y. Qian, *Nat. Commun.*, 2018, **9**, 2–10.
115 S. A. Cochrane and J. C. Vederas, *Med. Res. Rev.*, 2016, **36**, 4–31.
116 M. Navidinia, *J. Paramed. Sci.*, 2016, **7**, 43–57.
117 S. Pohle, C. Appelt, M. Roux, H. P. Fiedler and R. D. Süssmuth, *J. Am. Chem. Soc.*, 2011, **133**, 6194–6205.
118 K. Hayashi, M. Hashimoto, N. Shigematsu, M. Nishikawa, M. Ezaki, M. Yamashita, S. Kiyoto, M. Okuhara, M. Kohsaka and H. Imanaka, *J. Antibiot.*, 1992, **45**, 1055–1063.
119 M. Bae, H. Kim, K. Moon, S. J. Nam, J. Shin, K. B. Oh and D. C. Oh, *Org. Lett.*, 2015, **17**, 712–715.
120 J. Shi, C. L. Liu, B. Zhang, W. J. Guo, J. Zhu, C. Y. Chang, E. J. Zhao, R. H. Jiao, R. X. Tan and H. M. Ge, *Chem. Sci.*, 2019, **10**, 4839–4846.
121 R. W. Rickards and D. Skropeta, *Tetrahedron*, 2002, **58**, 3793–3800.
122 Z. Deng, J. Liu, T. Li, H. Li, Z. Liu, Y. Dong and W. Li, *Angew. Chem., Int. Ed.*, 2021, **60**, 153–158.
123 T. M. Wood and N. I. Martin, *MedChemComm*, 2019, **10**, 634–646.
124 D. Jung, A. Rozek, M. Okon and R. E. Hancock, *Chem. Biol.*, 2004, **11**, 949–957.
125 R. Sauermann, M. Rothenburger, W. Graninger and C. Joukhadar, *Pharmacology*, 2008, **81**, 79–91.
126 B. M. Hover, S.-H. Kim, M. Katz, Z. Charlop-Powers, J. G. Owen, M. A. Ternei, J. Maniko, A. B. Estrela, H. Molina, S. Park, D. S. Perlin and S. F. Brady, *Nat. Microbiol.*, 2018, **3**, 415–422.
127 J. G. Owen, B. V. B. Reddy, M. A. Ternei, Z. Charlop-Powers, P. Y. Calle, J. H. Kim and S. F. Brady, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 11797–11802.
128 M. Katz, B. M. Hover and S. F. Brady, *J. Ind. Microbiol. Biotechnol.*, 2016, **43**, 129–141.
129 S. Banala and R. D. Süssmuth, *ChemBioChem*, 2010, **11**, 1335–1337.
130 N. Mahanta, D. M. Szantai-Kis, E. J. Petersson and D. A. Mitchell, *ACS Chem. Biol.*, 2019, **14**, 142–163.
131 J. J. Lipsky and M. O. Gallego, *Drug Metab. Drug Interact.*, 1988, **6**, 317–326.
132 Y. Hayakawa, K. Sasaki, K. Nagai, K. Shin-ya and K. Furihata, *J. Antibiot.*, 2006, **59**, 6–10.
133 L. Frattaruolo, R. Lacret, A. R. Cappello and A. W. Truman, *ACS Chem. Biol.*, 2017, **12**, 2815–2822.
134 L. Kjaerulff, A. Sikandar, N. Zaburannyi, S. Adam, J. Herrmann, J. Koehnke and R. Müller, *ACS Chem. Biol.*, 2017, **12**, 2837–2841.
