# Phylogenetic distance and structural diversity directing a reclassification of glycopeptide antibiotics

Athina Gavriilidou[1], Martina Adamek[1,2], Jens-Peter Rodler[3], Noel Kubach[1], Susanna Kramer[1], Daniel H. Huson[4], Max J. Cryle[5,6], Evi Stegmann[3]*, Nadine Ziemert[1,2]*

1: Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany / Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen, Germany
2: German Centre for Infection Research (DZIF), Partnersite Tübingen, Tübingen, Germany
3: Microbial Bioactive Compounds, Interfaculty Institute of Microbiology and Infection Medicine Tübingen, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen,
4: Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, 72076, Germany / International Max Planck Research School "From Molecules to Organisms", Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, Tübingen, 72076, Germany / Cluster of Excellence 'Controlling Microbes to Fight Infections' (CMFI), University of Tübingen, Tübingen / Germany, Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany
5: Department of Biochemistry and Molecular Biology, The Monash Biomedicine Discovery Institute, Monash University, Clayton, VIC, 3800, Australia
6: EMBL Australia, Monash University, Clayton, Victoria 3800, Australia
* co-corresponding author

# Abstract

Antibiotics have been an essential part of modern medicine since their initial discovery. The continuous search for new antibiotic candidates remains a necessity given the increasing emergence of resistance to antimicrobial compounds among pathogens. The glycopeptide antibiotics (GPAs) represent an important group of last resort antibiotics which inhibit bacterial growth through non-covalent binding to the cell wall precursor lipid II. The so far reported GPAs exhibit an enormous diversity in the biosynthetic gene clusters that encode their production, which is in turn reflected in the variety of their structures. GPAs are typically composed of seven amino acids, which are highly crosslinked and decorated with a variable collection of sugar moieties as well as other modifications. Based on their structural characteristics, they have been classified into four main types. More recently, atypical GPAs have been identified that differ from type I-IV GPAs in both their structure and function and have consequently been classified as type V GPAs. Given these differences, we studied the phylogeny of all gene sequences related to the biosynthesis of the GPAs and observed a clear evolutionary diversification between the lipid II binding GPA classes and the so-called type V GPAs. Here we suggest the adoption of a phylogeny-driven reclassification and a separation of classical lipid II binding GPAs from type V GPAs, which we propose to identify instead as glycopeptide- related peptides (GRPs).

# Introduction

The glycopeptide antibiotics (GPA) are an important group of clinical antibiotics, with the first member - vancomycin - discovered in 1953. Since then, 27 natural GPAs have been identified and many semi-synthetic derivatives have been synthesised, some of which are in clinical use for the treatment of infections caused by multi-resistant Gram-positive bacterial pathogens [1] [2]. Both the biosynthesis (*in vivo* and *in vitro*) and the mode of action of GPAs have now been studied for decades [1], [3]–[10]. The name GPA itself summarises their structure: a peptidic backbone typically comprised of seven amino acids, to which one or more sugars are attached. Beyond this, GPAs are extensively crosslinked through their side chains of aromatic amino acid residues, which confers rigidity to the core and represents a hallmark of this class of natural products that is essential for bioactivity. The peptide backbone is further decorated by the addition of halogen atoms, sulphate moieties, sugar residues and methyl groups (**Figure 1**) [1], [11], [12].

### Existing GPA classification

To date, GPAs have been classified into five types (I-V) according to their structural characteristics. The backbone of type I GPAs consists of two aliphatic amino acids in positions 1 and 3 (Leu and Asn) of the peptide, and five non-proteinogenic aromatic amino acids ($\beta$-hydroxytyrosine (Bht); 4-hydroxyphenylglycine (Hpg) and 3,5-dihydroxyphenylglycine (Dpg)), which are linked through three phenolic/biaryl crosslinks [11]. Type II GPAs share a similar crosslinking pattern to the type I GPAs, although this class possesses aromatic amino acids, non-proteinogenic (Hpg) and proteinogenic (Phe) at positions 1 and 3 of the peptide core instead of aliphatic residues [11]. The peptide core of type III GPAs consists exclusively of aromatic amino acids ($Hpg^1$–$Bht^2$–$Dpg^3$–$Hpg^4$–$Hpg^5$-$Bht^6$–$Dpg^7$), which are all crosslinked in contrast to those of type II GPAs. Therefore, type III GPAs contain an additional crosslink compared to type I/II GPAs (4 vs 3) [11]. Type IV GPAs do not differ within the core peptide sequence from type III GPAs, but in addition they contain an acyl group attached to one of the pendant sugar residues. Despite these differences in structure, type I-IV GPAs all share a common mechanism of antibiotic action, which involves the sequestration of bacterial cell wall precursors (lipid II), thus preventing correct cell wall formation. With respect to type I-IV GPAs, type V GPAs form an outlier group [12], given that they all vary in the length of the peptide backbone (up to 10 amino acids), are not glycosylated [12], and share different peptide sequences. Despite this, type V GPAs do contain similarities to other GPAs, particularly in the crosslinking patterns between the aromatic residues.

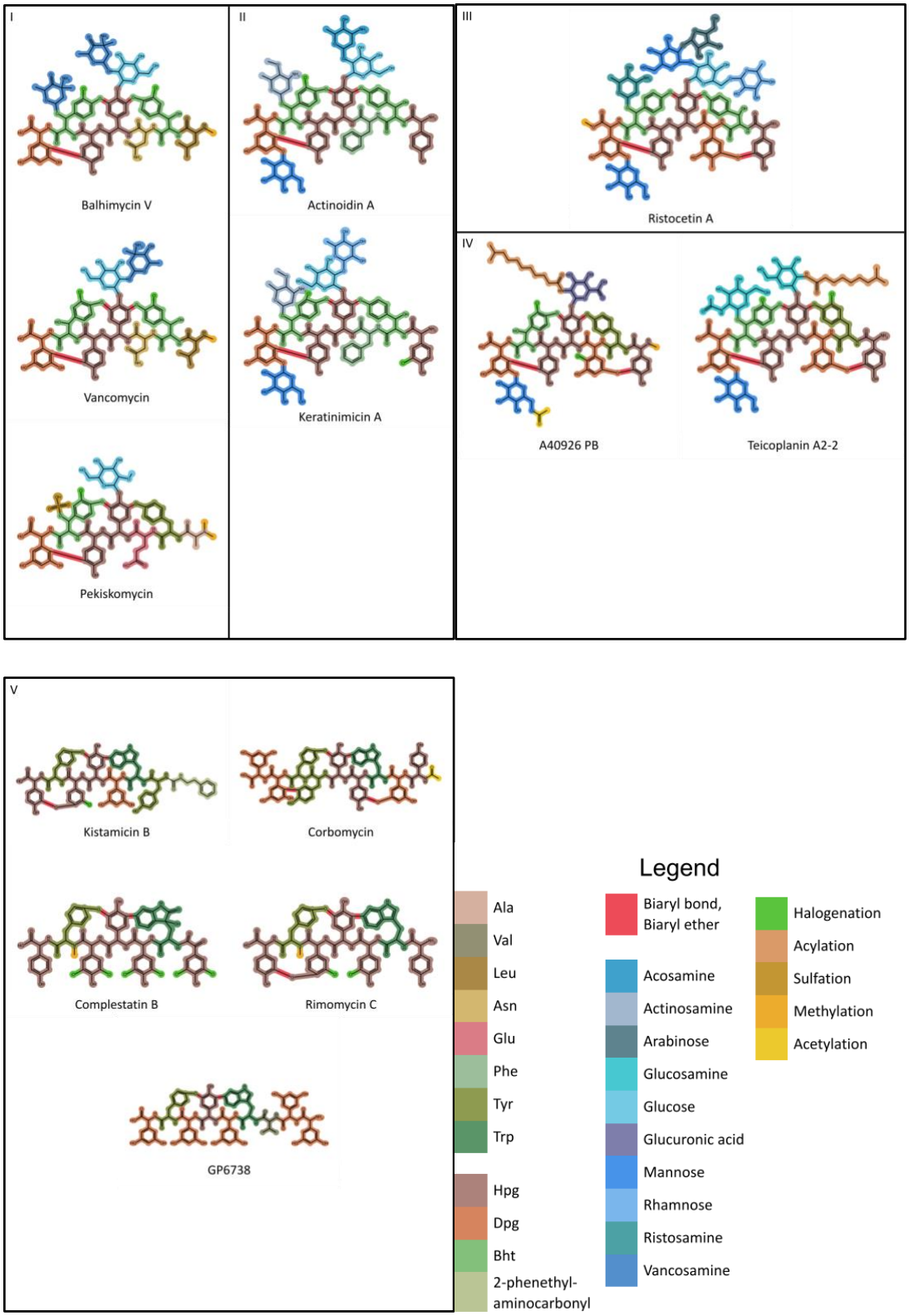



**Figure 1**: Representative structures of known GPA classes. The structures are visualised as straight backbones for comparison purposes and are separated into types I-V. The amino acids, biaryl bonds or biaryl ethers, sugar moieties and other decorations are coloured distinctly. Abbreviations: Tyr, tyrosine; Leu, leucine; Asn, asparagine; Bht, $\beta$-hydroxytyrosine; Hpg, 4-hydroxyphenylglycine; Dpg, 3,5-dihydroxyphenylglycine; Ala, alanine; Glu, glutamate; Phe, phenylalanine, Trp, tryptophane; Val, valine. The SMILES of the structures can be found in **Supplementary Table 1**

# GPA biosynthesis

The biosynthesis of GPAs has been extensively studied due to their importance, clinical application and the need for their production by *in vivo* fermentation of GPA producers [1], [3]–[10]. Understanding GPA biosynthesis is especially important if new derivatives of these compounds are to be produced at scale given the limitations of current chemical syntheses for these complex molecules [10], [13], [14].

Most biosynthetic enzymes necessary to produce GPAs are colocalized within the bacterial genome within so called biosynthetic gene clusters (BGCs), which also encode genes for export, regulation, and self-resistance. In principle, the biosynthesis of GPAs can be divided into three main steps: the synthesis of non-proteinogenic amino acids (1), the formation of the peptide backbone by the action of non-ribosomal peptide synthetases (NRPS), which are large multi-modular enzymes that assemble peptides without the involvement of the ribosome via the stepwise incorporation of single amino acids (2), and the modification of the peptide backbone (3).

### (1) Biosynthesis of the non-proteinogenic amino acids

All known GPAs contain non-proteinogenic amino acids, predominantly phenylglycine derivatives. [15]. Specific biosynthetic pathways for the formation of these residues are usually encoded as subclusters within the GPA-BGCs. The precursors of those non-proteinogenic amino acids namely hydroxyphenylpyruvate (4-HPP) and tyrosine are derived predominantly from the shikimate pathway. Since the shikimate metabolism is tightly regulated by feedback mechanisms, bacteria have developed alternative mechanisms to circumvent this regulation in order to ensure the supply of precursors for secondary metabolism. Producers of GPAs, for example, have acquired a second copy of the key shikimate pathway genes *dahp* and *pdh*, which are also found within GPA BGCs.

Hpg is biosynthesised from 4-HPP by the actions of a 4-hydroxymandelate oxidase (HmO) and a 4-hydroxymandelate synthase (HmaS). The amino group of tyrosine is transferred by the aminotransferase HptT/Pgat to form Hpg [19], [23],[24].

Dpg is biosynthesised from acetyl-coA through the actions of the DpgA,B,C,D enzymes, which encode a type III polyketide synthase as well as modifying enzymes that generate 3,5-dihydroxyphenylglyoxylate, before a final transamination reaction using Tyr performed by pgT/Pgat [17]–[20].

While the biosynthesis of phenylglycine residues is common to all GPAs, Bht is produced by two different mechanisms depending on the type of GPA. The first mechanism relies on the hydroxylation of Tyr by a non-heme iron oxygenase after selection/activation by the main nonribosomal peptide synthetase (NRPS) and loading of Tyr onto assembly line [23]. The correct modification of Tyr is then controlled by the atypical actions of a peptide bond forming domain within the main enzyme complex [24]. The second mechanism relies on the biosynthesis of Bht prior to its incorporation by the main assembly line [25], [26]. An additional minimal NRPS activates Tyr [29], which is subsequently hydroxylated by a cytochrome P450 monooxygenase [30]. Bht is then cleaved from the minimal NRPS module by a specific thioesterase [31], resulting in free Bht for subsequent activation by the main NRPS.

### (2) The formation of the backbone

The peptide backbone of GPAs is produced by NRPSs [17], [30], which within each module contain catalytic domains that perform specific functions in peptide biosynthesis. Adenylation domains (A-domains) recognize and activate specific amino acids for the incorporation into

the peptide chain [17], [30]. The amino acid is subsequently loaded as a thioester onto the thiol group of the phosphopantetheine arm of an adjacent peptidyl carrier protein (PCP) domain. This enables *trans* acting enzymes to further modify the PCP-tethered amino acid via halogenation [8] and hydroxylation [24]. The PCP domain then shuttles the amino acid to the acceptor pocket of the condensation (C) domain [33]. C-domains, usually located at the N-terminus of a module, catalyse amide bond formation between two typically PCP-bound substrates [33] via the addition of the amine group of the downstream acceptor amino acid to the thioester linkage of the upstream donor substrate [17], [30]. Some modules also include an epimerization domain (E domain), which converts the L-configured amino acid residue of the peptide C-terminal fragment into their D-form [34]. In this way, the peptide chain is progressively elongated until the final - for GPAs typically the seventh - module [19] [17].

