Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Adrian Thorsplass

Automatic Reconstruction of Metabolic Pathways for Ribosomally Synthesized and Post-translationally Modified Peptides

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Snorre Sulheim
May 2023

NTNU
Norwegian University of
Science and Technology

Adrian Thorsplass

Automatic Reconstruction of Metabolic Pathways for Ribosomally Synthesized and Post-translationally Modified Peptides

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Snorre Sulheim
May 2023

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Acknowledgement

I begin this thesis by first thanking my two supervisors, Eivind Almaas and Snorre Sulheim, for their invaluable guidance and vital feedback throughout this project. I would then thank my family: my mother and father for their support and for accommodating my curiosity since childhood, my brother for his interest in my work and for inspiring me, and my aunt for motivating me as she nears the end of her master's thesis. I would also like to thank my friends, both in- and outside of Norway, several of whom took the time to read through parts of my thesis, and all whom mean a great deal to me.

Abstract

Biosynthetic gene clusters (BGCs) facilitate the production of secondary metabolites in organisms. These compounds are known to have useful properties such as antibiotic, antiviral or anti-tumor activity. However, only a tiny fraction of identified BGCs have had their products and metabolic pathways experimentally verified, with the vast majority of them being identified *in silico* by genome mining tools. The manual experimental analysis of these uncharacterized BGCs is limited due to it being much slower than the discovery rate of BGCs and due to low production yields of their associated products in laboratory conditions. To effectively study BGCs and guide the experimental analysis of them, there is a need for a tool which can accurately model the metabolic production of their associated secondary metabolites.

This thesis presents a method for automatically reconstructing the metabolic pathways of biosynthetic gene clusters (BGCs) from annotated gene data. This method was developed as a piece of software (ARMRiPP) that uses data from the BGC genome mining tool, antiSMASH, and outputs metabolic pathway data that can easily be implemented into genome-scale metabolic models (GEMs). As a way of guiding the pathway reconstruction, ARMRiPP was developed to generate structure predictions of the BGC associated compounds. The focus was on the major BGC family, ribosomally synthesized and post-translationally modified peptides (RiPPs), and three major classes were implemented: lanthipeptides, thiopeptides and lasso peptides.

Both ARMRiPP's ability to predict correct structures and correct pathways was evaluated. Using the Tanimoto score as a metric of structural similarity, we arrived at an average score of 0.9 across 57 BGCs of different RiPP classes, indicating a high accuracy. Four of the reconstructed pathways were tested for their ability to correctly predict production yield of the associated RiPP compounds when compared to their pathways in literature, and here we arrived at an average error of 7%. Limitations of these results related in large part to factors such as missing implementations for various modification reactions, limited mechanistic knowledge of certain RiPP pathways, limitations in antiSMASH annotations and potential sources of data skewness. Structural prediction accuracy was also considered in light of the performance of a similar software, PRISM.

As a case study, the reconstructed pathways were used to estimate the metabolic burden of RiPP compound production in different organisms, using multiple GEMs. Slight differences in metabolic burden were observed between the RiPP classes, and larger differences between different phylogenetic groups. It was also observed that precursor length had a large impact on the associated metabolic burden of RiPPs. A general trend was observed that metabolic burden would be lower for RiPP pathways when put into GEMs of their native host. Investigating this further revealed statistically significant differences in the metabolic burden of pathways in their

native hosts and in heterologous hosts, and this difference was more pronounced for pathways put into hosts of different phylogeny than their native. The observations from the metabolic burden results were discussed from the perspective of microbial ecology, and how it potentially relates to BGC specific evolutionary effects.

Sammendrag

Biosyntetiske genkluster (BGC-er) fasiliterer produksjonen av sekundære metabolitter i organismer. Disse forbindelsene er kjent for å ha nyttige egenskaper som antibiotisk, antiviral eller anti-tumor aktivitet. Imidlertid har bare en liten brøkdel av identifiserte BGC-er hatt sine produkter og metabolske stier eksperimentelt verifisert, der flertallet av dem kun er identifisert gjennom *in silico* informasjonsutvinning av genomdata. Den manuelle eksperimentelle analysen av disse ikke-karakteriserte BGC-ene er begrenset av treghet og av lave produksjonsutbytter for deres tilknyttede produkter i laboratorieforhold. For å effektivt studere BGC-er og veilede den eksperimentelle analysen av dem så trengs det et verktøy som kan modellere produksjonen av deres tilknyttede sekundære metabolitter med høy presisjon.

Dette arbeidet presenterer en metode for å automatisk rekonstruere metabolske stier av biosyntetiske genkluster (BGC-er) fra annoterte gen-data. Denne metoden ble utviklet som en programvare (ARMRiPP) som bruker data fra BGC-genomdatautvinningsprogrammet, antiSMASH, og returnerer metabolske stier i en dataform som enkelt kan implementeres i genomskala metabolske modeller (GEM-er). For å veilede sti-rekonstruksjonen så ble ARMRiPP utviklet med evne til å generere prediksjoner av den kjemiske strukturen til BGC-assosierte forbindelser. Fokuset var på en av de større BGC-familiene, ribosomalt syntetiserte og post-translasjonelt modifiserte peptider (RiPP), og tre hovedklasser ble implementert: lanthipeptider, thiopeptider og lasso-peptider.

ARMRiPP sin evne til å forutsi riktige strukturer og riktige stier ble begge evaluert. Tanimoto-score ble brukt som et mål for strukturell likhet, og vi kom frem til en gjennomsnittlig score på 0,9 for prediksjon av riktige strukturer ved et utvalg av 57 strukturer fra forskjellige RiPP-klasser, som indikerer høy nøyaktighet. Fire av de rekonstruerte stiene ble testet for deres evne til å forutsi korrekt produksjonsutbytte av de tilknyttede RiPP-forbindelsene sammenlignet med de kjente stiene fra litteraturen, og her kom vi frem til en gjennomsnittlig feilmargin på 7%. Begrensningene ved disse resultatene skyldtes i stor grad faktorer som manglende implementering av ulike RiPP biosyntetiske reaksjoner, begrenset mekanistisk kunnskap om biosyntesen til visse RiPP, begrensninger i antiSMASH-annoteringer og potensiell dataforskyvning. Nøyaktigheten av den strukturelle prediksjonen til programvaren ble vurdert i lys av ytelsen til en lignende programvare, PRISM.

Som en casestudie så ble de rekonstruerte stiene brukt til å estimere den metabolske byrden for produksjonen av RiPP-forbindelser i forskjellige organismer, ved bruk av flere GEM-er. Små forskjeller i metabolsk byrde ble observert mellom RiPP-klassene, og større forskjeller mellom forskjellige fylogenetiske grupper. Lengden på RiPP-forgjengerpeptidet ble observert å ha en stor effekt på den metabolske byrden til RiPP-stien. Det ble også observert en generell trend at

metabolsk byrde ville være lavere for RiPP-stier som ble satt inn i GEM-er av sin naturlige vert. Ved videre undersøkelse ble det avdekket statistisk signifikante forskjeller i metabolsk byrde hos RiPP-stier i deres naturlige verter og i heterologe verter, og denne forskjellen var mer signifikant for stier som ble satt inn i verter av forskjellig fylogeni enn deres naturlige. Observasjonene fra disse resultatene ble diskutert fra perspektivet av mikrobiell økologi, og for hvorvidt de har sammenheng med BGC-spesifikke evolusjonære effekter.

Table of Contents

Abbreviations

General	
BGC	Biosynthetic gene cluster
RiPP	Ribosomally synthesized and post-translationally modified peptide
RRE	RiPP recognition element
NP	Natural product
FBA	Flux balance analysis
GEM	Genome-scale metabolic model
MIBiG	Minimum Information about a Biosynthetic Gene cluster
smCOG	Secondary metabolite Cluster of Orthologous Groups
antiSMASH	Antibiotics and Secondary Metabolite Analysis Shell
pHMM	Profile Hidden Markov Model
Pfam	Protein family database
TIGRFAM	The Institute for Genomic Research protein family database
BiGG	Biochemical, Genetic and Genomic database

Metabolites and Reacting Motifs	
Dha	Dehydroalanine
Dhb	Dehydrobutyrine
QA	Quinaldic acid
MIA	Methylindolic acid
SAM	S-Adenosyl-methionine
SAH	S-Adenosyl-L-homocysteine
AviCys	Aminovinyl Cysteine

1 Introduction

Natural Products (NPs) are compounds produced by organisms for a variety of purposes in nature, and many of them have been found to have uses in human applications (1): More than 60% of small-molecule FDA approved drugs are either NPs, NP derivatives or synthetic compounds that mimic NPs (2). However, only a tiny fraction of the natural product diversity has been experimentally characterized and exploited for medical or other purposes. The majority of NPs and their corresponding Biosynthetic Gene Clusters (BGCs) have only been identified *in silico* by computational mining of genomes and metagenomes (3). NPs are often referred to as secondary metabolites as they are not a part of the primary metabolism that is required for the growth, reproduction and proliferation of organisms. Although the purpose of many of these secondary metabolites in their natural environment is yet to be fully understood, they are often involved in more sophisticated tasks like communication and competition with other species or strains (4). A secondary metabolite can e.g. function as an antibiotic compound that facilitates the survival of the organism by the removal of competitors in its environment (5, 6). Other examples of known functions of secondary metabolites are anti-tumor or cytotoxic activity and regulatory activity through the inhibition or induction of ribosomal and transcriptional components (4, 7). Interest in secondary metabolites is a result of these compounds' potential applications, especially considering mankind's continued fight against cancer and the growing concern of multi-drug resistant pathogens that cannot be treated with the currently available antibiotics (8).

Biosynthesis of secondary metabolites requires several specialized enzymes, and the corresponding genes are usually co-expressed and co-localized in the organism's genome. The region of an organism's genome containing such genes in close proximity is what is referred to as a biosynthetic gene cluster (BGC) (9). There have been found BGCs in organisms from every domain of life (10). Features of BGCs such as homologous and/or orthologous genes, recurring genetic motifs and sequence similarity to reference gene clusters drive the computational discovery of new BGCs. Similarly, these features have made categorizing BGCs more useful and consistent, as each category and subcategory tend to have the same recurring structure and specific orthologous genes, as well as a degree of modularity between them (11, 12). The major families of BGCs are generally divided into polyketides (PKs), nonribosomal peptides (NRPs), ribosomally synthesized and post-translationally modified peptides (RiPPs), and several smaller families (13).

Due to currently available BGC genome mining tools, the discovery of novel BGCs is no longer a large undertaking, at least not within the most frequent and well-described BGC families. With tools such as antiSMASH, BGCs can be predicted from genetic sequence data and functionally annotated with a high degree of accuracy. However, the challenges lie in exploiting these

findings in a useful way, such as prioritizing findings for experimental characterization and functional assays and producing them efficiently at sufficient yields. As the native BGC host species is usually poorly characterized it is often easier to achieve efficient production in a heterologous host species that is easily cultivable and where genetic tools can be applied. However, manually choosing a BGC host is not optimal for several reasons. Firstly, the fast growing number of uncharacterized BGCs from genome mining outpaces efforts to analyze them manually (11). Secondly, the existence of "silent" BGCs among these uncharacterized BGCs, whose products have severely low yields under standard laboratory conditions, can make manual analysis even slower or even infeasible (14). Furthermore, many BGC products can have low solubility (15), which combined with the previously mentioned low yields can make product isolation and structural analysis very difficult. A solution to these challenges lies in finding ways to effectively automatize analyses of BGC compound production in its source organism or in heterologous hosts, and provide information on expected production yield. An efficient way of achieving this is through the use of genome-scale metabolic models (GEMs) and flux balance analysis (FBA). In fact, these tools have been utilized to construct and analyze the metabolic pathways associated with both PKs and NRPs, but not yet for RiPPs (16, 17).

The focus of this master thesis will be on the major BGC family called ribosomally synthesized and post-translationally modified peptides, commonly referred to as RiPPs. The RiPP family of BGCs produces secondary metabolites through the post-translational modification and tailoring of a ribosomally produced polypeptide. The original peptide before any modifications is called precursor peptide or prepeptide, and has properties analogous to other ribosomally produced polypeptides. Its amino acids can be divided into a core region, where the modifications take place, and a leader region, which binds to specific regions of the modification enzymes and is cleaved off afterwards. The modified peptides of RiPPs can have much more varied properties than traditional peptides, for example being very hydrophobic, having special chemical motifs such as pyridines, various azole and lanthionine groups, as well as uncharacteristic conformations due to other forms of cross-linking than di-sulfide bridges. This structural diversity is what gives rise to their large diversity of applications (18).

To bridge the gap between annotated BGC data and metabolic modeling, we have in this thesis developed a tool which can automatically and accurately reconstruct the metabolic pathways of RiPP products based on genetic data of the clusters. The accuracy of the developed software (ARMRiPP) is evaluated both on its ability to predict the structure of the end product and to correctly predict the required precursors and cofactors. Furthermore, we show that these metabolic pathways can be used to extend GEMs to analyze the production of the associated RiPP products. These GEMs may be manually curated reference models of well known bacterial producers, or they can be GEMs of the RiPP source organisms automatically reconstructed using genomic data. This can aid future screening of heterologous RiPP expression hosts or enable model-based strain design for enhanced production, as previously demonstrated (16). However,

it also allows us to study the production RiPPs in their native host in the context of microbial ecology, asking questions about the cost and metabolic dependencies for their production. It is this latter question we've addressed in this thesis as a use-case of the developed software. Specifically, we ask if there are specific trends in the metabolic burden (cost) of the different RiPP classes, and whether species' metabolism has evolved to lower the metabolic burden associated with the production of its secondary metabolites. We find that the metabolic burden of different RiPP classes has high variance, and that the precursor length has a large effect on the metabolic burden of RiPPs. It was also found the metabolic burden of RiPP pathways in their native hosts is significantly lower than in heterologous hosts, with the exception of more closely related ones.

2 Theory

2.1 Secondary Metabolites and Biosynthetic Gene Clusters

Biosynthetic gene clusters (BGCs) can be described as a set of genes physically close to each other, which together facilitate the production of a secondary metabolite (19). They contain genes for enzymes which catalyze the biosynthetic steps required for the production of specific secondary metabolite compounds. As these compounds are produced outside of the primary metabolism of organisms, not as energy storage nor building blocks, they tend to have more sophisticated functions that nonetheless ensure the organism's survival and proliferation. Such functions can include antimicrobial activity to deter competing organisms, signaling activity and even protagonistic activity for hosts in symbiotic relationships (4–7, 20). Other genetic components of BGCs can give rise to transporters to move the secondary metabolite around and out of the cell, regulatory proteins and immunity proteins, the lattermost being necessary for bacteria producing antibiotic compounds to protect themselves from their own product (20).

There is an underlying genetic efficiency to BGCs as they appear in nature. Across phylogenetic domains there can be found groups of BGCs with similar biosynthetic strategies, which distinguish different BGC families such as ribosomally synthesized and post translationally modified peptides (RiPPs) from non-ribosomal peptides (NRPs) (19). Within BGC families, differences in cluster gene functions and cluster composition give rise to compounds of high chemical diversity. If BGCs were structured inefficiently, one would expect that the diversity of BGC compounds would solely be the result of specialized and unique biosynthetic genes. However, in reality BGC families can contain functionally identical biosynthetic genetic domains conserved across larger groups of BGCs (21). The presence of certain central and conserved biosynthetic domains within a group of BGCs give rise to further classifications beyond the BGC families (22). Despite the similarities in biosynthetic mechanisms for these BGC classifications, a large degree of chemical diversity in the secondary metabolite products can still be found as a result of differences in overall cluster composition and in predictable genetic elements. With this in mind, a large motivation behind identifying gene functions in BGCs is that one can uncover information behind the biosynthesis of multiple secondary metabolites of diverse structures and functions (23).

Genome Mining of BGCs

There are many available tools for BGC genome mining, but currently the most widely utilized is antiSMASH (12). It employs profile Hidden Markov Model (pHMM) methods to identify gene clusters, and functionally annotate genes in BGCs. In short, pHMMs are probabilistic models

which are trained to associate certain conserved patterns of sequences with a known characterization present in a training set (Figure 2.1.1). Using a pHMM on a new unknown set of sequence data is a way of sorting through it to find the regions with similar patterns to those which were present in the pHMM training set. The pHMM assigns the characterization associated with the pattern, and ranks it with a measure of statistical confidence (24). Put concisely, a pHMM trained on a set of signature genes for a certain kind of BGC can be used to predict the presence of those BGC genes in a new set of sequence data above a determined statistical threshold. In the case of antiSMASH, a set of pHMMs was created containing models for each implemented group of BGCs as well as false-positive models for non-BGC regions containing known homology to BGCs. The BGC signature gene regions in the sequence data are then predicted through a filtering process of these BGC and false-positive models, and the cluster is limited to an area where no signature gene is present for a certain stretch of kilobases depending on the BGC type. As this is a greedy method, the predicted gene clusters may contain an overlap of multiple clusters (11).

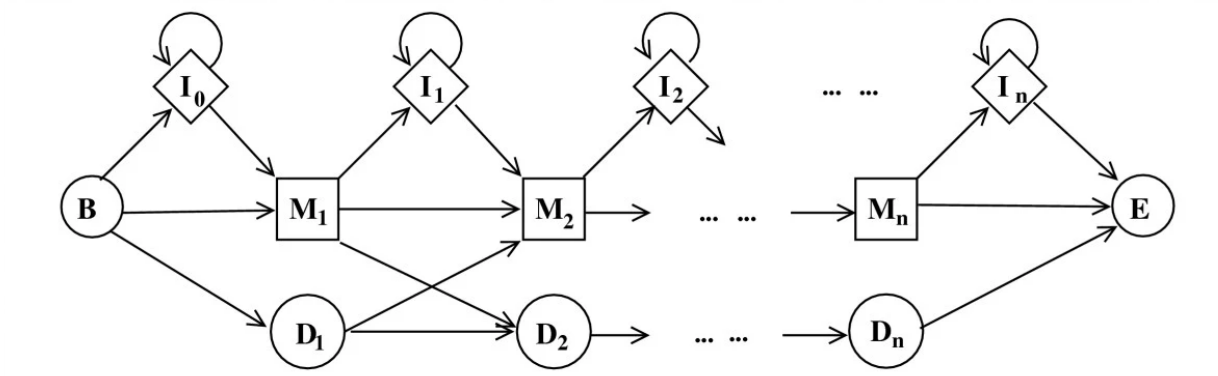


Figure 2.1.1: A schematic overview of a profile Hidden Markov Model (pHMM). Here the different shapes represent different states, and the arrows represent the transition from one state to another. The squared M states are the match states of the model, representing symbols which match to the pattern. The diamond I states are the insert states, and represent symbols which do not match with the pattern. The circular D states are the deletion states, and represent empty positions in the alignment of the patterns, while the circular B and E states simply represent the beginning and end state of the model. Each match and insertion state contains a probability distribution of possible symbols, which for the match states represents the sequence pattern. Additionally, the transition out of a state has a probability distributed among every outward pointing arrow. All probability parameters are estimated using a training set of sequences and a statistical method, which fit the model to the patterns in the training set (25). Figure adapted from Figure 1 in (26)

Beyond the cluster detection from predicted signature BGC genes, antiSMASH can also functionally annotate genes present in the cluster through a similar method of using pHMMs. A library of pHMMs trained to detect conserved functional regions is applied to the genes found in the predicted cluster, returning several annotations based on the pattern similarity and threshold for each model. For a set of pHMMs in this library antiSMASH leverages the orthologous nature of BGC genes to functionally annotate them. This set is referred to as secondary metabolite clusters of orthologous groups (smCOG), and was created by investigating signature biosynthetic genes of different BGCs, clustering them and manually assigning annotations based on common features shared within the orthologous clusters. The smCOG annotations are given in the form of "SMCOG*XXXX*", where the X-es are digits assigning it to one of the clustered orthologous groups, followed by a descriptor for its function (11). For example, SMCOG1155, which has the descriptor "Lantibiotic_dehydratase_domain_protein", is consistently annotated to both the lanB genes of class I lanthipeptide BGCs as well as its functionally identical counterpart present in thiopeptide BGCs. Outside of smCOGs, antiSMASH also has functional annotation pHMMs for clusters where the annotation is directed by certain rules. These rules can be if the gene matches one or more specific profiles, if the gene matches one profile and specifically not another one and so on. This rule-based annotation usually accounts for gene groups with high homology and significant functions in the biosynthesis of a specific BGC class (12, 27). An example of this rule-based clustering annotation is for the thiopeptide YcaO gene, which is simply annotated as "thiopeptide: YcaO". Additionally, functional annotation may be done by pHMMs in the library which are trained on specific Pfam or TIGRFAM groups, returning the Pfam or TIGRFAM protein ID of the group (27).

A challenge with BGCs found through genome mining is that a large number of them are expressed in very low amounts using their native hosts in standard laboratory conditions (14). Furthermore, the great majority of microorganisms will not grow efficiently or at all with current laboratory methods, and so the BGC products of these organisms cannot be immediately exploited (28). To be able to efficiently produce secondary metabolites from BGCs of interest, it may be desirable to express them in heterologous hosts with more rigorous production ability in laboratory conditions (14, 28).

2.2 Ribosomally Synthesized and Post-translationally Modified Peptides (RiPP)

Ribosomally Synthesized and Post Translationally Modified Peptides (RiPPs) are one of the major BGC families, accounting for about 11% of secondary metabolite records present in the MIBiG database as of writing. They have been discovered in every domain of life, and due to advancements in genome sequencing there is a growing number of them being identified (29). Several RiPP products have been found to have antibiotic, antiviral or anticancer properties (30).

As the name suggests, RiPP compounds are distinguished by having a ribosomal origin. This means that the initial substrate of the biosynthetic pathway is structurally limited by the 20 proteinogenic amino acids, the same as any other ribosomal polypeptide. Unlike conventional polypeptides however, the peptides of RiPP clusters go through more diverse and extensive post-translational modifications. This typically involves changing the chemical structure of their amino acid residues, non-canonical cross-linking between residues or certain motifs, as well as specific tailoring. These modifications can turn the initial peptide into highly cyclical compounds with chemical properties very different from conventional polypeptides (18).

General Structure of RiPP Biosynthesis

The structure of RiPP biosynthesis can be generalized into three main steps; peptide synthesis, modification and cleavage (Figure 2.2.1). RiPP BGCs conventionally contain a gene coding for a precursor peptide, which is translated canonically in a ribosome like any other polypeptide. Precursor peptides can be divided into a leader and core region. The leader region directs proteins and enzymes related to the biosynthesis, as well as transport and immunity proteins, while modification reactions take place in the core region (18).

There are motifs present in the leader region which bind with regions that can be found in certain proteins, called RiPP precursor peptide recognition elements (RRE) (31). Proteins which contain RREs tend to have genes present in the same cluster as its precursor peptide, and tend to be related to central steps of the RiPP biosynthetic pathway, such as enzymes that catalyze peptide modification. However, not all peptide modification enzymes are leader specific, meaning they contain RREs and require a certain leader peptide for activity, and some may be substrate specific, where they only have activity in core peptides with a specific peptide sequence (32).

After modifications of the core peptide are either finished or inevitable, the leader peptide is cleaved off from the core. This is usually done by a protease, but may also be done by specific enzymes as part of the core modifications (33). The fully modified core peptide is now a finished

product of the RiPP BGC. Following this, the product is usually transported either within or outside the cell (18).

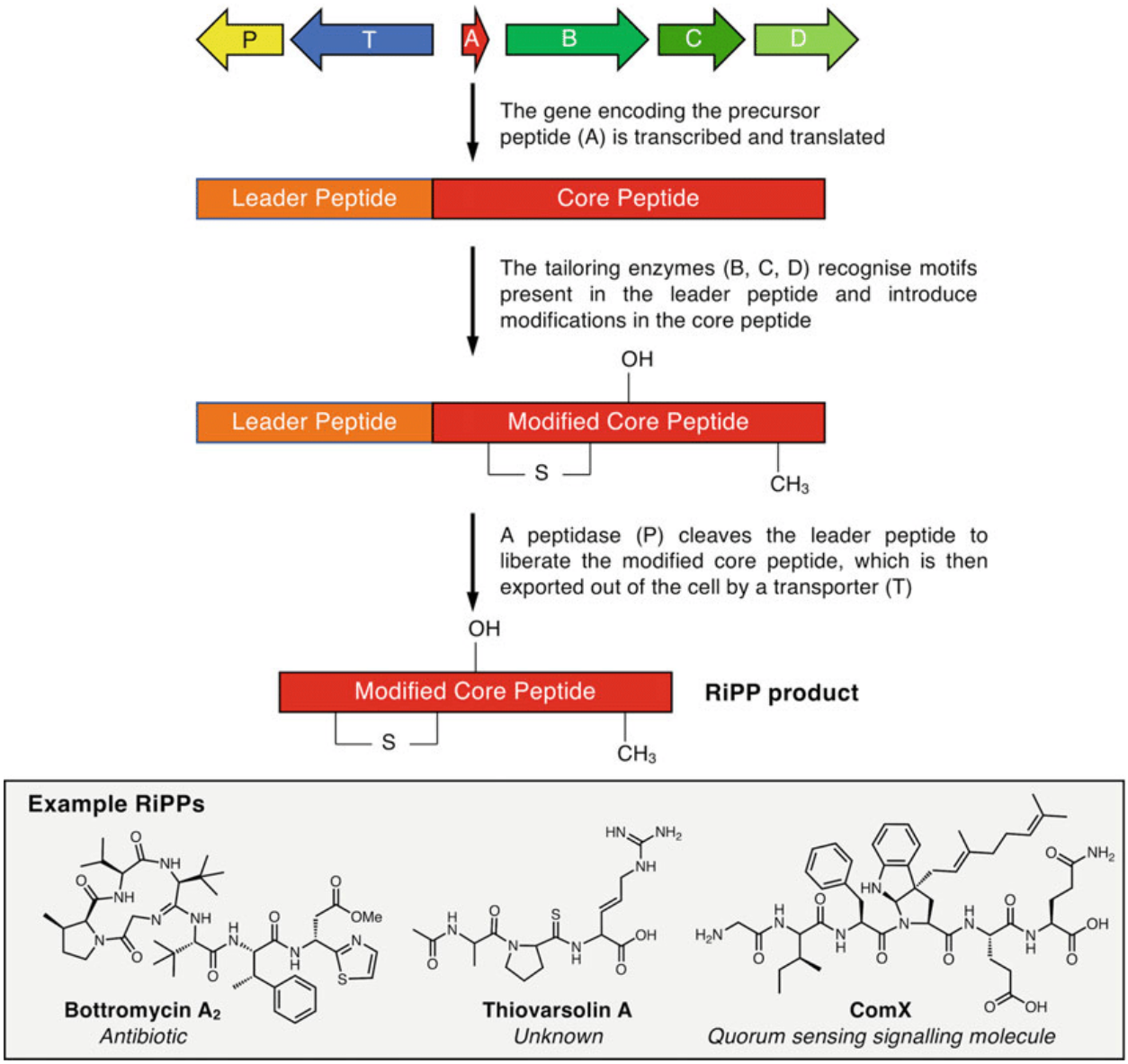


Figure 2.2.1: Schematic of a general RiPP biosynthetic pathway. In this RiPP gene cluster, genes A-D play a part in the maturation of the RiPP product, gene A encoding the precursor peptide and genes B-D performing the modification reactions needed for its maturation. Additional genes such as P and T may have activities relating to the mature RiPP product. Gene P can encode an immunity protein which protects the producing cell from the effects of the RiPP product, while gene T may be for a transporter enzyme that carries the RiPP product to its destination. Adapted from Figure 1 of (34).

RiPP Classification

Despite the general mechanism of biosynthesis described for RiPP BGCs, there is a huge variety in the structure of RiPP products. This owes to the different forms of modification reactions which can happen to the core peptide as well as differences in the core peptide sequence. A primary factor which distinguishes the different RiPP classes is the presence of specific chemical motifs in the end product of the RiPP BGC. Since these chemical motifs are installed by modification enzymes found in the RiPP BGC, it is reasonable to classify the RiPP BGC through the presence of the genes for such an enzyme or combination of enzymes. However, it is worth to note that due to the orthology between BGCs across evolutionary time, two RiPP BGCs may contain genes for modification enzymes with identical functions despite the two clusters being classified differently on a whole. This is also generally the case for the different subclasses of a RiPP class (18).

Although specific enzymatic modifications define the different RiPP classes, there can be additional modification enzymes which can cause structural diversity both inside the class and across subclasses. Usually these additional modification reactions within specific RiPP classes and subclasses are referred to as tailoring reactions. These tailoring reactions often involve N- and C-terminal modification to confer stability in the end product, but can also facilitate larger structural variations. Nonetheless, when it comes to the diversity of the overall chemical structure of the end product for the same RiPP class or subclass, the major differentiating factor is the core peptide sequence. There are currently over 20 classifications of RiPPs, many with several subclasses. Of these 20 RiPP classes, lanthipeptides, thiopeptides and lasso peptides are some of the more well described in literature (18, 32–35). Due to this, these three RiPP classes will be the focus of this thesis.

Genome Mining of RiPPs

In the specific case of RiPP BGC genome mining, antiSMASH uses many of the same methods as described for general BGC mining, and in recent years it has gotten improved RiPP gene annotation through expanded RiPP specific pHMMs and detection rules (12, 27). The most important element of RiPP biosynthesis could be considered the precursor peptide synthesis, and so correctly predicting the peptide sequence for the precursor is of high interest. Additionally, knowing the proteolytic cleavage site of the precursor peptide allows us to distinguish between core and leader domains, which is also of high interest since every biosynthetic modification reaction in RiPPs takes place in its core domain and its structure can reflect the mature RiPP product to a larger degree. Both precursor peptide prediction and cleavage site prediction is performed by antiSMASH each through pHMM libraries containing models specific to different

RiPP classes or groups. Additionally, rule based annotation with other cluster elements is used to annotate which RiPP class and subclass the precursor belongs to (11, 12, 27, 36–39).

Although antiSMASH can predict both the precursor and core peptide sequence of RiPPs, it cannot on its own assess the chemical structure of mature RiPP compounds. PRISM is a genome analysis tool which similarly to antiSMASH employs hidden markov models to predict BGCs and cluster gene functions. Unlike antiSMASH, PRISM is designed with the intent of predicting secondary metabolite chemical structures, including those of RiPP compounds, from BGC data. The methods it uses involve using a graph model for the substrate and connecting functional annotations of modification enzyme genes to reactions which act on subgraphs of the substrate model, such as a hydroxyl group or a specific amino acid. For a given set of annotated biosynthetic genes in the cluster, it will in effect produce every possible structure for the products. In principle, if the BGC annotation and reaction mechanisms are precise enough, one could expect that one of these produced structures would be identical with the actual compound structure (13).