135 M. Izawa, T. Kawasaki and Y. Hayakawa, *Appl. Environ. Microbiol.*, 2013, **79**, 7110–7113.
136 B. J. Burkhart, C. J. Schwalen, G. Mann, J. H. Naismith and D. A. Mitchell, *Chem. Rev.*, 2017, **117**, 5389–5456.
137 T. Hamada, T. Sugawara, S. Matsunaga and N. Fusetani, *Tetrahedron Lett.*, 1994, **35**, 609–612.
138 T. Hamada, T. Sugawara, S. Matsunaga and N. Fusetani, *Tetrahedron Lett.*, 1994, **35**, 719–720.
139 T. Hamada, S. Matsunaga, G. Yano and N. Fusetani, *J. Am. Chem. Soc.*, 2005, **127**, 110–118.
140 M. F. Freeman, C. Gurgui, M. J. Helf, B. I. Morinaka, A. R. Uria, N. J. Oldham, H.-G. Sahl, S. Matsunaga and J. Piel, *Science*, 2012, **338**, 387–390.
141 M. F. Freeman, M. J. Helf, A. Bhushan, B. I. Morinaka and J. Piel, *Nat. Chem.*, 2017, **9**, 387–395.
142 T. Hamada, S. Matsunaga, M. Fujiwara, K. Fujita, H. Hirota, R. Schmucki, P. Güntert and N. Fusetani, *J. Am. Chem. Soc.*, 2010, **132**, 12941–12945.
143 S. Oiki, I. Muramatsu, S. Matsunaga and N. Fusetani, *Folia Pharmacol. Jpn.*, 1997, **110**, 195–198.
144 A. Bhushan, P. J. Egli, E. E. Peters, M. F. Freeman and J. Piel, *Nat. Chem.*, 2019, **11**, 931–939.
145 A. Renevey and S. Riniker, *Eur. Biophys. J.*, 2017, **46**, 363–374.
146 M. Gozari, M. Alborz, H. R. El-Seedi and A. R. Jassbi, *Eur. J. Med. Chem.*, 2021, **210**, 112957.
147 L. A. M. Murray, S. M. K. McKinnie, B. S. Moore and J. H. George, *Nat. Prod. Rep.*, 2020, **37**, 1334–1366.
148 R. Geris and T. J. Simpson, *Nat. Prod. Rep.*, 2009, **26**, 1063–1094.
149 R. Teufel, in *Methods in Enzymology*, Elsevier Inc., 1st edn, 2018, vol. 604, pp. 425–439.

150 X. Zhang, T. T. Wang, Q. L. Xu, Y. Xiong, L. Zhang, H. Han, K. Xu, W. J. Guo, Q. Xu, R. X. Tan and H. M. Ge, *Angew. Chem., Int. Ed.*, 2018, **57**, 8184–8188.
151 Y. Matsuda and I. Abe, *Nat. Prod. Rep.*, 2016, **33**, 26–53.
152 G. P. Horsman and D. L. Zechel, *Chem. Rev.*, 2017, **117**, 5704–5783.
153 J. L. Marquardt, E. D. Brown, W. S. Lane, T. M. Haley, Y. Ichikawa, C. H. Wong and C. T. Walsh, *Biochemistry*, 1994, **33**, 10646–10651.
154 T. Kuzuyama, T. Shimizu, S. Takahashi and H. Seto, *Tetrahedron Lett.*, 1998, **39**, 7913–7916.
155 E. Bowman, M. McQueney, R. J. Barry and D. Dunaway-Mariano, *J. Am. Chem. Soc.*, 1988, **110**, 5575–5576.
156 X. Yu, J. R. Doroghazi, S. C. Janga, J. K. Zhang, B. Circello, B. M. Griffin, D. P. Labeda and W. W. Metcalf, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 20759–20764.
157 K. S. Ju, J. Gao, J. R. Doroghazi, K. K. A. Wang, C. J. Thibodeaux, S. Li, E. Metzger, J. Fudala, J. Su, J. K. Zhang, J. Lee, J. P. Cioni, B. S. Evans, R. Hirota, D. P. Labeda, W. A. van der Donk and W. W. Metcalf, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12175–12180.