### (2.1) Oxidative crosslinks and aglycone release

An essential step in the NRPS assembly of a GPA is the recruitment of up to four cross-linking enzymes to the bound peptide on the final NRPS module, where these specific cytochrome P450s, also known as Oxy enzymes, generate the crosslinks between the aromatic side chains of residues within the peptide chain. Recruitment of the Oxy enzymes to the peptide is facilitated by a recruitment domain unique to GPA biosynthesis: the X-domain. This domain, which is structurally related to a C-domain, uses a conserved interface to sequentially recruit P450s to the NRPS-bound peptide via a shuffling mechanism [41]. For type V GPAs, cyclisation activity retains the use of an X-type domain for Oxy recruitment, although individual Oxy enzymes have been identified that install multiple rings [6]. Additional enzymes may also act to reduce cyclic intermediates in the biosynthesis of some type V GPAs [40], although the timing and mechanism of this process remain unresolved. Completion of peptide cyclisation is catalysed by the terminal thioesterase (TE) domain, which releases the fully cyclised peptide products from the enzyme complex.

## (3) Modification of the peptide backbone

### (3.1) Sugar biosynthesis and decoration of the peptide chain

In addition to the cross-linked peptide backbone, another key feature of type I-IV GPAs is the presence of glycosyl moieties. These sugars are biosynthesised by several enzymes including epimerases, transaminases, dehydratases, and methyltransferases, mostly encoded within the GPA BGC. [40].

Numerous and different sugar moieties have been detected in the structures of GPAs, which is a major reason for their structural diversity [41]. After synthesis, the sugar moieties are linked to the cyclic peptide (aglycone) by the action of specific glycosyltransferases [17] .

### (3.2) Other modification reactions: acylation, methylation, and sulfation

Many GPAs contain optional modifications including methylation [17], [42], sulfation of amino acid side chains [17], [43] and the acylation of sugar moieties [17].

## (4) Export and mode of action

Once the GPA biosynthesis is complete, the active molecule is transported out of the cell by a specific ATP-binding cassette (ABC) transporter [44]. Type I-IV GPAs lead to the death of the producer's bacterial competitors by binding to the D-alanyl-D-alanine (D-Ala-D-Ala) terminus of lipid II, a key precursor of peptidoglycan [45]. In contrast, type V GPAs use a different, recently identified mode of action: they inhibit autolysins, i.e., enzymes responsible for remodelling peptidoglycan[46].

## How do the newly discovered GPAs fit into the classification?

Based on our understanding of the biosynthesis of GPAs and their mode of action, we can legitimately conclude that the type V GPAs that have been identified over the last few years differ considerably from the others, being of variable length, non-glycosylated and having a distinct mode of action [12], [14], [46], [47]. These differences raise the question as to whether these compounds truly belong to the class of GPAs or form a separate class. To address this question, we analysed an extensive dataset of GPA BGCs. Through comprehensive phylogenetic analyses conducted on multiple levels (whole BGCs, specific genes and domains) we confirmed a large "distance" between type V and all other types of GPAs, which is not only structural but also evolutionary in nature. Based on these results, we re-evaluated the classification of the types of GPAs and herein propose a new classification system that is guided by phylogeny and by chemical structure, which more appropriately defines these peptide natural products.

.

# Results

**Type V GPAs are a diverse group of compounds that differ significantly from other GPAs in structure and gene content.**

We started by exploring the evolutionary history of GPA encoding BGCs. To this end, we generated a dataset of GPA BGCs from previously published BGCs [12], [38], [46], [58]–[78] expanded through an extensive search of public databases for the presence of the GPA signature X-domain. Potential candidates were further filtered by manual investigation of BGC completeness. The 90 BGCs comprising the final dataset (**Supplementary Table 2**) were classified into type I-V GPAs based on similarity to the BGC of prototypical GPAs of each type. It became immediately apparent that BGCs of type V GPAs are far more diverse and distinct compared to those of types I-IV GPAs. Many genes commonly found in the BGCs of type I-IV GPAs were missing in type V BGCs, while additional genes encoding further enzymatic functions were present. Given these differences, we undertook further investigations based on the publicly available structures of GPAs.

Reported GPA structures were compared with the Tanimoto similarity matrix, and a network was visualised based on their pairwise similarity values (**Figure 2**: left panel). Although there is a high degree of similarity between the structures of type I-IV GPAs, there is a significant distance with those of type V GPAs. The type V GPAs appear to occupy a distinct (and broad) chemical space. These observations are reflected in the phylogenetic tree of GPA BGCs, based on the protein sequences of the encoded enzymes (**Figure 2** right panel). Here, the BGCs of type I-IV GPAs form distinct clades but are all part of the same subtree. On the other hand, the BGCs of type V GPAs take up a much larger part of the tree surface area and are clearly distinct from the BGCs of type I-IV GPAs.

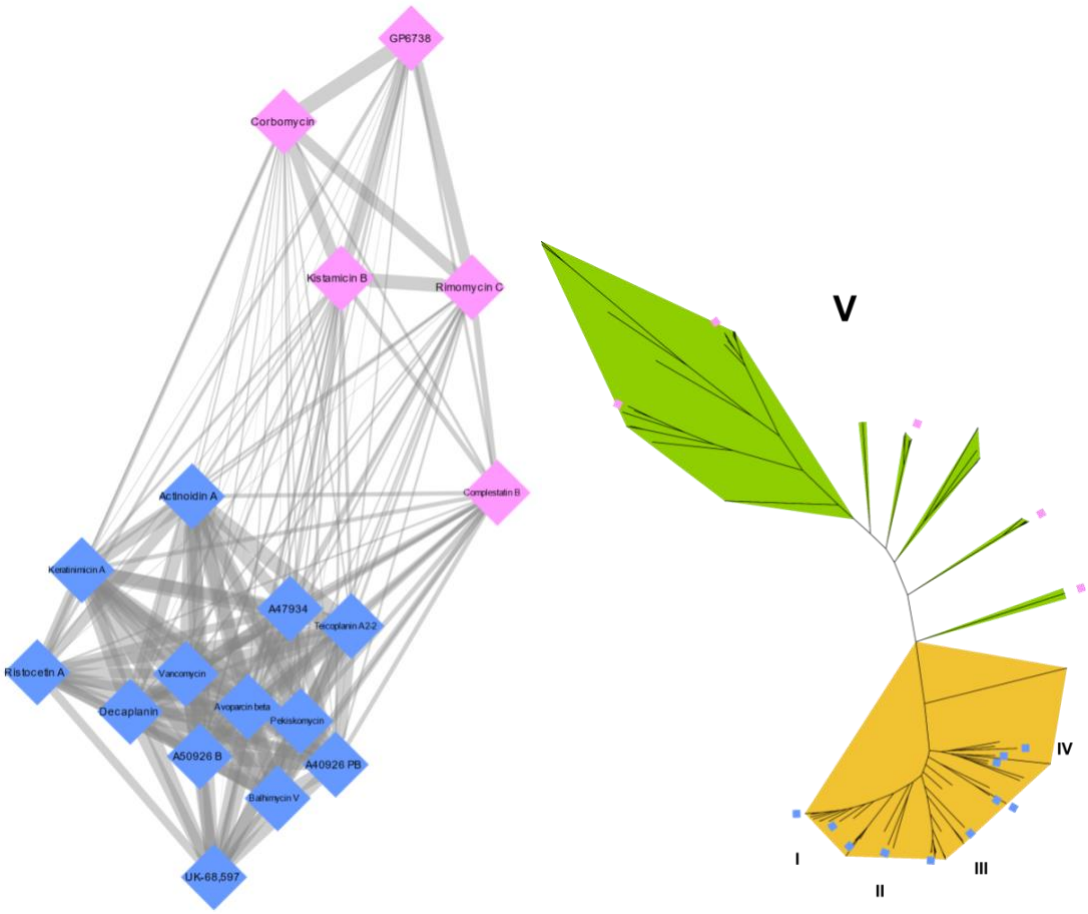



**Figure 2: Structural and phylogenetic distance of type V GPAs**. Left panel: Tanimoto similarity network of selected GPA structures (see Methods, **Supplementary Table 1**). The type V GPAs are coloured pink, while the type I-IV GPAs are shown in blue. The width of the edges represents the similarity value. Right panel: Unrooted concatenated phylogeny of all GPA BGCs in our dataset. BGCs of type V GPAs are highlighted in green, the others in yellow. The BGCs encoding the structures included in the left panel are marked with a coloured square on the corresponding leaves of the tree (**Supplementary Table 2**).

The type V GPAs, as mentioned earlier, have accumulated several peculiar features. The most notable of these is the absence of sugar moieties. In view of this, the term “glycopeptide” is inaccurate for molecules in the type V class. For this reason, and inspired by their separate, but not unrelated phylogeny, we propose a dichotomous reclassification of the current GPAs into true GPAs (types I-IV) and glycopeptide-related peptides (GRPs).

## Reclassification based on predicted backbone structure

Our dataset included many BGCs with no connection to known natural products and with very low similarities to known clusters. To gain some insight into their possible structures and hence their resemblance to known compounds, the sequences of the A-domains from the NRPS-encoding genes were extracted and an analysis of their Stachelhaus selection code was conducted (**Supplementary Table 3**). The predicted amino acid specificity of these A-domains, though not always conclusive, was used as a structure-based starting point for the initial classification of compounds and their BGCs into the GPA types and GRPs.

## Phylogenetic analysis of full clusters supports new naming convention

Once all homologous groups of genes (and domains, see Methods) had been identified and their trees calculated, they were used to build a representative phylogeny. A graphical summary of their phylogenetic trees can be seen in **Figure 3**. The difference in phylogenetic diversity between GPAs and GRPs is once again evident. The network illustrates the extensive gene flow that takes place between the BGCs of the two classes - the level of conservation between different genes and domains is not equal, increasing the complexity of their evolutionary history. However, the separation of GRPs from GPAs is still discernible. Moreover, also the presence of different subtypes within type I GPAs as well as the proposed class of GRPs. Additionally, due to the wide range of structural and genomic characteristics encompassed in the class of glycopeptide-related peptides (GRPs), we suggest the categorization of these into types A-E.

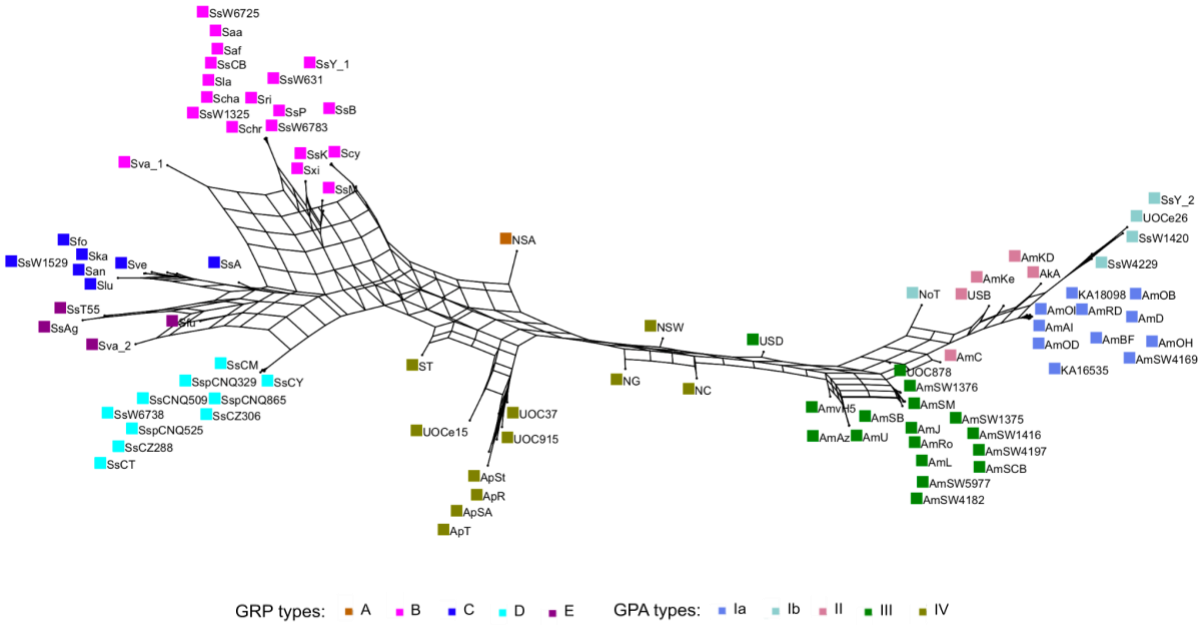



**Figure 3: Phylogenetic network of GPA BGCs**. Super network constructed from ML trees of all genes and domains from GPA and GRP encoding BGCs (**Supplementary Table 2**) (seed=0), computed with the SplitsTree program, using "greedy weak compatibility" filtering to reduce visual complexity. Each suggested new type is coloured differently next to the BGC ID. The network is built in three dimensions, which is why some branches are not completely clear.