2.3 Lanthipeptides

One of the largest and most well studied classes of RiPPs are the lanthipeptides. Named after the presence of lanthionine in their end products, a non-proteinogenic amino acid consisting of two amino acid groups connected by a thioether bridge, as can be seen in Figure 2.2.1 (32). Many lanthipeptides are well known for having antimicrobial activity, encompassing the historic compound class lantibiotics (32, 40). The mode of action of antibiotic lanthipeptides is generally disrupting the proton motive force over the plasma membrane of sensitive cells through the formation of pores, increasing membrane permeability and decreasing the rate of energy-requiring reactions in the cell (41). These antibiotic lanthipeptides achieve this through highly specific and selective interaction with membrane components, which relate to the specific structure of the thio-ether crosslinks in them. For example, the lanthipeptide Nisin targets the peptidoglycan Lipid II (32, 42–44), while another lanthipeptide, Cinnamycin, targets the membrane phospholipid phosphatidyl-ethanolamine (32, 45, 46). However, through the increased discovery of lanthipeptide RiPP clusters, there have been found to be lanthipeptides with even more diverse functions such as antifungal, morphogenetic, antiviral, antinociceptive and antiallodynic activity (32).

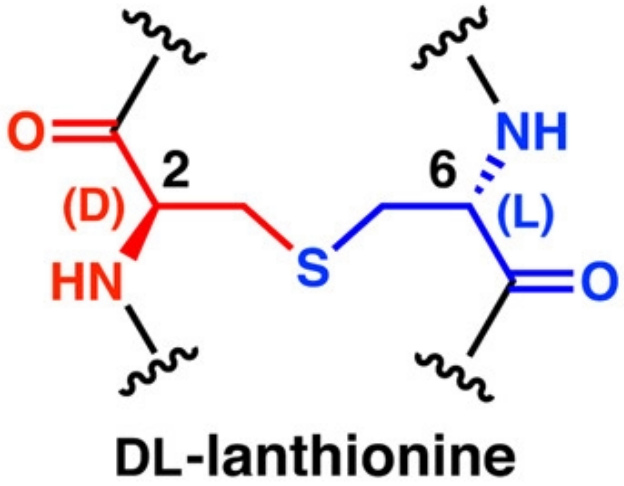


Figure 2.3.1: Chemical structure of the lanthionine group as it appears in lanthipeptides. The red and blue coloring indicates different amino acids in the lanthipeptide core. Adapted from Figure 1 in (32).

General Biosynthetic Mechanism of Lanthipeptides

Formation of lanthionine groups in lanthipeptides (figure 2.3.1), their defining feature, is mainly facilitated through two enzymatic modification reactions; dehydration of amino acid residues and thio-ether crosslinking (figure 2.3.2). First, a dehydratase enzyme dehydrates serine and threonine residues in the core into dehydroalanine (Dha) and dehydrobutyrine (Dhb) respectively. Then, a cysteine residue in the core reacts with one of the dehydrated amino acids in a 1,4-conjugate addition, producing an enolate. The enolate quickly reacts further with two separate outcomes. In one outcome the enolate is protonated, resulting in the formation of either lanthionine, if the reacting residue was Dha, or methyllanthionine, if it was Dhb. For the other outcome a second conjugate addition happens by leveraging the enolate to react with another Dha. Conversely as in the first outcome, the second enolate is protonated, causing the formation of either labionin or methyllabionin. The labionin chemical motif is currently present only in the class III lanthipeptides. Generally, the thio-ether cross-linking reaction proceeds until there are either no cysteine or no dehydrated amino acid residues left in the core peptide, transforming it into a highly polycyclical compound given by its amount of serine/threonine and cysteine residues (32).

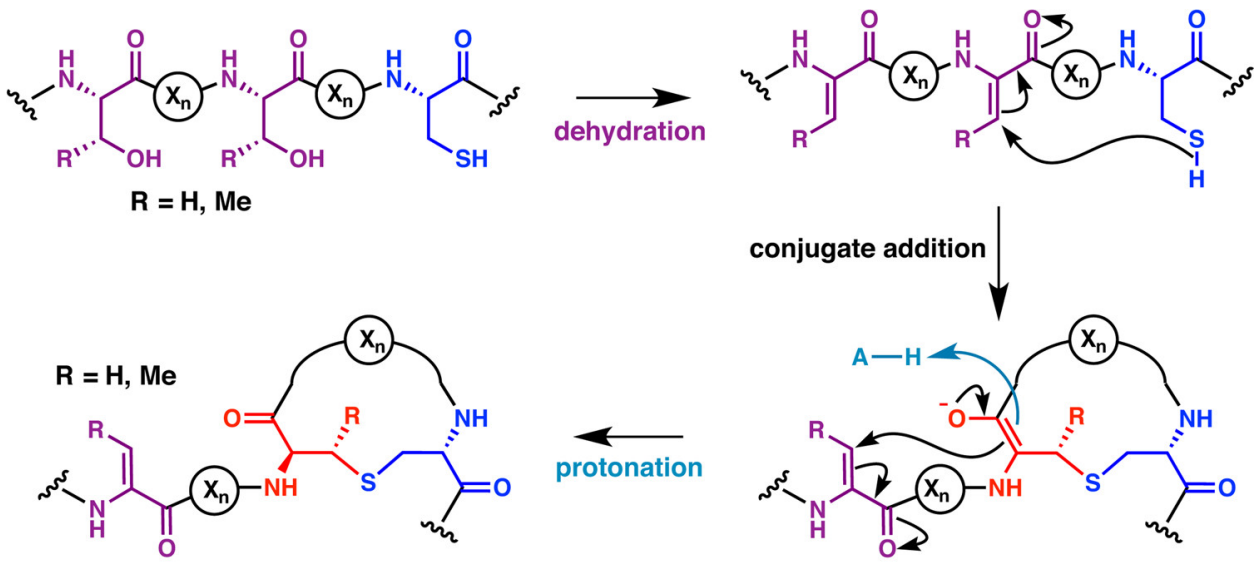


Figure 2.3.2: Biosynthesis of lanthionine groups in lanthipeptides. Serine and threonine residues (purple) are dehydrated into dehydroalanine (Dha) and dehydrobutyrine (Dhb) respectively. These dehydrated residues then react in a conjugate addition reaction with cysteine (blue), which forms an enolate with a thio-ether cross-link between the residues. The enolate is then protonated into the lanthionine group. Adapted from Figure 2 of (32).

Lanthipeptide Classifications and Class Specific Biosynthesis

Classification of lanthipeptides is mainly determined by differences in the mechanism behind the amino acid dehydration and thio-ether cross-linking. These differences can relate to the metabolic requirements and cross-linking pattern of the biosynthesis, and are caused by altered versions of the central lanthipeptide modification enzymes. To discern these different versions of the modification enzymes in the lanthipeptide classes, a lanthipeptide-specific nomenclature for the enzyme names is often used. Following this nomenclature, a general lanthipeptide polypeptide is named on the form "Lan*X*", where the *X* is one or more capital letters. The capital letters can denote polypeptides with both class specific and unspecific roles, so for example LanA is the general name of the precursor peptide for all lanthipeptide classes, while LanB refers to the serine and threonine dehydratase specific to class I lanthipeptides. Furthermore, the nomenclature can be used for polypeptides in a specific cluster. For example, in the case of Nisin A, a class I lanthipeptide, the precursor peptide is named NisA, and the class I specific dehydratase is named NisB, following from LanA and LanB respectively (18, 32).

Class I Lanthipeptides

For class I lanthipeptides the two main defining core modification enzymes are LanB, responsible for serine (Ser) or threonine (Thr) dehydration, and LanC, which facilitates correct thio-ether cyclization. LanB achieves dehydration of Ser/Thr residues through a process called glutamylation elimination (Figure 2.3.3). In this process, LanB enzymes catalyze a transesterification reaction where the glutamyl of a glutamyl-tRNA^{Glu} molecule forms an ester bond with the -OH group of either a serine or threonine sidechain. Afterwards, the glutamyl along with the Ser/Thr sidechain oxygen is β-eliminated, forming the double bond between the α- and β-carbon present in the dehydroalanine (Dha) and dehydrobutyrine (Dhb) residues (32).

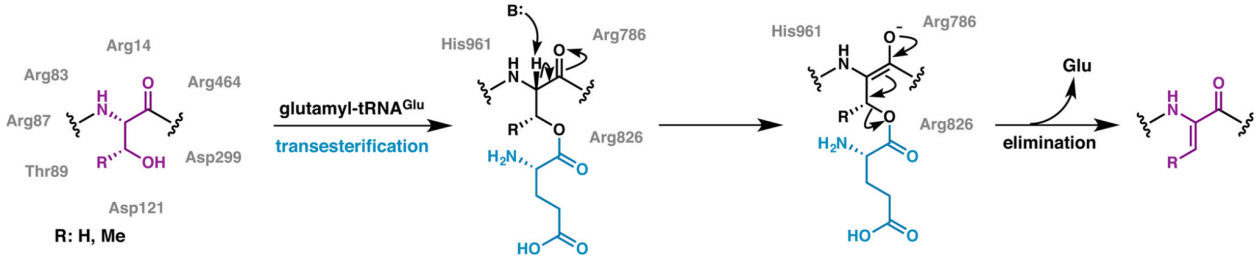


Figure 2.3.3: The glutamylation elimination reaction dehydrates serine and threonine residues in two steps. First a glutamyl from an activated tRNA^{Glu} molecule forms an ester bond with the hydroxyl group of serine/threonine, and is subsequently eliminated as glutamate causing the formation of a double bond between the serine/threonine α- and β-carbon. Figure adapted from Figure 8 of (32).

Due to the known reactivity between thiolates and the dehydrated amino acids, it is in principle possible for the thio-ether reaction to happen spontaneously between dehydrogenated cysteine and Dha/Dhb under basic conditions (32, 47–49), but an attempt to synthesize class 1 lanthipeptides without LanC resulted in various products with wrong ring topologies, indicating the importance of this enzyme (32, 50). Non-enzymatically, the reaction between Cys and Dha producing lanthionines is faster than the one between Cys and Dhb producing methyllanthionines (32, 51), and so it is reasoned that LanC counteracts this effect to ensure that a single product is produced with the correct ring topology, while also under cell conditions and with a specific stereochemistry. How exactly LanC achieves this is currently not known, and when considering how a single lanthipeptide core substrate can lead to 10^3s different ring topologies (10^6s taking stereochemistry into account), it is impressive that a single enzyme would achieve such a high precision (32).

A proposed mechanism of the thio-ether crosslinking achieved by LanC under cell-conditions involves the coordination from an active site in LanC (figure 2.3.4). The LanC active site consists of two Cys and one His residue with a zinc-ion ligand, weakly bound to molecular water. The thio-ether crosslinking reaction is initiated by the zinc-ion ligand of the LanC active site binding with the sulfur of a Cys residue in the core peptide, displacing the zinc-ion's

molecular water and lowering the pK_a of the sulfur. As a result of this, the sulfur is dehydrogenated into a thiolate, which subsequently attacks the β-carbon of either Dha or Dhb, forming an enolate. The enolate is then protonated, possibly by an active site acid in LanC, resulting in the finished lanthionine or methyllanthionine (32).

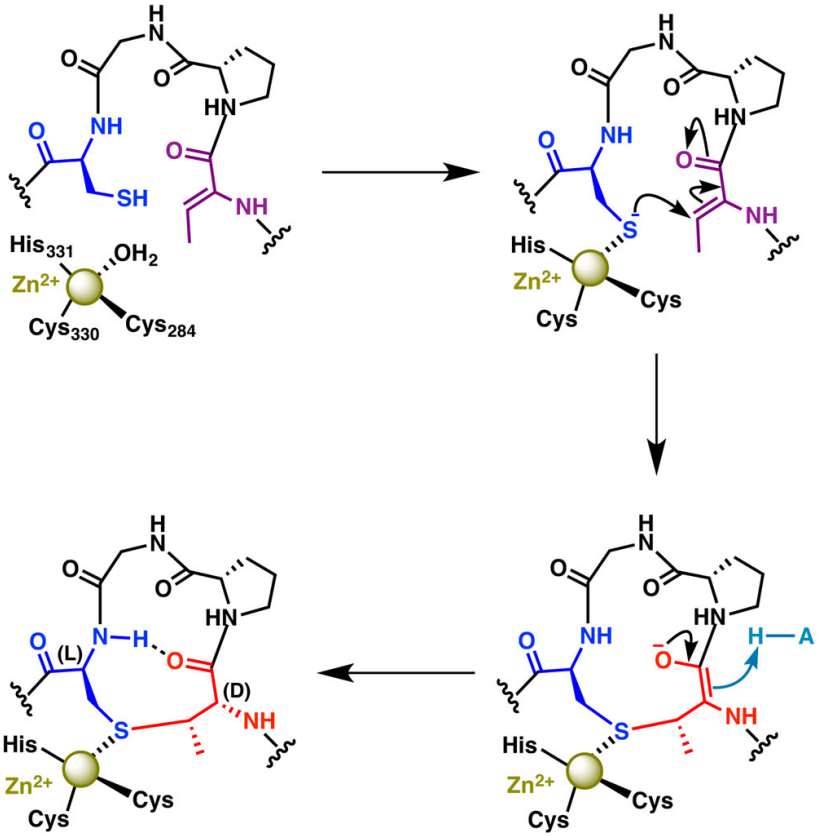


Figure 2.3.4: The LanC thio-ether crosslinking reaction which forms lanthionine groups occurs in three main reaction steps. First the active site zinc ion of LanC (yellow) dehydrogenates the sulfur of a cysteine residue (blue). The resulting thiolate attacks the β-carbon of Dha/Dhb (purple) creating an enolate. A proton donor (light blue) then protonates the enolate, resulting in the complete lanthionine. Figure adapted from Figure 13 of (32).

For simplicity, the activity of the LanB dehydration and LanC cyclization have been presented as two separate stages of the lanthipeptide biosynthesis, but in reality the enzymes function in tandem with each other to produce the correct product. Evidence suggests that LanB and LanC interact and co-localize in the plasma membrane of the cell, and a similar effect is present between LanC and the lanthipeptide secretion transporter LanT (32, 52, 53). Studies have found that the Ser/Thr dehydration by LanB and transport by LanT are non-contingent on LanC cyclization. However, when LanC is present, the ring formation between Cys and Dha/Dhb can occur before the Ser/Thr dehydration of LanA is complete (32, 54–56). This results in certain Ser or Thr residues becoming sterically protected from the LanB activity. Pattern-analysis of the position of these protected Ser/Thr residues, generally being toward the C-terminal of the

thio-ether rings, suggested an N-C direction for the LanB processing of LanA (32, 57). Consequently, LanC would also follow this N-C directionality, as it is dependent on the presence of Dha/Dhb for the ring formation to occur. The implied order of dehydration and ring formation from this N-C processing direction fits with the observed ring topologies of many of the class I lanthipeptides (32).

Two common tailoring reactions in class I lanthipeptides are done by the modification enzymes LanO, a dehydrogenase, and LanD, a flavin decarboxylase. The LanO tailoring reaction involves the reduction of an N-terminal Dha residue in the class I lanthipeptide core into a lactate group (Figure 2.3.5) (58). The LanD tailoring reaction involves the decarboxylation of a C-terminal cysteine residue in the class I lanthipeptide core (Figure 2.3.6). After which, LanC uses the decarboxylated cysteine residue to form a special thio-ether crosslinking group called aminovinyl cysteine (AviCys) (59).

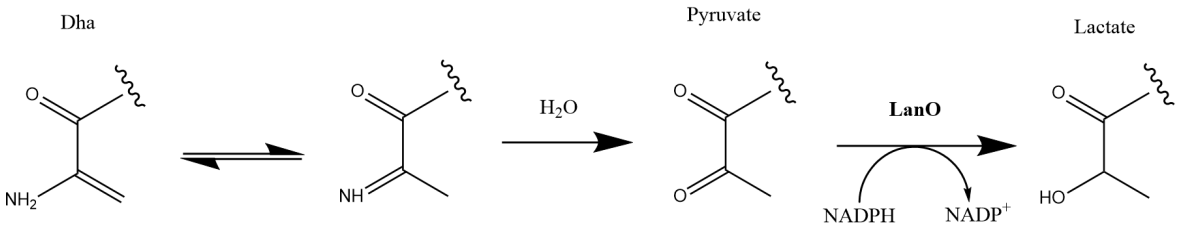


Figure 2.3.5: Reduction of N-terminal Dha to lactate by LanO. Occurs through a series of reactions where the amine group of the Dha is first dehydrogenated, and then a hydrolysis reaction turns the Dha into a pyruvate group which is subsequently reduced by NADPH into lactate (58).

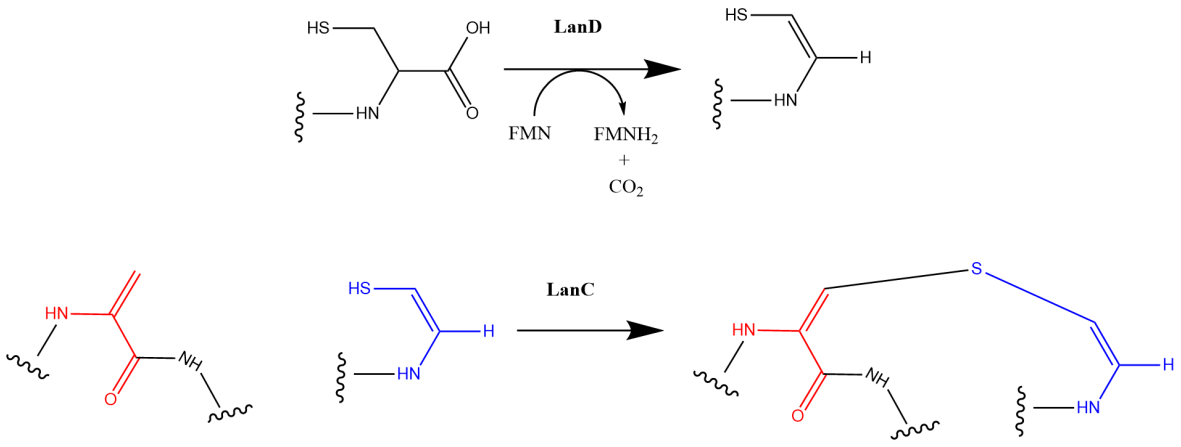


Figure 2.3.6: LanD tailoring reaction. First the C-terminal cysteine of a class I lanthipeptide core is decarboxylated with FMN as cofactor by LanD. The decarboxylated cysteine (blue) then participates in the LanC thio-ether crosslinking reaction with Dha (red), but produces an AviCys group instead of lanthionine (59).

Class II Lanthipeptides

Class II lanthipeptides are defined by the presence of the core modification enzyme termed LanM, which is responsible for both the dehydration of serine and threonine as well as the thio-ether crosslinking. LanM has a C-terminal domain with partial sequence homology to LanC, however it contains an N-terminal domain that does not have homology with LanB. Unlike the glutamylation dehydration reaction present in class I lanthipeptides by LanB, LanM instead achieves the same result through a phosphorylation reaction (Figure 2.3.7). In this reaction, LanM catalyzes the binding of an ATP phosphate group to either serine or threonine, displacing it from its ATP molecule and creating ADP. Afterwards, in the presence of ADP and Mg^{2+} LanM mediates the elimination of the phosphorylated amino acid, yielding the dehydrated amino acid as well as free phosphate and ADP. Both the phosphorylation and elimination is believed to be catalyzed by the same active site in LanM. The LanC-like domain of LanM is thought to catalyze thio-ether ring formation through the same mechanism as the class I LanC enzyme (32).

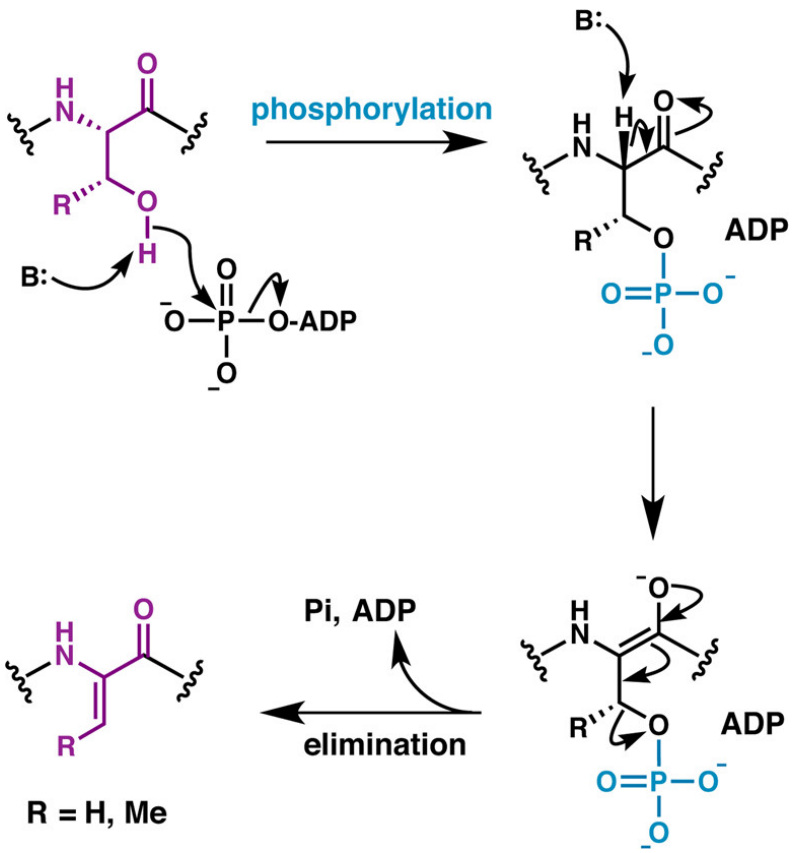


Figure 2.3.7: Reaction mechanism for serine/threonine dehydration by phosphorylation. A nucleophile (B:) binds to the hydrogen of the serine/threonine sidechain -OH group, which is subsequently phosphorylated by ATP. A second nucleophile binds to the hydrogen of the α-carbon of serine/threonine, causing a chain of reactions which eliminates the phosphate group and establishes a double bond between the α- and β-carbon. Adapted from Figure 28 of (32).

Despite the different dehydration mechanisms of LanM and LanB, one could expect the products of either LanM or LanBC for a specific LanA to be identical, yet this expectation does not always hold. The general pattern of ring formation observed in class I lanthipeptides, which is inferred from the dehydration happening in N-C direction and cyclization following it closely, is present only in some class II lanthipeptides. A group of class II lanthipeptides called Prochlorosins have been studied extensively and have ring topologies which do not follow the same formation pattern as observed for class I lanthipeptides. Prochlorosins originate from 30 different LanA substrates that are modified by the same central enzyme called ProcM (32, 60), and as such the study of the Prochlorosins has given a lot of insight into ProcM as well as the huge diversity of ring topologies which can originate from a single LanM enzyme. Through analysis of the dehydration event of several Prochlorosins by using α-carbon deuterium labeling, it was found that ProcM appears to perform Ser/Thr dehydration in the C-N direction, opposite to the class I lanthipeptides (32, 61). Additionally, a kinetic analysis of several ProcA substrates showed that unlike class I lanthipeptides, ProcM finishes dehydration before cyclization

commences, which could explain how the ring formation in ProcM substrates appears non-directional (32, 60). Studies of ProcM ring formation also seem to indicate that it favors the formation of smaller rings first (32, 61). Taking all of this into account, it becomes much more challenging to describe the reaction order of the cyclization event for ProcM and similar enzymes than it was for the class I lanthipeptide enzymes.

In essence, it appears that specific LanC enzymes or LanC-like domains favor the formation of certain ring structures during the cyclization event of a LanA substrate. What ring structures these enzymes would be able to achieve are limited by the conformational and chemical restrictions by the LanA peptide sequence, which can in some cases lead to the formation of the same mature end product. As such, a difference in the regioselectivity between LanC and LanC-like enzymes is an important factor determining the end-product ring topology as long as the LanA accommodates the possibility of such differences (32).

Class III and IV Lanthipeptides

Class III and IV lanthipeptides are distinguished by the presence of the modification enzymes LanKC and LanL respectively. Similarly to the class II LanM modification enzyme, both LanKC and LanL are large multi-domain enzymes which catalyze both serine/threonine dehydration as well as thio-ether cyclization reactions in the lanthipeptide precursor peptide core. What distinguishes them from LanM is the presence of separate kinase and lyase domains (Figure 2.3.8). The kinase domain catalyzes the phosphorylation of the hydroxyl group in serine/threonine and the lyase domain catalyzes the elimination of phosphate and consequent dehydration of the residue. Although not fully understood, these differences in dehydration strategy have apparent consequences that further distinguish class III and IV from class I and II lanthipeptides. It has been observed from studies of different LanKC enzymes that other nucleoside triphosphates than ATP can be required for the phosphorylation step, with preferences for GTP and even ambiguous preference towards multiple NTPs. Furthermore, studies of the directionality of LanKC and LanL dehydration found a much larger variance than the class II and particularly class I dehydration event, with less of a uniform pattern or direction. As we saw with the tight coupling between dehydration and cyclization events in class I lanthipeptides, we can expect that these differences in dehydration order will have a large effect on the expected structural outcome, making it challenging to infer a specific reaction order (32).

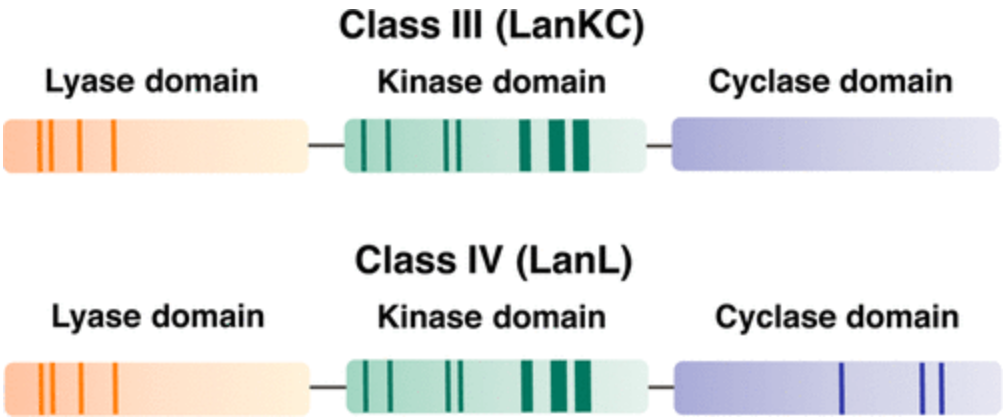


Figure 2.3.8: Overview of common genetic similarities and differences between the domains of LanKC and LanL genes. We may notice the lack of blue lines in the cyclase domain of LanKC indicating its lack of the zinc-binding active site. Figure adapted from Figure 46 of (32).

LanKC and LanL enzymes can be distinguished from each other from the lack of a zinc-binding active site in the LanC-like cyclase domain of LanKC. The consequences of this is not fully understood, however a feature which can be found exclusively in class III lanthipeptides is the presence of labionin groups. Labionine forms as a result of the enolate from the conventional LanC-like catalyzed thioether reaction having another conjugate addition with an upstream Dha residue before protonation (Figure 2.3.9) (32).

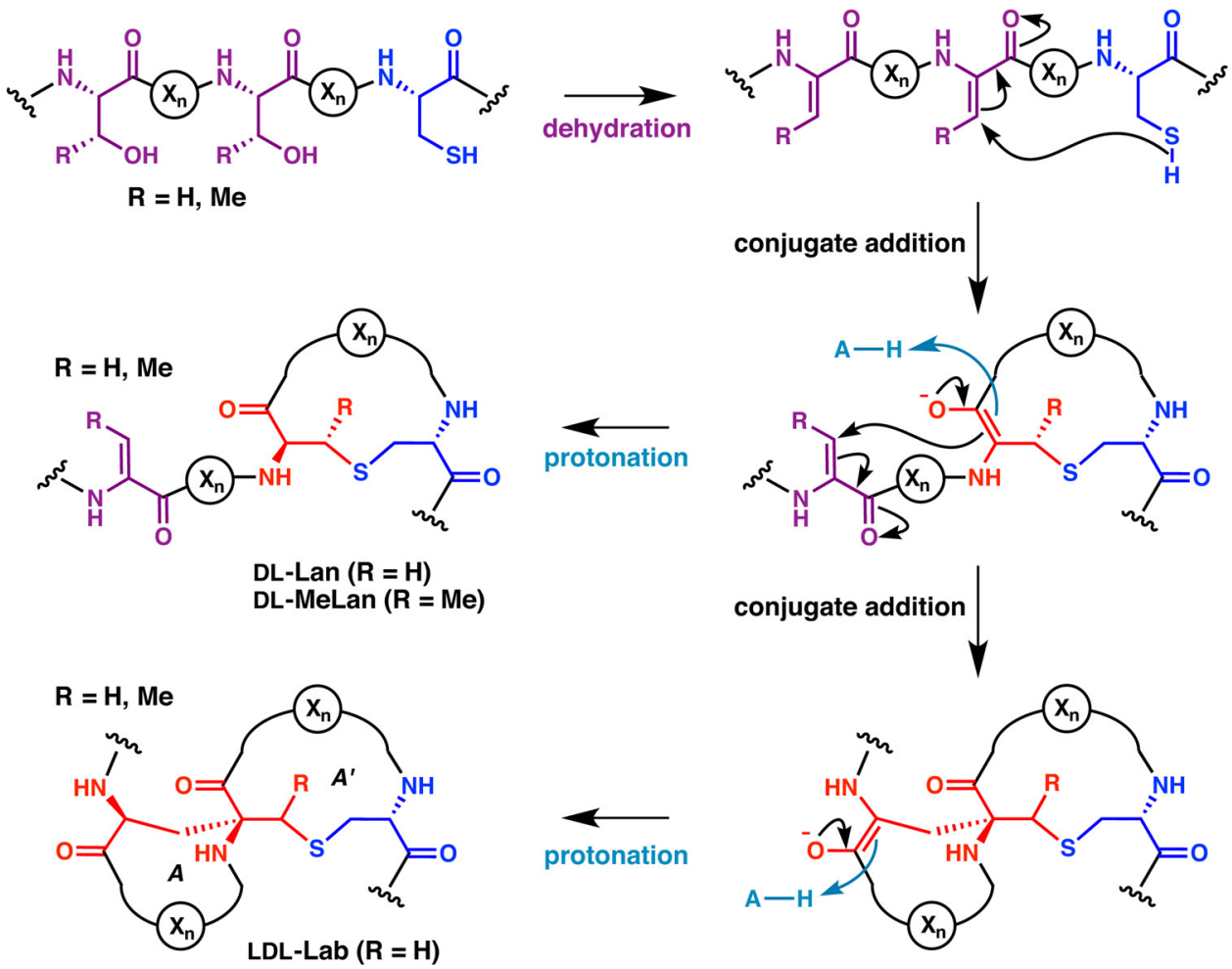


Figure 2.3.9: Overview of the possible routes in the thio-ether cyclization reaction of class III lanthipeptides. In certain cases, the enolate created from the first conjugate addition may react with a downstream Dha residue, and after protonation this forms the class III specific labionine group. Adapted from Figure 2 of (32).

Lanthipeptide Genome Mining

The most essential components in lanthipeptide gene clusters are the domains responsible for serine/threonine dehydration and thio-ether crosslinking. It is also important to discern between the different types of domains between the lanthipeptide classes. Version 6 of antiSMASH, the newest as of writing, contains trained pHMMs for functionally annotating these central lanthipeptide domains (12). Determining the lanthipeptide class annotation is done by filtering through a set of rules which first checks if the cluster contains the class I characteristic LanB domain, then the class II characteristic LanM domain, both class III and IV lanthipeptides are then picked if the characteristic kinase domain is present. To discern between class III and IV, a separate algorithm checks for the presence of the zinc-binding active site in the cyclase domain

characteristic for class I, II and IV lanthipeptides. For the LanO and LanD tailoring reactions, hits against pHMMs trained on specific Pfam profiles for the protein families they belong to were used to annotate them (27).