158 C. M. Kayrouz, Y. Zhang, T. M. Pham and K. S. Ju, *ACS Chem. Biol.*, 2020, **15**, 1921–1929.
159 E. C. O'Neill, M. Schorn, C. B. Larson and N. Millán-Aguiñaga, *Crit. Rev. Microbiol.*, 2019, **45**, 255–277.
160 J. Davies and D. Davies, *Microbiol. Mol. Biol. Rev.*, 2010, **74**, 417–433.
161 X. Tang, J. Li, N. Millán-Aguiñaga, J. J. Zhang, E. C. O'Neill, J. A. Ugalde, P. R. Jensen, S. M. Mantovani and B. S. Moore, *ACS Chem. Biol.*, 2015, **10**, 2841–2849.
162 V. M. D'Costa, K. M. McGrann, D. W. Hughes and G. D. Wright, *Science*, 2006, **311**, 374–377.
163 M. N. Thaker, W. Wang, P. Spanogiannopoulos, N. Waglechner, A. M. King, R. Medina and G. D. Wright, *Nat. Biotechnol.*, 2013, **31**, 922–927.
164 J. B. Parsons and C. O. Rock, *Curr. Opin. Microbiol.*, 2011, **14**, 544–549.
165 Y. M. Zhang, S. W. White and C. O. Rock, *J. Biol. Chem.*, 2006, **281**, 17541–17544.
166 K. Young, H. Jayasuriya, J. G. Ondeyka, K. Herath, C. Zhang, S. Kodali, A. Galgoci, R. Painter, V. Brown-Driver, R. Yamamoto, L. L. Silver, Y. Zheng, J. I. Ventura, J. Sigmund, S. Ha, A. Basilio, F. Vicente, J. R. Tormo, F. Pelaez, P. Youngman, D. Cully, J. F. Barrett, D. Schmatz, S. B. Singh and J. Wang, *Antimicrob. Agents Chemother.*, 2006, **50**, 519–526.
167 R. M. Peterson, T. Huang, J. D. Rudolf, M. J. Smanski and B. Shen, *Chem. Biol.*, 2014, **21**, 389–397.
168 H. Oishi, T. Noto, H. Sasaki, K. Suzuki, T. Hayashi, H. Okazaki, K. Ando and M. Sawada, *J. Antibiot.*, 1982, **35**, 391–395.
169 H. Sasaki, H. Oishi, T. Hayashi, I. Matsuura, K. Ando and M. Sawada, *J. Antibiot.*, 1982, **35**, 396–400.
170 M. D. Mungan, M. Alanjary, K. Blin, T. Weber, M. H. Medema and N. Ziemert, *Nucleic Acids Res.*, 2020, **48**, W546–W552.
171 O. Sozinova, Y. Jiang, D. Kaiser and M. Alber, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17255–17259.
172 J. Herrmann, A. A. Fayad and R. Müller, *Nat. Prod. Rep.*, 2017, **34**, 135–160.
173 F. Panter, D. Krug, S. Baumann and R. Müller, *Chem. Sci.*, 2018, **9**, 4898–4908.
174 S. Baumann, J. Herrmann, R. Raju, H. Steinmetz, K. I. Mohr, S. Hüttel, K. Harmrolfs, M. Stadler and R. Müller, *Angew. Chem., Int. Ed.*, 2014, **53**, 14605–14609.
175 H. H. Yeh, M. Ahuja, Y. M. Chiang, C. E. Oakley, S. Moore, O. Yoon, H. Hajovsky, J. W. Bok, N. P. Keller, C. C. C. Wang and B. R. Oakley, *ACS Chem. Biol.*, 2016, **11**, 2275–2284.
176 W. Fenical, P. R. Jensen, M. A. Palladino, K. S. Lam, G. K. Lloyd and B. C. Potts, *Bioorg. Med. Chem.*, 2009, **17**, 2175–2180.
177 G. Lin, D. Li, T. Chidawanyika, C. Nathan and H. Li, *Arch. Biochem. Biophys.*, 2010, **501**, 214–220.
178 Y. Yan, Q. Liu, X. Zang, S. Yuan, U. Bat-Erdene, C. Nguyen, J. Gan, J. Zhou, S. E. Jacobsen and Y. Tang, *Nature*, 2018, **559**, 415–418.