To further support the previous analyses, an additional phylogenetic analysis was performed by concatenation of core genes and domains within all GPA and GRP BGCs. The "core" genes and domains were determined as genes or domains found in at least 90% of the clusters in the dataset. The concatenated phylogeny turned out to be in good agreement with the previous evolutionary picture. A combined visualisation of the representative phylogeny together with the gene content of each BGC (**Figure 4**) revealed several patterns that are characteristic for each suggested new type.

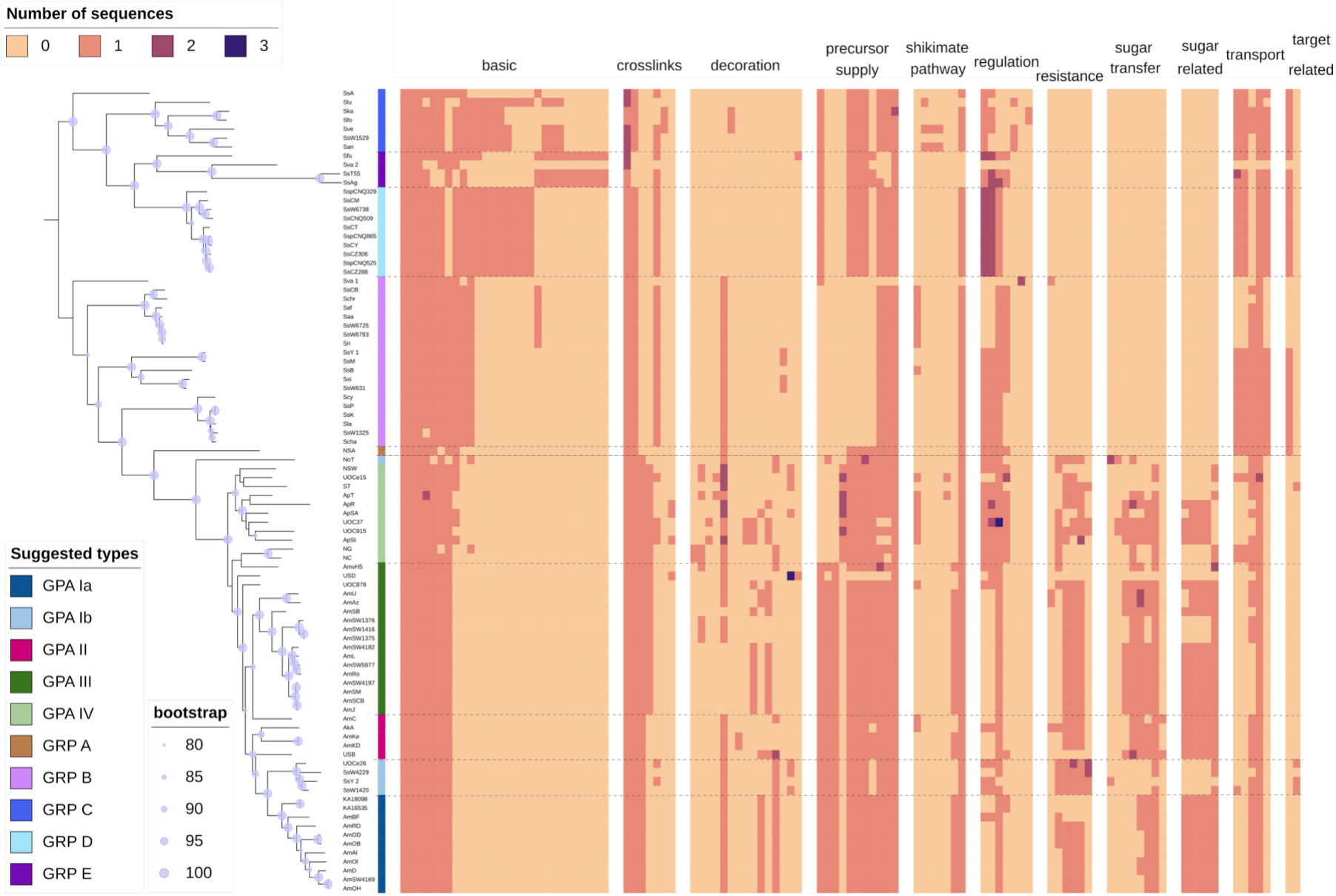

Number of sequences
0
1
2
3
basic
crosslinks
decoration
precursor supply
shikimate pathway
regulation
resistance
sugar transfer
sugar related
transport
target related
Suggested types
GPA Ia
GPA Ib
GPA II
GPA III
GPA IV
GRP A
GRP B
GRP C
GRP D
GRP E
bootstrap
80
85
90
95
100

**Figure 4: Phylogeny and gene patterns in GPA BGCs**. Left: concatenated phylogeny of GPA and GRP encoding BGCs (**Supplementary Table 2**). The types are coloured differently next to the tree labels (BGC IDs). Right: presence/absence heatmap of genes and domains, organised per general function.

## Reclassification of true glycopeptide antibiotics (GPAs)

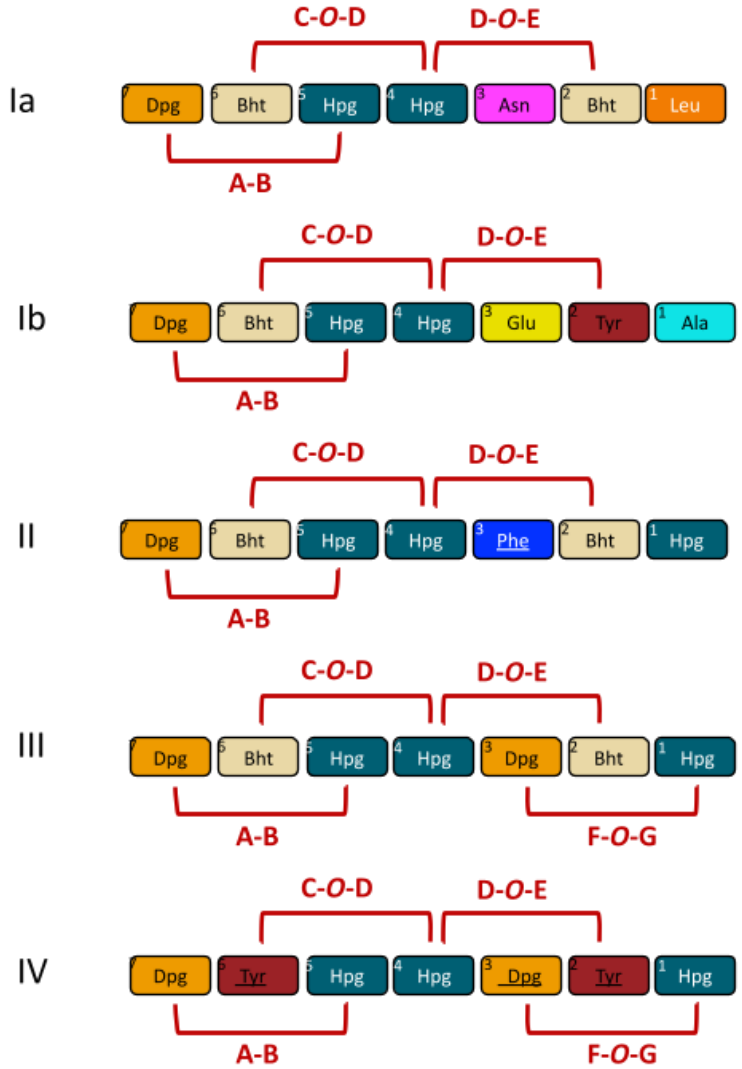



**Figure 5: GPA types I-IV backbone and cross-links.** For each type, the common amino acid composition and the positions of the oxidative cross links are shown. Underlined amino acids represent possible variations (β-hydroxytyrosine, Bht, or tyrosine, Tyr) within the type. Amino acids are determined based on either confirmed or predicted A-domain specificity (Stachelhaus code [79] analysis, **supplementary table 3**). Ala, alanine; Asn, asparagine; Bht, β-hydroxytyrosine; Dpg, dihydroxyphenylglycine; Glu, glutamate; Hpg, hydroxyphenylglycine; Leu, leucine; Phe, phenylalanine; Tyr, tyrosine.

We suggest dividing type I GPAs into two subtypes type Ia and type Ib, this is supported both by their phylogeny and their gene patterns (**Figure 5**). Type Ib GPAs differ from type Ia GPAs mainly by the different (non-aromatic) amino acids in positions 1 and 3 of the core peptide **(Figure 5)**. Unlike those of type Ia, their BGCs do not encode SAM-dependent

methyltransferases **(Figure 4)**. Type II to IV have not been changed, their gene patterns and phylogeny agree mostly with the previous classification.

## Classification of glycopeptide-related peptides (GRPs)

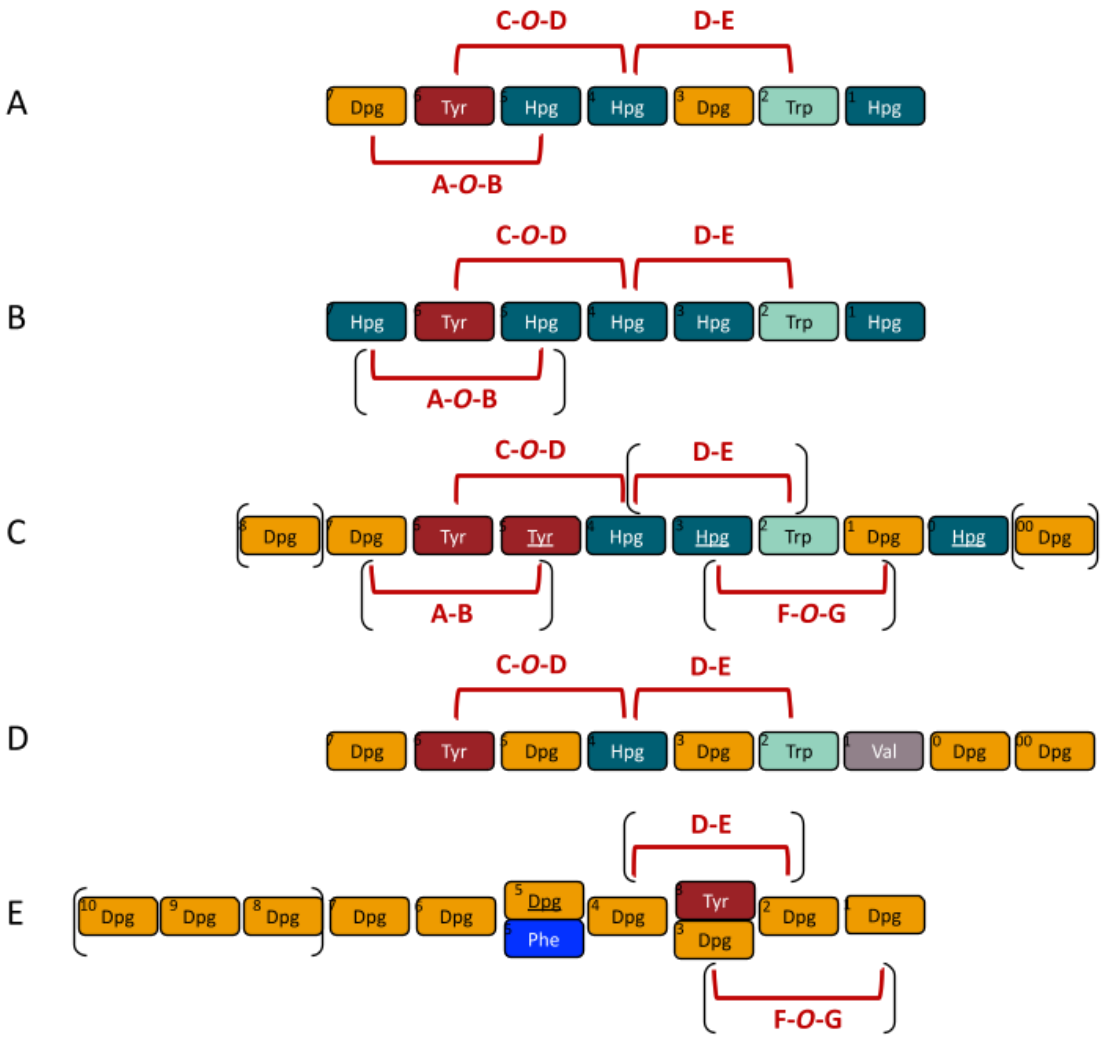



**Figure 6: GRP types A-E backbone and cross-links.** For each type, the representation of common amino acid composition and the positions of the oxidative cross links are shown. The schematics of the structures are in alignment with the structural centre of most compounds, which is the fourth amino acid. Underlined amino acids represent possible variations within the type. Elements in parenthesis are not always present in the type or are unconfirmed structural characteristics. Amino acids are determined based on either confirmed or predicted A-domain specificity (Stachelhaus code [79] analysis, **supplementary table 3**).