2.4 Thiopeptides

Thiopeptides are a large RiPP class consisting of structurally complex compounds that commonly have strong antibacterial activity against Gram-positive bacteria (33, 62). The main defining chemical motif of thiopeptides is the presence of a macrocycle with a central nitrogen heterocycle, which is formed through a reaction between dehydrated amino acid residues in the core peptide (figure 2.4.1). Additionally, thiopeptides generally consist of several thiazoles and sometimes oxazoles, which are five membered heterocyclic groups derived from cysteine and serine/threonine residues respectively, as well as dehydrated serine and threonine residues. Different groups of thiopeptides may also have an additional macrocycle or sidechain. There are also structures made by tailoring reactions which can be highly specific to certain thiopeptide compounds (33).

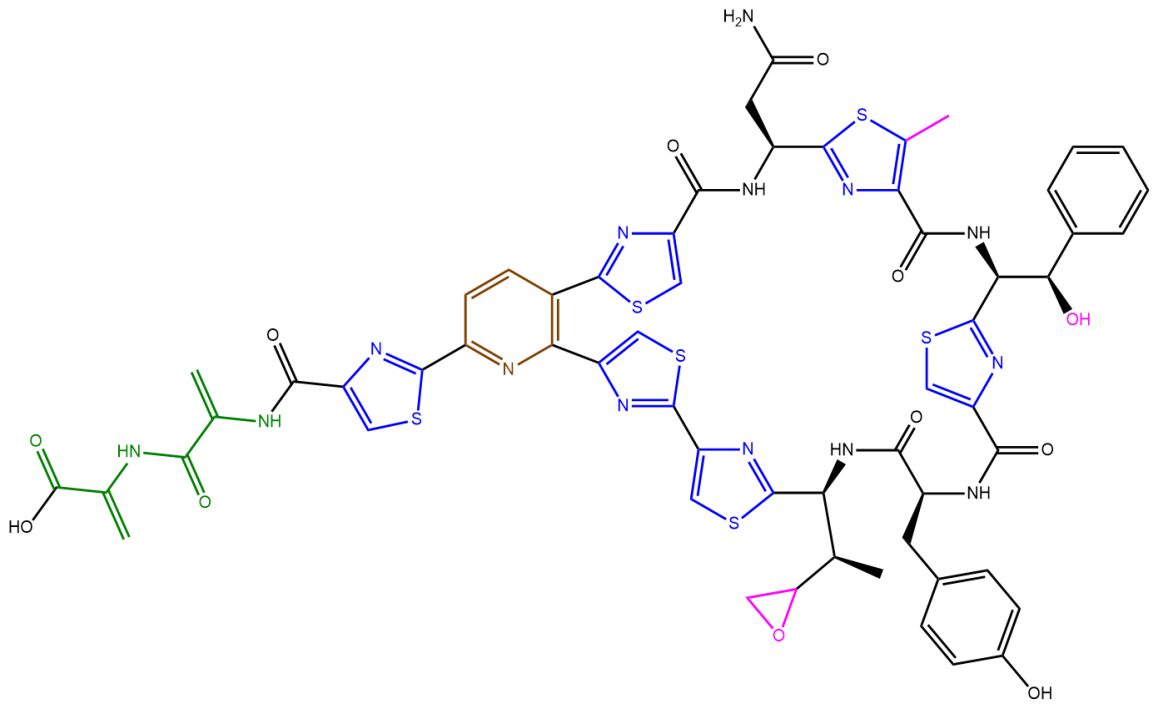


Figure 2.4.1: Chemical structure of the thiopeptide Thiomuracin A. The central nitrogen heterocycle is colored brown, azoles are colored blue, dehydrated residues green and structures from various tailoring reactions are pink.

There is a large degree of structural variation between thiopeptides, and for thiopeptides with antibacterial activity there are many unique modes of action. Nonetheless, the target of all known thiopeptide antibiotics is protein synthesis, where they act as inhibitors (33, 63–68). Several groups of thiopeptides achieve this by binding to the bacterial ribosome, particularly the guanosine triphosphatase associated center (GAC) responsible for peptide elongation (33, 69). These groups can sterically block the association of translation elongation factors with the GAC, prevent the ribosomal assembly by blocking initiation factors associated with the GAC, and even lock ribosomal proteins in a continuous unproductive ATP consuming state. Some groups can also induce autophagy of the cell through ER-stress (33). These modes of action distinguish themselves from those of conventional pharmacological antibiotics (70), which are becoming less potent, and therefore the potential for application of thiopeptide antibiotics is of great interest (33, 62). Unfortunately, due to problems with thiopeptide solubility and stability, there have been challenges both in extracting thiopeptides as well as using them in therapeutic applications (33).

Biosynthetic Mechanism of Thiopeptides

Thiopeptide biosynthesis occurs through a complex series of chemical events requiring multiple modification enzymes. Six genes are homologically conserved across every thiopeptide BGC, and are essential for their central biosynthesis as well as the primary macrocyclization which distinguishes thiopeptides as a RiPP class. The enzymes produced by these six genes catalyze reactions in three main chemical events of the biosynthesis: azole formation, serine and threonine dehydration, and primary [4+2] macrocyclization. Beyond the six central modification enzymes, thiopeptide BGCs can have as many as 11 additional modification enzymes that may participate in secondary macrocyclization reactions and other supplementary tailoring reactions (33). Both the central and additional biosynthetic modifications can cause large alterations in the overall structure of the peptide, as well as of multiple residues in the peptide, and so it becomes apparent why there is such a large structural diversity between thiopeptides.

Central Biosynthesis of Thiopeptides

Azole formation is the first event in the chain of modifications to the thiopeptide core for the majority of thiopeptides, and is a central step that is crucial for the later primary [4+2] macrocyclization event. Azoles are five membered aromatic heterocyclic chemical groups consisting of at least one nitrogen and one other non-carbon atom (33, 71). In thiopeptides two types of azoles can be found: thiazoles, which are formed from cysteine residues, and oxazoles, which are formed from either serine or threonine residues. However, the vast majority of thiopeptides only contain thiazoles, and generally thiopeptides will have every cysteine residue modified into thiazoles (33, 62, 72). On the other hand, when oxazoles are present they rarely

appear for the majority of available serine/threonine residues, and can mostly be found in two small thiopeptide groups, berninamycins and lactazoles (33, 72).

Two reactions make up the thiazole formation in thiopeptides, requiring three proteins. The enzyme that catalyzes the first reaction is from the YcaO superfamily, which consists of enzymes that through phosphorylation can activate the peptide backbone and open up a nucleophilic attack on the peptide bond (73). In the case of the thiopeptide YcaO, a partner protein referred to as a Thif-Ocin-like protein which contains RREs is generally required to successfully bring the catalytic site of the YcaO enzyme close enough to the thiopeptide core. The thiopeptide YcaO then uses a molecule of ATP to phosphorylate the carboxyl group in the N-terminal peptide bond of a cysteine residue, causing it to be attacked with the cysteine sidechain as the nucleophile. The cysteine's sulfur atom binds to the N-terminal peptide bond's carbon, displacing a molecule of water and creating a group called a thiazoline. This thiazoline group is then dehydrogenated into a thiazole through an FMN-oxidation catalyzed by a thiopeptide specific dehydrogenase, whose enzyme group is not fully characterized (33).

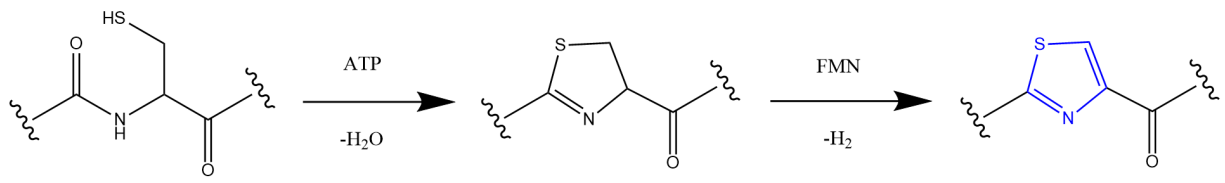


Figure 2.4.2: Thiazole formation performed by YcaO enzyme and Thif-Ocin-like protein. This reaction occurs through two steps: first ATP is used to form a thiazoline intermediate, then FMN acts as a proton acceptor to aromatize the thiazoline group into a thiazole (blue).

In the case of oxazole formation, the same general reaction as in thiazole formation happens but with the serine/threonine sidechain as a nucleophile. Oxazolines dehydrogenate much slower than thiazolines due to their lower aromaticity, and some thiopeptide dehydrogenases are incapable of dehydrogenating oxazolines at all. A different structure of YcaO enzymes and thiopeptide dehydrogenases can be observed in thiopeptide BGCs with known oxazole groups in their products. YcaO/dehydrogenase complexes and fused proteins, as well as both multiple YcaO enzymes and thiopeptide dehydrogenases seem to enable some thiopeptide BGCs to form oxazoles in their products. Additionally, in the case of the lactazole dehydrogenase, oxazoline dehydrogenation appears to be facilitated by conjugation with a neighboring Dha residue. This also indicates that the oxazole formation can slip over into the next chemical event in the central thiopeptide biosynthesis, serine and threonine dehydration (33).

Having read the biosynthetic mechanism of lanthipeptides, the dehydration of serine and threonine should sound familiar. In fact, the serine and threonine dehydration event in

thiopeptides uses the same mechanism as the class I lanthipeptide LanB enzymes, glutamylation elimination. Indeed, the associated genes for serine/threonine dehydration in thiopeptides share homology with the glutamylation and elimination domains found in the LanB enzymes (32, 33, 74). This is a good example of the orthologous nature of genes in BGCs, where one could expect thiopeptides and class I lanthipeptides to share a common ancestral BGC containing the glutamylation elimination genes. Over evolutionary time this ancestral BGC would evolve into the class I lanthipeptide and thiopeptide BGCs that can be seen today.

After the formation of azoles and Dha/Dhb residues in the thiopeptide core is finished, the primary macrocyclization event occurs through the presence of cycloaddition enzymes. Thiopeptide primary macrocyclization happens through a [4+2]-cycloaddition reaction between two Dha residues in the core peptide, producing a central pyridine/piperidine-esque core where each end of the macrocycle is attached (Figure 2.4.3). Formally, this is an intramolecular aza-Diels-Alder reaction, where the 2π "alkene" component is the α- and β-carbon of the furthest N-terminal Dha, while the 4π "diene" component consists of another Dha α- and β-carbon connected to a base dehydrogenated N-terminal peptide bond, which gives a conjugated chemical structure (33, 75). It also appears that the 4π-component is always flanked by an azole group at its C-terminal, and often also at its N-terminal. In the peptide sequence of thiopeptides the 2π- and 4π-components commonly appear from motifs such as $X_{(N)}$-S-$X_{(N)}$-S-C-$X_{(N)}$ and $X_{(N)}$-S-$X_{(N)}$-C-S-C-$X_{(N)}$, where the subscript N is a variable number of residues. Additionally, the length of thiopeptide cores is relatively short compared to for example the lanthipeptide cores, and generally we can expect there to be only one set of 2π- and 4π-components in a single thiopeptide core. Additionally, the modification enzyme that coordinates this macrocyclization reaction appears to favor the formation of the same ring size for the same thiopeptide group, similarly to how different LanC-domains favor certain ring topologies in different groups of lanthipeptides. The macrocyclization enzymes of several thiopeptides have been found to be highly specific to the 4π-component and surrounding structures (33).

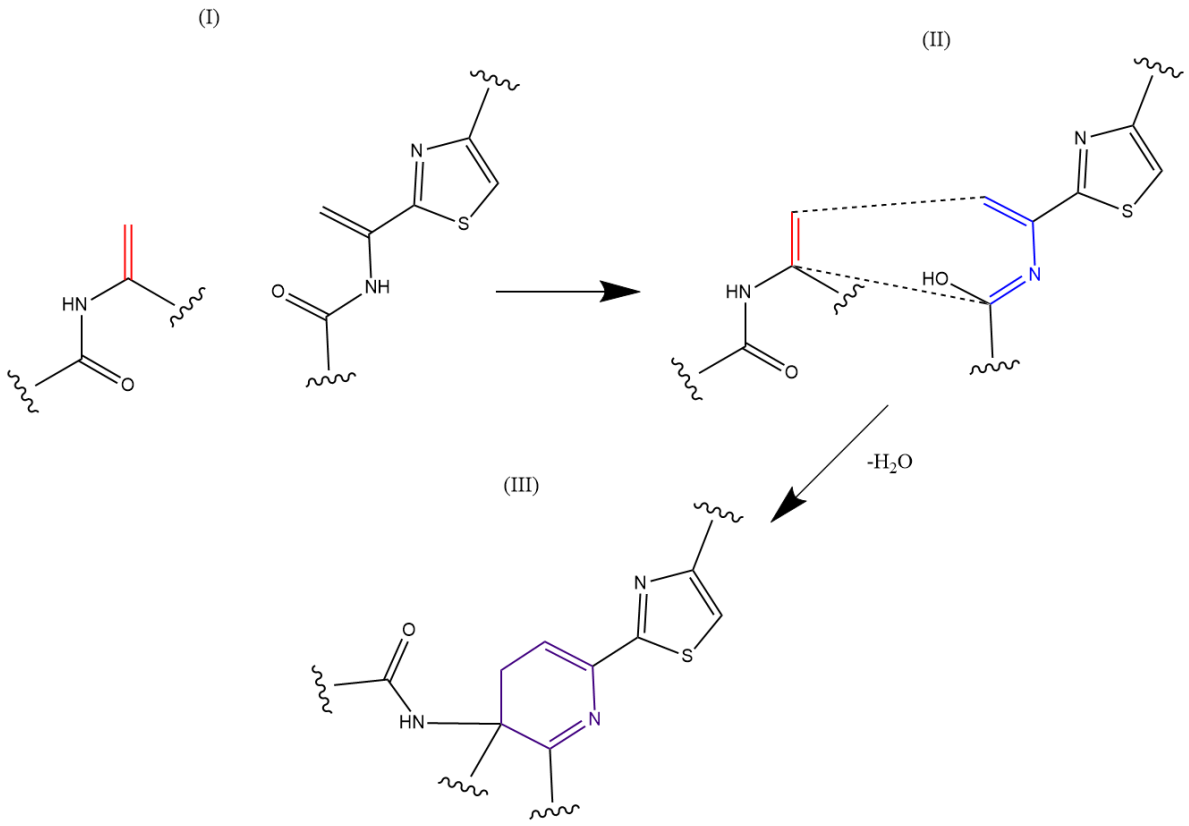


Figure 2.4.3: Aza-Diels-Alder conjugation reaction for primary macrocyclization in thiopeptides. For the first set of reactants (I), we see the 2π-component. Keep in mind that these are two different motifs within the thiopeptide core. For the first reaction step the 4π-component (red) is formed through dehydrogenation of the Dha nitrogen, and the conjugation then happens between the carbon positions pictured in the second set of reactants (II). For the second reaction step, the central nitrogen heterocycle (purple) forms as an intermediate (III), which then leads into subsequent modification reactions.

The first step of the primary macrocyclization reaction is the dehydrogenation of the peptide-bond nitrogen yielding the 4π-component, which leads right into the second step which is cycloaddition. In this second step chemical bonds are formed between the 2π-component Dha β-carbon and 4π-component Dha β-carbon, and between the 2π-component Dha α-carbon and 4π-component N-terminal peptide bond carbon (Figure 2.4.3). The cycloaddition produces what is referred to as a Bycroft-Gowland Intermediate, a six membered heterocycle where the peptide chain of the 2π- and 4π-component is attached. This intermediate is quickly dehydrogenated, displacing a water molecule and aromatizing the six-membered central cycle into a second intermediate. This second intermediate is then either reduced or oxidized (Figure 2.4.4), yielding two separate routes with further possible reductions and oxidations of the six membered central cycle. The oxidation of the thiopeptide central heterocycle is initiated by a proton acceptor, and the leader peptide chain is disattached as the central heterocycle is aromatized into a pyridine

(75). The central pyridine group can be further enzymatically oxidized to have a substituted alcohol group, using NADPH and O_2. The reduction mechanism for the thiopeptide central heterocycle is poorly understood, but it is implied that the first reduction requires some kind of proton donor (33). Unlike the oxidation stages, the reduction stages of the central heterocycle retain every attached peptide chain from the Bycroft-Gowland Intermediate (33, 75).

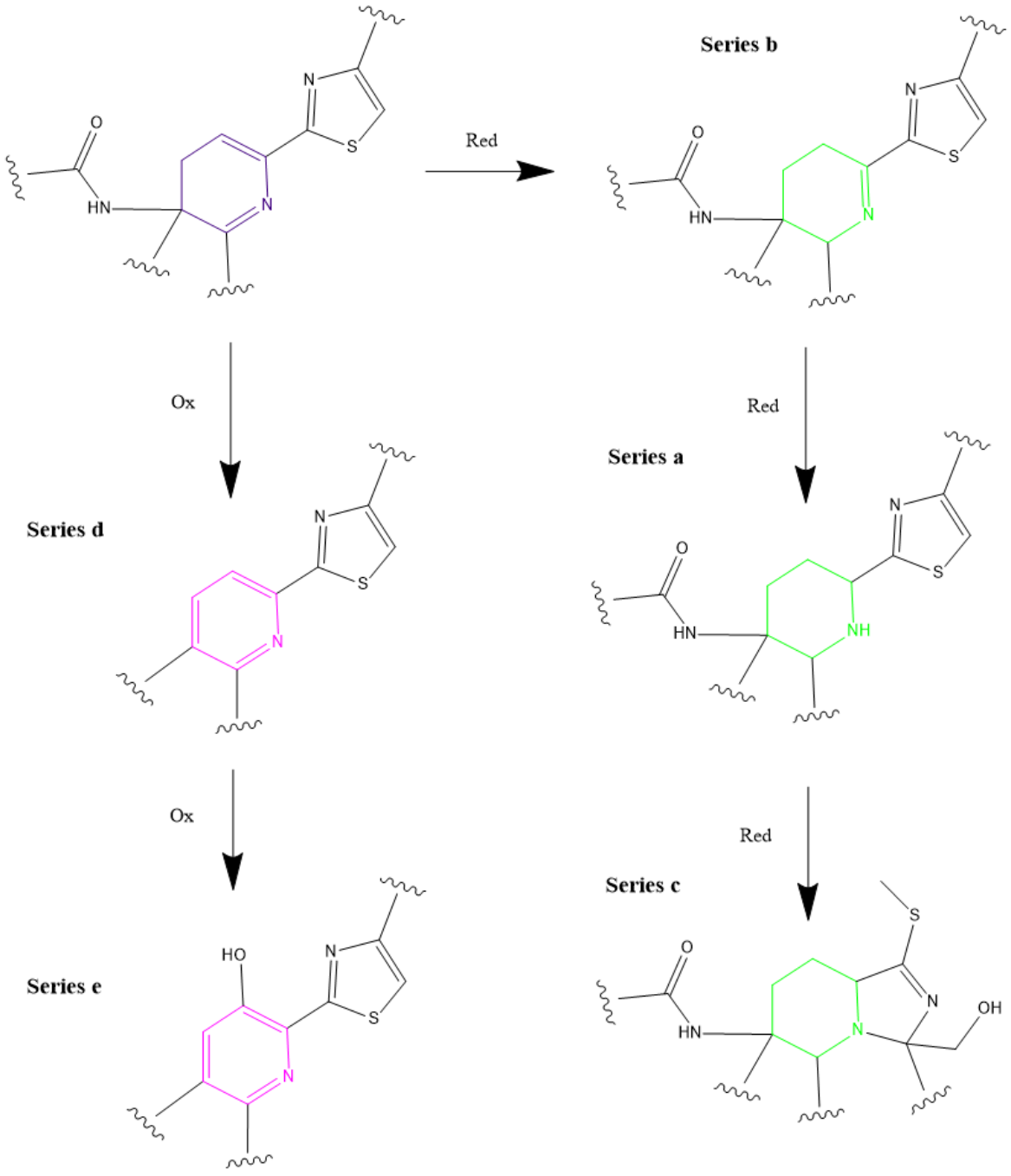


Figure 2.4.4: Central heterocycle modification reactions in thiopeptides. The central heterocycle intermediate (purple) can be reduced in several steps of piperidine products (green) or oxidized into pyridine products (pink), based on the presence of specific thiopeptide modification enzymes.

Thiopeptide Classification

Currently there is more than one way to classify thiopeptides. The three reduction stages of the thiopeptide central cycle give rise to three thiopeptide groups referred to as series b, series a and series c in order of reduction. Likewise, the two oxidation stages give rise to the thiopeptide groups series d and series e, in order of oxidation. The categories defined by the reduction/oxidation stage of the thiopeptide central cycle are commonly used in classifying thiopeptides (Figure 2.4.4), as they can in some cases relate to differences caused by further thiopeptide modification reactions. Another classification method is based on the primary macrocycle member size. Empirical evidence suggests there is a correlation between this and the activity of thiopeptides in biology, and so it may be useful to use this classification when considering biological activity (33).

As mentioned earlier there are also thiopeptide groups, such as the berninamycins and lactazoles that are known for containing oxazoles. Thiopeptide groups are generally determined by chemical similarities, such as the substitution pattern around the central heterocycle and different additional modifications like secondary macrocycles and siderings. The substitution pattern around the central heterocycle is in large part determined by the flanking residues of the 2π- and 4π-components, and so motifs present in the peptide sequence of the thiopeptide core could in part be used to determine thiopeptide groups. However, the substitution pattern is also related to the thiopeptide series classification, as the reduction stages of the central heterocycle (series a-c) can have an additional peptide chain attachment to the central heterocycle while the last oxidation stage (series e) contains a substituted alcohol group. Some thiopeptide groups which are of particular interest are the thiostrepton and nosiheptide groups. Each group contains conserved genes in their cluster for secondary macrocyclization, in the case of thiostrepton, and sidechain modification, in the case of nosiheptide. Furthermore, all currently characterized series a-c thiopeptides contain features of the thiostrepton group, while all of the nosiheptide-like thiopeptides are series e (33).

Secondary Macrocyclization of Thiopeptides

A major modification reaction outside of the central thiopeptide biosynthesis is secondary macrocyclization, which is present in slightly different ways in the thiostrepton and nosiheptide thiopeptide groups. Secondary macrocyclization involves an extensive set of additional biosynthetic genes in the thiopeptide cluster, which facilitate the production of an addition reactant, the binding of this reactant to the thiopeptide core and the intramolecular reaction to form the secondary macrocycle. Both the thiostrepton and nosiheptide groups start off this reaction chain with free L-tryptophan, though in the case of thiostrepton the addition reactant is a quinaldic acid compound (QA) (Figure 2.4.5), while in the case of nosiheptide it is a 3-methylindolic acid compound (MIA) (Figure 2.4.6) (33, 76, 77). Both pathways involve radical SAM (rSAM) enzyme mediated reactions, some of which are not fully understood. The timing of the secondary macrocyclization in the thiostrepton group is not fully elucidated.

However it has been found that the side-ring formation in the nosiheptide group occurs before primary macrocyclization (33, 76).

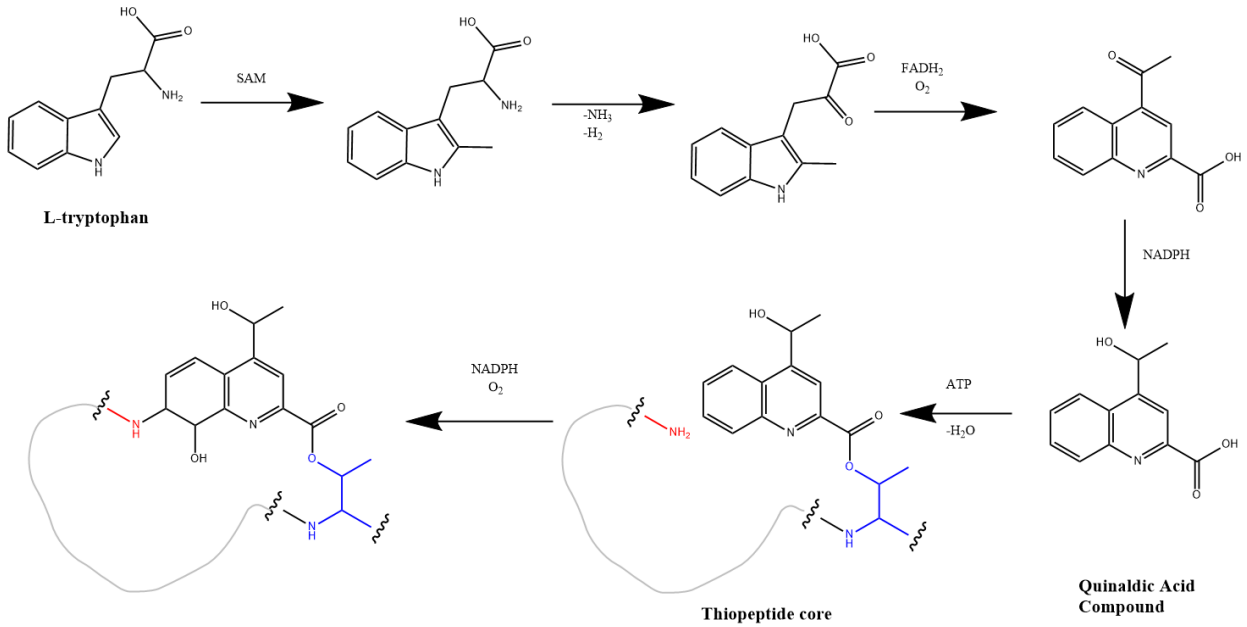


Figure 2.4.5: Pathway of secondary macrocyclization in the thiostrepton-like thiopeptides. Through a series of reactions L-tryptophan is converted to a QA compound, which is subsequently attached to a threonine (blue) in the thiopeptide core. The QA moiety then reacts with the N-terminal (red) of the thiopeptide core to complete the secondary macrocyclization. In the conventional thiostrepton compound pathway, the reacting threonine is at the 12th position in the thiopeptide core sequence (33).

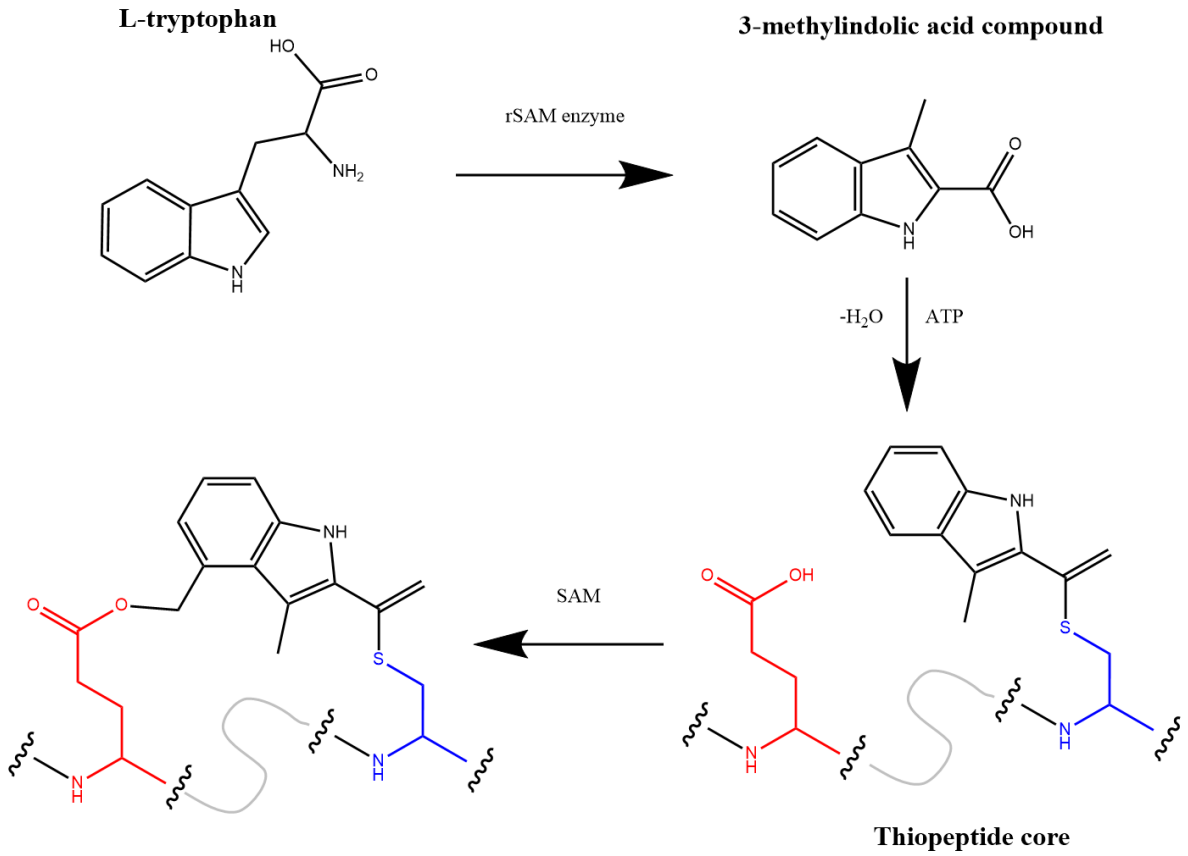


Figure 2.4.6: Pathway of secondary macrocyclization in the nosiheptide-like thiopeptides. Through a series of reactions L-tryptophan is converted to a MIA compound, which is first attached to a carrier protein and then transferred to a cysteine in the thiopeptide core. The formation of the nosiheptide ring is completed when the MIA moiety binds to a glutamate in thiopeptide core. In the conventional nosiheptide compound pathway, the reacting cysteine is at the 8th position, and the reacting glutamate is at the sixth position in the thiopeptide core sequence (33, 76).

Thiopeptide Genome Mining

The detection rules for thiopeptide clusters that antiSMASH uses involves checking for the presence of thiopeptide YcaO enzyme domains or its associated TIGRFAM ID using pHMMs, and at least one other of a series of thiopeptide characteristic domains. This series of domains includes the class I lanthipeptide dehydratase domain, LanB, and cyclase, LanC, certain thiopeptide specific Pfam domains or a predicted thiostrepton-like precursor sequence. The thiostrepton functional annotation of the precursor also indicates that the cluster has secondary macrocyclization genes. Thiopeptide series classification is done by checking for the presence of either the thiostrepton-like quinaldic acid related genes, for series a-c, or the nosiheptide-like

methylindolic acid related genes, for series e. If neither are found, series d classification is presumed (12, 33). The exact prediction rules that antiSMASH uses for classification was unclear, and it is important to note that it does not appear to be based on the presence of the genes for the redox enzymes that actually facilitate the different central nitrogen heterocycle modification reactions, relying only on the presence of the secondary macrocyclization genes (11, 12, 36–39).