179 T. M. Amorim Franco and J. S. Blanchard, *Biochemistry*, 2017, **56**, 5849–5865.
180 Y. Tsuda, M. Kaneda, A. Tada, K. Nitta, Y. Yamamoto and I. Yoichi, *J. Chem. Soc., Chem. Commun.*, 1978, 160–161.
181 G. M. Dill, *Pest Manage. Sci.*, 2005, **61**, 219–224.
182 A. Milshteyn, D. A. Colosimo and S. F. Brady, *Cell Host Microbe*, 2018, **23**, 725–736.
183 J. Chu, X. Vila-Farres, D. Inoyama, M. Ternei, L. J. Cohen, E. A. Gordon, B. V. B. Reddy, Z. Charlop-Powers, H. A. Zebroski, R. Gallardo-Macias, M. Jaskowski, S. Satish, S. Park, D. S. Perlin, J. S. Freundlich and S. F. Brady, *Nat. Chem. Biol.*, 2016, **12**, 1004–1006.
184 C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben and D. H. Huson, *Nucleic Acids Res.*, 2005, **33**, 5799–5808.
185 M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch and O. Kohlbacher, *Nucleic Acids Res.*, 2011, **39**, 362–367.
186 J. Chu, X. Vila-Farres and S. F. Brady, *J. Am. Chem. Soc.*, 2019, **141**, 15737–15741.
187 L. Du and L. Lou, *Nat. Prod. Rep.*, 2010, **27**, 255–278.
188 J. Chu, B. Koirala, N. Forelli, X. Vila-Farres, M. A. Ternei, T. Ali, D. A. Colosimo and S. F. Brady, *J. Am. Chem. Soc.*, 2020, **142**, 14158–14168.
189 M. A Hostetler, C. Smith, S. Nelson, Z. Budimir, R. Modi, I. Woolsey, A. Frerk, B. Baker, J. Gantt and E. I. Parkinson, *ACS Chem. Biol.*, 2021, DOI: 10.1021/acschembio.1c00641.
190 A. H. Russell and A. W. Truman, *Comput. Struct. Biotechnol. J.*, 2020, **18**, 1838–1851.
191 X. Yang, K. R. Lennard, C. He, M. C. Walker, A. T. Ball, C. Doigneaux, A. Tavassoli and W. A. van der Donk, *Nat. Chem. Biol.*, 2018, **14**, 375–380.
192 S. Schmitt, M. Montalbán-López, D. Peterhoff, J. Deng, R. Wagner, M. Held, O. P. Kuipers and S. Panke, *Nat. Chem. Biol.*, 2019, **15**, 437–443.
193 A. J. van Heel, T. G. Kloosterman, M. Montalban-Lopez, J. Deng, A. Plat, B. Baudu, D. Hendriks, G. N. Moll and O. P. Kuipers, *ACS Synth. Biol.*, 2016, **5**, 1146–1154.

194 R. Liu, Y. Zhang, G. Zhai, S. Fu, Y. Xia, B. Hu, X. Cai, Y. Zhang, Y. Li, Z. Deng and T. Liu, *Adv. Sci.*, 2020, **7**, 2001616.
195 L. M. Repka, J. R. Chekan, S. K. Nair and W. A. van der Donk, *Chem. Rev.*, 2017, **117**, 5457–5520.
196 K. A. Wetterstrand, *The Cost of Sequencing a Human Genome*, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, accessed 1 September 2021.
197 I. Paterson and E. A. Anderson, *Science*, 2005, **310**, 451–453.
198 M. H. Medema, P. Cimermancic, A. Sali, E. Takano and M. A. Fischbach, *PLoS Comput. Biol.*, 2014, **10**, e1004016.
199 M. H. Medema, E. Takano and R. Breitling, *Mol. Biol. Evol.*, 2013, **30**, 1218–1223.