Type V GPAs have a different backbone, no post-aglycone modifications, may consist of more than seven amino acids, and usually contain a tryptophan (Trp) cross-linked to the central 4-hydroxyphenylglycine (Hpg). Although all GPAs affect cell wall biosynthesis in the target cell, the mode of action is different. Type I to IV GPAs bind the terminal D-alanyl-D-alanine (D-Ala-D-Ala) terminus of lipid II and thereby inhibit transglycosylation of the N-acetyl-glucosamine

pentapeptide from lipid II to the backbone glycans of the bacterial cell wall and transpeptidation of the peptidoglycan strands (Sheldrick et al., 1978). Type V GPAs, instead, bind to peptidoglycan and thereby inhibit autolysins (Culp et al., 2020). (**Figure 4**).

### Type A GRPs

The proposed type A GRPs include only the BGC encoding kistamicin biosynthesis [6]. This BGC contains four genes (*dpgA-D*) encoding Dpg biosynthesis (**Figure 4**). Notably, its NRPS genes are composed of 7 modules (**Figure 6**), but the first module is missing the A-domain, a phenomenon still unresolved.

### Type B GRPs

Type B GRPs differ from type A in the absence of genes associated with Dpg biosynthesis (**Figure 4**), which is reflected in the absence of Dpg-specific A domains within the NRPS (**Figure 6**). Based mainly on differences in their regulatory genes, they can be further subdivided into three subtypes. The complestatin-like B1 subtype BGCs include only the vanRS two-component system and a StrR-like regulator, the rimomycin-like B2 subtype BGCs contain only a StrR-like and a LuxR-like regulator, and the misaugamycin-like B3 subtype BGCs carry a copy of all of these genes (**Figure 4**).

### Type C GRPs

The number of amino acids in the backbone of type C GRPs (e.g., corbomycin) varies between 8 and 10 and is reflected by an equal number of modules in the NRPS genes of the corresponding BGCs (**Figure 6**). Other distinctive features of these BGCs are the presence of only one $\alpha/\beta$ hydrolase gene from the three-enzyme Bht biosynthetic cassette and the presence of a vanR-like but not a StrR-like regulator (**Figure 4**).

### Type D GRPs

The GRP encoding BGCs of type D are very uniform, as evidenced t by their clearly defined monophyletic clade in all phylogenetic representations (**Figures 3, 4**) and their gene patterns (**Figure 4**). These GRPs always contain nine amino acids in their backbone, with the third residue of the peptide likely to be a valine (Val) residue (**Figure 6**). Like type C GRPs, they also contain only the $\alpha/\beta$ hydrolase from the Bht-biosynthesis-related cassette, their unique characteristic is the presence of two copies of the vanR/vanS two-component system (**Figure 4**).

### Type E GRPs

Type E comprises GRPs that are either heptapeptides or decapeptides. Their backbone contains neither Trp, nor Hpg (**Figure 6**). As with type C and D GRPs, the BGCs of these compounds contain only the $\alpha/\beta$ hydrolase of the Bht cassette with an additional unique characteristic being the lack of an oxyC gene (**Figure 4**).

# Discussion

Following up on the blatant differences of type V GPAs with all other types both in structure and mode of action, an extensive phylogenetic analysis of the encoding BGCs was conducted. We were able to confirm those discrepancies on the evolutionary level and to underline the contrast in variety within this type compared to the rest. These observations were the basis for the suggested reclassification of GPAs, dichotomizing them first into true GPAs and glycopeptide-related peptides (GRPs) and then further categorising them into types. The original types I-IV were retained, with the exception of type I which was split into type Ia and type Ib. Type V GPAs were rebranded as GRPs, and their divergent characteristics compelled their division into the new types A-E.

A milestone of our analysis was the separation of the genes involved in GPA BGCs into homologous groups, suitable for alignment and for generating a biologically meaningful phylogenetic tree. The use of suitable bioinformatics tools for solving this problem, in our case OrthoFinder [82], [83], only helped partially because the different kinds of genes have different degrees of sequence similarity. For example, there are five types of p450 monooxygenase genes found in GPA BGCs: *oxyA-E* [6]. The algorithm was able to separate *oxyD* and place it in its own group, which is reasonable considering the unique function it has among GPA oxygenases, but the remaining ones were all pooled together into one group. To separate the rest of the other oxy genes, their phylogeny was visualised (**SFigure 1**) and their classification was based on their cladding in proximity to genes with already resolved functions. The same method was followed also for a group of glycosyltransferase-encoding genes (**SFigure 2**).

The classification into types I-V was made based on the presence of genes homologous to those identified so far in GPA BGCs [58]. Consequently, one temporal constraint of this analysis was the fact that the BGCs needed to be manually and thoroughly checked, in order to both confirm that they are, in fact, GPA encoding BGCs and to determine their precise borders. These were in a first step defined with the antiSMASH algorithm [57] which, though very efficient, is known to be charitable with the proposed length of a cluster, often containing neighbouring genes that do not really belong [80]. Even though this is still preferable to excluding genes that should be part of the cluster, it still demands a manual inspection of the BGCs, if the exact limits are important for further analyses, like in our case. Therefore, each antiSMASH-predicted BGC in our dataset had to be extensively investigated gene by gene (and sometimes domain by domain) to locate the most likely borders of the BGC based on the predicted function of the genes/domains within the cluster. This implies that the detection of the exact BGC limits was heavily reliant on the existing knowledge on the biosynthesis of known compounds and may be improved should new results come to light, for example, on the (so far unknown) role of certain enzymes in the biosynthesis.

Furthermore, our analysis was naturally constrained by the availability and quality of public genomic information. In fact, some of the BGCs in our dataset came from lower quality or incomplete assemblies of the producer genomes. Whenever possible, a better-quality sequence containing the BGC was extracted from more recently sequenced and assembled genomes (and in one case the producer strain was resequenced). Unfortunately, the 20 orphan clusters, whose full genome was not made public, contained low quality sequences, thus limiting an accurate definition of these clusters [81].

Another challenging aspect of GPA BGC phylogeny is the fact that evolution is acting on different levels: on the separate genes, but also on the whole BGCs and on distinct domains [52], [84]–[87]. As explained earlier, the backbone of GPAs and GRPs is assembled by NRPS enzymes, which can contain from 7 to 10 modules. In each module there are always A and T domains, very often C domains and sometimes also E domains. This means that for each position, the corresponding domain sequences needed to be collected to form a homologous group that could be used for constructing phylogeny [88]. The variation in the number of modules introduced an additional difficulty, since a “centre” position had to be chosen around which the rest of the domains would be aligned. Eventually, the Dpg in position 4 was chosen for this purpose, as it is the only one participating in two oxidative cross-links and is also present in most GPA and GRP types (**Figure 5**, **Figure 6**). As type E GRPs miss this central Dpg, we cannot rule out the possibility that their domains are aligned in the wrong positions. This is an inevitable inaccuracy that can only be rectified when a member of this type has its structure elucidated. We still considered it useful to include those BGCs in the full phylogenetic analyses.

Inference of the evolutionary history of a full BGC was complicated by the known fact that many GPA encoding BGCs have been, in part or in their entirety, horizontally transferred [40], [86], [89]. The potential presence of horizontal gene transfer (HGT) is always an obstacle to the construction of a biologically sound, concatenated phylogeny. To exclude the use of sequences of suspicious origin, we conducted a congruence analysis [90] prior to the choice of the sequences to include in the concatenated phylogeny of the GPA and GRP BGCs. While reducing the sensitivity of the final phylogenetic tree, since fewer genes and domains could be involved, this process, however, guaranteed a reduction of the “noise” that HGTed sequences would have introduced.

Considered all the challenges we faced in the reclassification process, we struggled in some instances to choose a level of dissimilarity as a criterion for declaring the new types, especially for the GRPs. And there were debatable cases, especially when the phylogeny of the BGCs with an uneven number of modules placed them closely together. Such was the case for type E GRPs (**Figure 6**), where no compound has been isolated and thereby no structure could be used as a guide for those classifications.

Additionally, there were two cases of BGCs (NG and NC) whose encoded GPAs should be classified as GPAs type IV but they displayed an overwhelming amount of sequences cladding closer to the type III GPAs. This was to be expected, since the main structural difference between compounds of type III and IV is the presence of an acyl group bound to a sugar moiety (**Figure 1**) [12]. The two types are closely related and are not always forming distinct monophyletic clades (**Figure 2, Figure 3**). But since the presence of the acyl group affects the potency of the produced compounds [91], we considered it a factor important enough to retain the original structure-based classification, to which the phylogeny also largely agrees. In the concatenated phylogeny of the full BGCs (**Figure 4**) the type A GRPs are also not forming monophyletic clades - though their subtypes 1-3 actually do.
A quite surprising exception to our types was the existence of GPAs without sugars, namely ST (A47934) which clades clearly with type IV GPAs and USD (TEG), also a type III GPA, though for the latter there is limited information. Based on the phylogenetic relation to their respective types, it can be assumed that they originally did include sugar-related genes (and

possibly also an acyltransferase for ST), which were lost later in their evolutionary history. A more focused study on these clusters is needed to elucidate such events.

Finally, there was one BGC whose classification could not be resolved. The NoT BGC appears to bear characteristics of multiple true GPAs: It has glycosyltransferases and specific resistance genes (like GPAs), has 3 P450 monooxygenases (like GPA types I and II), its Bht biosynthesis genes are mostly missing (unlike any GPA), has no *dahp* gene (like GPA type Ib), and has a SAM dependent methyltransferase gene type I but not type II (like GPA type Ib). Depending on the gene studied, this BGC falls into different clades. A possible explanation is that this cluster represents an intermediate type of true GPAs, whose history involved several recombination events.

Though the path to calculating a reliable phylogeny of GPA and GRP BGCs was fraught with challenges, the current analysis does provide the - to our knowledge - most comprehensive study of the evolution of these BGCs. The evidence supporting our suggested reclassification system is present and compelling. After all, the distinct structural characteristics of GPAs prompted their classification into different types, [11], [12] and it promoted exchange between scientists from many different disciplines (e.g., biochemistry, synthetic chemistry, microbiology, bioinformatics) studying them from a different perspective. Putting together in the same group compounds and BGCs that are so different would have only hindered progress in the corresponding field. It was of course a necessity that was born from the fact that type V GPAs were the last ones discovered [92], and every new and differently looking GPA BGC has been sorted there since [12], [38], [46], [93], [94]. However, the new classification system we suggest, backed by an exhaustive phylogenetic analysis of a large dataset of BGCs, is expected to fuel new, type-specific research of GPAs and GRPs. The use of both phylogeny and structure-based criteria for the declaration of the new types will ensure the reconciliation of scientific communities focused on evolution and chemical structure, respectively. We hope that this new classification system will get adopted by the community and will improve with new insights and feedback from the community.

# Methods

## BGC dataset creation

In order for the evolutionary analysis of GPAs BGCs to be as complete as possible, we aimed to create a dataset of all sequenced clusters. The initial dataset contained all known clusters found in the literature [12], [38], [46], [58]–[78] and was extended by searching in all publicly available sequence databases (NCBI RefSeq [95], [96], JGI IMG DB [97]) and some published projects [98]–[100]. The target of the search, accommodated via HMMER (v. 3.3.1, accessed from http://hmmer.org/), was the X domain in the last NRPS module (its HMM was extracted from antiSMASH v6 [57]), which is found only in GPA encoding BGCs so far [31]. However, the X domain has evolved from a condensation domain (C domain) and their sequences remain quite similar [7]. To avoid false positives, the known clusters were searched with the X domain HMM and the lowest local score of the confirmed X domains (300) was used as a minimal threshold. Lower local scores (up to 250) were manually checked and indeed no X domains were found below the chosen threshold.

The sequences that were a hit in the search were used as input for antiSMASH v6 [57] for the detection of the full BGC. At this stage, after manual inspection, some candidates were dropped due to low quality assemblies or incomplete clusters. In some cases there were BGCs with obvious sequencing errors or gaps introducing frame shifts, which was indicated by visibly fragmented genes and was made clear through a six-frame alignment of the BGC sequence to itself with Geneious Prime 2022.0.2 (https://www.geneious.com). In those cases, the sequences were corrected if possible via simple nucleotide additions. This seemed to improve the quality of our analysis but it does introduce the possibility for errors, since only an improved sequencing and assembly of the producer genome can elucidate the true sequences. For the RefSeq sequences, their Genbank counterparts were also checked and in some cases the quality of the BGC was better, so that assembly was the one used for this project. The assortment of genomes was re-annotated with prokka (v1.14.16, default parameters with specified genus when known) [101] to ensure homogeneity.
Knowing this and to ensure the quality of the evolutionary reconstruction, which would be affected by the accidental inclusion of unrelated sequences, we meticulously manually curated each cluster by investigating every single gene for its function and possible role in the biosynthesis. Sequencing errors resulting in frameshifts were manually corrected. Then the BGCs were trimmed to their true limits. Thus, the final dataset of 90 trimmed clusters from 7 different genera was created (**Supplementary Table 2**).