2.5 Lasso Peptides

Lasso peptides are a class of RiPP compounds distinguished by their “lasso” shaped chemical structure, where a tail segment of the peptide is threaded through a macrocyclic lactam segment and locked in place. This unusual conformation confers a high degree of stability, resisting degradation in low and high pH environments, during heat treatment and from protease activity. Furthermore, lasso peptides have a wide array of biological activities and potential applications. Lasso peptides with antibacterial activity have shown to be effective against large groups of pathogens. Further, there are lasso peptides shown to have antifungal activity against human-pathogenic fungi in drug combinations, and even suppressive activity against retroviral proteases such as the HIV protease (78).

Generally the modes of action for lasso peptides involve binding to and disrupting or inhibiting specific proteins, and a single lasso peptide may have several such bioactivities. As it has been found that several of these target proteins are part of pathogens and other human antagonistic biological processes, it is of great interest to discover novel lasso peptides through genome mining, as well as researching potential therapeutic applications for lasso peptides (78). One such application could be epitope grafting the lasso peptide onto proteins with useful reactive functions, and using the modified peptide as a probe or for drug targeting. Lasso peptides are good epitope candidates due to their high stability (78, 79). Despite the promising potential uses for them, the library of heterologous-expression systems for lasso peptides is limited, their biosynthetic regulation is poorly understood, and there currently has not been great success in upscaling lasso peptide production (78).

Biosynthetic Mechanism of Lasso Peptides

Compared to the biosynthesis of lanthipeptides and thiopeptides, lasso peptides have a much less complex biosynthetic mechanism that appears highly conserved between different lasso peptides. The lower complexity is in large part due to the lower amount of modifications needed for the central lasso peptide biosynthesis. The central genes for biosynthesis in lasso peptide BGCs are

an RRE-containing protein, a cysteine protease and a lasso cyclase. In some clusters the RRE-containing protein is fused to the protease, and in some the protease may be split up in several genes (78, 80). Biosynthesis is initiated by the RRE-containing protein binding to the leader domain of the lasso peptide precursor and recruiting the cysteine protease, or in the case of the fusion protein it only needs to bind to the leader domain (Figure 2.5.1). The protease then cleaves the leader domain from the precursor, spending one molecule of ATP. The lasso peptide is pre-folded in the correct conformation before the lasso cyclase employs a molecule of ATP to AMP-activate an acid residue within the core peptide, either aspartate or glutamate. Lasso cyclase then guides the formation of a macrolactam ring in the core peptide through the N-terminus of the peptide reacting with the activated acid residue in a condensation reaction (35, 80, 81). Like thiopeptides, lasso peptide compounds are subject to a diverse and specific types of tailoring reactions, often by unknown mechanisms and genes. However, these modifications are generally very structurally minor, with few alterations to the largely intact polypeptide backbone of lasso peptides (81).

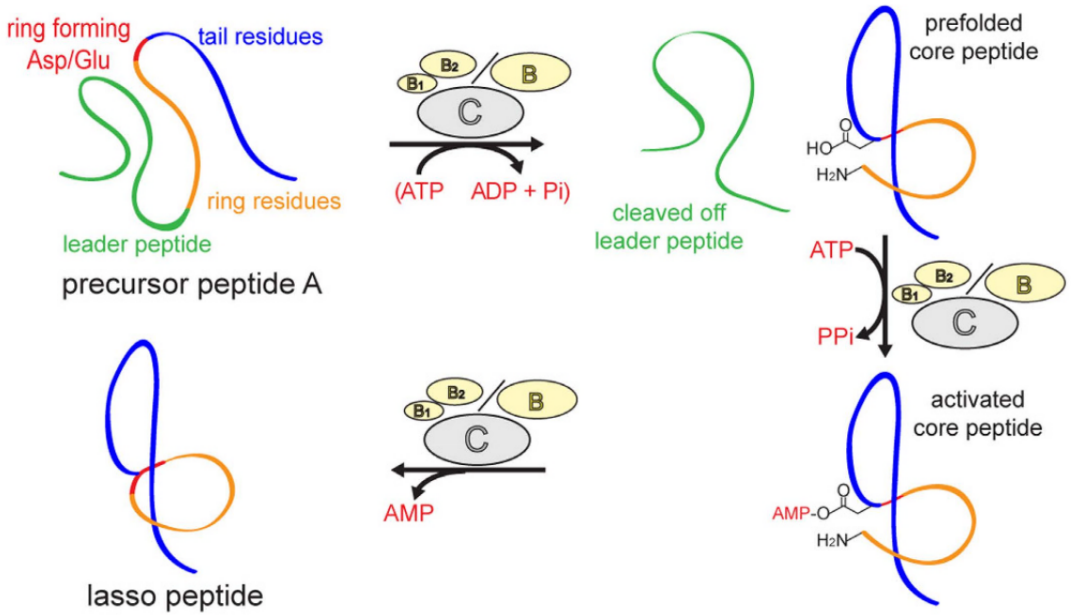


Figure 2.5.1: Overview of central lasso peptide biosynthesis. A lasso peptide protease (B) cleaves the leader peptide together with a lasso peptide cyclase (C), which prefolds the core peptide into its correct conformation. The cyclase then AMP-activates an acidic residue within the core peptide, either aspartate or glutamate, which then reacts with the N-terminus of the core peptide and closes the macrolactam ring. Adapted from Figure 1 of (80).

Lasso Peptide Classifications

Lasso peptides are classified by different strategies of retaining its conformation (Figure 2.5.2). With the exception of Class II lasso peptides, lasso peptides use polypeptide canonical disulfide bridging between cysteine residues to stabilize the conformation, but the three classes do this in

different ways. Class I lasso peptides are recognized by having four cysteine residues in its core, two in the tail domain and two in the ring domain. Here the cysteine residue closest to the N-terminal binds with a cysteine residue close to the middle of the tail domain of the core, while the cysteine residue closest to the C-terminal binds with a cysteine residue closer to the condensation reaction site (aspartate/glutamate location). Class III lasso peptides have only two cysteine residues, with a disulfide bridge at similar locations as the latter disulfide bridge in Class I lasso peptides. Class IV lasso peptides also only have two cysteine residues, but these can be found close together near the C-terminal in the tail domain of the core. This strategy involves creating something like a disulfide bridge plug at the end of the threaded tail domain in the core, which keeps the conformation stable through steric hindrance. A similar strategy is in fact employed by the Class II lasso peptides, but instead of using a disulfide bridge as a plug there is a residue close to the C-terminal with a large sidechain that creates high steric hindrance (35).

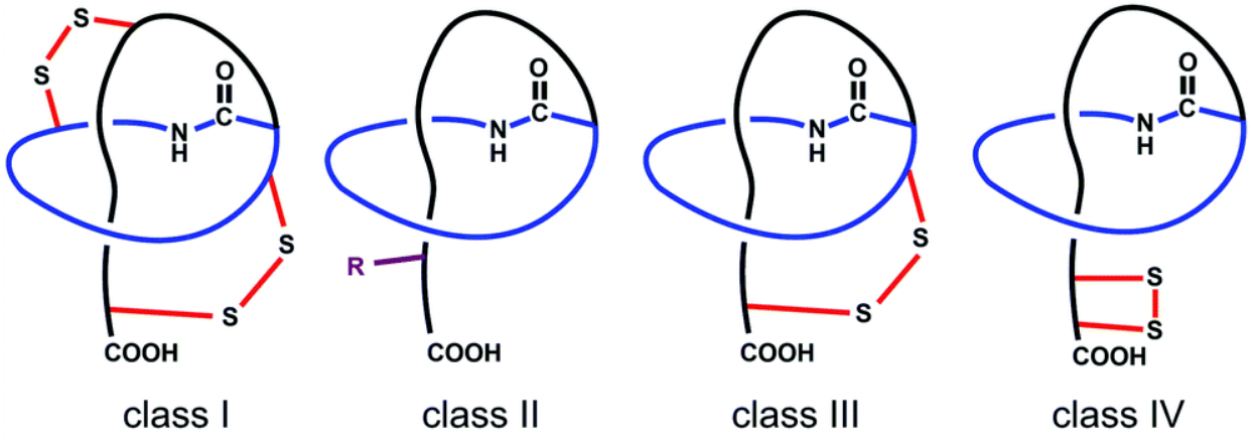


Figure 2.5.2: Different strategies of maintaining the lasso peptide characteristic conformation give rise to the different lasso peptide classifications. Class I lasso peptides employ two disulfide bridges between the tail and macrolactam ring, class II uses a C-terminal positioned residue with a large side group as a "steric plug", class III uses a single disulfide bridge between the tail and macrolactam ring and class IV uses a disulfide bridge as a "steric plug". Adapted from Figure 4 of (35).

Lasso Peptide Genome Mining

Lasso peptide clusters are determined in antiSMASH by the presence of a lasso peptide characteristic Pfam domain for a transglutaminase-like superfamily and either a lasso peptide characteristic asparagine synthetase or domains with homology to genes of a specific lasso peptide, microcin J25 (39). It appears that antiSMASH assigns class II to predicted cores without cysteine and class I to cores with two pairs of cysteine. It is however unclear how or whether antiSMASH distinguishes between class III and IV (11, 12, 36–39).

2.6 Metabolic Modeling

Genome-scale Metabolic Models (GEMs) and Their Reconstruction

Genome-scale metabolic models (GEMs) are model representations of the metabolites, reactions and associated genes of an organism's metabolism. They allow the simulation of both the metabolic phenotype and genotype, and can therefore be utilized to study things like gene-knockout effects on the metabolism of an organism (82). The metabolites and reactions of a GEM are structured in a stoichiometric matrix, which mathematically describes the stoichiometry of each reaction as a set of columns where each row is a metabolite. Genes can then be mathematically associated with one or more of these reactions (83).

The reconstruction of GEMs involves using reactions identified from genetic elements in an organism's genome to build a model that reflects the organism's metabolism. A common method of achieving this is to choose a specific organism with an annotated genome and use its associated reactions to create a draft model, which is then manually curated by adding missing reactions, removing blocked pathways and generally "fixing" the model until the GEM is accurate enough (84). Through continuous manual curation and experimental verification, this method allows for creating high quality GEMs with accurate predictions of the organism's phenotype, such as Sco-GEM (85). This process is however very time-consuming, and in cases where one wishes to reconstruct the GEMs of multiple organisms it becomes impractical to scale up. For this purpose, there exists tools which can automatically reconstruct GEMs that can perform close to manually curated GEMs. CarveMe is one such tool (84).

Similar to the method described earlier, the development of CarveMe also involved manually curating a draft model, but instead of using a draft for the metabolism of a single organism, a draft model was created from every reaction present in the BiGG database (84, 86). This universal model was developed to be a fully functional representation of bacterial metabolism on a whole, through manual curation in a similar way as the first method. However, with the developed universal model, CarveMe automates the reconstruction of organism-specific GEMs by "carving" them out of the already curated universal model. This process involves matching the genome annotation of the specific organism with the reactions present in the universal model. Additionally, if certain reactions appear missing in the organism GEM they can be added from the universal model using a scoring system (84).

Flux Balance Analysis (FBA)

Flux balance analysis (FBA) is one of many methods which can be used to analyze GEMs. It is performed by calculating the metabolic fluxes through a given metabolic network that maximizes or minimizes a pre-defined objective in a given environment (87). Maximization of the growth rate is the most common objective, based on the assumption that bacteria has evolved towards optimizing growth (88). Each reaction in the stoichiometric matrix of the GEM is given a flux, describing the rate at which metabolites are moving through it. An assumption of FBA is that the metabolic network is in a steady state, where the mass of each of the metabolites going into the network is balanced with the mass of metabolites going out of the network, meaning that no metabolite accumulates or depletes. The justification for this is that the metabolic processes represented in the metabolic network are fast enough compared to other cellular processes like transcription and translation, that on a large enough time-scale FBA will represent a good approximation of an organism's metabolic phenotype (87, 89).

Through this steady-state assumption, the range of the reaction fluxes are constrained so that there is a finite range of flux values that FBA can arrive at that fulfills the assumption. Further constraints can be arrived at by restricting the lower and upper bounds of the reaction fluxes, which can be done for reactions with known ranges (such as ATP maintenance), reactions that are known to be irreversible in a certain direction under intracellular conditions and reactions which describe the exchange of resources from the environment. From these constraints we get what is commonly called the solution space of the GEM. Growth is maximized by using a reaction representing the associated metabolic cost of the organism's growth and finding the point within the solution space where this reaction has its highest possible flux value. Likewise, the maximum flux value of any other reaction in the GEM can be identified through FBA. By analyzing the ranges of different flux values, and the relationship between growth and different reactions, we can get a better understanding of an organism's metabolic phenotype and its range of feasible states when subjected to environmental pressures (87, 89).

2.7 Data Models for Molecules and Reactions

The digital modeling of molecular structures and reactions involves the use of specific forms of molecular notation as well as software which can process them. Examples of commonly used notations for these purposes are the Simplified Molecular-Input Line-Entry System (SMILES) and the SMILES Arbitrary Target Specification (SMARTS) (90, 91). Both involve the description of a molecular structure, but are used in different situations. A commonly used software for processing models created using these notations is RDKit, which has a wide array of cheminformatic functionalities. Among many other functions, RDKit can be used to model and

visualize molecular structures, search for substructures within larger molecules, simulate reactions and assess the chemical similarity between two molecules (92). Molecular similarity is computed using the Tanimoto similarity coefficient, which calculates the amount of shared features between two structures in a range from 0 to 1, where the higher the value the more similar the two structures are (93, 94).

SMILES is a human writeable/readable line notation commonly used to give a precise description of a molecular structure. Text strings for the notation are written using the standard ASCII text format, and are structured in such a way as to be more easily understandable by a human reader. As an example, the SMILES string for methane can be written as [CH4], and for ethanol it can be [CH3][CH2][OH]. The system built around SMILES does allow the input to be less specific, such as simply writing [C][C][O], as it infers the presence of hydrogen. It also does not require stereochemistry to be specified, but allows it through use of certain ASCII symbols. RDKit uses SMILES as an input to construct stereochemically specific molecular models (90).

SMARTS builds upon SMILES, and is a notation commonly used for describing molecular structures in a more ambiguous way, such as describing structural patterns. With SMARTS one could describe a molecule where the first atom is carbon, the second is any atom, and the third is either nitrogen or oxygen with the code “[C:1][*:2][N,O:3]”. Additionally, these ambiguous molecules can be used to represent patterns which are present within specific molecules, and so SMARTS derived molecular objects can be used for substructure searching. In the case of RDKit, substructure searching is done using the molecular model created from SMARTS as a search query for patterns in other molecular models (91).

There is a variant of the SMARTS notation called reaction SMARTS, which is used to describe chemical reactions. It does so by expressing reactions using multiple SMARTS queries in one string of text. As an example, if one wished to create a representation of the reaction of methane to methanol using reaction SMARTS, one would write “[CH4:1] >> [CH3:1][OH]”. Likewise, one could use more ambiguous SMARTS queries to represent more general reactions such as “[C:1] >> [C:1][OH]”, which not only represents the reaction of methane to methanol but the reaction of any carbon to a carbon with a hydroxyl group. Keep in mind that since the valence of the carbon in the reaction is not specified, this reaction can also apply to any carbon moiety of a larger molecule as long as it does not conflict with the valence. Similarly to substructure searching, reaction modeling in RDKit leverages the reaction SMARTS notation (91).

3 Methods

This section is organized according to the sequence that these methods were used to ultimately create and analyze GEMs with RiPP biosynthesis pathways. Hence, the first section describes how RiPP datafiles were obtained from antiSMASH and MIBiG. Then I describe the python program I developed to convert these datafiles into metabolic pathways, named ARMRiPP. Finally I describe how I analyzed GEMs extended with the reconstructed RiPP biosynthesis pathways using additional Python and R scripts. The description of the ARMRiPP script is accompanied by a pseudocode while the actual code is publicly available online at https://github.com/AlmaasLab/BiGMeC/tree/Adrian.

3.1 Data Gathering

RiPP data was gathered from the MIBiG repository in the form of antiSMASH annotated genbank files. Although MIBiG has functionality for mass-downloading annotated BGC genbank files, the latest batch available at time of writing was not annotated with the newest antiSMASH version, while the individual genbank files for each BGC in the repository were. Using a Python script, the folder of the old version antiSMASH annotated genbank files was iterated through and a list of every BGC ID which had at least one RiPP-associated gene in it was created. Every BGC ID present in this list was then downloaded by individually accessing them from the MIBiG repository and downloading them. These were annotated using antiSMASH version 6. The folder of these genbank files can be found online at https://github.com/AlmaasLab/BiGMeC/tree/Adrian.

3.2 Description of ARMRiPP

Automatic Reconstruction of Metabolic pathways in RiPPs (ARMRiPP) is a Python script that was developed within this thesis to achieve the aim of reconstructing RiPP pathways. From a general overview, the ARMRiPP script works by taking in antiSMASH annotated genbank files of RiPPs, determining the pathways of their compounds, and returning the pathways in a metabolic model format along with a summary file (Figure 3.2.1). The ARMRiPP pathway reconstruction can be divided into three parts: gathering information from the genbank file, modeling reactions using the RDKit package and constructing metabolic pathways using the COBRApy package. A section will be devoted to each, with a final section giving an overview of the ARMRiPP script functions, structure and pseudocode.

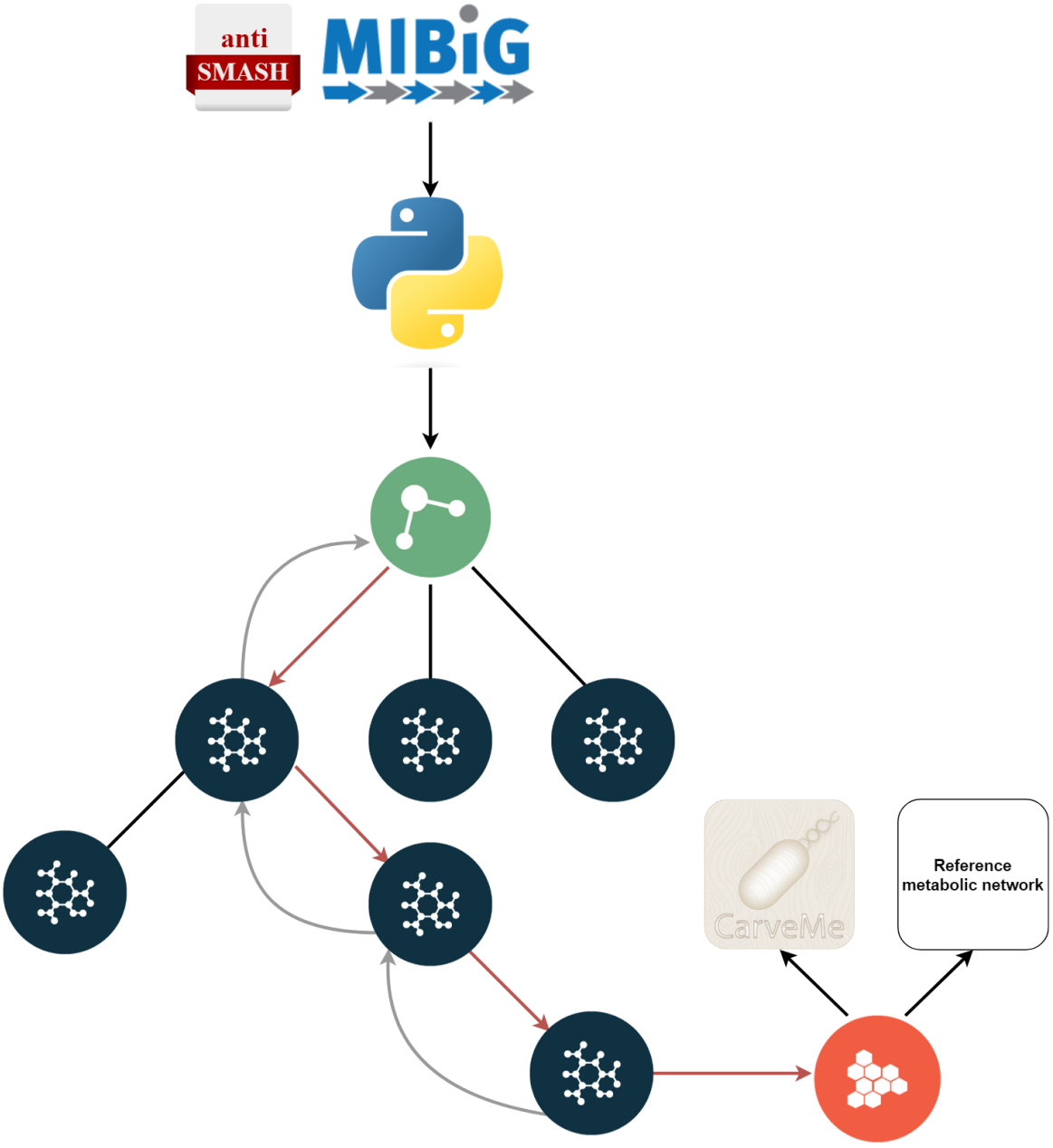


Figure 3.2.1: Overview of ARMRiPP. Annotated genbank files for RiPPs are accessed from MIBiG and read by the ARMRiPP Python code. After the reaction environment of the RiPP is determined, the precursor peptide is identified (green circle). Reaction modeling and the reaction environment data is then used to guide the pathway construction from initial metabolite to mature product (blue circles). The finished pathway (red circle) is then saved as a metabolic model, which can be used to add the pathway to GEMs. Figures of precursor peptide, intermediates and final pathway are adapted from (95).

BGC Data Parsing

During the development of ARMRiPP, the various antiSMASH annotations present in the RiPP genbank files acquired from MIBiG were investigated. The motivation for this was in large part to associate the functional annotations with specific enzymes or enzyme groups to determine whether a specific reaction in the RiPP pathway can take place.

When reading the RiPP genbank files, ARMRiPP creates a temporary set of data items referred to as the BGC dictionary. As these items are important for determining both reaction modeling and pathway construction, an overview of the items in the BGC dictionary will be detailed here. One set of these items are various useful bits of information describing the BGC and its cores, such as the MIBiG ID, source organism, BGC name, core number, core name, predicted core class, and predicted core subclass. This information is accessed when parsing through the BGC genbank file, and for each core in the BGC the information is added to a new row in the summary csv file. For the data analysis, this information is used for grouping data, particularly of the end product structure results. Additionally, the predicted core class and subclass information is used in part to guide the pathway construction.

A second set of items in the BGC dictionary consists of information relating to the reaction environment. Among these items are various smCOG codes, relevant biosynthetic gene names, and other shorthand names for functional annotations from antiSMASH. Initially, all of these items are given a “False” value in the initial dictionary. Then, when parsing through the BGC genes in the genbank file, they are given a “True” value if they are present in the antiSMASH annotations. The ARMRiPP script will check every gene present in the cluster as determined by antiSMASH. Therefore a generalized assumption was made that as long as the identified functional annotation of the genes in the cluster is associated with reactions of a specific precursor RiPP class or subclass, they will allow those reactions to happen. To clarify, the code does not take into account leader-peptide specificity from RREs or genes from different merged clusters, and instead presumes any gene in the cluster with the appropriate functional annotation will enable its reaction for precursors of the same class or subclass. An overview of these reaction environment items can be seen in Table 3.2.1. In addition to these items, the BGC dictionary also includes the core and leader peptide sequence information, which is assigned when analyzing the core in the ARMRiPP reconstruction script.

Certain conditions had to be met for the item presence to be set as “True”, and these conditions were determined by investigating well characterized clusters and checking their unique function annotations. For the smCOGs, it simply checks if the smCOG code is present in the functional annotation text of any of the genes. For the LanD item, it checks for gene names present in the BGC which follow the correct lanthipeptide nomenclature and has a “D” identifier. For the LanO item, it likewise checks if there is an “O” identifier in the gene names as well as at least one

short-dehydrogenase associated smCOG code (SMCOG1007, SMCOG1001 or SMCOG1251). For the YcaO item, the presence of the "YcaO" functional annotation identifier in any of the BGC genes was checked. For the Thiostrepton item, it checks for the presence of "Thiostrepton" in the functional annotation of the peptide core. This was found to be associated with thiopeptide clusters which have end products with secondary macrocyclization or side-rings.

Table 3.2.1: Overview of items in BGC dictionary which are relevant for the reaction environment, their function and associated reactions.

Item Name	**Function**
SMCOG1155	Indicates the presence of a LanB-like gene in the BGC. Associated with the dehydration by glutamylation-elimination reaction,
SMCOG1030	Indicates the presence of a LanKC-like gene in the BGC. Associated with phosphorylation dehydration reaction and LanC-like thio-ether cyclization.
SMCOG1070	Indicates the presence of a LanM-like gene in the BGC. Associated with phosphorylation dehydration reaction and LanC-like thio-ether cyclization.
SMCOG1140	Indicates the presence of a LanC-like gene in the BGC. Associated with LanC-like thio-ether cyclization.
LanD	Indicates the presence of a LanD-like gene in the BGC. Associated with the oxidative decarboxylation reaction of peptide C-terminal for subsequent AviCys installation in class I lanthipeptides.
LanO	Indicates the presence of a LanO-like gene in the BGC. Associated with the dehydrogenation reaction of peptide N-terminal into a lactate in class I lanthipeptides.
Thiostrepton	Indicates the presence of multiple genes related to secondary macrocyclization and side-ring formation in thiopeptides in the BGC. The specific reaction, thiostrepton secondary macrocyclization or nosiheptide side-ring, is dependent on the assigned thiopeptide type.
YcaO	Indicates the presence of a YcaO-like gene in the BGC. Associated with azole formation reaction in thiopeptides.

The last set of items in the BGC dictionary all relate to the results of the reconstruction, and like the first set is added to a row on the summary for each core the ARMRiPP runs on. These items include the BGC initial product SMILES, the BGC end product SMILES, a list of the metabolite IDs for each metabolite in the reconstructed pathway and the initial product and end product matching scores.

Reaction Modeling

In this thesis, reaction modeling refers to the modeling of RiPP metabolite molecular structures and reactions. It was performed using functionalities from the RDKit Python package, version 2023.03.1 (Table 3.2.2) (96). To create the molecular model of the initial substrate present in every RiPP pathway, the RDKit function called *MolFromSequence* was used. The *MolFromSequence* function takes in the peptide sequence for the core peptide of the RiPP BGC, and constructs a model of its chemical structure in an RDKit molecule object. This molecule object is then used in later reaction modeling. Constructing a model only for the RiPP core peptide is done for two reasons: the first is due to the theory based assumption that all modification reactions happen within the core peptide (18), and the second is that it simplifies later reaction modeling as the leader peptide area would have to be continuously protected from participating in reactions. This does however mean that the molecular models for the intermediate metabolites are technically inaccurate to their real world structures, but as every RiPP compound has its leader region cleaved before maturation, the final metabolite structure should be unaffected by this.

To compare the final metabolite structure with its known chemical structure, RDKit molecular objects of both were used to generate Tanimoto similarity scores. To generate RDKit molecule objects of the known RiPP compound structures, their SMILES notation were generally accessed from the same MIBiG page as their BGC ID indicates, which contains structural information from PubChem, while some were directly accessed from PubChem using their RiPP compound name either due to errors in the structures present in MIBiG or due to the structures not being present.

Reaction modeling in RDKit was performed by using the reaction SMARTS notation. Reaction SMARTS representations were created for every reaction to be modeled, using literature as a guide. In ARMRiPP, the *RunReactants* function was used to generate molecular models for the product of the reactions in each reconstructed RiPP pathway. Specifically, only the modification reactions of the core peptide were modeled in this way, and the modeling was done only using SMARTS notation of the reacting motif of the core peptide reactant and the product motif. Additionally, transient reaction intermediates were not modeled, and so the only required structures to be used in the reaction SMARTS were the initial reacting motif pre-reaction and the final motif post-reaction.

Not every modification reaction is possible even if the necessary enzymes are present, as some RiPP substrates may be missing the necessary reacting motifs. If *RunReactants* is run on these substrates it produces an empty list. To ensure that every modeled reaction will have products, as it is necessary for the pathway construction, RDKit functions were used to check for the presence of the reacting motifs required for the reactions to happen. *HasSubstructMatch* was used in cases

where the reaction requires at least one of the reacting motifs, and *GetSubstructMatches* was used if a larger number of a certain motif was required to be present.

Table 3.2.2: Overview of RDKit objects and functions and their associated utility.

RDKit objects	**Utility**
Molecule	Contains the information about a molecule's structure. Every non-hydrogen atom in the structure is indexed with a unique integer. RDKit molecule objects may be ambiguous, meaning they do not have a completely defined structure, or non-ambiguous, meaning they do.
Reaction	Contains information about a chemical reaction.
RDKit functions	**Utility**
MolFromSequence	Uses peptide sequence information to generate a molecular model of the polypeptide structure as an RDKit molecule object.
MolFromSmiles	Uses SMILES notation to generate a molecular model of the described structure as an RDKit molecule object.
MolFromSmarts	Uses SMARTS notation to generate an ambiguous molecular model of the described structure as an RDKit molecule object.
RDKFingerprint	Generates molecular fingerprint data for an RDKit molecule object.
FingerprintSimilarity	Compares the molecular fingerprint data of two molecules and generates a Tanimoto score of their similarity.
HasSubstructMatch	Uses an ambiguous RDKit molecule object as a search query for a non-ambiguous RDKit molecule object. Returns "True" if the molecular pattern in the query is present, returns "False" otherwise.
GetSubstructMatches	Uses an ambiguous RDKit molecule object as a search query for a non-ambiguous RDKit molecule object. Returns a list of the index positions for each occurrence of the query molecular pattern in the target molecule.The length of this list indicates the amount of matches that the search query has in the molecule.
ReactionFromSmarts	Uses reaction SMARTS notation to generate an RDKit reaction object.
RunReactants	Takes in RDKit molecular objects as substrates, and uses the information contained in an RDKit reaction object to generate molecular models of the reaction products as new RDKit molecular objects. If the reaction can occur multiple places in the substrates, it will generate sets of products for each single possible reaction.

As is apparent in the theory section for the biosynthesis of the different RiPP classes, there are reactions that technically could result in several possible products, but RiPP modification enzymes will generally heavily favor only one of these products per step of the pathway (33, 35, 61). Likewise, the *RunReactants* function will return every feasible product from the reaction SMARTS, but will not favor one over the other. To emulate the product selectivity done by the RiPP enzymes, several reaction specific rules were created for selecting the correct molecule objects of the many potential products generated by *RunReactants*. Creating these rules involved assumptions based on the literature as well as analysis of well characterized RiPP compounds. These rules are explained in detail for every RiPP class in the sections below.

Some general assumptions were made for the reaction modeling that applies for every considered RiPP compound. If the identified core peptide sequence in the RiPP cluster contains unknown amino acids, X, then the code replaces these with G for glycine. The altered code is then used for the molecular model of the initial RiPP substrate. The choice of glycine was due to it not participating in any of the reactions. Additionally, stereochemistry and stereoselection was not considered during reaction modeling, and as such any generated structures will have the standardized stereochemistry that RDKit automatically assigns when generating molecules. Since the focus of the thesis is on the metabolic stoichiometry of these pathways, stereochemistry is irrelevant as none of the considered reactions are dependent on a specific previous stereochemistry in the metabolite. Furthermore, calculating the Tanimoto coefficient for molecular similarity is unaffected by stereochemistry.

Lanthipeptides

For lanthipeptides, reaction modeling was done for the major biosynthetic events present for all lanthipeptides, Ser/Thr dehydration and thio-ether cyclization, as well as for two common class I lanthipeptide tailoring reactions. As Ser/Thr dehydration is performed by two different mechanisms, glutamylation elimination for class I and phosphorylation for class II-IV lanthipeptides. However, as the core peptide reactant and product is the same in both mechanisms, the reaction modeling SMARTS notation was the same for both reactions. Based on observations from literature that most serine and threonine residues in the RiPP precursor core tend to be dehydrated (32), the generalized assumption was made that these two reactions fully dehydrate the RiPP core peptide. Due to this the reaction modeling for both phosphorylation and glutamylation elimination was made so that the reaction would repeat on the substrate molecular model until every Ser/Thr residue had been dehydrated into Dha/Dhb, and then output the final fully dehydrated molecular model as the product. This was achieved by looping the *RunReactants* until only one product was present in the output.

Thio-ether cyclization had to be modeled in a more complex way. The reaction was modeled around the following general assumptions: that thio-ether cyclization proceeds until there are either no cysteine or no Dha/Dhb residues left, that there is a certain minimum ring size for the

cyclization, and that it occurs with a certain directionality. Similarly to the dehydration reactions, the first assumption meant that the reaction would have to repeat the thio-ether reaction step until no further products could be made. However, given the second and third assumptions these repeated reaction steps would have to be guided in order to arrive at the best fitting product. This was achieved by checking each reaction product and picking the one which best followed the assumptions by using the RDKit substructure searching. During development it was found that the directionality of the thio-ether cyclization steps appeared to fit well with a C-N terminus processing, and so specifically the second assumption refers to this. The third assumption of minimal ring size was achieved by using the molecular indexing of the RDKit molecule objects for substructures, and creating an average value for their position. During development a value of 9 was arrived at for the spacing of the reacting motifs in the thio-ether reaction. In addition to the lanthionine thio-ether cyclization reaction, labionin formation for class III lanthipeptides was also modeled. It was done using the same logic as with lanthionine, but checking whether the specific chemical motifs for the labionin reaction is present, meaning both a second and third Dhx residue to the N-terminus of the reacting Cys residue. If present, a reaction with separate SMARTS notation specific for labionin is used, otherwise the lanthionine reaction notation is used.

The last two lanthipeptide reactions modeled were of two class I specific tailoring reactions, referred to as LanO and LanD tailoring. LanO tailoring involves the reduction of an N-terminus Dha residue in the precursor core into a lactate group in a single reaction step. As such, the reaction modeling for it is simply to check if there is an N-terminus Dha in the molecular model of the substrate, and the reaction function then uses the notation for turning the Dha motif into lactate to generate the single product. LanD tailoring is responsible for the installation of AviCys, but its associated reaction only involves one step, which is oxidative decarboxylation of a C-terminus cysteine residue in the precursor core. Similarly to LanO, the correct checks are made and a single product is generated. However, AviCys formation is then performed by the thio-ether reaction function using the decarboxylated cysteine residue as a reacting motif.

Thiopeptides

Of the central biosynthesis in thiopeptides, three major reaction events are present: azole formation, Ser/Thr dehydration and macrocyclization. According to literature, oxazoles are uncommon in thiopeptides and require the presence of specific and poorly understood dehydrogenation enzymes (33). It was also difficult to determine their presence by the antiSMASH annotation and to determine which serines/threonines get dehydrated and which form oxazoles. It was therefore decided that for the first event, azole formation, only the thiazole reaction was to be modeled. Furthermore, a general assumption that every cysteine in the thiopeptide core reacts into thiazoles was made, as observed in literature (33, 75). An exception to this is when the N-terminal amino acid is a cysteine, as azole formation from cysteine requires an N-terminal flanking amino acid. This was however taken into account in the SMARTS

reaction notation. The azole forming reaction was then modeled by a single reaction step for thiazole formation, with SMARTS notation of cysteine into thiazole, for every cysteine present in the core peptide molecular model. For the dehydration reaction the same reaction modeling as in lanthipeptides for glutamylation elimination was used.

Thiopeptide macrocyclization reaction modeling involved determining the 2π- and 4π-components of the aza-Diels Alder reaction, performing the reaction and subsequent heterocycle modification reaction. Based on literature, the assumption was made that the 2π-component is the furthest N-terminal Dha residue in the thiopeptide precursor core, while the 4π-component is any Dha flanked by both a C-terminal and N-terminal thiazole group (33). In the case of multiple possible 4π-components, the one furthest to the C-terminal would be chosen. To follow these assumptions, the reaction modeling for thiopeptide macrocyclization was made with specific SMARTS notation describing the 2π- and 4π-components. After the reaction the product is picked by checking through the potential products and using substructure searching to determine if it was correct.

Immediately following the macrocyclization reaction, the heterocycle modification reaction occurs. For the modeling of the heterocycle modification reaction, the thiopeptide subclass information from the BGC dictionary is used to determine which modification reaction is performed. For type I thiopeptides, the series e modification is modeled, for type III thiopeptides the series d is modeled, while for type II thiopeptides only the initial reaction step for series b thiopeptides is modeled. The SMARTS notation for each of these reactions reflects the mechanism described in the theory section, and each of them consist of a single step reaction with only one possible product (75).

Outside of the central thiopeptide biosynthetic reactions, the two types of secondary macrocyclization present in nosiheptide-like and thiostrepton-like compounds were also modeled. In each of them the pathways for the addition-products were modeled only stoichiometrically with no RDKit molecular models based on the assumption that these pathways are the same in every case. The structural reaction modeling therefore only takes place in the step where these addition-products react with the thiopeptide precursor core.

In the case of the nosiheptide side-ring reaction, the reaction between the free methyl-indolic acid (MIA) molecule and the cysteine residue in the thiopeptide core, and subsequently the reaction between the MIA moiety and a glutamate residue in the core peptide, was modeled. The assumptions around which cysteine and glutamate residue is reacted with is based on the conventional nosiheptide pathway described in the theory section (33, 76), and so we presume that cysteine is close to the 8th position in the peptide sequence and the glutamate is close to the sixth. In each reaction step, products from the RDKit reaction are picked by using substructure

searching and the assumptions. If no residue is found at the exact position, it instead checks for the closest one.

Likewise for the thiostrepton secondary macrocyclization reaction, the reaction between a free quinaldic acid (QA) molecule and a threonine residue in the thiopeptide core, as well as the following reaction between the QA moiety and core peptide N-terminus, was modeled. The threonine assumption was that the reaction would be close to the conventional thiostrepton reaction's threonine placement, with a threonine residue close to the 12th position (33, 77). Products were picked from the initial RDKit reaction to check for this, and the last step would only produce one possible product.

Lasso Peptides

Only the two central reactions of the lasso peptide biosynthetic pathway were modeled, macrolactam formation and disulfide bridging. The macrolactam reaction was modeled around the assumption that lasso peptides generally only contain aspartate/glutamate residues around the area where the macrolactam formation takes place, which was made by observing lasso peptide core sequences. In the case of multiple Asp/Glu residues, the largest one is to be picked. In the macrolactam reaction modeling this is done by using RDKit functions for checking the ring sizes, and picking the product molecular model with the largest. For the disulfide bridge reaction function the general assumption was made that for class III and IV lasso peptides there are always two cysteine residues in the core peptide, and for class I there are always four. In the case of class III and IV lasso peptides the reaction modeled is simply the binding of the two cysteine sulfur moieties, using SMARTS notation, and producing one product through the RDKit reaction functionality. For class I lasso peptides, an assumption was made based on its structure known from literature, where the first cysteine in the peptide sequence is bound with the third and the second with the fourth, in N-C direction (35). Code similar to the thio-ether cyclization reaction function was made to choose the correct combination of cysteine residues from the products generated by the same RDKit reaction performed for the class III and IV lasso peptides.

Pathway Construction

In the ARMRiPP script, pathway construction is achieved by using functionality from the COBRApy package (Table 3.2.3) (97). COBRApy metabolite objects for metabolite cofactors and by-products were gained by using their BiGG metabolite ids, and copied from the Sco-GEM. Custom metabolite objects were created for the central RiPP peptide metabolites. The stoichiometry of RiPP modification reactions was described using COBRApy reaction objects, which describe the consumption and production of the different metabolite objects. Using a chain of these reaction objects, entire RiPP pathways can be added to metabolic models. In the case of ARMRiPP, these COBRApy reaction objects were stringed together guided by both the reaction

modeling and the reaction environment in the BGC dictionary, and each reaction was then saved together in the format of a metabolic model.

Table 3.2.3: Overview of COBRApy objects and descriptions of their utility in the ARMRiPP script.

COBRApy object	**Utility**
Metabolite	Contains information on a metabolite's name and chemical formula.
Reaction	Contains information on reaction name, minimum and maximum fluxes as well as the metabolite stoichiometry of the reaction. It takes in metabolite objects either as reactants or products.
Model	Represents a metabolic model. Contains information of all the reactions present in the model and their associated genes.

Every RiPP pathway has the ribosomal polypeptide synthesis of its precursor as its first pathway step. A COBRApy reaction is created for this translation reaction, and using a dictionary containing the BiGG metabolite IDs associated with each amino acid letter code, every amino acid present in the leader and core peptide sequence is added to the reaction object as reactants. Based on the theory on ribosomal polypeptide synthesis, it was presumed that for each amino acid two GTPs and one ATP would be consumed producing two GDPs, two phosphates, one AMP and two pyrophosphates (98). The amount of peptide bonds in the polypeptide is equivalent to the precursor peptide sequence length minus one, and so this number determines the stoichiometry of the cobrapy reaction describing the assembly of the core peptide (Reaction 1). Note that the amino acid metabolites on their own are used in this reaction, instead of their tRNA-activated form which in reality participates in the reaction. However, the net cost with the aminoacyl activation of amino acids onto their tRNA-carrier is taken into account with the ATP → AMP part of the reaction. The reasoning for this was to allow the reactions to take place in metabolic models where tRNA aminoacyl activation reactions are not present.

$$2n\ GTP + n\ ATP + n\ AA \rightarrow RiPP\ precursor_n + 2n\ GDP + n\ AMP + 2n\ P_i + n\ PP_i + (n - 1)\ H_2O \qquad (1)$$

For FBA purposes, a demand reaction was created for the final RiPP metabolite (Reaction 2). This reaction simply takes in whichever RiPP metabolite was produced in the last pathway reaction, and irreversibly produces nothing. The assumption behind this was to emulate the practical consumption ceiling of the RiPP compounds. Every RiPP pathway will have this reaction as its last.

$$Mature\ RiPP\ metabolite \rightarrow \quad (2)$$

Lanthipeptide Pathways

Although the two different types of lanthipeptide dehydration reactions were modeled with the same SMARTS notation in the reaction modeling, they have different reaction stoichiometry and therefore separate COBRApy reactions were created for each in the ARMRiPP script. Based on the theory, the glutamylation elimination reaction object was created to consume one glutamate activated $tRNA^{Glu}$ and produce one free glutamate and one $tRNA^{Glu}$ for each serine or threonine dehydration in the RiPP precursor core. For phosphorylation the reaction object was created to consume one ATP and produce one ADP and one phosphate for each serine or threonine dehydration. In both cases the amount of dehydration reactions was calculated by counting the amount of times the RDKit reaction function had to run on the substrate molecular model. Furthermore, even though some class IV lanthipeptides are known for having dehydration enzymes which can allow multiple nucleoside triphosphates in the reaction, it was instead assumed that every phosphorylation reaction uses ATP. The reaction stoichiometry for the total reaction of glutamylation elimination (3) and phosphorylation (4) can be seen below.

$$RiPP\ core + n\ Glu\text{-}tRNA^{Glu} \rightarrow Dehydrated\ RiPP\ core + n\ Glu + n\ tRNA^{Glu} \quad (3)$$

$$RiPP\ core + n\ ATP \rightarrow Dehydrated\ RiPP\ core + n\ ADP + n\ P_i \quad (4)$$

The thio-ether cyclization was found to be metabolically balanced. The water molecule and hydrogen which is displaced by the initial binding of the LanC are both assumed from literature to be regenerated after the reaction, and so the total reaction stoichiometry of the thio-ether reaction object is simply the consumption of RiPP metabolites to produce thio-ether cyclic RiPP metabolites (Reaction 5).

$$Dehydrated\ RiPP\ core \rightarrow Cyclic\ RiPP\ metabolite \quad (5)$$

The remaining two class I lanthipeptide tailoring reactions are both single step reactions, and therefore have the same stoichiometry in every case. For the first reaction, LanO N-terminus reduction, one molecule of water displaces one molecule of ammonium from the RiPP core peptide N-terminus Dha forming a pyruvate group, and then NADPH reduces the pyruvate group into lactate (Reaction 6).

$$RiPP\ metabolite + H_2O + NADPH \rightarrow LanO\ tailored\ RiPP\ metabolite + NH_4^+ + NADP^+ \quad (6)$$

For the second reaction, LanD C-terminus decarboxylation, one molecule of FMN is reduced while cleaving off one molecule of CO_2 from the RiPP core peptide C-terminus (Reaction 7). Reaction objects were created for both reactions, however the later AviCys forming reaction of the LanD tailored C-terminus is performed by the same mechanism as any other LanC reaction, and therefore is not considered in its own reaction object.

$$RiPP\ metabolite + FMN \rightarrow LanD\ tailored\ RiPP\ metabolite + FMNH_2 + CO_2 \quad (7)$$

Thiopeptide Pathways

For the first event in the central thiopeptide biosynthesis, azole formation, a COBRApy reaction was created for the total reaction event. The reaction object was created so that for each non N-terminal cysteine in the thiopeptide core, one ATP and one FMN is consumed and one water molecule, one ADP, one phosphate and one $FMNH_2$ is produced (Reaction 8). To determine the amount of azoles, the loops of the azole formation RDKit reaction function were counted. Since oxazole formation was not implemented in the reaction modeling, only thiazole formation was taken into account.

$$RiPP\ metabolite + ATP + FMN \rightarrow Azole\text{-}containing\ RiPP\ metabolite + H_2O + ADP + P_i + FMNH_2 \quad (8)$$

The second event, dehydration by glutamylation elimination, follows the exact same mechanism and stoichiometry as it does for lanthipeptides, and so the same COBRApy reaction object was repurposed (Reaction 3)

In the final event, COBRApy reactions were created for both the macrocyclization reaction and for each central nitrogen heterocycle modification reaction. The macrocyclization reaction is the same for all thiopeptides, is single step and produces one molecule of water and the macrocyclic thiopeptide intermediate (Reaction 9). Each of the heterocycle modification reactions are specific to different thiopeptide series classifications, and COBRApy reactions were created for the series d, e and b modification reactions (Reactions 10-12). Since the exact mechanism of these modification reactions is not well understood, reactants and products were inferred from the studied structures of the different thiopeptide series central heterocycles (75). For the two oxidation reactions, a proton product was presumed for the series d and a water reactant and proton product was presumed for series e (Reaction 11 and 12). Although series e follows series d modification in literature (75), the reaction modeling was done for the total reaction from the macrocyclic intermediate to the series e thiopeptide, and so the COBRApy reaction also was created for the total reaction. For the reduction reaction, series b, a proton reactant was presumed (Reaction 12). As explained in the reaction modeling section for these reactions, the subsequent oxidation reactions for series a and c thiopeptides were not modeled, and therefore COBRApy reaction objects were not created for them.

$$\textit{RiPP metabolite} \rightarrow \textit{Macrocyclic intermediate RiPP metabolite} + H_2O \qquad (9)$$

$$\textit{Macrocyclic intermediate RiPP metabolite} \rightarrow \textit{Series d thiopeptide} + H^+ \qquad (10)$$

$$\textit{Macrocyclic intermediate RiPP metabolite} + H_2O \rightarrow \textit{Series e thiopeptide} + H^+ \qquad (11)$$

$$\textit{Macrocyclic intermediate RiPP metabolite} + H^+ \rightarrow \textit{Series b thiopeptide} \qquad (12)$$

The secondary macrocyclization present in the thiostrepton-like and nosiheptide-like groups of thiopeptides both required multiple COBRApy reaction objects for: the reaction pathway of the addition reactant, the addition reaction and the secondary macrocyclization reaction. Starting with the thiostrepton-like quinaldic acid (QA) pathway, L-tryptophan is first given a methyl group through a radical SAM methyltransferase reaction, consuming SAM and producing SAH (Reaction 13). This creates a QA-intermediate, which subsequently loses an amine group as ammonium (Reaction 14). The QA-intermediate then undergoes a conjugation addition reaction which consumes oxygen, oxidizes $FADH_2$ and produces water (Reaction 15). Following this, the QA addition reactant is produced by a reduction reaction of the previous QA-intermediate, which oxidizes NADPH (Reaction 16). The finished QA then reacts in an addition reaction with threonine in the thiopeptide core, consuming ATP and producing water (Reaction 17). The subsequent intramolecular reaction between the thiopeptide core QA-moiety and its N-terminus consumes oxygen and oxidizes NADPH (Reaction 18). The cofactors were assumed using the literature structures and known mechanisms of the thiostrepton pathway (33).

$$Trp + SAM \rightarrow QA\text{-}intermediate + SAH \quad (13)$$

$$QA\text{-}intermediate \rightarrow QA\text{-}intermediate + NH_4^+ \quad (14)$$

$$QA\text{-}intermediate + FADH_2 + O_2 \rightarrow QA\text{-}intermediate + FAD + H_2O \quad (15)$$

$$QA\text{-}intermediate + NADPH \rightarrow QA + NADP^+ \quad (16)$$

$$RiPP\ metabolite + QA + ATP \rightarrow RiPP\ metabolite\ (QA) + ADP + P_i \quad (17)$$

$$RiPP\ metabolite\ (QA) + NADPH + O_2 \rightarrow Secondary\ macrocyclic\ RiPP\ metabolite + NADP^+ \quad (18)$$

For the nosiheptide-like methylindolic acid (MIA) pathway, L-tryptophan first undergoes a radical SAM reaction into MIA (Reaction 19). The cofactors here were not assigned due to lacking mechanistic information from the literature (33, 76). In the next reaction step MIA is attached to a carrier protein, which consumes ATP and produces ADP, phosphate and water (Reaction 20). The MIA is then attached to a cysteine in the thiopeptide core (Reaction 21), and it was presumed from theory that this reaction consumes ATP as well (76). After MIA is attached to the thiopeptide core it is given a methyl group through a radical SAM methyltransferase reaction (Reaction 22), and spontaneously reacts with glutamate to finish the secondary macrocyclization. The cofactors were assumed using the literature structures and known mechanisms of the nosiheptide pathway (33).

$$Trp \rightarrow MIA \quad (19)$$

$$MIA + ATP \rightarrow Protein\text{-}MIA + ADP + P_i + H_2O \quad (20)$$

$$Protein\text{-}MIA + ATP + RiPP\ metabolite \rightarrow RiPP\ metabolite\ (MIA) + ADP + P_i \quad (21)$$

$$RiPP\ metabolite\ (MIA) + SAM \rightarrow Secondary\ macrocyclic\ RiPP\ metabolite + SAH \quad (22)$$

Lasso Peptide Pathways

Although not covered in the reaction modeling section, the initial reaction for the lasso peptide central biosynthesis is the proteolysis of the lasso peptide leader, and so a COBRApy reaction was created for it. From literature this reaction is known to consume ATP and produce ADP and phosphate (80), but within the thesis it was presumed to require water as well (Reaction 23). The presumption was made due to water generally being required to break peptide bonds. Additionally, only the RiPP core was considered as a product, since no further reaction modeling was done or pathways were created for the RiPP leader by-product.

$$RiPP\ precursor + ATP + H_2O \rightarrow RiPP\ core + ADP + P_i \quad (23)$$

The next central lasso peptide reaction is macrolactam formation, and the COBRApy reaction object created for it directly followed its literature stoichiometry (80). Macrolactam formation consumes one ATP and produces AMP and pyrophosphate (Reaction 24).

$$RiPP\ metabolite + ATP \rightarrow Macrolactam\ intermediate\ RiPP\ metabolite + AMP + PP_i \quad (24)$$

Lastly, a COBRApy reaction was created for the disulfide bridging reaction present in class I, III and IV lasso peptides. It is known from literature that a redox enzyme from the lasso peptide cluster can mediate the disulfide bridging reaction of the lasso peptide core (99), but as this mechanism is not well understood the reaction stoichiometry was instead assumed to be based on the total reaction requirements of polypeptide canonical disulfide bridging (100). The reactants were therefore assumed to be one GSSG and the products two GSH per cysteine pair in the lasso peptide core (Reaction 25).

$$RiPP\ metabolite + n\ GSSG \rightarrow Disulfide\ bridged\ RiPP\ metabolite + 2n\ GSH \quad (25)$$

Overview of ARMRiPP Functions and Pseudocode

Structurally the ARMRiPP script consists of a large number of functions and an object class custom to ARMRiPP. The main functions are listed in Table 3.2.4, and these are used in each RiPP pathway reconstruction. The intermediate RiPP metabolites in the biosynthesis pathway are encoded as *python class* objects (called *Metabolite_node*) and hold metabolite key information such as the COBRApy metabolite object, the RDKit molecular model and the SMILES formula. Each metabolite node also has a parent attribute, which associates each metabolite node with the previous metabolite node in the pathway. Below are explanations of the different ARMRiPP functions and pseudocode related to its general logic.

Table 3.2.4: Overview of central ARMRiPP functions with their associated utility.

ARMRiPP function	**Utility**
run	Main run code, sets up BGC dictionary and summary dataframe. Runs pathway reconstruction for every .gbk file if a folder is used. Returns summary when finished.
_run	Run code for every identified core in the .gbk file.
analyse_bgc	Checks for the presence of RiPP prepeptide cores in .gbk record features, runs pathway reconstruction for each.
parse_gbk	Iterates through .gbk file features and updates BGC dictionary (reaction environment). Partially adapted from the main BiGMec script (17).
model_select	Depending on run mode, either chooses an appropriate CarveMe model associated with the BGC source organism, uses Sco-GEM reference model, or uses an empty model.
add_to_model	Main pathway reconstruction part of script. Chooses model based on mode, creates chain of metabolite nodes using the reaction functions. Iterates through metabolite node chain adding reactions to the model object, returns pathway-extended model to output folder

The pseudocode of ARMRiPP's general logic is as follows:

1. The script calls *run*, which sets up a BGC dictionary for the reaction environment of the annotated BGC genbank file and then calls *analyse_bgc*.
2. The *analyse_bgc* function first calls the *parse_gbk* function, which iterates through the BGC genbank file updating information in the reaction environment dictionary related to presence of enzymes and their functional annotation.
3. The *analyse_bgc* function then updates information in the reaction environment dictionary pertaining to the precursor peptide, such as core and leader peptide sequence, identified class or subclass. Using this updated dictionary, the *_run* function is called. If no core is found the script skips the reconstruction of the BGC, and if multiple cores are found steps 3-9 is repeated for each core.
4. The *_run* function checks the reaction environment dictionary then calls the *add_to_model* function.
5. The *add_to_model* function calls the *model_select* function if specific GEMs are to be used for the BGCs, such as the input models or default reference models. Otherwise, an empty metabolic model is created.
6. The *add_to_model* function initializes the starting metabolite node by using a function which models the translation reaction of the precursor peptide.
7. The *add_to_model* function proceeds with a series of conditional statements which each precedes an associated reaction, creating a new node. The conditions contain checks for the presence of certain enzymes, RiPP classes or functions in the reaction environment dictionary, as well as checking whether the reaction can happen using the chemical model of the metabolite (This process is explained in detail in the Reaction Modeling chapter).
8. Using the final metabolite node object, the *add_to_model* function updates the reaction environment dictionary with information for the summary, and then iterates back through the parents of the node adding the reactions to the metabolic model for each step of the pathway.
9. Returning back to the *run* function, the summary is updated with the information from the BGC dictionary.
10. After every *_run* is complete a .csv file of the summary is created, containing the BGC IDs, core name, source organism, list of metabolites, and end product SMILES structure.

Pathway reconstruction takes place within the *add_to_model* ARMRiPP function, and contains functions that represent the different reaction steps of the pathway. The metabolite node object is used here, passing through these reaction functions and connecting to the next metabolite node of the pathway. The reaction functions are structured in a series of conditional statements which check whether the necessary reacting motifs are present using the information saved in the metabolite, and checking the reaction environment using the BGC dictionary. Explanations of the different ARMRiPP reaction functions (Table 3.2.5) and structure of the reconstruction logic can be found below.

Table 3.2.5: Overview of ARMRiPP objects and reaction functions with their associated utility.

ARMRiPP objects	**Utility**
Metabolite_node	Objects used for associating COBRApy metabolite and reaction objects as well as RDKit molecule object, SMILES string and metabolite name with a given metabolite in the RiPP pathway.
ARMRiPP reaction functions	**Utility**
translate	Function for the translation reaction of the precursor peptide, returns the initial metabolite node.
core_cleave	Function for the proteolysis reaction of core and leader peptides.
glutamylation_elimination	Function that dehydrates serine and threonine per the glutamylation elimination reaction.
phosphorylation_dehydration	Function that dehydrates serine and threonine per the phosphorylation reaction.
cyclization	Function for the thioether crosslinking reaction between serine/threonine and cysteine.
lan_o_tailoring	Function for the LanO tailoring reaction.
lan_d_tailoring	Function for the LanD tailoring reaction.
heterocyclization	Function for the thiazole formation reaction.
macrocyclization	Function for the thiopeptide macrocyclization reaction. Passes metabolite node to *macrocycle_modification*.

macrocycle_modification	Function for the central nitrogen heterocycle modification reaction of thiopeptide macrocyclic intermediates.
thiostrepton_sec_macrocyc_qa_pathway	Function that adds the pathway for the production of quinaldic acid for the thiostrepton-like secondary macrocyclization.
thiostrepton_qa_addition	Function for the addition reaction of quinaldic acid to thiostrepton-like thiopeptide core.
thiostrepton_sec_macrocyclization	Function for the thiostrepton-like secondary macrocyclization.
nosiheptide_sec_macrocyc_mia_pathway	Function that adds the pathway for the production of methylindolic acid for the nosiheptide-like secondary macrocyclization.
nosiheptide_mia_addition	Function for the addition reaction of methylindolic acid to nosiheptide-like thiopeptide core.
nosiheptide_sec_macrocyclization	Function for the nosiheptide-like secondary macrocyclization.
macrolactam_ring_formation	Function for the macrolactam ring formation in lasso peptides
disulfide_bridging	Function for the disulfide bridge reaction in lasso peptides
maturation	Function that passes the final metabolite into one with “mature” in its metabolite id

Below is the structure of the pathway construction part of *add_to_model*:

1. Initialize first metabolite with *translate.*
2. If RiPP core class is thiopeptide and subclass is type II and thiostrepton is true:
    a. Run *thiostrepton_qa_addition* on metabolite, which also runs *thiostrepton_sec_macrocyc_qa_pathway.*
3. If RiPP core class is thiopeptide and subclass is type I and thiostrepton is true:
    a. Run *nosiheptide_mia_addition* on metabolite, which also runs *nosiheptide_sec_macrocyc_mia_pathway.*
    b. Run *nosiheptide_sec_macrocyclization* on metabolite.
4. If cysteine is present and YcaO is true or cysteine is present and core class is thiopeptide:
    a. Run *heterocyclization* on metabolite.
5. If serine/threonine is present and SMCOG1155 is true or serine/threonine is present and core class is thiopeptide:
    a. Run *glutamylation_elimination* on metabolite.
6. If serine/threonine is present and SMCOG1030 is true or serine/threonine is present and SMCOG1070 is true:
    a. Run *phosphorylation_dehydration.*
7. If C-terminal cysteine is present and LanD is true and core class is lanthipeptide:
    a. Run *lan_d_tailoring* on metabolite.
8. If Dha/Dhb and cysteine is present and at least one of the following items is true: SMCOG1140, SMCOG1070, SMCOG1030, core class is thiopeptide and type III:
    a. Run *cyclization* on metabolite.
9. If N-terminal Dha is present and LanO is true and core class is lanthipeptide:
    a. Run *lan_o_tailoring* on metabolite.
10. If 2π- and 4π-components are present and core class is thiopeptide:
    a. Run *macrocyclization* on metabolite.
11. If core class is thiopeptide and subclass is type II and thiostrepton is true:
    a. Run *thiostrepton_sec_macrocyclization* on metabolite.
12. If aspartate or glutamate is present and core class is lasso peptide:
    a. Run *macrolactam_ring_formation* on metabolite.
13. If more than one cysteine is present and core class is lasso peptide:
    a. Run *disulfide_bridging* on metabolite.
14. If core class is lanthipeptide or core class is lasso peptide:
    a. Run *core_cleave* on metabolite.
15. Run *maturation* on metabolite.
16. Create demand reaction for product of *maturation.*

The ARMRiPP script also has different modes that slightly alter the logic and output. The first mode adds an additional step near the end of the *add_to_model* function where the generated SMILES structure of the initial and end product for a BGC is matched with its known structure, given that one is provided. These two resulting Tanimoto scores are then added to two new columns in the summary on the same row as its matched BGC. The model selection in the first mode uses the provided reference Sco-GEM model. This mode was used to generate the data present in the prediction of end product structure results for the 57 BGCs with known product

structures. It was also used on a subset of around half of these 57 BGCs during development, to guide the implementation of reaction functions. In the second mode the code does not output extended GEMs, but instead mini-models containing only the reconstructed pathways. Using auxiliary Python code, the reaction pathway within these mini-models can be used to extend GEMs without requiring ARMRiPP to run every time. For the default mode, the code runs using the associated Sco-GEM reference GEM with no alterations from the pseudocode.

3.3 GEM Reconstruction

CarveMe GEMs

Draft metabolic models for all the species associated with one or more RiPPs analyzed in this work were reconstructed using CarveMe version 1.5.1 with default settings. Taxonomy IDs for each of these species were found in the RiPP GenBank files and assembled into a list that was used to download the NCBI RefSeq accession IDs from NCBI´s Entrez database. CarveMe can use these accession IDs to gain the protein FASTA of the genome of the associated taxonomy id, and provides a GEM reconstruction of as output. This is done from the command line using the function *carve* and inputting a list of the accession ids. However, as not every taxonomy ID has an associated genome protein FASTA file in the NCBI database, not every RiPP BGC source organism will have its GEM reconstructed.

Predicting the Production Yield and Cost of RiPPs

The constructed pathways were inserted into 1) The Sco-GEM model of *S. coelicolor* and 2) reconstructed GEMs of their native hosts to evaluate the reconstructed pathways. The Sco-GEM model was chosen as a reference model because it is a well tested model, and due to *S. coelicolor* being known as an efficient heterologous host for secondary metabolites (85). For both the extended Sco-GEM based models and the reconstructed GEMs, the Sco-GEM complex media present within the Sco-GEM model file was used, which has minimal glucose as its only carbon source. In cases where the exchange reaction was not present in the reconstructed GEM, it would be omitted from the media.