## Chemical structure similarity network

The structures of the glycopeptides (**Figure 1**) were collected from the original publications [12], [38], [46], [64], [65], [67], [71], [94], [102]–[111] (**Supplementary Table 1**). In order to generate proper depictions and to align all the molecules in a unified way, a python script was created using the cheminformatics toolkit RDKit (v. 2022.3.4, accessed from

https://www.rdkit.org/). Since the biaryl and biaryl ether bonds often cause problems while generating 2D coordinates, the script removes these bonds, generates a 2D representation for the remaining structure and finally adds the bonds again. After manual curation with MarvinSketch (v. 22.19.0, ChemAxons, accessed from https://www.chemaxon.com), all the glycopeptides had clear 2D representations with the backbones aligned horizontally, the carboxyl-terminus to the left and the Hpg with two biaryl ethers on top of the structure (as is the common consensus in the literature).
The examination of the structural variation within and among GPA types was done by use of the Tanimoto similarity metric [112]. The SMILES of known GPA compounds were inserted into the Cytoscape program (v. 3.9.1) [113], which accommodates the calculation and visualisation of a structure similarity network based on Tanimoto with *chemViz2*, a cheminformatics app for Cytoscape. The nodes of the resulting network were annotated by GPA type (I-IV or V) and the edge width was adjusted to represent the degree of similarity among the connecting parts.

## Stachelhaus code analysis

A-domain specificity was predicted (**Supplementary Table 3**) using the Stachelhaus-code prediction from NRPSpredictor2, implemented in antiSMASH (v. 5.1.2) [79], [114], [115]. Known amino acid specificities for glycopeptides, which sometimes differed from the predicted ones, were taken into account. A Stachelhaus-code prediction with 70% similarity or less was considered as weak and the amino acid was classified as unknown. A summary of A domain specificity was visualised in **Figure 5** and **Figure 6**.

## Detection of homologous genes

The identification of homologous groups of genes from the full dataset (**Supplementary Table 2**) was mostly carried out by the OrthoFinder tool (v. 2.3.11) [82], which infers phylogenetic orthology for comparative genomics. Its original intended input is proteomes, but for the purposes of this study the protein sequences of the clusters were used. The platform (run with default parameters) generated 94 (n>1) so-called Orthogroups. However, there were cases of related genes with distinct functions being placed in the same Orthogroup, like with the oxy genes and a glycosyltransferase group. Those proteins were used for multiple sequence alignments (MSAs) and their separation into groups was guided by their phylogenetic placement compared to proteins of known function. Similar cases with only a few clusters containing multiple copies of a gene were dealt with via a custom script, discarding some copies based on similarity criteria [116]. Finally, the NRPS domains were extracted from the genes and split into groups based on the position of their module compared to the rest.

## Visualising complete evolutionary history

All MSAs were performed with the mafft tool (v7.490, 2021/Oct/30) [117]. Phylogenetic trees for every occasion in this study were built with iqtree (multicore v2.2.0.3 COVID-edition for Linux 64-bit built Aug 2 2022) [118], [119] (parameters used: -nt AUTO -m TEST -bb 1000) and visualised with iTOL [120].
A graphical summary of the 125 separate (partial) gene and domain trees was calculated and visualised by the super network algorithm [121] (default options) implemented in the SplitsTree

CE tool (version 6.0.10-beta) [122], [123] (**Figure 3**) after greedily selecting a weakly compatible set of splits of maximum support (GreedyWeaklyCompatible splits filter). The super network summarises the set of input trees, taking into account that many of the trees are incomplete. It is a splits network in which each band of parallel edges represents one of the splits or branches found in the set of input trees. Incompatibilities among the input trees give rise to parallelograms in the network. Edges in the network are scaled to represent the average relative length of the corresponding edges in the input trees. The set of splits computed by the super network method was greedily filtered by decreasing support (number of trees that contain a given split), so as to obtain a subset of "weakly compatible splits" that maintains major incompatibilities, while avoiding higher-dimensional edge configurations in the network, thus avoiding visual clutter. The super network displayed in Figure 3 indicates that the set gene trees largely agree on the partitioning of the GPAs BGCs into the described types, despite an otherwise high level of incompatibility among the trees.

## Concatenated phylogeny

A species phylogeny can be constructed from concatenated sequences of core genes, as long as their separate trees are congruent [124]. Following this concept on the BGCs, 30 genes or domains found in at least 90% of the clusters in the dataset were identified. Cases where there were duplications in max 5% of the cases were acceptable but otherwise these groups were single-copy genes. These limits come from the methodology of a wide-scale phylogenetic study [124]. However, it was necessary to ensure the congruence of the participating genes before concatenating them. The underlying MSAs were first trimmed with trimAl (v1.4.rev15 build[2013-12-17]) [125] with default parameters and then the trees were recalculated. The latter were used in a congruence analysis as performed by Parks and colleagues [124]: The well supported splits were calculated by checking their existence in random subsampled concatenated phylogenies and then for each gene tree, their presence was used to calculate 'normalised compatible split length' ,a metric that reflects congruence of this tree to the rest in the group (**Supplementary Table 4**). This value was computed with a python script, implemented with Biopython [126] , FastTree v2.1.11 [127] and a function from the (still in development) GeneTreeTk toolbox (accessed from https://github.com/dparks1134/GeneTreeTk). 11 genes or domains passed a specific threshold (0.67) and were used for the concatenated phylogeny representing the evolutionary history of the GPA and GRP BGCs. This tree, together with an absence/presence heatmap of the various genes and domains present in the clusters was visualised with iTOL [120] (**Figure 4**).

# References


[1] E. Stegmann, H. J. Frasch, and W. Wohlleben, “Glycopeptide biosynthesis in the context of basic cellular functions,” *Current Opinion in Microbiology*, vol. 13, no. 5, pp. 595–602, 2010, doi: 10.1016/j.mib.2010.08.011.

[2] E. van Groesen, P. Innocenti, and N. I. Martin, “Recent Advances in the Development of Semisynthetic Glycopeptide Antibiotics: 2014–2022,” *ACS Infect Dis*, vol. 8, no. 8, pp. 1381–1407, Aug. 2022, doi: 10.1021/acsinfecdis.2c00253.

[3] M. S. Butler, K. A. Hansford, M. A. T. Blaskovich, R. Halai, and M. A. Cooper, “Glycopeptide antibiotics: Back to the future,” *J Antibiot*, vol. 67, no. 9, Art. no. 9, Sep. 2014, doi: 10.1038/ja.2014.111.

[4] M. H. Mccormick, J. M. Mcguire, G. E. Pittenger, R. C. Pittenger, and W. M. Stark, “Vancomycin, a new antibiotic. I. Chemical and biologic properties,” *Antibiot Annu*, vol. 3, pp. 606–611, 1956 1955.

[5] M. H. Hansen, E. Stegmann, and M. J. Cryle, “Beyond vancomycin: recent advances in the modification, reengineering, production and discovery of improved glycopeptide antibiotics to tackle multidrug-resistant bacteria,” *Curr Opin Biotechnol*, vol. 77, p. 102767, Oct. 2022, doi: 10.1016/j.copbio.2022.102767.

[6] A. Greule *et al.*, “Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics,” *Nature Communications*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-10384-w.

[7] M. Schoppet *et al.*, “The biosynthetic implications of late-stage condensation domain selectivity during glycopeptide antibiotic biosynthesis,” *Chemical Science*, vol. 10, no. 1, pp. 118–133, 2019, doi: 10.1039/C8SC03530J.

[8] T. Kittilä *et al.*, “Halogenation of glycopeptide antibiotics occurs at the amino acid level during non-ribosomal peptide synthesis,” *Chemical Science*, vol. 8, no. 9, pp. 5992–6004, 2017, doi: 10.1039/c7sc00460e.

[9] W. Wohlleben, E. Stegmann, and R. D. Süssmuth, “Chapter 18. Molecular genetic approaches to analyze glycopeptide biosynthesis,” *Methods Enzymol*, vol. 458, pp. 459–486, 2009, doi: 10.1016/S0076-6879(09)04818-6.

[10] A. W. Truman, Q. Fan, M. Röttgen, E. Stegmann, P. F. Leadlay, and J. B. Spencer, “The Role of Cep15 in the Biosynthesis of Chloroeremomycin: Reactivation of an Ancestral Catalytic Function,” *Chemistry & Biology*, vol. 15, no. 5, pp. 476–484, May 2008, doi: 10.1016/j.chembiol.2008.03.019.

[11] S. Chen, Q. Wu, Q. Shen, and H. Wang, “Progress in Understanding the Genetic Information and Biosynthetic Pathways behind Amycolatopsis Antibiotics, with Implications for the Continued Discovery of Novel Drugs,” *ChemBioChem*, vol. 17, no. 2, pp. 119–128, Jan. 2016, doi: 10.1002/cbic.201500542.

[12] M. Xu, W. Wang, N. Waglechner, E. J. Culp, A. K. Guitor, and G. D. Wright, “GPAHex-A synthetic biology platform for Type IV–V glycopeptide antibiotic production and discovery,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–12, Oct. 2020, doi: 10.1038/s41467-020-19138-5.

[13] C. Kegler and B. Helge B, “Artificial Splitting of a Non-Ribosomal Peptide Synthetase by Inserting Natural Docking Domains,” *Angew Chem Int Ed Engl.*, vol. 10, 2020, doi: 10.1002/anie.201915989.

[14] K. A. J. Bozhüyük *et al.*, “De novo design and engineering of non-ribosomal peptide synthetases,” *Nature Chemistry 2017 10:3*, vol. 10, no. 3, pp. 275–281, Dec. 2017, doi: 10.1038/nchem.2890.

[15] R. S. A. Toma, C. Brieke, M. J. Cryle, and R. D. Süssmuth, “Structural aspects of phenylglycines, their biosynthesis and occurrence in peptide natural products,” *Nat. Prod. Rep.*, vol. 32, no. 8, pp. 1207–1235, Jul. 2015, doi: 10.1039/C5NP00025D.

[16] J. Thykaer *et al.*, “Increased glycopeptide production after overexpression of shikimate pathway genes being part of the balhimycin biosynthetic gene cluster,” *Metabolic Engineering*, vol. 12, no. 5, pp. 455–461, Sep. 2010, doi: 10.1016/j.ymben.2010.05.001.

[17] G. Yim, M. N. Thaker, K. Koteva, and G. Wright, “Glycopeptide antibiotic biosynthesis,” *Journal of Antibiotics*, vol. 67, no. 1, pp. 31–41, Jan. 2014, doi: 10.1038/ja.2013.117.

[18] H. Chen, C. C. Tseng, B. K. Hubbard, and C. T. Walsh, “Glycopeptide antibiotic biosynthesis: Enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 26, pp. 14901–14906, Dec.