FBA was run using COBRApy version 0.26.0. For each extended GEM, FBA was performed with biomass as the objective function and then RiPP compound production as the objective function. The gradient was calculated by dividing the RiPP compound objective value with the biomass objective value. For ease of interpretation, the gradient had its value sign shifted so that a higher value means a steeper downward slope (Figure 3.3.1). This means that the higher the

gradient value, the higher the RiPP compound production rate relative to a decrease in growth rate, and the lower the metabolic burden of the RiPP compound. The Model Analysis script returns a summary data frame of these results, and can be found at https://github.com/AlmaasLab/BiGMeC/tree/Adrian.

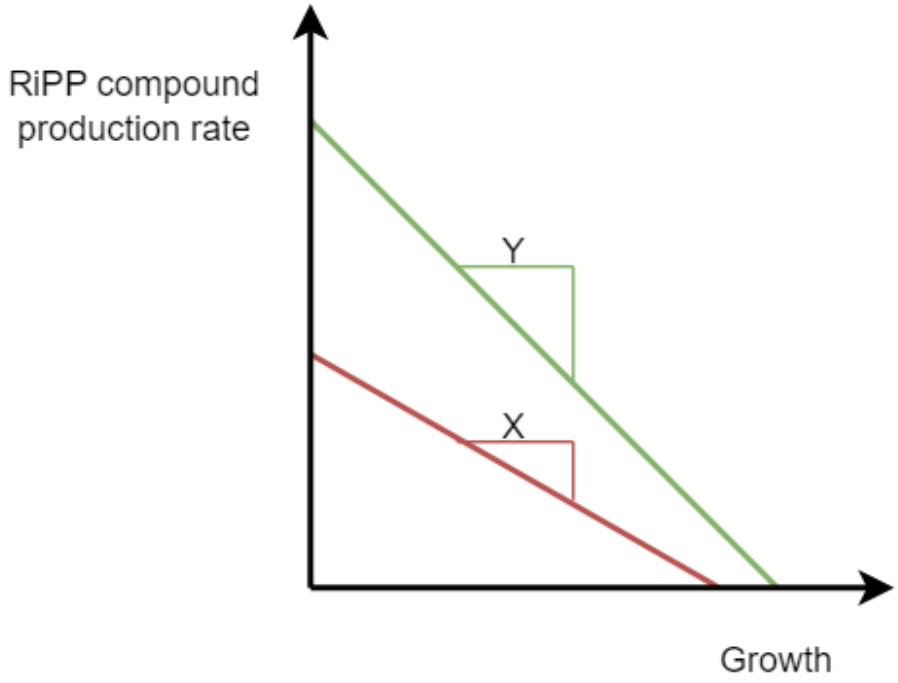


Figure 3.3.1: With RiPP compound production on the y-axis and growth on the x-axis, a steeper downward slope means a lower tradeoff between maximizing growth and maximizing RiPP production, indicating lower metabolic burden. With their signs shifted, this means the gradient value for Y is greater than X, indicating that it has a lower metabolic burden.

3.4 Data Analysis

To map the phylogenetic relationship between the different source organisms of the RiPPs, a common tree was created using NCBI's common tree tool (101). A list of the taxonomy IDs in every genbank file present in the accessed data from MIBiG was used as input for the common tree tool, and a phylogenetic tree for the associated organisms was created in PHYLIP format. The PHYLIP tree was then converted to Newick format and with BioPython's phyloXML module the clades of each organism were mapped out in a data frame. This was done layer by layer from the tree's root, so that each clade level represents a lower group of branches in the tree.

Data analysis and visualization was performed with several R scripts (Table 3.4.1), using the summary files generated from the ARMRiPP and Model Analysis scripts for the various sets of RiPP pathway-extended GEMs, as well as the data frame of organism clades. The t-tests and ANOVA were performed using integrated R functions. All R scripts and associated results data can be found online at https://github.com/AlmaasLab/BiGMeC/tree/Adrian.

Table 3.4.1: Overview of R scripts used for data analysis and their associated utility.

R script	**Utility**
Struct_Prediction.R	Used to generate histograms, calculate means and percentages and filter results in the "Accuracy of End Product Structure Predictions" chapter of the results.
Ref_Exp.R	Used to generate the pie donut chart, boxplot, bar chart and scatter plots in the "Metabolic Burden in Reference GEM" section of the results.
Phy_Exp.R	Used to generate the pie chart, pie donut chart and bar charts in the "Metabolic Burden of Phylogenetic Groups" section of the results.
Ref_CM_Comp.R	Used to generate the scatter plots, calculate means and perform t-tests for the "Comparison of Reference and CarveMe Reconstructed GEMs" section of the results.
Het_Exp.R	Used to generate the scatter plots, calculate means and perform t-tests for the "Metabolic Burden in Heterologous Hosts" section of the results.

4 Results

To assess the accuracy of the metabolic pathway reconstruction by ARMRiPP, two tests were performed. The first was a test of ARMRiPP's ability to predict the correct structure of the final RiPP product associated with each RiPP. The test was performed by using a subset of the RiPPs with known chemical structures from the MIBiG repository. Predicted structures from ARMRiPP for this BGC subset were matched with the MIBiG structures, generating a Tanimoto score of chemical similarity for each BGC compound. Additionally, the same subset was then used to generate chemical structure models through PRISM. These structures were also compared with the structures obtained from MIBIG, and provide a "gold standard" for chemical structure prediction from BGCs. The second test was to check the ARMRiPP's ability to correctly predict the metabolic pathway structure and the stoichiometry of its cofactors for each reaction. Using a smaller subset of RiPPs from the MIBiG repository, the generated pathways were curated by comparing reaction steps of the reconstruction with the reaction steps described in literature for each BGC in the subset. To quantify the accuracy of the reconstructed pathways the production yield values of the reconstructed and literature pathways were compared.

Keeping the accuracy of the pathway reconstruction in mind, several sets of GEMs were extended with one reconstructed RiPP pathway for each GEM, and then analyzed. These extended models were of two groups, either reference GEMs or CarveMe reconstructed GEMs . Since the metabolic environment is identical between all of the extended reference GEMs, this analysis focuses on the differences between classes of RiPPs in secondary metabolite metabolic burden. Unlike the reference GEMs, the CarveMe reconstructed GEMs simulate different metabolic environments given from the different source organisms of the RiPPs. Additionally, the CarveMe reconstructed GEMs were a smaller set of models than the reference GEMs, as not every RiPP source organism had accessible annotated genomic data for genome-scale reconstruction, and not every reconstructed GEM had growth. Analyses of the CarveMe reconstructed GEMs were focused either on the phylogeny of the source organisms or on heterologous expression. The phylogenetic analysis involved grouping the models by the clade different clade levels associated with the BGC source organism and comparing metabolic burden. Heterologous expression analysis involved extending the same reconstructed RiPP pathways in all of the CarveMe reconstructed GEMs with growth, and then analyzing metabolic burden.

4.1 Accuracy of End Product Structure Predictions

To evaluate the performance of ARMRiPP to predict RiPP end product structures, we collected known compound structures from PubChem corresponding to the end product of 57 different RiPPs from MIBiG with an identified core (out of 261 RiPPs). These 57 RiPPs covered multiple subclasses from lanthipeptides, thiopeptides and lasso peptides. We then used ARMRiPP to reconstruct the corresponding biosynthetic pathways and predict the end product structure, and compared predicted vs known compound structure based on their SMILES formula using the Tanimoto score. We also predicted and compared end product structures from the unmodified precursor peptides of these 57 RiPPs (by ignoring all modification reactions) with PubChem structures as a "worst-case" comparison. The difference between the Tanimoto scores of structures from unmodified and modified end products shows how much of the end product structure is determined by the modification reactions. Based on literature, it is common to consider a Tanimoto score above 0.85 as indicating high molecular similarity (93, 94, 102), and so we will use this delineation when considering the following results.

Unmodified Core Peptide Tanimoto Score Comparison

We find a clear improvement (38%) for the structural prediction accuracy of ARMRiPP compared to the unmodified structures, with the mean Tanimoto score increasing from 0.65 to 0.9. We also find that for the ARMRiPP results the majority of Tanimoto scores are above 0.85, which would indicate high structural similarity to the PubChem structures (Figure 4.1.1). Although the threshold of 0.85 might seem a bit arbitrary, this has historically been used to define "high structural similarity" (102). Additionally, the spread of the Tanimoto scores for the ARMRiPP predicted structures is lower compared to the core peptide structures, where many Tanimoto scores are below 0.85.

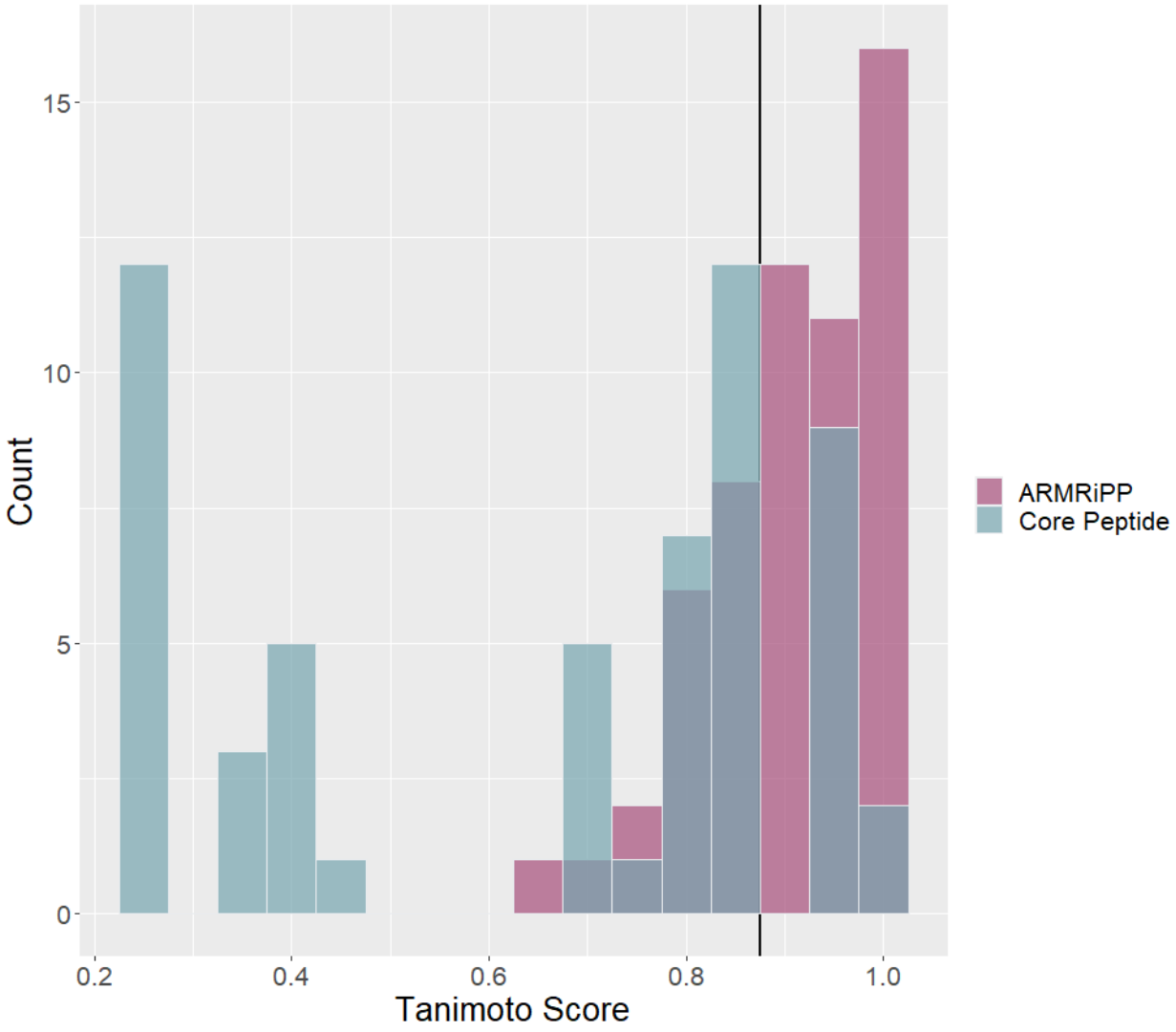


Figure 4.1.1: Histograms of structure similarities shown as Tanimoto similarity scores for 57 RiPPs obtained from MIBiG. The purple bars show the similarity between structures obtained from PubChem and ARMRIiPP predicted structures, while the light blue bars show the scores obtained from comparing PubChem structures with structures predicted from unmodified core peptides. Bin width is 0.05. Line intersects where the bins contain values above 0.85 Tanimoto score.

We then asked if ARMRiPP predicts the end product structures of certain RiPP classes better or worse than other classes. Our test set of 57 RiPPs is composed of 22 lantipeptides, 14 lasso peptides and 21 thiopeptides, and by plotting the Tanimoto scores for these classes individually we see clear differences between the RiPP classes, although mostly in the unmodified structures (Figure 4.1.2). The group in the Tanimoto score range of 0.2-0.5 fits with the thiopeptide distribution, the one in the 0.6-0.9 range fits with the lanthipeptides, and the one in the 0.9-1.0 range fits for the most part with the lasso peptide distribution. As the Tanimoto score of these

RiPP core peptides is based on the molecular similarity to the PubChem assigned structure of their mature product, we would expect that a lower value indicates a higher degree of structural modification from the RiPP maturation pathway, and that mature RiPPs with more unmodified amino acids from its core will have a higher Tanimoto score. The results appear to coincide with this expectation. Of these three, the RiPP class with the most and largest structural modifications, thiopeptides, have the lowest core peptide T-values, while the RiPP class with the least modification reactions, lasso peptides, have the highest core peptide Tanimoto scores. Lanthipeptides, as expected, lie in between them.

Observing the distance between the pair of distributions for each RiPP class, we get an indication of differences in the improvement in structural prediction by ARMRiPP according to the RiPP class. Thiopeptides appear to have the largest improvement from ARMRiPP, lanthipeptides with less of an improvement and lasso peptides with the smallest improvement (Figure 4.1.2 and Table 4.1.1). However, considering that there is in a sense less room for improvement for the RiPP classes where fewer structural modifications happen, such as lasso peptides, this is as expected. Furthermore, even though thiopeptides may have the largest improvement, they also have the lowest share of predicted structures with a Tanimoto score above 0.85, with only 12 out of the 21 structures. To compare, lanthipeptides had 19 out of 21 over 0.85 and lasso peptides had 12 out of 14.

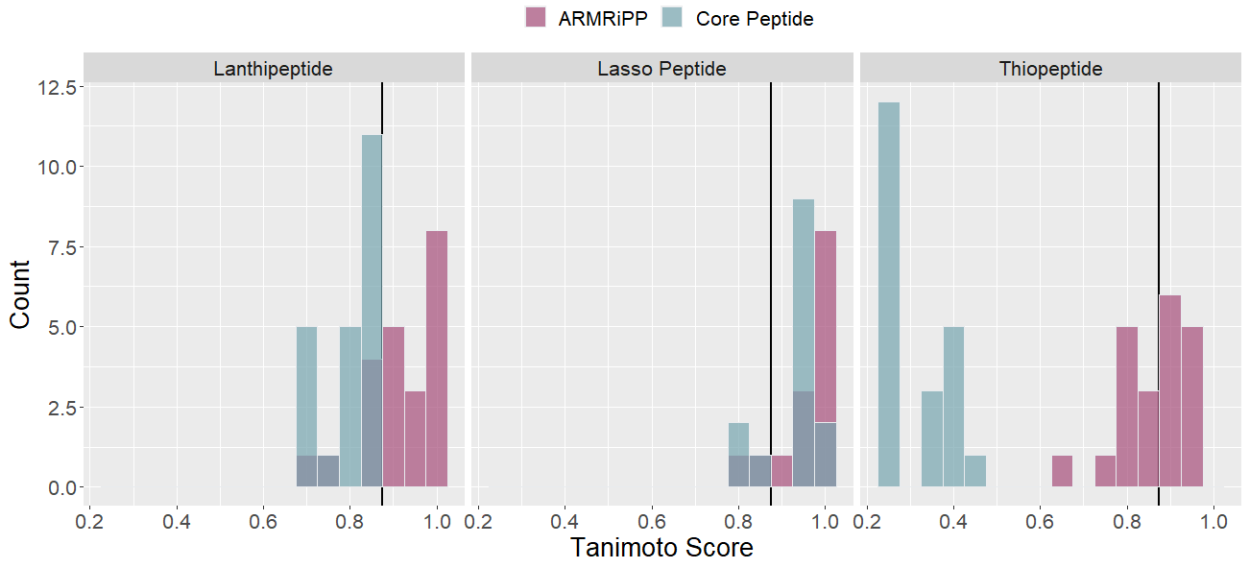


Figure 4.1.2: Three pairs of histograms of structure similarities shown as Tanimoto similarity scores for each RiPP class of the 57 RiPPs obtained from MIBiG. Note that there are 22 lanthipeptide, 14 lasso peptide and 21 thiopeptide RiPPs. The purple bars show the similarity between structures obtained from PubChem and ARMRiPP predicted structures, while the light blue bars show the scores obtained from comparing PubChem structures with structures predicted from unmodified core peptides. Bin width is 0.05. Line intersects where the bins contain values above 0.85.

Table 4.1.1: Mean Tanimoto scores for core peptide and ARMRiPP predicted structure similarity in each RiPP class, with mean percentage improvement for the predicted structures.

RiPP class	Mean core peptide structure Tanimoto score	Mean ARMRiPP Tanimoto score	Mean improvement
Lanthipeptides	0.7999	0.9174	14.69%
Thiopeptides	0.3154	0.8529	170.40%
Lasso peptides	0.9285	0.9552	2.88%

PRISM Tanimoto Score Comparison

Having established that ARMRiPP predicts RiPP end product structures with high accuracy, we aimed to benchmark ARMRiPP against PRISM, a state-of-the-art software for identification of BGCs and end-product structure predictions (13). Tanimoto scores were generated from the structure prediction results of PRISM for the same 57 RiPP clusters compared to the PubChem structures. Note that PRISM can predict a range of possible end product structures, while ARMRiPP only provides a single structure prediction. Both the median and max Tanimoto score from PRISM of each set of structures was used to get values to compare with ARMRiPP.

As noted before, ARMRiPP's Tanimoto scores appear grouped together with the majority above a Tanimoto score of 0.85, while the PRISM median Tanimoto scores appear to have two distinct groups (Figure 4.1.3). However, 42 out of the 57 ARMRiPP structures and 41 of the median PRISM structures are above 0.85 Tanimoto score, and so by this metric they are similar. The rightmost group of the median PRISM distribution appears in a tight range, with few Tanimoto scores below 0.8 and the majority above 0.85, while the leftmost group have all their Tanimoto scores in the 0 to 0.05 range. The cause of the high amount of low Tanimoto scores present in the leftmost group for the median PRISM distribution is due to PRISM not always managing to predict structures from the BGC data used in this section. For some BGCs it returned no structures, and for others it returned a minimal structure, such as only a pair of oxygen atoms. In both cases the Tanimoto score was calculated as 0. The mean Tanimoto score of the ARMRiPP predicted structures was 0.90, while for the median value of the PRISM predicted structures it was 0.77, which gives a positive difference of 17% for the ARMRiPP Tanimoto score. This is in large part affected by the left group in the PRISM's Tanimoto score distribution.

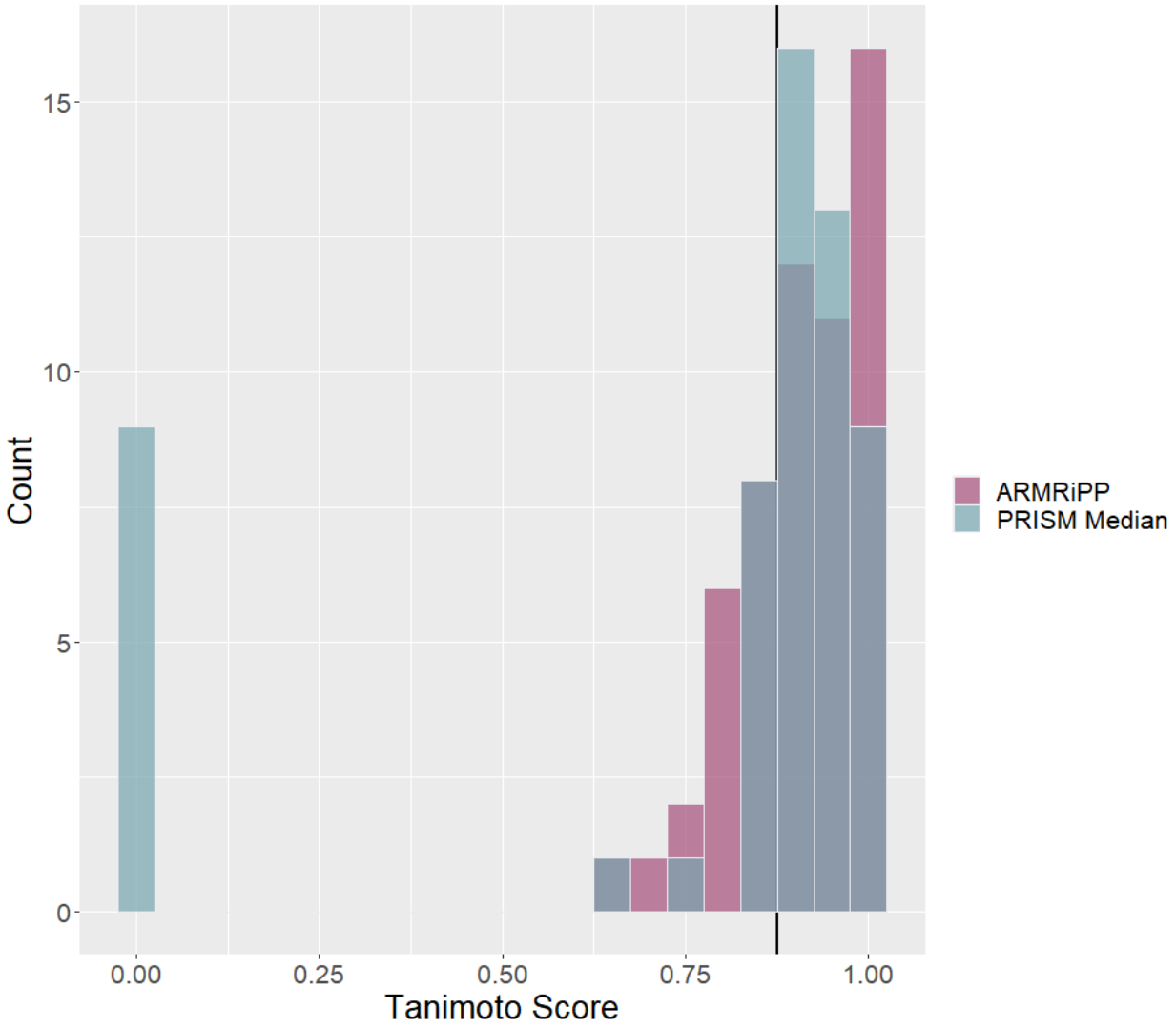


Figure 4.1.3: Histograms of structure similarities shown as Tanimoto similarity scores for 57 RiPPs obtained from MIBiG. The purple bars show the similarity between structures obtained from PubChem and ARMRIiPP predicted structures, while the light blue bars show the median scores obtained from comparing PubChem structures with PRISM predicted structures. Bin width is 0.05. Line intersects where the bins contain values above 0.85 Tanimoto score.

Moving on to the max Tanimoto scores for the PRISM structures we can observe that they are distributed similarly as before, in two distinct groups, but with the rightmost group lying much closer to a Tanimoto score of 1 (Figure 4.1.4). Using the 0.85 Tanimoto score metric we find that 45 of the max score PRISM structures are above it, out of the total 57 structures. For the max Tanimoto score of each set of PRISM structures the mean was 0.80, which gives a positive difference of 13% for ARMRiPP. Same as before, this is skewed by the large group of values with a Tanimoto score of 0.

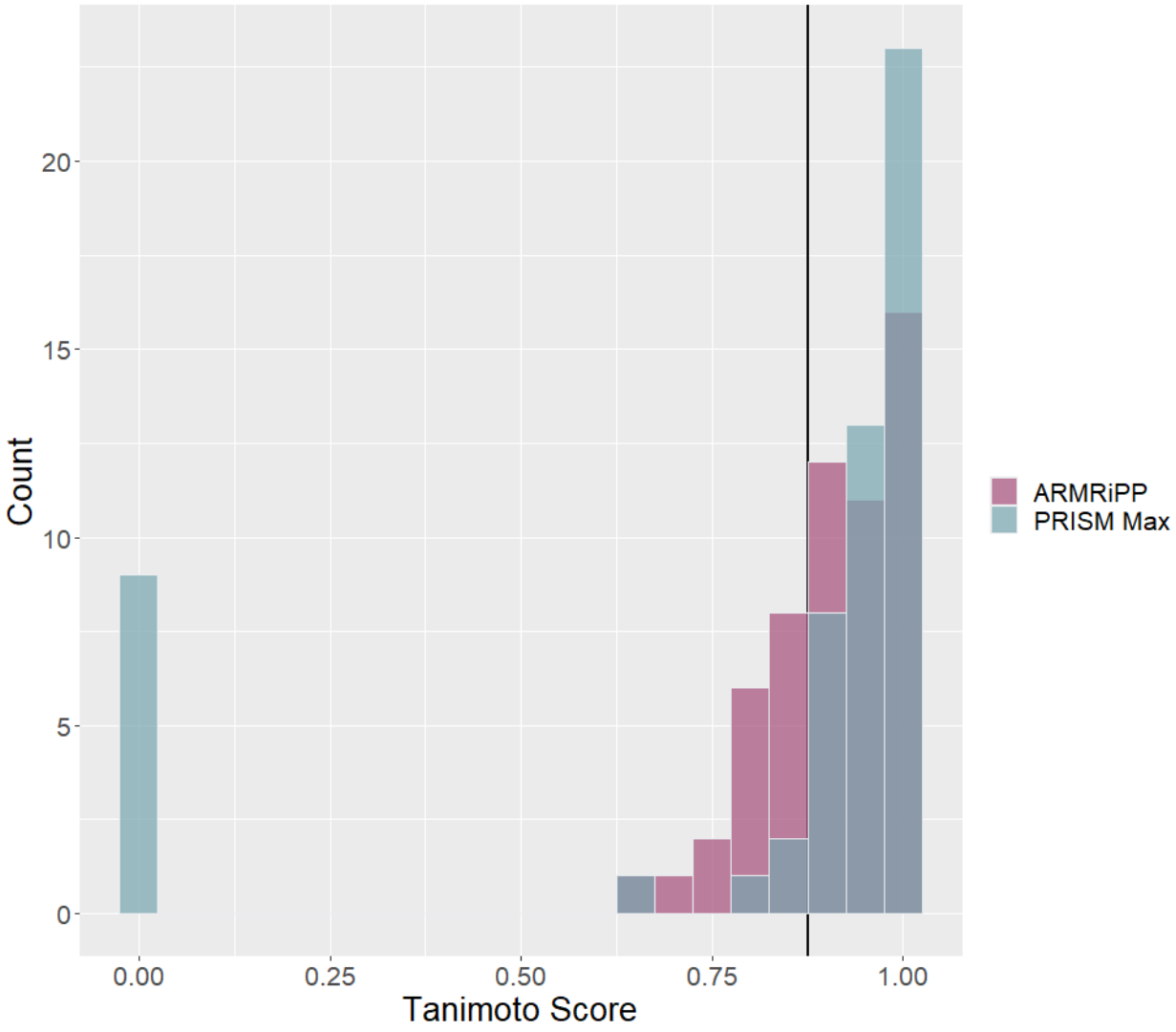


Figure 4.1.4: Histograms of structure similarities shown as Tanimoto similarity scores for 57 RiPPs obtained from MIBiG. The purple bars show the similarity between structures obtained from PubChem and ARMRIiPP predicted structures, while the light blue bars show the maximum scores obtained from comparing PubChem structures with PRISM predicted structures. Bin width is 0.05. Line intersects where the bins contain values above 0.85 Tanimoto score.

To check whether there are differences in the performance of ARMRiPP and PRISM for different RiPP classes, each RiPP class was considered. There are 22 lanthipeptide, 14 lasso peptide and 21 thiopeptide RiPPs. We can observe that the rightmost group of PRISM median Tanimoto scores has a similar distribution as the ARMRiPP Tanimoto scores for lasso peptides, has a distribution more towards slightly lower values for lanthipeptides, and has a distribution more towards higher values for thiopeptides (Figure 4.1.5). The rightmost groups of the PRISM max scores again lie further towards a Tanimoto score of 1, particularly the PRISM lanthipeptide

distribution is now pushed towards higher values when compared to ARMRiPP (Figure 4.1.6). Additionally, by considering the leftmost group of both the PRISM median and max Tanimoto scores we can see the amount of failed structure prediction from PRISM for the different classes (Figure 4.1.5 and 4.1.6). Despite there being a similar number of lanthipeptide as thiopeptide products, and a lower number of lasso peptides, PRISM had the lowest number of failed structure predictions in the thiopeptide distribution. Using the 0.85 Tanimoto index cutoff, of the 22 lanthipeptides ARMRiPP had 19 over 0.85 while the median PRISM structures had 15 and the max PRISM structures had 17. For the 21 thiopeptides ARMRiPP had 11, PRISM median had 18 and max had 19. Lastly, for the 14 lasso peptides ARMRiPP had 12, PRISM median had 8 and max had 9. We can see the mean Tanimoto scores of the median PRISM values in Table 4.1.2, and the max PRISM values in Table 4.1.3, compared with the ARMRiPP values.

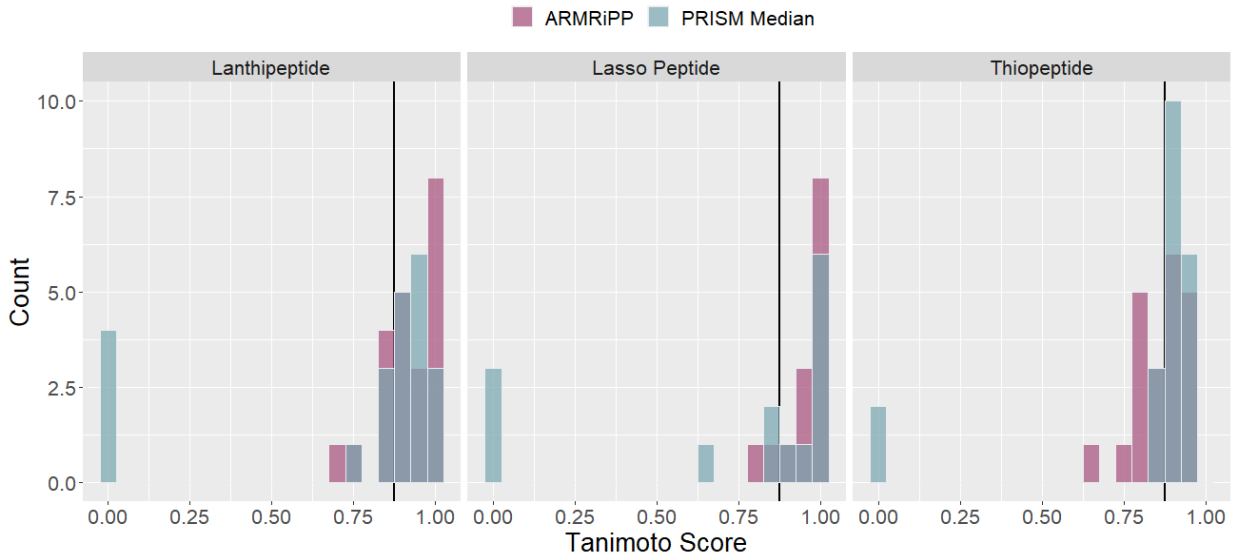


Figure 4.1.5: Three pairs of histograms of structure similarities shown as Tanimoto similarity scores for each RiPP class of the 57 RiPPs obtained from MIBiG. Note that there are 22 lanthipeptide, 14 lasso peptide and 21 thiopeptide RiPPs. The purple bars show the similarity between structures obtained from PubChem and ARMRIiPP predicted structures, while the light blue bars show the median scores obtained from comparing PubChem structures with PRISM predicted structures. Bin width is 0.05. Line intersects where the bins contain values above 0.85.

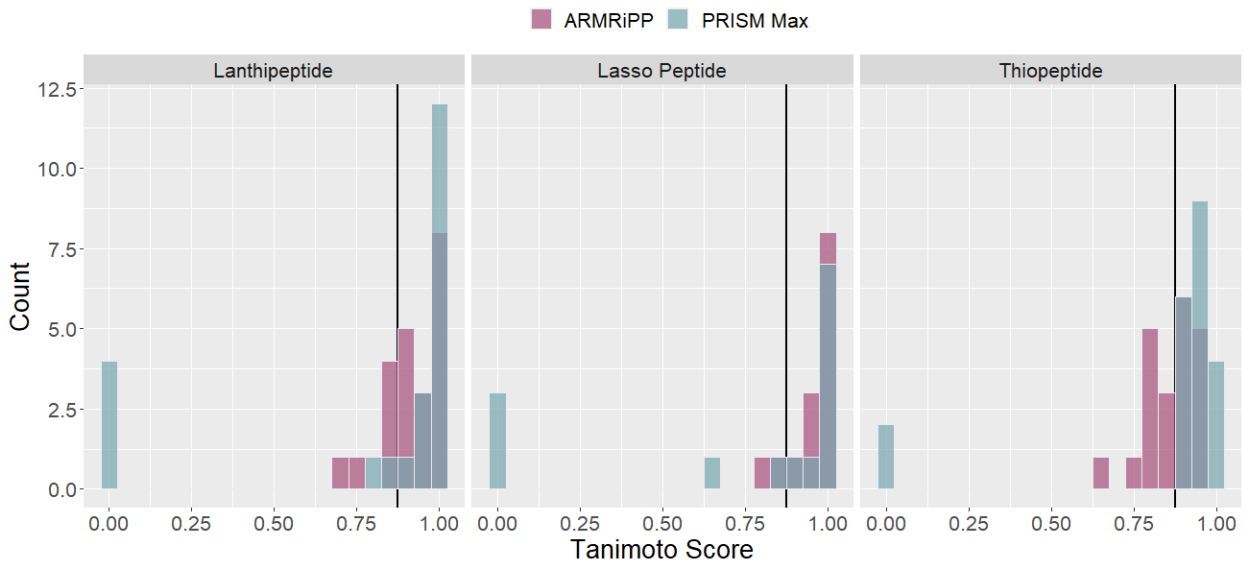


Figure 4.1.6: Three pairs of histograms of structure similarities shown as Tanimoto similarity scores for each RiPP class of the 57 RiPPs obtained from MIBiG. Note that there are 22 lanthipeptide, 14 lasso peptide and 21 thiopeptide RiPPs. The purple bars show the similarity between structures obtained from PubChem and ARMRIiPP predicted structures, while the light blue bars show the max scores obtained from comparing PubChem structures with PRISM predicted structures. Bin width is 0.05. Line intersects where the bins contain values above 0.85.

Table 4.1.2: Mean Tanimoto scores for predicted structure similarity for PRISM (median) and ARMRiPP in each RiPP class, with mean percentage difference for the predicted structures.

RiPP class	Median PRISM predicted structure Tanimoto score	ARMRiPP predicted structure Tanimoto score	Difference
Lanthipeptides	0.7531	0.9174	21.81%
Thiopeptides	0.8240	0.8529	3.50%
Lasso peptides	0.7274	0.9552	31.32%

Table 4.1.3: Mean Tanimoto scores for predicted structure similarity for PRISM (max) and ARMRiPP in each RiPP class, with mean percentage difference for the predicted structures.

RiPP class	Max PRISM predicted structure Tanimoto score	ARMRiPP predicted structure Tanimoto score	Difference
Lanthipeptides	0.7910	0.9174	15.98%
Thiopeptides	0.8550	0.8529	-0.24%
Lasso peptides	0.7386	0.9552	29.33%

4.2 Accuracy of Metabolic Pathway Prediction and Cofactor Usage

An important difference from PRISM is that ARMRiPP also predicts the metabolic pathway required to make the end products, including all required cofactors and by-products. This is a crucial feature for the tool to be applicable for genome-scale metabolic modeling and constraint-based analyses. To evaluate ARMRiPP's ability to accurately predict the metabolic pathway and cofactor usage we chose 3 RiPPs which had their biosynthesis accurately described in the literature, namely Lacticin 481, Thiomuracin A, Felipeptin A1 and Felipeptin A2.

We then compared differences from the output of the COBRApy reaction chain from each reconstructed RiPP cluster with its literature defined pathway. In the case of such differences, FBA was performed in the Sco-GEM reference model using the entire reconstructed and literature pathway in one-step reactions as the objective functions, to determine how large of an effect the difference makes on the estimated production potential of the RiPP compound in the reference GEM. Production yield was chosen as the metric for the accuracy of ARMRiPP's pathway prediction. The carbon source for the media was glucose, and the production yield was determined by mmol RiPP production divided by mmol carbon source intake. Similarly, FBA was performed using only the ribosomal synthesis reaction of the reconstructed pathway, to check if the following reactions of the reconstructed pathway consistently improve the accuracy in predicting production yield.

Across the four tested pathways, the average absolute error from the expected production yield from literature was 6.96%.

Lacticin 481 Biosynthetic Pathway

Lacticin 481 is a class II lanthipeptide, and has a pathway which closely follows the central lanthipeptide biosynthesis with no additional modification reactions (103, 104). The biosynthesis pathway consists of four steps, the ribosomal peptide synthesis, two tailoring reactions and finally proteolysis and cleavage of the leader peptide (Table 4.2.1). We see that the major difference between the two pathways is that the pathway predicted by ARMRiPP wrongly assigns the number of reacting ATPs and produced ADPs and P_is. This is expected as it follows from the assumption made in the method section that every serine and threonine of the lanthipeptide core is dehydrated. In the case of Lacticin 481 however, only its four serine residues are dehydrated, while its two threonine residues are not. The mechanism of the proteolysis reaction of Lacticin 481 is not well known, but it is presumed within this thesis that it requires a molecule of water.

Table 4.2.1: Comparison between the metabolites taking part in the reconstructed pathway and literature pathway of Lacticin 481. Leftmost column denotes the name of the reaction step. Differences in the reconstruction are colored in red, and if the reaction is not present the cell is empty.

Reaction	**ARMRiPP predicted Pathway**	**Literature Pathway (103, 104)**
Ribosomal polypeptide synthesis	Ala + 4 Asn + Asp + 51 ATP + 3 Cys + 3 Gln + 5 Glu + 4 Gly + 102 GTP + 2 His + 3 Ile + 5 Leu + 2 Lys + 2 Met + 3 Phe + 6 Ser + 3 Thr + Trp + 3 Val → Lacticin 481 prepeptide + 51 AMP + 102 GDP + 50 H_2O + 102 P_i + 51 PP_i	Ala + 4 Asn + Asp + 51 ATP + 3 Cys + 3 Gln + 5 Glu + 4 Gly + 102 GTP + 2 His + 3 Ile + 5 Leu + 2 Lys + 2 Met + 3 Phe + 6 Ser + 3 Thr + Trp + 3 Val → Lacticin 481 prepeptide + 51 AMP + 102 GDP + 50 H_2O + 102 P_i + 51 PP_i
Phosphorylation dehydration	Lacticin 481 prepeptide + 6 ATP → Lacticin 481 dehydrated intermediate + 6 ADP + 6 P_i	Lacticin 481 prepeptide + 4 ATP → Lacticin 481 dehydrated intermediate + 4 ADP + 4 P_i
Cyclization	Lacticin 481 dehydrated intermediate → Lacticin 481 cyclic intermediate	Lacticin 481 dehydrated intermediate → Lacticin 481 cyclic intermediate
Proteolysis	Lacticin 481 cyclic intermediate + H_2O → Lacticin 481 mature product	Lacticin 481 cyclic intermediate + H_2O → Lacticin 481 mature product

Performing FBA using the reaction stoichiometry of the literature pathway gives a production yield of $1.893*10^{-2}$ mmol product/mmol carbon source, while the reconstructed pathway gives a production yield of $1.892*10^{-2}$. This gives an error of -0.10% from the reconstructed to literature pathway values. As expected, the reconstructed pathway has a lower production yield, likely due to its higher ATP consumption. Furthermore, it is reasonable that the error in objective values would be small given how small the difference in general between the pathways are. To check if the reconstructed pathway subsequent to ribosomal synthesis improves the error, FBA was performed only using the reconstructed ribosomal reaction stoichiometry. This gave a production yield of $1.897*10^{-2}$, which gives an error of 0.21% between ribosomal synthesis and the literature pathway value. It is expected that the production yield would be higher only taking the ribosomal synthesis into account, as it consumes less high energy metabolites than the pathway on a whole. Here we can see the reconstructed pathway lowered the absolute error for the objective value by 0.11 percentage points when compared to using only the ribosomal synthesis reaction.

Thiomuracin A Biosynthetic Pathway

Thiomuracin A is a series d thiopeptide with a pathway which closely follows the central biosynthesis of thiopeptides (33). The pathway has only a couple of additional tailoring reactions, and no secondary macrocyclization. Looking at Table 4.2.2 we may notice several differences between the reconstructed and literature pathways. Firstly, we see that the reconstructed pathway has one less alanine residue in its ribosomal polypeptide synthesis reaction, and as a result the net energy cost for the synthesis is lowered by 2 GTP + 2 ATP (two high energy ATP bonds are required to regenerate the AMP) (98), as well as producing one less molecule of water. To investigate this the antiSMASH page for the associated Thiomuracin A BGC was checked, and it was found that the C-terminal alanine residue was not present in the predicted core peptide sequence. However, according to the end product structure of Thiomuracin A from literature (105), it appears that the C-terminal alanine residue is in fact missing there as well. It can be presumed that at some point during biosynthesis the C-terminal alanine is proteolytically cleaved, but no proposed mechanism for this was found from the literature. On one hand, it seems that antiSMASH correctly informed about the C-terminal alanine missing from the end structure, but due to it not being present in the initial core peptide sequence ARMRiPP did not account for its ribosomal production. Similarly to before, we add a final proteolysis reaction with the thesis assumption that proteolysis reactions generally require a molecule of water.

The next major differences between the reconstructed and literature pathways were two missing tailoring reactions. These are not present in the reconstructed pathway simply due to the functionality for these reactions not being implemented in ARMRiPP. As for why this functionality is missing, it is in part due to limitations of the antiSMASH annotation, but also due to thiopeptide tailoring reactions being highly group or compound specific. In either case, we can see in Table 4.2.2 that the first missing tailoring reaction is a methylation reaction on a thiazole in the thiopeptide core performed by an rSAM-enzyme methyltransferase pre-macrocyclization, which consumes SAM and produces SAH. The second is an epoxidation reaction on a thiopeptide core isoleucine residue, which is mediated by a Cytochrome P450 enzyme and is assumed to follow the standard reaction mechanism of the CYP enzyme superfamily for reactive oxygen species (106).

The final difference between the reconstructed and literature pathways is from the central heterocycle modification reaction. Here the reconstructed pathway contains the reaction for the formation of series b thiopeptides, while it is known from literature that Thiomuracin A is a series d thiopeptide (33). This was caused by antiSMASH incorrectly assigning the Thiomuracin A cluster precursor as a type II thiopeptide (series a-c), while it should have been assigned as a type III thiopeptide (series d). The reason behind this is due to antiSMASH categorizing thiopeptide classifications by presence of genes generally responsible for group specific

secondary macrocyclization reactions, such as type I (series e) containing methyl-indolic acid pathway genes, type II (series a-c) containing quinaldic acid pathway genes and type II (series d) containing no such genes. As antiSMASH appears to have discovered quinaldic acid pathway genes in the Thiomuracin A cluster, it wrongly categorized it as a type II (series a-c) thiopeptide, for which ARMRiPP reconstructs the series b modification reaction.

Table 4.2.2: Comparison between the metabolites taking part in the reconstructed pathway and literature pathway of Thiomuracin A. Leftmost column denotes the name of the reaction step. Differences in the reconstruction are colored in red, and if the reaction is not present the cell is empty.

Reaction	**ARMRiPP predicted Pathway**	**Literature Pathway (33, 75, 105–107)**
Ribosomal polypeptide synthesis	4 Ala + Asn + 5 Asp + 48 ATP + 6 Cys + 3 Glu + 4 Gly + 96 GTP + His + Ile + 4 Leu + 3 Met + 2 Phe + Pro + 6 Ser + 2 Thr + Tyr + 4 Val → Thiomuracin A prepeptide + 48 AMP + 96 GDP + 47 H_2O + 96 P_i + 48 PP_i	5 Ala + Asn + 5 Asp + 49 ATP + 6 Cys + 3 Glu + 4 Gly + 98 GTP + His + Ile + 4 Leu + 3 Met + 2 Phe + Pro + 6 Ser + 2 Thr + Tyr + 4 Val → Thiomuracin A prepeptide + 49 AMP + 98 GDP + 48 H_2O + 98 P_i + 49 PP_i
Azole formation	Thiomuracin A prepeptide + 5 ATP + 5 FMN → Thiomuracin A thiazole intermediate + 5 ADP + 5 $FMNH_2$+ 5 H_2O + 5 P_i	Thiomuracin A prepeptide + 5 ATP + 5 FMN → Thiomuracin A thiazole intermediate + 5 ADP + 5 $FMNH_2$+ 5 H_2O + 5 P_i
Dehydration by glutamylation elimination	Thiomuracin A thiazole intermediate + 4 Glu-$tRNA^{Glu}$ → Thiomuracin A dehydrated intermediate + 4 Glu + 4 $tRNA^{Glu}$	Thiomuracin A thiazole intermediate + 4 Glu-$tRNA^{Glu}$ → Thiomuracin A dehydrated intermediate + 4 Glu + 4 $tRNA^{Glu}$
Tailoring: Methylation		Thiomuracin A dehydrated intermediate + SAM → Thiomuracin A dehydrated intermediate + SAH
Tailoring: Epoxidation		Thiomuracin A dehydrated intermediate + O_2 + 2 H^+ → Thiomuracin A dehydrated intermediate + H_2O

Macrocyclization	Thiomuracin A dehydrated intermediate → Thiomuracin A macrocyclic intermediate + H_2O	Thiomuracin A dehydrated intermediate → Thiomuracin A macrocyclic intermediate + H_2O
Central heterocycle modification	Thiomuracin A macrocyclic intermediate + H^+ → Thiomuracin A series b product	Thiomuracin A macrocyclic intermediate → Thiomuracin A series d product + H^+
Unknown C-terminal proteolysis		Thiomuracin A series d product + H_2O → Thiomuracin A series d product

The production yield for the literature pathway was $2.125 * 10^{-2}$, while for the reconstructed pathway it was $2.169 * 10^{-2}$. This gives an error of 2.08% between the two values, which is reasonable given that the reconstructed pathway consumed fewer high energy bonds in its ribosomal synthesis reaction. To check if the reconstructed pathway subsequent to ribosomal synthesis improves the error, FBA was performed only using the reconstructed ribosomal reaction stoichiometry. The ribosomal reaction had a production yield of $2.161 * 10^{-2}$, which gave an error of 1.70% when compared to the literature pathway. Interestingly this is a lower error than for the reconstructed pathway, and indicates that the reactions subsequent to ribosomal synthesis in the reconstruction are on a whole metabolically favorable in comparison.

Felipeptin A1 and Felipeptin A2 Biosynthetic Pathways

Felipeptin A1 and Felipeptin A2 are two class IV lasso peptides present in a single two-component RiPP cluster (108), although their pathways will be considered separately. As is known from the theory section, the central lasso peptide biosynthesis is generally less complex than that of lanthipeptides and thiopeptides (33, 61, 80). Starting with the Felipeptin A1 pathway, there are large differences between the reconstructed pathway and literature pathway in the ribosomal polypeptide synthesis reaction. Observing Table 4.2.3, we see that the reconstructed pathway has more amino acids of several kinds when compared to the literature pathway. Investigating the predicted precursor peptide sequence for Felipeptin A1 from antiSMASH, it appears that antiSMASH has assigned a methionine start codon much farther to the 5' end of the ORF than in the literature peptide sequence, which then includes 13 additional predicted amino acids in the leader region. In the literature article (108), they appear to assign a start methionine codon at what would be the 14th amino acid in the antiSMASH sequence, which antiSMASH instead assigns as a valine.

Table 4.2.3: Comparison between the metabolites taking part in the reconstructed pathway and literature pathway of Felipeptin A1. Leftmost column denotes the name of the reaction step. Differences in the reconstruction are colored in red, and if the reaction is not present the cell is empty.

Reaction	**ARMRiPP predicted Pathway**	**Literature Pathway (80, 108)**
Ribosomal polypeptide synthesis	3 Ala + 2 Arg + 2 Asn + 3 Asp + 55 ATP + 2 Cys + 7 Glu + 5 Gly + 110 GTP + 2 Ile + 4 Leu + 2 Lys + Met + 4 Phe + 4 Pro + 4 Ser + 2 Thr + 2 Trp + 6 Val → Felipeptin A1 prepeptide + 55 AMP + 110 GDP + 54 H_2O + 110 P_i + 55 PP_i	2 Ala + 2 Arg + 2 Asn + 3 Asp + 42 ATP + 2 Cys + 4 Glu + 5 Gly + 84 GTP + 2 Ile + 3 Leu + Lys + Met + 3 Phe + 2 Pro + 2 Ser + 2 Thr + 2 Trp + 4 Val → Felipeptin A1 prepeptide + 42 AMP + 84 GDP + 41 H_2O + 84 P_i + 42 PP_i
Leader cleavage	Felipeptin A1 prepeptide + ATP + H_2O → Felipeptin A1 core peptide + ADP + P_i	Felipeptin A1 prepeptide + ATP + H_2O → Felipeptin A1 core peptide + ADP + P_i
Macrolactam formation	Felipeptin A1 core peptide + ATP → Felipeptin A1 macrolactam peptide + AMP + PP_i	Felipeptin A1 core peptide + ATP → Felipeptin A1 macrolactam peptide + AMP + PP_i
Disulfide bridge redox	Felipeptin A1 macrolactam peptide + GSSG → Felipeptin A1 mature peptide + 2 GSH	Felipeptin A1 macrolactam peptide + GSSG → Felipeptin A1 mature peptide + 2 GSH

The literature pathway of Felipeptin A1 gave a production yield of $2.350 * 10^{-2}$, while the reconstructed pathway had a production yield of $1.781 * 10^{-2}$. This gives an error of -24.22% from the literature to the reconstructed pathway yield values. This is expected as the much larger predicted precursor length of the reconstructed pathway not only requires more amino acids but also directly requires more high energy metabolites to drive the ribosomal synthesis. Running FBA only for the reconstructed ribosomal synthesis gave a production yield of $1.785 * 10^{-2}$, which has an error from the literature of -24.07%. It is reasonable that the error in the yield for the ribosomal synthesis reaction is lower than for the entire reconstructed pathway, since the remaining reactions consume more high energy metabolites and ribosomal synthesis is already predicting a too high consumption.

The Felipeptin A2 pathway also had differences in the ribosomal synthesis reaction, but unlike the Felipeptin A1 pathway this was due to the presence of one missing alanine from the antiSMASH predicted precursor peptide sequence. We can see in Table 4.2.4 that even though

this is only a difference of only one amino acid, it changes the stoichiometry of multiple cofactor reactants and products.

Table 4.2.4: Comparison between the metabolites taking part in the reconstructed pathway and literature pathway of Felipeptin A2. Leftmost column denotes the name of the reaction step. Differences in the reconstruction are colored in red, and if the reaction is not present the cell is empty.

Reaction	**ARMRiPP predicted Pathway**	**Literature Pathway (80, 108)**
Ribosomal polypeptide synthesis	Ala + Arg + 2 Asn + 2 Asp + 41 ATP + 2 Cys + Gln + 5 Glu + 6 Gly + 82 GTP + 2 Ile + 2 Leu + Lys + 2 Met + 2 Phe + 2 Pro + Ser + 3 Thr + 3 Tyr + 3 Val → Felipeptin A2 prepeptide + 41 AMP + 82 GDP + 40 H_2O + 82 P_i + 41 PP_i	2 Ala + Arg + 2 Asn + 2 Asp + 42 ATP + 2 Cys + Gln + 5 Glu + 6 Gly + 84 GTP + 2 Ile + 2 Leu + Lys + 2 Met + 2 Phe + 2 Pro + Ser + 3 Thr + 3 Tyr + 3 Val → Felipeptin A2 prepeptide + 42 AMP + 84 GDP + 41 H_2O + 84 P_i + 42 PP_i
Leader cleavage	Felipeptin A2 prepeptide + ATP + H_2O → Felipeptin A2 core peptide + ADP + P_i	Felipeptin A2 prepeptide + ATP + H_2O → Felipeptin A2 core peptide + ADP + P_i
Macrolactam formation	Felipeptin A2 core peptide + ATP → Felipeptin A2 macrolactam peptide + AMP + PP_i	Felipeptin A2 core peptide + ATP → Felipeptin A2 macrolactam peptide + AMP + PP_i
Disulfide bridge redox	Felipeptin A2 macrolactam peptide + GSSG → Felipeptin A2 mature peptide + 2 GSH	Felipeptin A2 macrolactam peptide + GSSG → Felipeptin A2 mature peptide + 2 GSH

The literature pathway of Felipeptin A2 gave a production yield of 2.301 * 10^{-2}, while the reconstructed pathway had a production yield of 2.334 * 10^{-2}. This gives an error of 1.43% from the literature to the reconstructed pathway yield values. This is expected due the lower cost of the ribosomal synthesis in the reconstructed pathway. Running FBA only for the reconstructed ribosomal synthesis gave a production yield of 2.340 * 10^{-2}, which has an error from the literature of 1.69%. As we can see, this error is slightly larger than for the reconstructed pathway, which is expected as it would have less of a metabolic cost.

4.3 Prediction of RiPP Metabolic Burden

RiPPs have important functions not only in human medicine, but also in shaping natural microbiomes (109). For example, as many RiPPs have an antibiotic effect it is not unlikely that these specialized molecules are used to inhibit other species that compete for the same resources. However, the production of these compounds comes with a cost, but the magnitude of this cost (metabolic burden) has not been thoroughly investigated. Combined with other model reconstruction tools, ARMRiPP allows us to estimate these costs, and we therefore reconstructed 262 RiPP pathways from 155 BGCs (some BGCs have multiple cores) and introduced them into either 1) a curated reference GEM (Sco-GEM) or 2) a draft GEM of the native host, as created using CarveMe.

Of these 262 cores, 249 had reconstructed pathways which yielded mature products of either lanthipeptide, thiopeptide or lasso peptide. The definition for mature product here is the presence of metabolites in the pathway with either thioether cyclization, a central heterocycle, or a macrolactam ring in the case of lanthipeptides, thiopeptides and lasso peptides respectively. We then computed the production/growth-rate gradient as a proxy of the metabolic burden. A lower gradient means higher metabolic burden as the growth rate is more heavily reduced upon production of the RiPP.

Using the results of the FBA, several analyses were performed for the different sets of pathway-extended GEMs. In the case of the extended reference GEMs, the focus was on analyzing the different RiPP classes and subclasses to look for significant differences between the groups. For the extended reconstructed GEMs, the focus was on using the phylogeny of the BGC source organisms to assess differences between phylogenetic groups of different clade levels. The results of the extended reference and reconstructed GEMs were also compared to check if the different groups are consistent for both sets. Finally, the results of the heterologous host set of extended reconstructed GEMs were analyzed to observe group trends for increases or decreases in the growth-production gradient between the source-organism GEM and its heterologous host GEMs.

Metabolic Burden in Reference GEM

By introducing the reconstructed pathways into the same reference GEM we can compare the metabolic burden of the different RiPPs in a way that is not affected by metabolic differences between different hosts. Investigating the RiPP classes of these reconstructed pathways we find that over 70% were of the lanthipeptide class, and with lasso peptides having about twice as many pathways as thiopeptides, which had the fewest (Figure 4.3.1). For the growth-production gradient we can observe that each RiPP class has its median line at about a gradient of 0.25, that lanthipeptides seems to have a group of outliers around 0.5 and thiopeptide has outliers around 4 (Figure 4.3.2). We can also observe that thiopeptides have a larger range of gradient values, indicating a larger variance.

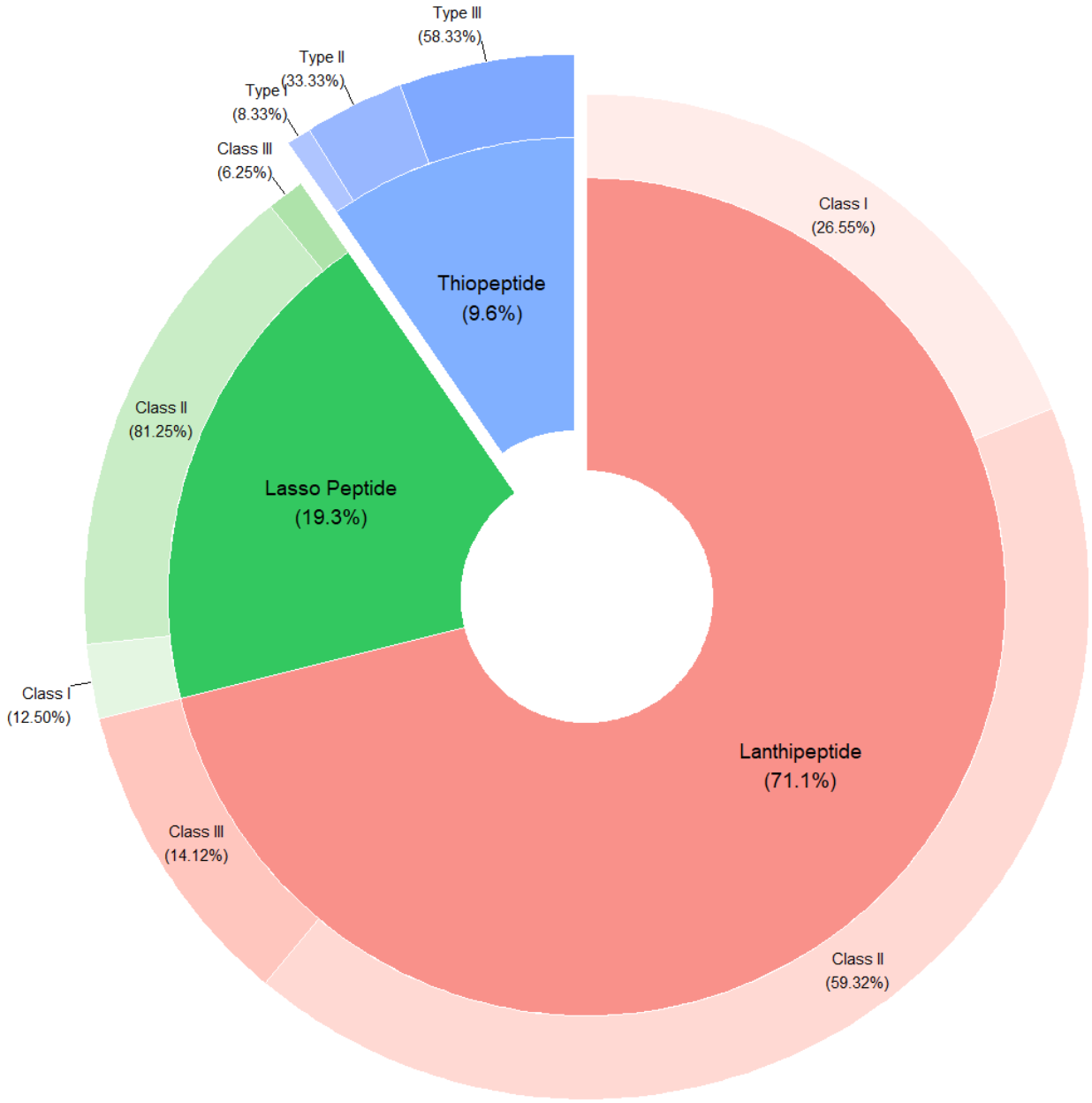


Figure 4.3.1: Pie donut chart of the distribution of RiPP classes and subclasses for the 249 pathways with mature products. Lanthipeptides make up the majority of pathway RiPP classes with 177, while lasso peptides and thiopeptides make up 24 and 48 respectively.

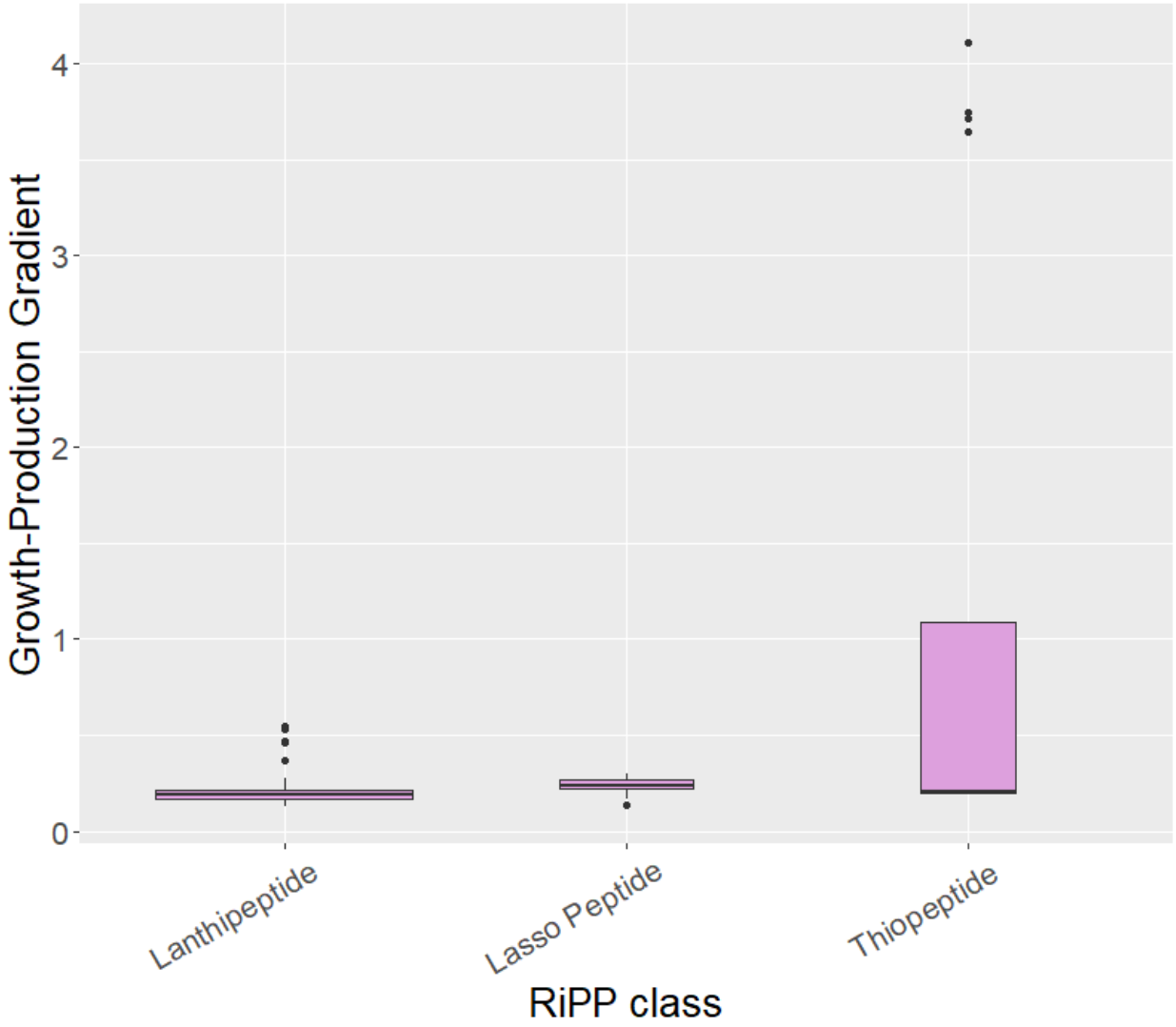


Figure 4.3.2: Box plot of growth-production gradient values in reference GEMs, grouped by RiPP class. The box plot shows the distribution of the gradient values with the following parts: a box representing the range between the first and third quartile of the values, a line within the box representing the median, lines extending above and below representing minimum and maximum values for accepting outliers, and dots representing potential outliers.