2001, doi: 10.1073/pnas.221582098.
[19] C. C. Tseng, S. M. McLoughlin, N. L. Kelleher, and C. T. Walsh, “Role of the Active Site Cysteine of DpgA, a Bacterial Type III Polyketide Synthase,” *Biochemistry*, vol. 43, no. 4, pp. 970–980, Feb. 2004, doi: 10.1021/bi035714b.
[20] V. Pfeifer *et al.*, “A Polyketide Synthase in Glycopeptide Biosynthesis: THE BIOSYNTHESIS OF THE NON-PROTEINOGENIC AMINO ACID (S)-3,5-DIHYDROXYPHENYLGLYCINE*,” *Journal of Biological Chemistry*, vol. 276, no. 42, pp. 38370–38377, Oct. 2001, doi: 10.1074/jbc.M106580200.
[21] T.-L. Li, O. W. Choroba, E. H. Charles, A. M. Sandercock, D. H. Williams, and J. B. Spencer, “Characterisation of a hydroxymandelate oxidase involved in the biosynthesis of two unusual amino acids occurring in the vancomycin group of antibiotics,” *Chem. Commun.*, no. 18, pp. 1752–1753, Jan. 2001, doi: 10.1039/B103548G.
[22] “Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics - ScienceDirect.” https://www.sciencedirect.com/science/article/pii/S1074552100000430?via%3Dihub (accessed Feb. 02, 2023).
[23] S. Stinchi *et al.*, “A derivative of the glycopeptide A40926 produced by inactivation of the β-hydroxylase gene in Nonomuraea sp. ATCC39727,” *FEMS Microbiology Letters*, vol. 256, no. 2, pp. 229–235, 2006, doi: 10.1111/j.1574-6968.2006.00120.x.
[24] M. Kaniusaite, J. Tailhades, E. A. Marschall, R. J. A. Goode, R. B. Schittenhelm, and M. J. Cryle, “A proof-reading mechanism for non-proteinogenic amino acid incorporation into glycopeptide antibiotics,” *Chemical Science*, vol. 10, no. 41, pp. 9466–9482, 2019, doi: 10.1039/C9SC03678D.
[25] H. Chen and C. T. Walsh, “Coumarin formation in novobiocin biosynthesis: beta-hydroxylation of the aminoacyl enzyme tyrosyl-S-NovH by a cytochrome P450 NovI,” *Chem Biol*, vol. 8, no. 4, pp. 301–312, Apr. 2001, doi: 10.1016/s1074-5521(01)00009-6.
[26] H. Chen, B. K. Hubbard, S. E. O’Connor, and C. T. Walsh, “Formation of beta-hydroxy histidine in the biosynthesis of nikkomycin antibiotics,” *Chem Biol*, vol. 9, no. 1, pp. 103–112, Jan. 2002, doi: 10.1016/s1074-5521(02)00090-x.
[27] J. Recktenwald *et al.*, “Nonribosomal biosynthesis of vancomycin-type antibiotics: a heptapeptide backbone and eight peptide synthetase modules,” *Microbiology (Reading)*, vol. 148, no. Pt 4, pp. 1105–1118, Apr. 2002, doi: 10.1099/00221287-148-4-1105.
[28] M. J. Cryle, A. Meinhart, and I. Schlichting, “Structural Characterization of OxyD, a Cytochrome P450 Involved in β-Hydroxytyrosine Formation in Vancomycin Biosynthesis,” *The Journal of Biological Chemistry*, vol. 285, no. 32, p. 24562, Aug. 2010, doi: 10.1074/JBC.M110.131904.
[29] S. Mulyani *et al.*, “The thioesterase Bhp is involved in the formation of beta-hydroxytyrosine during balhimycin biosynthesis in Amycolatopsis balhimycina,” *Chembiochem*, vol. 11, no. 2, pp. 266–271, Jan. 2010, doi: 10.1002/cbic.200900600.
[30] B. R. Miller and A. M. Gulick, “Structural Biology of Non-Ribosomal Peptide Synthetases,” *Methods Mol Biol*, vol. 1401, pp. 3–29, 2016, doi: 10.1007/978-1-4939-3375-4_1.
[31] K. Haslinger, M. Peschke, C. Brieke, E. Maximowitsch, and M. J. Cryle, “X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis,” *Nature*, vol. 521, no. 7550, pp. 105–109, 2015, doi: 10.1038/nature14141.
[32] E. Stegmann, C. Rausch, S. Stockert, D. Burkert, and W. Wohlleben, “The small MbtH-like protein encoded by an internal gene of the balhimycin biosynthetic gene cluster is not required for glycopeptide production,” *FEMS Microbiol Lett*, vol. 262, no. 1, pp. 85–92, Sep. 2006, doi: 10.1111/j.1574-6968.2006.00368.x.
[33] T. Izoré *et al.*, “Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity.,” *Nature communications*, vol. 12, no. 1, p. 2511, May 2021, doi: 10.1038/s41467-021-22623-0.
[34] L. Luo, R. M. Kohli, M. Onishi, U. Linne, M. A. Marahiel, and C. T. Walsh, “Timing of Epimerization and Condensation Reactions in Nonribosomal Peptide Assembly Lines: Kinetic Analysis of Phenylalanine Activating Elongation Modules of Tyrocidine Synthetase B,” *Biochemistry*, vol. 41, no. 29, pp. 9184–9196, Jul. 2002, doi: 10.1021/bi026047+.
[35] K. Woithe *et al.*, “Oxidative Phenol Coupling Reactions Catalyzed by OxyB: A Cytochrome P450 from the Vancomycin Producing Organism. Implications for Vancomycin Biosynthesis,” *J. Am. Chem. Soc.*, vol. 129, no. 21, pp. 6887–6895, May 2007, doi: 10.1021/ja071038f.
[36] E. Stegmann *et al.*, “Genetic analysis of the balhimycin (vancomycin-type) oxygenase genes,” *Journal of Biotechnology*, vol. 124, no. 4, pp. 640–653, Aug. 2006, doi: 10.1016/J.JBIOTEC.2006.04.009.

[37] K. Zerbe, K. Woithe, D. B. Li, F. Vitali, L. Bigler, and J. A. Robinson, “An Oxidative Phenol Coupling Reaction Catalyzed by OxyB, a Cytochrome P450 from the Vancomycin-Producing Microorganism,” *Angewandte Chemie International Edition*, vol. 43, no. 48, pp. 6709–6713, 2004, doi: 10.1002/anie.200461278.
[38] M. Xu, W. Wang, N. Waglechner, E. J. Culp, A. K. Guitor, and G. D. Wright, “Phylogeny-Informed Synthetic Biology Reveals Unprecedented Structural Novelty in Type V Glycopeptide Antibiotics,” vol. 46, p. 18, doi: 10.1021/acscentsci.1c01389.
[39] “Regulation of the P450 Oxygenation Cascade Involved in Glycopeptide Antibiotic Biosynthesis | Journal of the American Chemical Society.” https://pubs.acs.org/doi/10.1021/jacs.6b00307 (accessed Feb. 02, 2023).
[40] S. Donadio, M. Sosio, E. Stegmann, T. Weber, and W. Wohlleben, “Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis,” *Molecular Genetics and Genomics*, vol. 274, no. 1, pp. 40–50, Aug. 2005, doi: 10.1007/s00438-005-1156-3.
[41] K. C. Nicolaou, C. N. C. Boddy, S. Bräse, and N. Winssinger, “Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics,” *Angewandte Chemie International Edition*, vol. 38, no. 15, pp. 2096–2152, 1999, doi: 10.1002/(SICI)1521-3773(19990802)38:15<2096::AID-ANIE2096>3.0.CO;2-F.
[42] C. Brieke, G. Yim, M. Peschke, G. D. Wright, and M. J. Cryle, “Catalytic promiscuity of glycopeptide N-methyltransferases enables bio-orthogonal labelling of biosynthetic intermediates †,” *Chem. Commun*, vol. 52, p. 13679, 2016, doi: 10.1039/c6cc06975d.
[43] L. Kalan, J. Perry, K. Koteva, M. Thaker, and G. Wright, “Glycopeptide sulfation evades resistance,” *Journal of Bacteriology*, vol. 195, no. 1, pp. 167–171, 2013, doi: 10.1128/JB.01617-12.
[44] R. Menges, G. Muth, W. Wohlleben, and E. Stegmann, “The ABC transporter Tba of Amycolatopsis balhimycina is required for efficient export of the glycopeptide antibiotic balhimycin,” *Appl Microbiol Biotechnol*, vol. 77, no. 1, pp. 125–134, Nov. 2007, doi: 10.1007/s00253-007-1139-x.
[45] A. Müller, A. Klöckner, and T. Schneider, “Targeting a cell wall biosynthesis hot spot,” *Natural Product Reports*, vol. 34, no. 7, pp. 909–932, Jul. 2017, doi: 10.1039/c7np00012j.
[46] E. J. Culp *et al.*, “Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling.,” *Nature*, no. May, 2020, doi: 10.1038/s41586-020-1990-9.
[47] M. A. T. Blaskovich, K. A. Hansford, M. S. Butler, Z. Jia, A. E. Mark, and M. A. Cooper, “Developments in Glycopeptide Antibiotics,” *ACS Infectious Diseases*, vol. 4, no. 5, pp. 715–735, May 2018, doi: 10.1021/acsinfecdis.7b00258.
[48] R. M. Shawky *et al.*, “The border sequence of the balhimycin biosynthesis gene cluster from Amycolatopsis balhimycina contains bbr, encoding a StrR-like pathway-specific regulator,” *Journal of Molecular Microbiology and Biotechnology*, vol. 13, no. 1–3, pp. 76–88, 2007, doi: 10.1159/000103599.
[49] K. H. Almabruk, L. K. Dinh, and B. Philmus, “Self-Resistance of Natural Product Producers: Past, Present, and Future Focusing on Self-Resistant Protein Variants,” *ACS Chem Biol*, vol. 13, no. 6, pp. 1426–1437, Jun. 2018, doi: 10.1021/acschembio.8b00173.
[50] R. D. Firn and C. G. Jones, “The evolution of secondary metabolism – a unifying model,” *Molecular Microbiology*, vol. 37, no. 5, pp. 989–994, 2000, doi: 10.1046/j.1365-2958.2000.02098.x.
[51] G. Gallo *et al.*, “Differential proteomic analysis reveals novel links between primary metabolism and antibiotic production in Amycolatopsis balhimycina,” *Proteomics*, vol. 10, no. 7, pp. 1336–1358, 2010, doi: 10.1002/pmic.200900175.
[52] N. Waglechner, A. G. McArthur, and G. D. Wright, “Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance,” *Nature Microbiology*, vol. 4, no. 11, pp. 1862–1871, 2019, doi: 10.1038/s41564-019-0531-5.
[53] L. Horbal *et al.*, “The pathway-specific regulatory genes, tei15* and tei16*, are the master switches of teicoplanin production in Actinoplanes teichomyceticus,” *Applied Microbiology and Biotechnology*, vol. 98, no. 22, pp. 9295–9309, 2014, doi: 10.1007/s00253-014-5969-z.
[54] R. Alduina *et al.*, “A Two-Component regulatory system with opposite effects on glycopeptide antibiotic biosynthesis and resistance,” *Scientific Reports*, vol. 10, no. 1, p. 6200, 2020, doi: 10.1038/s41598-020-63257-4.
[55] M. Adamek, M. Spohn, E. Stegmann, and N. Ziemert, “Mining bacterial genomes for secondary metabolite gene clusters,” in *Methods in Molecular Biology*, vol. 1520, Humana Press Inc., 2017, pp. 23–47. doi: 10.1007/978-1-4939-6634-9_2.

[56] N. Ziemert, M. Alanjary, and T. Weber, “The evolution of genome mining in microbes-a review,” *Natural Product Reports*, vol. 33, no. 8, pp. 988–1005, Aug. 2016, doi: 10.1039/c6np00025h.
[57] K. Blin *et al.*, “antiSMASH 6.0: improving cluster detection and comparison capabilities,” *Nucleic Acids Res*, vol. 49, no. W1, pp. W29–W35, Jul. 2021, doi: 10.1093/nar/gkab335.
[58] S. Pelzer *et al.*, “Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in Amycolatopsis mediterranei DSM5908,” *Antimicrob Agents Chemother*, vol. 43, no. 7, pp. 1565–1573, Jul. 1999, doi: 10.1128/AAC.43.7.1565.
[59] J. M. Wink *et al.*, “Three new antibiotic producing species of the genus Amycolatopsis, Amycolatopsis balhimycina sp. nov., A. tolypomycina sp. nov., A. vancoresmycina sp. nov., and description of Amycolatopsis keratiniphila subsp. keratiniphila subsp. nov. and A. keratiniphila subsp. nogabecina subsp. nov,” *Syst Appl Microbiol*, vol. 26, no. 1, pp. 38–46, Mar. 2003, doi: 10.1078/072320203322337290.
[60] L. Xu *et al.*, “Complete genome sequence and comparative genomic analyses of the vancomycin-producing Amycolatopsis orientalis,” *BMC Genomics*, vol. 15, no. 1, May 2014, doi: 10.1186/1471-2164-15-363.
[61] A. W. Truman, M. J. Kwun, J. Cheng, S. H. Yang, J.-W. Suh, and H.-J. Hong, “Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel amycolatopsis strain producing ristocetin,” *Antimicrob Agents Chemother*, vol. 58, no. 10, pp. 5687–5695, Oct. 2014, doi: 10.1128/AAC.03349-14.
[62] D. P. Labeda, “Amycolatopsis coloradensis sp. nov., the avoparcin (LL-AV290)-producing strain,” *International Journal of Systematic Bacteriology*, vol. 45, no. 1, pp. 124–127, Jan. 1995, doi: 10.1099/00207713-45-1-124/CITE/REFWORKS.
[63] J. Wink *et al.*, “Amycolatopsis decaplanina sp. nov., a novel member of the genus with unusual morphology,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, no. 1, pp. 235–239, doi: 10.1099/ijs.0.02586-0.
[64] F. Xu *et al.*, “A genetics-free method for high-throughput discovery of cryptic microbial metabolites,” *Nat Chem Biol*, vol. 15, no. 2, pp. 161–168, Feb. 2019, doi: 10.1038/s41589-018-0193-2.
[65] G. Yim *et al.*, “Harnessing the synthetic capabilities of glycopeptide antibiotic tailoring enzymes: characterization of the UK-68,597 biosynthetic cluster,” *Chembiochem*, vol. 15, no. 17, pp. 2613–2623, Nov. 2014, doi: 10.1002/cbic.201402179.
[66] M. Sosio, H. Kloosterman, A. Bianchi, P. de Vreugd, L. Dijkhuizen, and S. Donadio, “Organization of the teicoplanin gene cluster in Actinoplanes teichomyceticus,” *Microbiology (Reading)*, vol. 150, no. Pt 1, pp. 95–102, Jan. 2004, doi: 10.1099/mic.0.26507-0.
[67] O. Yushchuk *et al.*, “Genomic-Led Discovery of a Novel Glycopeptide Antibiotic by Nonomuraea coxensis DSM 45129,” *ACS Chemical Biology*, vol. 16, no. 5, pp. 915–928, May 2021, doi: 10.1021/acschembio.1c00170.
[68] M. Sosio, S. Stinchi, F. Beltrametti, A. Lazzarini, and S. Donadio, “The gene cluster for the biosynthesis of the glycopeptide antibiotic A40926 by nonomuraea species,” *Chem Biol*, vol. 10, no. 6, pp. 541–549, Jun. 2003, doi: 10.1016/s1074-5521(03)00120-0.
[69] M. R. Bardone, M. Paternoster, and C. Coronelli, “Teichomycins, new antibiotics from Actinoplanes teichomyceticus nov. sp. II. Extraction and chemical characterization,” *J Antibiot (Tokyo)*, vol. 31, no. 3, pp. 170–177, Mar. 1978, doi: 10.7164/antibiotics.31.170.
[70] H. T. Chiu *et al.*, “Molecular cloning and sequence analysis of the complestatin biosynthetic gene cluster,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 15, pp. 8548–8553, Jul. 2001, doi: 10.1073/PNAS.151246498.
[71] M. N. Thaker *et al.*, “Identifying producers of antibacterial compounds by screening for antibiotic resistance,” *Nature Biotechnology*, vol. 31, no. 10, pp. 922–927, Sep. 2013, doi: 10.1038/nbt.2685.
[72] J. Pootoolal *et al.*, “Assembling the glycopeptide antibiotic scaffold: The biosynthesis of A47934 from Streptomyces toyocaensis NRRL15009.” [Online]. Available: www.pnas.orgcgidoi10.1073pnas.102285099
[73] M. J. Kwun and H.-J. Hong, “Genome Sequence of Streptomyces toyocaensis NRRL 15009, Producer of the Glycopeptide Antibiotic A47934,” *Genome Announc*, vol. 2, no. 4, pp. e00749-14, Jul. 2014, doi: 10.1128/genomeA.00749-14.
[74] J. J. Banik, J. W. Craig, P. Y. Calle, and S. F. Brady, “Tailoring Enzyme-Rich Environmental DNA Clones: A Source of Enzymes for Generating Libraries of Unnatural Natural Products,” *J. Am. Chem. Soc.*, vol. 132, no. 44, pp. 15661–15670, Nov. 2010, doi: 10.1021/ja105825a.
[75] J. G. Owen *et al.*, “Mapping gene clusters within arrayed metagenomic libraries to expand the

structural diversity of biomedically relevant natural products," *Proc Natl Acad Sci U S A*, vol. 110, no. 29, pp. 11797–11802, Jul. 2013, doi: 10.1073/pnas.1222159110.
[76] B. Nazari, C. C. Forneris, M. I. Gibson, K. Moon, K. R. Schramma, and M. R. Seyedsayamdost, "Nonomuraea sp. ATCC 55076 harbours the largest actinomycete chromosome to date and the kistamicin biosynthetic gene cluster," *Medchemcomm*, vol. 8, no. 4, pp. 780–788, Apr. 2017, doi: 10.1039/c6md00637j.
[77] M. J. Kwun, J. Cheng, S. H. Yang, D.-R. Lee, J.-W. Suh, and H.-J. Hong, "Draft Genome Sequence of Ristocetin-Producing Strain Amycolatopsis sp. Strain MJM2582 Isolated in South Korea," *Genome Announc*, vol. 2, no. 5, pp. e01091-14, Oct. 2014, doi: 10.1128/genomeA.01091-14.
[78] J. J. Banik and S. F. Brady, "Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary," *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, pp. 17273–17277, Nov. 2008, doi: 10.1073/pnas.0807564105.
[79] T. Stachelhaus, H. D. Mootz, and M. A. Marahiel, "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases," *Chem Biol*, vol. 6, no. 8, pp. 493–505, Aug. 1999, doi: 10.1016/S1074-5521(99)80082-9.
[80] A. K. Chavali and S. Y. Rhee, "Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites," *Brief. Bioinform.*, vol. 19, no. 5, pp. 1022–1034, 2018, doi: 10.1093/bib/bbx020.
[81] S. S. Mantri *et al.*, "Metagenomic Sequencing of Multiple Soil Horizons and Sites in Close Vicinity Revealed Novel Secondary Metabolite Diversity," *mSystems*, vol. 6, no. 5, p. e0101821, 2021, doi: 10.1128/mSystems.01018-21.
[82] D. M. Emms and S. Kelly, "OrthoFinder: Phylogenetic orthology inference for comparative genomics," *Genome Biology*, vol. 20, no. 1, p. 238, Nov. 2019, doi: 10.1186/s13059-019-1832-y.
[83] B. T. L. Nichio, J. N. Marchaukoski, and R. T. Raittz, "New tools in orthology analysis: A brief review of promising perspectives," *Frontiers in Genetics*, vol. 8, no. OCT, Oct. 2017, doi: 10.3389/fgene.2017.00165.
[84] M. Baunach, S. Chowdhury, P. Stallforth, and E. Dittmann, "The Landscape of Recombination Events That Create Nonribosomal Peptide Diversity," *Molecular Biology and Evolution*, Jan. 2021, doi: 10.1093/molbev/msab015.
[85] M. G. Chevrette, A. Gavrilidou, S. Mantri, N. Selem-Mojica, N. Ziemert, and F. Barona-G ´ Omez, "The confluence of big data and evolutionary genome mining for the discovery of natural products," 2021, doi: 10.1039/d1np00013f.
[86] M. H. Medema, P. Cimermancic, A. Sali, E. Takano, and M. A. Fischbach, "A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis," *PLoS Computational Biology*, vol. 10, no. 12, Dec. 2014, doi: 10.1371/journal.pcbi.1004016.
[87] H. Jenke-Kodama and E. Dittmann, "Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges," *Natural Product Reports*, vol. 26, no. 7, pp. 874–883, Jun. 2009, doi: 10.1039/b810283j.
[88] N. Ziemert, S. Podell, K. Penn, J. H. Badger, E. Allen, and P. R. Jensen, "The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity," *PLoS ONE*, vol. 7, no. 3, Mar. 2012, doi: 10.1371/journal.pone.0034064.
[89] M. Adamek *et al.*, "Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species," *BMC Genomics*, vol. 19, no. 1, Jun. 2018, doi: 10.1186/s12864-018-4809-4.
[90] R. A. Barco, G. M. Garrity, J. J. Scott, J. P. Amend, K. H. Nealson, and D. Emerson, "A genus definition for bacteria and archaea based on a standard genome relatedness index," *mBio*, vol. 11, no. 1, pp. 1–20, 2020, doi: 10.1128/MBIO.02475-19.
[91] M. N. Thaker and G. D. Wright, "Opportunities for synthetic biology in antibiotics: Expanding glycopeptide chemical diversity," *ACS Synthetic Biology*, vol. 4, no. 3, pp. 195–206, Mar. 2015, doi: 10.1021/sb300092n.
[92] H. Seto, T. Fujioka, K. Furihata, I. Kaneko, and S. Takahashi, "Structure of complestatin, a very strong inhibitor of protease activity of complement in the human complement system," *Tetrahedron Letters*, vol. 30, no. 37, pp. 4987–4990, Jan. 1989, doi: 10.1016/S0040-4039(01)80562-1.
[93] N. Naruse *et al.*, "New antiviral antibiotics, kistamicins A and B. I. Taxonomy, production, isolation, physico-chemical properties and biological activities," *J Antibiot (Tokyo)*, vol. 46, no. 12, pp. 1804–1811, Dec. 1993, doi: 10.7164/antibiotics.46.1804.
[94] N. Naruse, M. Oka, M. Konishi, and T. Oki, "New antiviral antibiotics, kistamicins A and B. II.

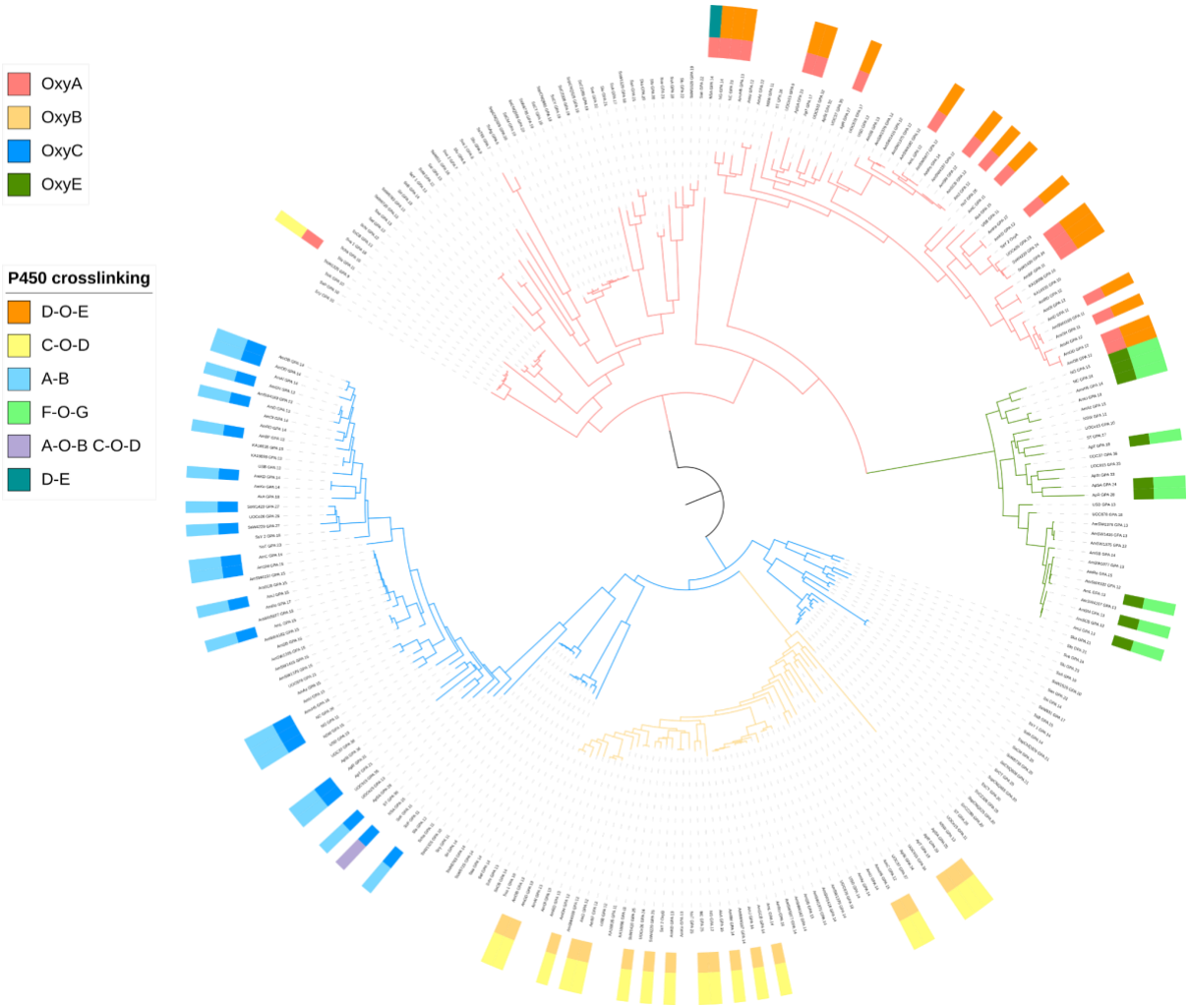
Structure determination," *J Antibiot (Tokyo)*, vol. 46, no. 12, pp. 1812–1818, Dec. 1993, doi: 10.7164/antibiotics.46.1812.
[95] T. Tatusova *et al.*, "NCBI prokaryotic genome annotation pipeline," *Nucleic Acids Res*, vol. 44, no. 14, pp. 6614–6624, Aug. 2016, doi: 10.1093/nar/gkw569.
[96] D. H. Haft *et al.*, "RefSeq: An update on prokaryotic genome annotation and curation," *Nucleic Acids Research*, vol. 46, no. D1, pp. D851–D860, 2018, doi: 10.1093/nar/gkx1068.
[97] I.-M. A. Chen *et al.*, "IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes," *Nucleic Acids Res*, vol. 47, no. Database issue, pp. D666–D677, Jan. 2019, doi: 10.1093/nar/gky901.
[98] A. M. Sharrar, A. Crits-Christoph, R. Méheust, S. Diamond, E. P. Starr, and J. F. Banfield, "Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type," *mBio*, vol. 11, no. 3, pp. e00416-20, May 2020, doi: 10.1128/mBio.00416-20.
[99] "Microbial communities across a hillslope-riparian transect shaped by proximity to the stream, groundwater table, and weathered bedrock - Lavy - 2019 - Ecology and Evolution - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.5254 (accessed Jan. 31, 2023).
[100] S. Nayfach *et al.*, "Author Correction: A genomic catalog of Earth's microbiomes," *Nat. Biotechnol.*, vol. 39, no. 4, p. 521, 2021, doi: 10.1038/s41587-021-00898-4.
[101] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.
[102] S. L. Heald, L. Mueller, and P. W. Jeffs, "Actinoidins A and A2: structure determination using 2D NMR methods," *J Antibiot (Tokyo)*, vol. 40, no. 5, pp. 630–645, May 1987, doi: 10.7164/antibiotics.40.630.
[103] L. Vértesy, H. W. Fehlhaber, H. Kogler, and M. Limbert, "New 4-oxovancosamine-containing glycopeptide antibiotics from Amycolatopsis sp. Y-86,21022," *J Antibiot (Tokyo)*, vol. 49, no. 1, pp. 115–118, Jan. 1996, doi: 10.7164/antibiotics.49.115.
[104] G. M. Sheldrick, P. G. Jones, O. Kennard, D. H. Williams, and G. A. Smith, "Structure of vancomycin and its complex with acetyl-D-alanyl-D-alanine," *Nature*, vol. 271, no. 5642, pp. 223–225, Jan. 1978, doi: 10.1038/271223a0.
[105] S. B. Christensen *et al.*, "Parvodicin, a novel glycopeptide from a new species, Actinomadura parvosata: discovery, taxonomy, activity and structure elucidation," *J Antibiot (Tokyo)*, vol. 40, no. 7, pp. 970–990, Jul. 1987, doi: 10.7164/antibiotics.40.970.
[106] F. Sztaricskai, C. M. Harris, A. Neszmelyi, and T. M. Harris, "Structural studies of ristocetin A (ristomycin A): carbohydrate-aglycone linkages," *J. Am. Chem. Soc.*, vol. 102, no. 23, pp. 7093–7099, Nov. 1980, doi: 10.1021/ja00543a035.
[107] M. J. Zmijewski, B. Briggs, R. Logan, and L. D. Boeck, "Biosynthetic studies on antibiotic A47934," *Antimicrob Agents Chemother*, vol. 31, no. 10, pp. 1497–1501, Oct. 1987, doi: 10.1128/AAC.31.10.1497.
[108] F. Parenti, "Structure and mechanism of action of teicoplanin," *Journal of Hospital Infection*, vol. 7, pp. 79–83, Mar. 1986, doi: 10.1016/0195-6701(86)90011-3.
[109] S. B. Singh *et al.*, "The complestatins as HIV-1 integrase inhibitors. Efficient isolation, structure elucidation, and inhibitory activities of isocomplestatin, chloropeptin I, new complestatins, A and B, and acid-hydrolysis products of chloropeptin I," *J Nat Prod*, vol. 64, no. 7, pp. 874–882, Jul. 2001, doi: 10.1021/np000632z.
[110] W. J. McGahren *et al.*, "Structure of avoparcin components," *J. Am. Chem. Soc.*, vol. 102, no. 5, pp. 1671–1684, Feb. 1980, doi: 10.1021/ja00525a036.
[111] M. L. Sanchez, R. P. Wenzel, and R. N. Jones, "In vitro activity of decaplanin (M86-1410), a new glycopeptide antibiotic," *Antimicrob Agents Chemother*, vol. 36, no. 4, pp. 873–875, Apr. 1992, doi: 10.1128/AAC.36.4.873.
[112] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics*, vol. 7, no. 1, p. 20, May 2015, doi: 10.1186/s13321-015-0069-3.
[113] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.
[114] M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch, and O. Kohlbacher, "NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity," *Nucleic Acids Res*, vol. 39, no. Web Server issue, pp. W362-367, Jul. 2011, doi: 10.1093/nar/gkr323.
[115] K. Blin *et al.*, "AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline," *Nucleic Acids Research*, vol. 47, no. W1, pp. W81–W87, 2019, doi: 10.1093/nar/gkz310.
[116] D. Megrian, N. Taib, A. L. Jaffe, J. F. Banfield, and S. Gribaldo, "Ancient origin and constrained

evolution of the division and cell wall gene cluster in Bacteria," *Nat Microbiol*, pp. 1–14, Nov. 2022, doi: 10.1038/s41564-022-01257-y.
[117] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Mol Biol Evol*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.
[118] B. Q. Minh *et al.*, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.
[119] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin, "ModelFinder: fast model selection for accurate phylogenetic estimates," *Nat Methods*, vol. 14, no. 6, Art. no. 6, Jun. 2017, doi: 10.1038/nmeth.4285.
[120] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W293–W296, 2021, doi: 10.1093/nar/gkab301.
[121] D. H. Huson, T. Dezulian, T. Klopper, and M. A. Steel, "Phylogenetic super-networks from partial trees," *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 1, no. 4, pp. 151–158, Oct. 2004, doi: 10.1109/TCBB.2004.44.
[122] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data.," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, Jan. 1998, doi: 10.1093/bioinformatics/14.1.68.
[123] D. H. Huson and D. Bryant, "Application of Phylogenetic Networks in Evolutionary Studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, Feb. 2006, doi: 10.1093/molbev/msj030.
[124] D. H. Parks *et al.*, "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life," *Nat Microbiol*, vol. 2, no. 11, pp. 1533–1542, 2017, doi: 10.1038/s41564-017-0012-7.
[125] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009, doi: 10.1093/bioinformatics/btp348.
[126] P. J. A. Cock *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.
[127] "FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix | Molecular Biology and Evolution | Oxford Academic." https://academic.oup.com/mbe/article/26/7/1641/1128976 (accessed Nov. 17, 2022).

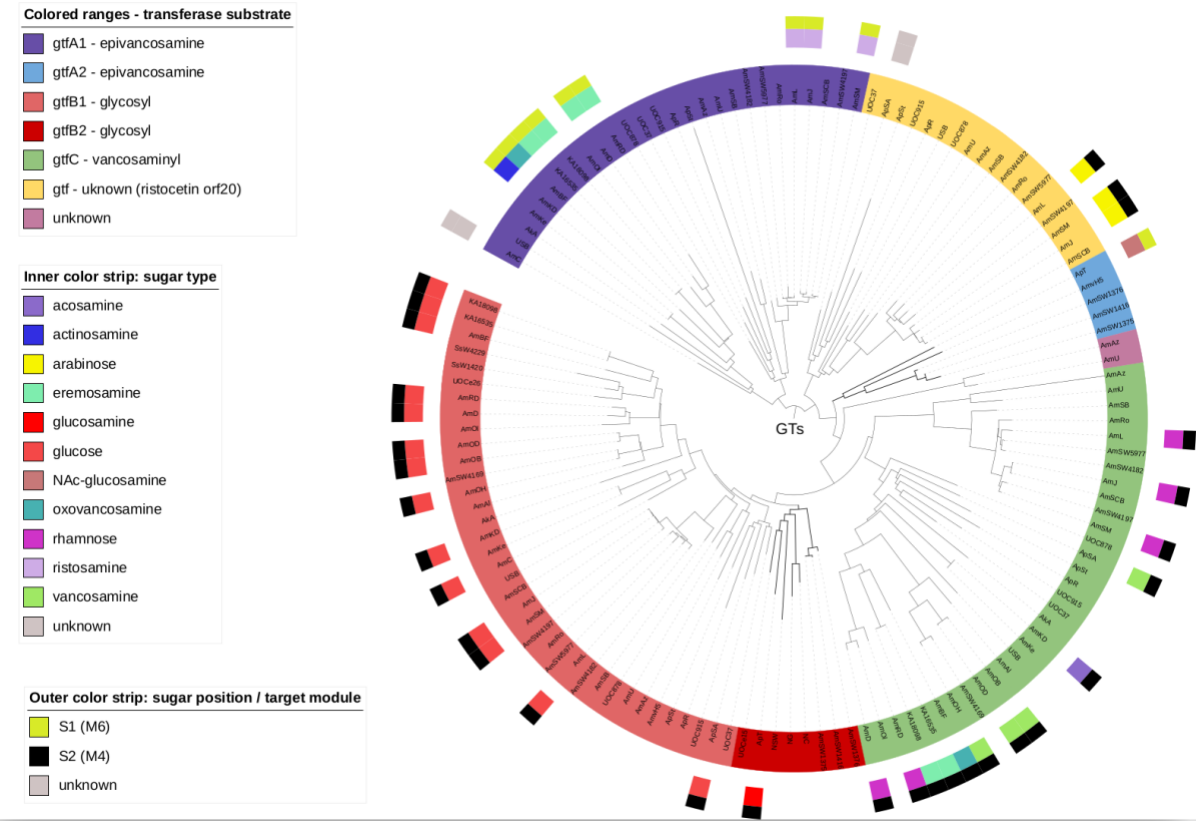
# Supplementary Data

Supplementary Table 1: SMILES of selected known GPA structures.
Supplementary Table 2: dataset of GPA BGCs and related metadata.
Supplementary Table 3: results of Stachelhaus code analysis of GPA BGC A-domains.
Supplementary Table 4: results of congruence analysis.

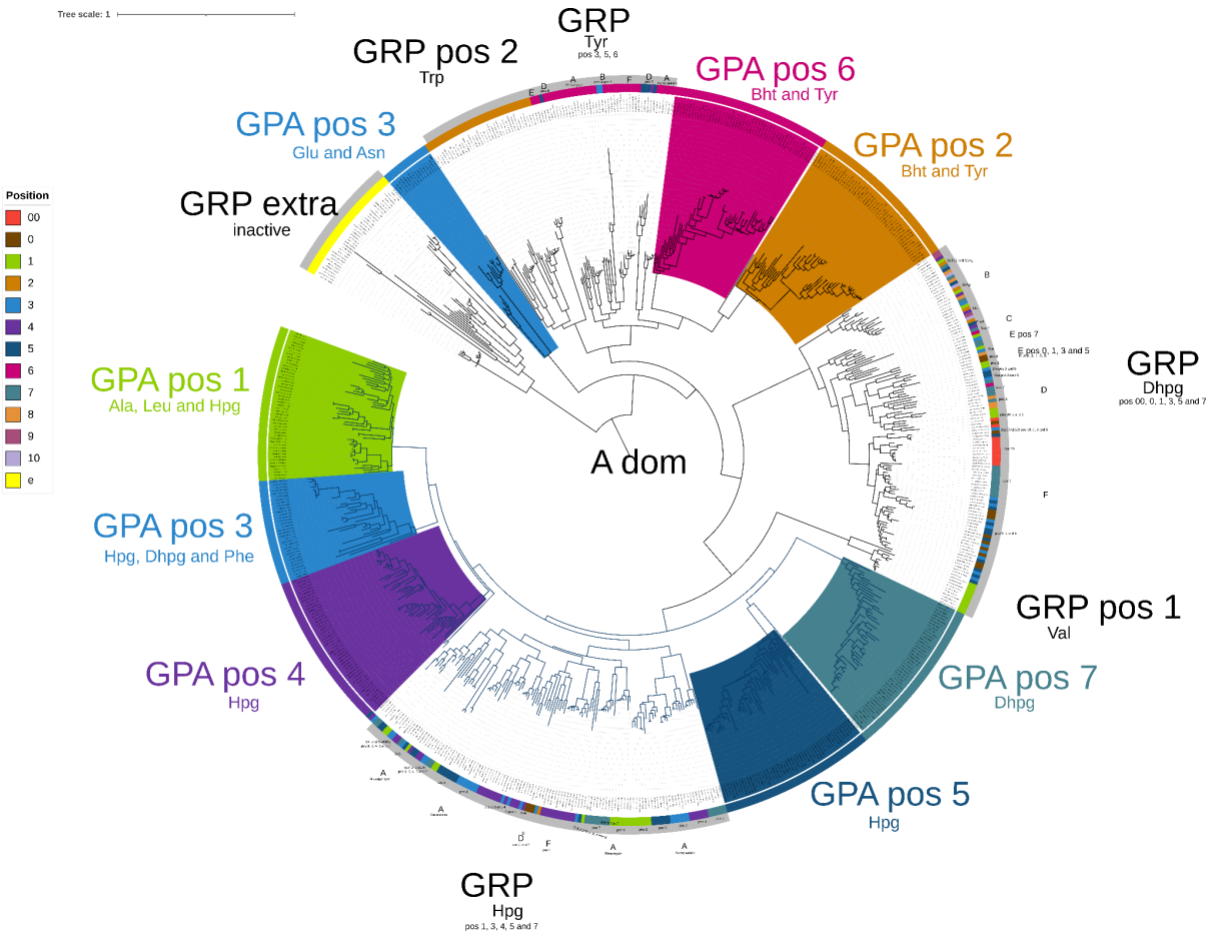
# Supplementary Figures




**SFigure 1:** Phylogenetic tree of orthogroup 1 (OG0000001) - p450 monooxygenases. The tree includes 4 types (oxyA, oxyB, oxyC, oxyE), whose clades can be separated after annotation of the enzymes with known function.




**SFigure 2:** Phylogenetic tree of orthogroup 2 (OG0000002) - glycosyltransferases. The tree includes multiple types (gtfA, gtfB, gtfC, gtf with unknown substrate), whose clades can be separated after annotation of the enzymes with known transferase substrate.




**SFigure 3:** Phylogenetic tree of all GPA A domains annotated with their predicted amino acid specificity.