To investigate the high variance in the thiopeptide gradient values, we looked at the thiopeptide subclasses. One source of this variance was found to be from the type I thiopeptides, which had a much higher average gradient and lower variance than other subclasses (Figure 4.3.3). It is important to note that there were only two type I thiopeptide pathways with mature products. We can also observe that there is a large variance of gradient values for the type II thiopeptides,

while the remaining type III subclass appears to have more similar average and standard deviant values when compared to the other RiPP classes. The dataset of growth-production gradient values was investigated by filtering for gradient values above 1, and it was discovered that the BGCs which matched the filter were all thiopeptides reconstructed with secondary macrocyclization pathways. Both type I thiopeptide BGCs were present here, BGC0000610 and BGC0001707, each with an identical gradient value of 4.11. Accessing them on the MIBiG repository it was found that they are both BGCs for the same thiopeptide, nosiheptide, but from different genetic sources. The other 4 BGCs found through the filter were all type II thiopeptides with the quinaldic acid secondary macrocyclization pathway and gradient values around 3.7.

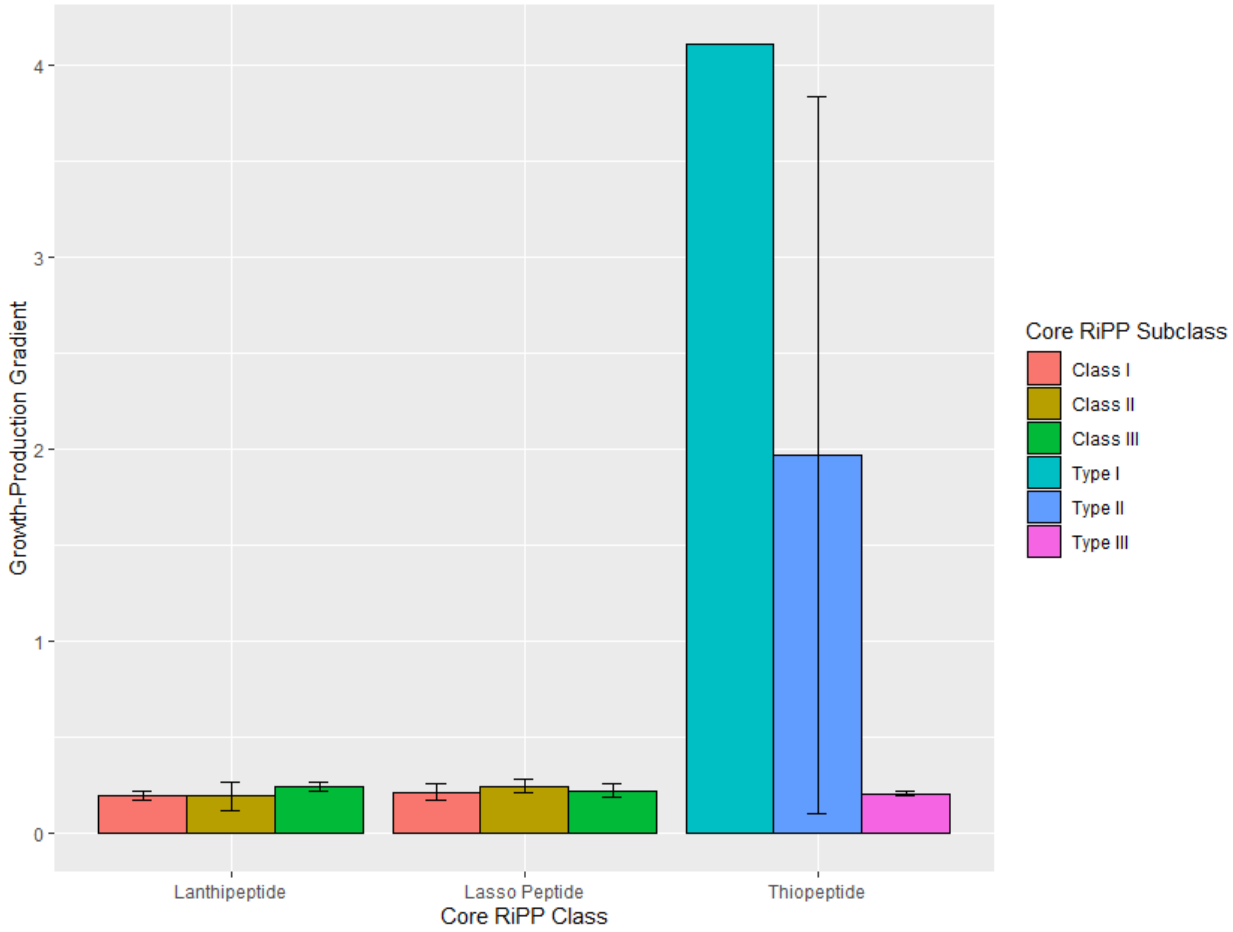


Figure 4.3.3: Bar chart of growth-production gradient averages from the extended reference GEMs, grouped by each RiPP subclass with standard deviation error bars. Note that there are only two type I thiopeptides.

From here, we considered the thiopeptides with secondary macrocyclization as outliers and treated only the remaining pathways. We can observe that the other subclasses appear to have gradient values in similar ranges as each other (Figure 4.3.4). Class II lanthipeptides appear to have a larger variance of gradient values than other subclasses, while thiopeptides without secondary macrocyclization generally seem to have lower variances. Using a pairwise t-test and adjusting p-value with the Bonferroni method, statistically significant group differences were found between class I and III lanthipeptides, class II and III lanthipeptides, class I lanthipeptides and class II lasso peptides, and class II lanthipeptides and class II lasso peptides. The remaining pairs of subclass groups all had p-values far above 0.05 after adjusting with the Bonferroni method.

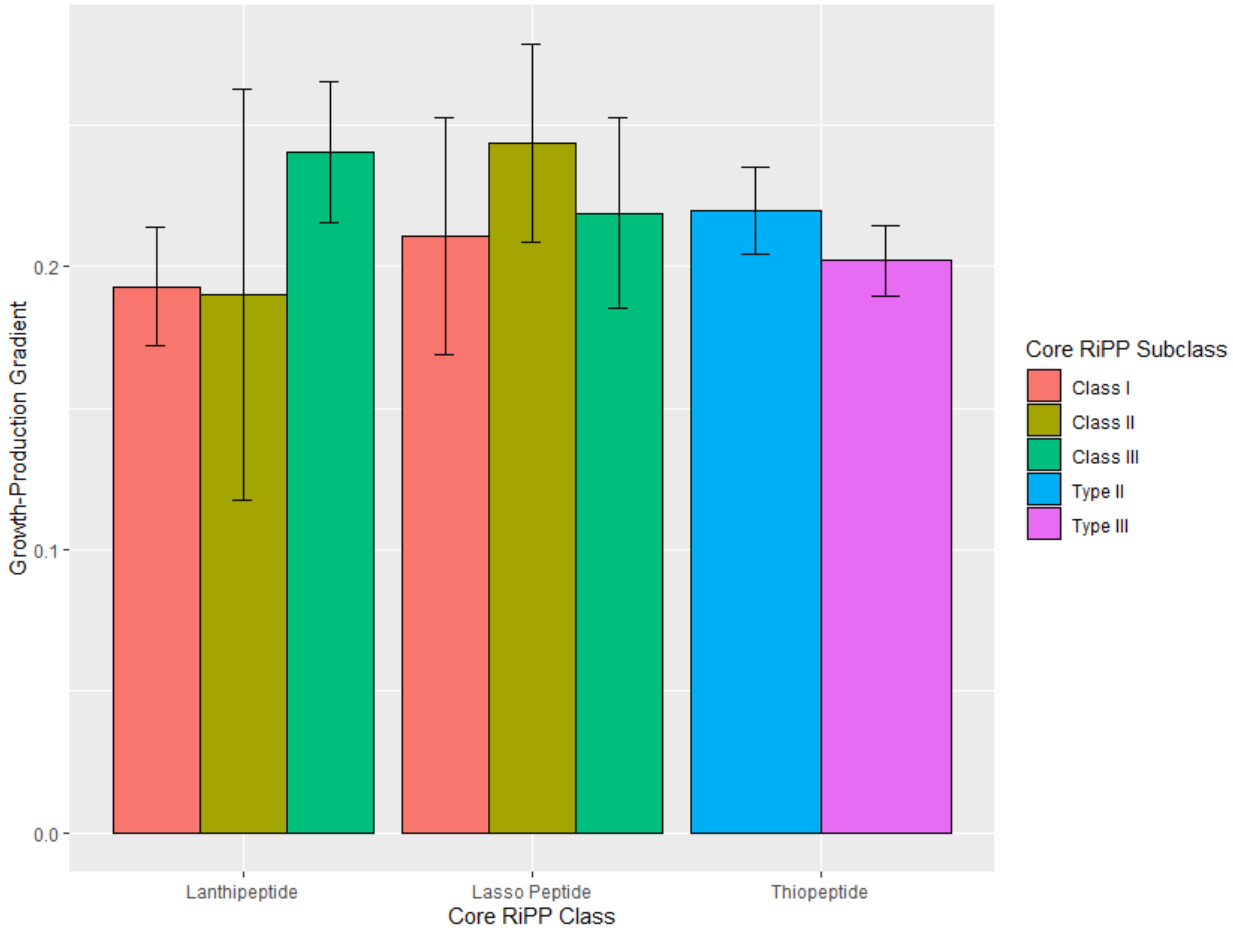


Figure 4.3.4: Bar chart of growth-production gradient averages from the extended reference GEMs, grouped by each RiPP subclass with standard deviation error bars. Gradient results above 1 have been filtered out.

To check the effect of the precursor peptide length on the growth-production gradient for the pathways in the Sco-GEM, the peptide length values of each pathway's precursor were compared against their gradient values. We can observe a negative correlation for peptide length on the

gradient value, with somewhat of a negative logarithmic trend (Figure 4.3.5). Observing the grouping of different RiPP classes and subclasses we see that most lasso peptides have a precursor peptide length of around 30-50, thiopeptides have about 50-60 and lanthipeptides appear to have a much larger range of 20-80. Particularly class II lanthipeptides appear to have a wider range of possible precursor lengths, which may explain why they had such a high variance in gradient values. We can also observe that thiopeptides appear to have higher gradients for the same precursor lengths when compared to the other RiPP classes. This also applies for the thiopeptides without the secondary macrocyclization reaction, which we saw earlier had abnormally high gradient values.

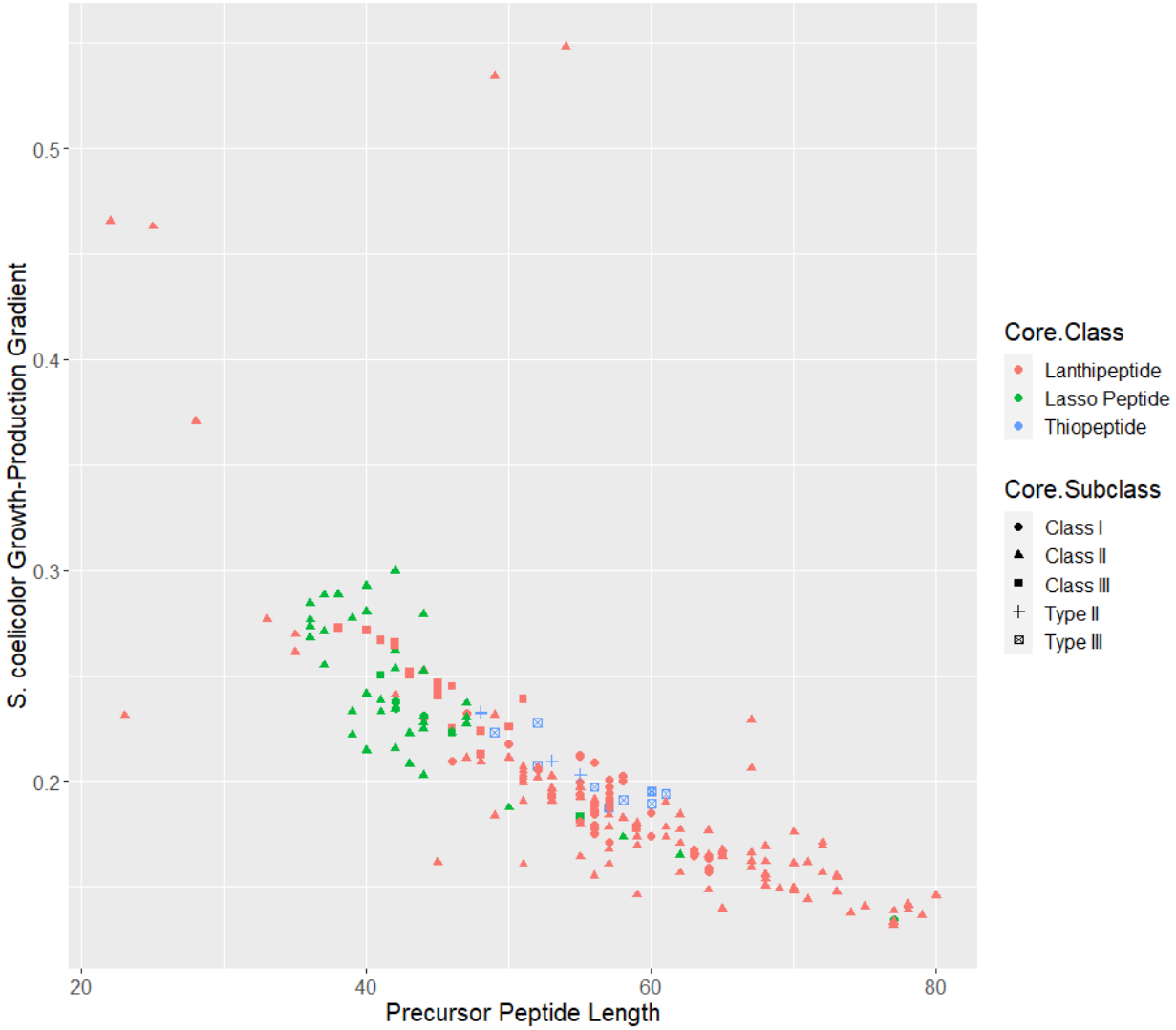


Figure 4.3.5: Scatter plot of pathway growth-production gradient in Sco-GEM plotted against precursor peptide length. Only pathways with mature products were included and type I and II thiopeptides with secondary macrocyclization were omitted.

In addition to peptide length, amino acid composition of the precursor peptide may also affect the metabolic burden as different amino acids have different total metabolic costs. To see whether the variance in RiPP product gradient values can be explained by this, the ARMRiPP *translation* reaction function was run for peptides of varying length, either composed entirely of glycine or methionine, and the associated gradient was calculated using the outputted reaction's objective value. Glycine was picked as the least metabolically costly amino acid, and methionine as the most (110). If the gradient values for the RiPP products lie outside the range between the "best case" glycine value and the "worst case" methionine value, then it would indicate that the metabolic burden of RiPP pathways is not only explained by the variances in amino acid composition of the precursor peptide. We can observe that the majority of RiPP pathway gradient values lie within this range, skewed towards the methionine values (Figure 4.3.6). The thiopeptides with secondary macrocyclization clearly have values outside of the methionine-glycine range, but besides them there was also found one pathway with a value below the range and one above.

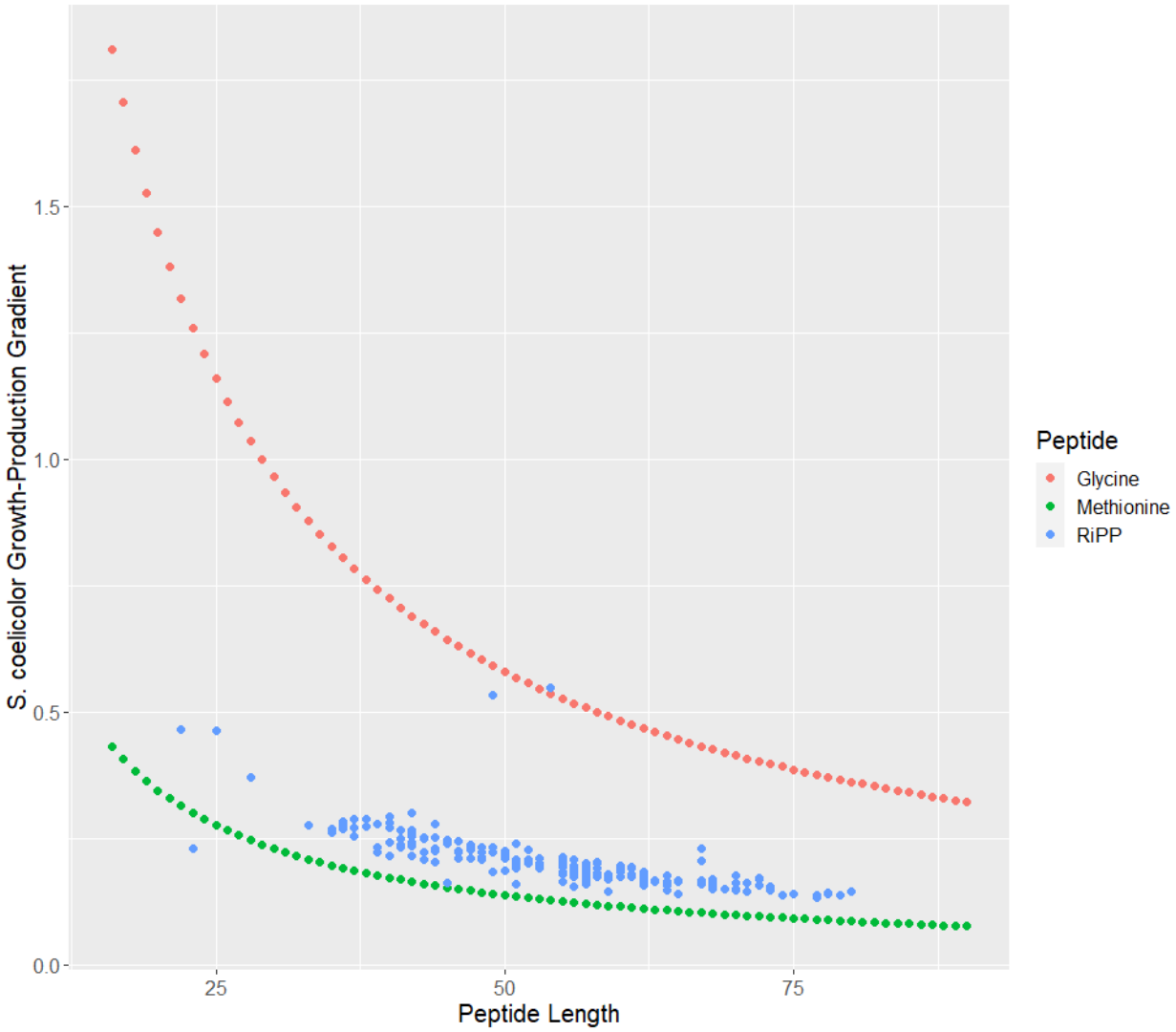


Figure 4.3.6: Scatter plot of growth-production gradient for mature RiPP products plotted against precursor peptide length. For comparison, the growth-production gradient for the ribosomal synthesis of peptides consisting of only glycine (red) and only methionine (green) are shown. For the RiPP values type I and II thiopeptides with secondary macrocyclization were omitted.

Metabolic Burden in GEMs of Different Phylogenetic Groups

Using CarveMe, 91 GEMs from different organism's genomes were created for which 113 BGCs have as their source organism. From these 113 BGCs the ARMRiPP script identified 172 RiPP precursors and reconstructed pathways for them. Each pathway was inserted into a copy of their source organism CarveMe reconstructed GEM, and FBA was performed on each extended GEM for RiPP product maximization and growth maximization. Of the 172 pathway-extended GEMs only 32 could grow using the Sco-GEM media, and 30 of those could produce the RiPP product. Using the definition for mature RiPP products, either containing a thio-ether cyclization step, macrolactam step or macrocyclization step, the 30 GEMs were checked for mature products and it was found that they all fit the definition. However, as these 30 GEMs can contain pathways for different cores of the same BGCs, and therefore the same organism, there were only 16 distinct CarveMe GEMs. Observing Figure 4.3.7, we see the RiPP subclass distribution for these 30 extended CarveMe GEMs. Similarly to the extended reference model distribution, over ¾ of the RiPP products were lanthipeptides, but in this case more evenly distributed among its subclasses. We can also see that no GEMs with thiopeptides in them are present, either due to the model not being able to grow or due to not being able to produce the thiopeptide product.

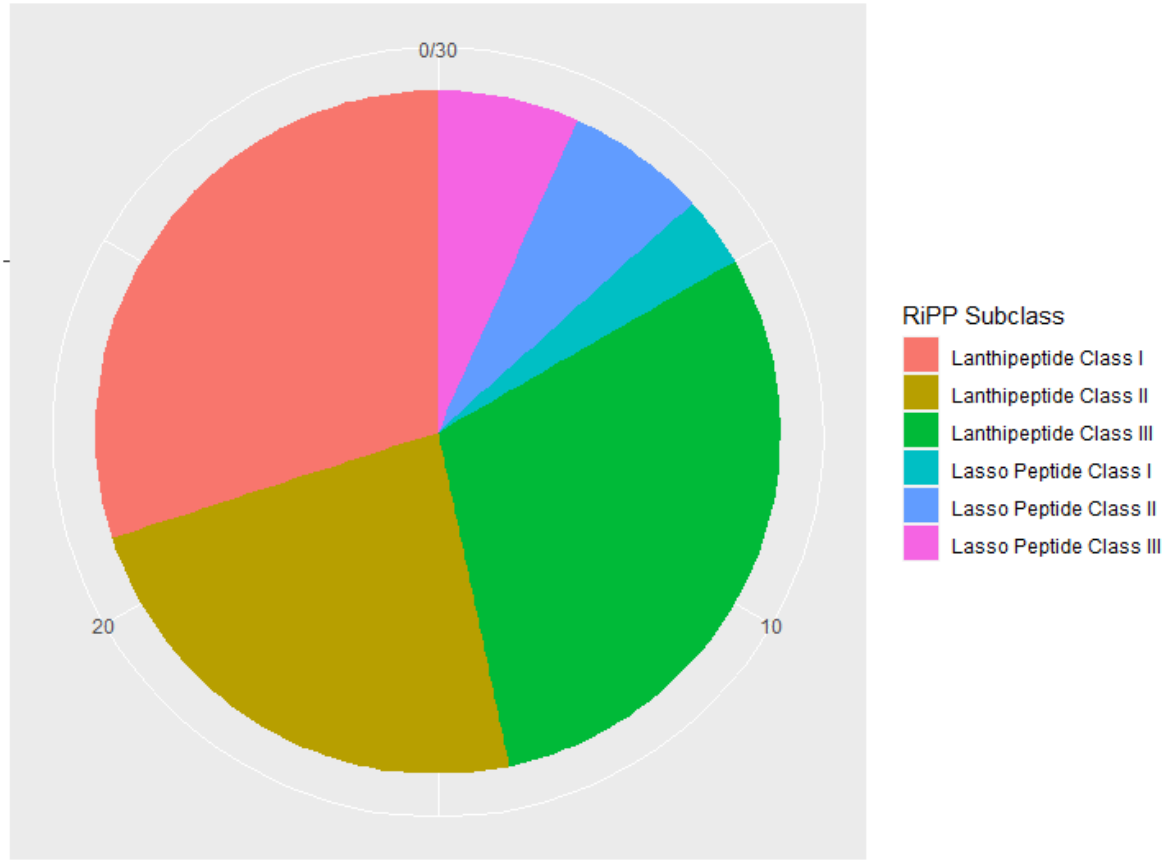


Figure 4.3.7: Pie chart of RiPP product subclass distribution for pathways in extended CarveMe reconstructed GEMs. Only pathways which were in models that had growth and managed to produce the RiPP product have been included.

Investigating the generated phylogenetic tree for the organisms of these 30 pathway-extended reconstructed GEMs, it was found that they all belong to the Terrabacteria taxon at the first clade level. At the second level it was observed that around ⅔ of the organisms were of the Bacillota phylum, a little under ⅓ were of the Actinomycetes phylum and one organism was of the Cyanophyceae class (Figure 4.3.8). At the third level the organisms are grouped by a mix of classes, orders and families, the largest groups being the Bacilli class, the Eubacteriales order and the Pseudonocardiaceae family. One organism, *Streptomyces cattleya*, was not able to be categorized into the common tree, but is formally considered part of the Terrabacteria taxon as well as the Streptomycetaceae family. Additionally, the organism *Catenulispora acidiphila* was only categorized until clade level 2, and belongs to the Actinomycetes phylum.

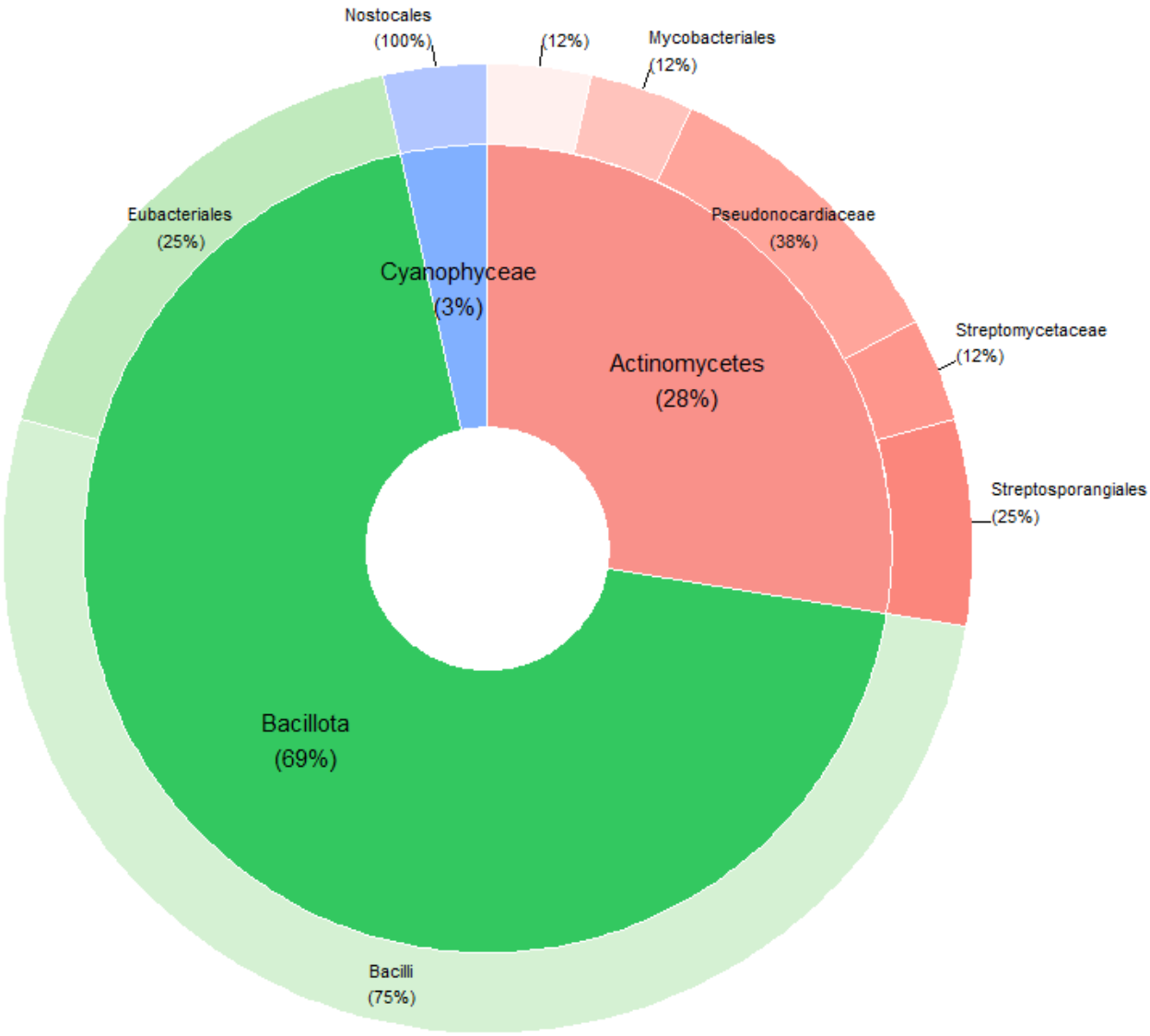


Figure 4.3.8: Pie-Donut chart for clade level 2 and 3 of the organisms used in the 30 pathway-extended CarveMe GEMs. Areas without names are organisms which were not classified at that clade level.

The growth-production gradient values of the pathway-extended CarveMe GEMs were grouped by the second and third clade level. Of the major groups in the second clade, Bacillota had a slightly higher average gradient than the Actinomycetes, but they both had a large amount of variance (Figure 4.3.9). The Cyanophyceae group only consisted of one organism, and had a gradient value at about the same as the Actinomycetes average. The unclassified *S. cattleya* had a much lower gradient when compared with the other groups. The major groups at the third clade level, Bacilli, Eubacteriales and Pseudonocardiaceae, all had similar and high gradient values (figure 4.3.10). Bacilli however had a much larger variance than the others and Eubacteriales a

much lower. We may also notice that three groups had much lower gradient values than the others, which includes *S. cattleya* again but also the Streptomycetaceae and the Streptosporangiales groups.

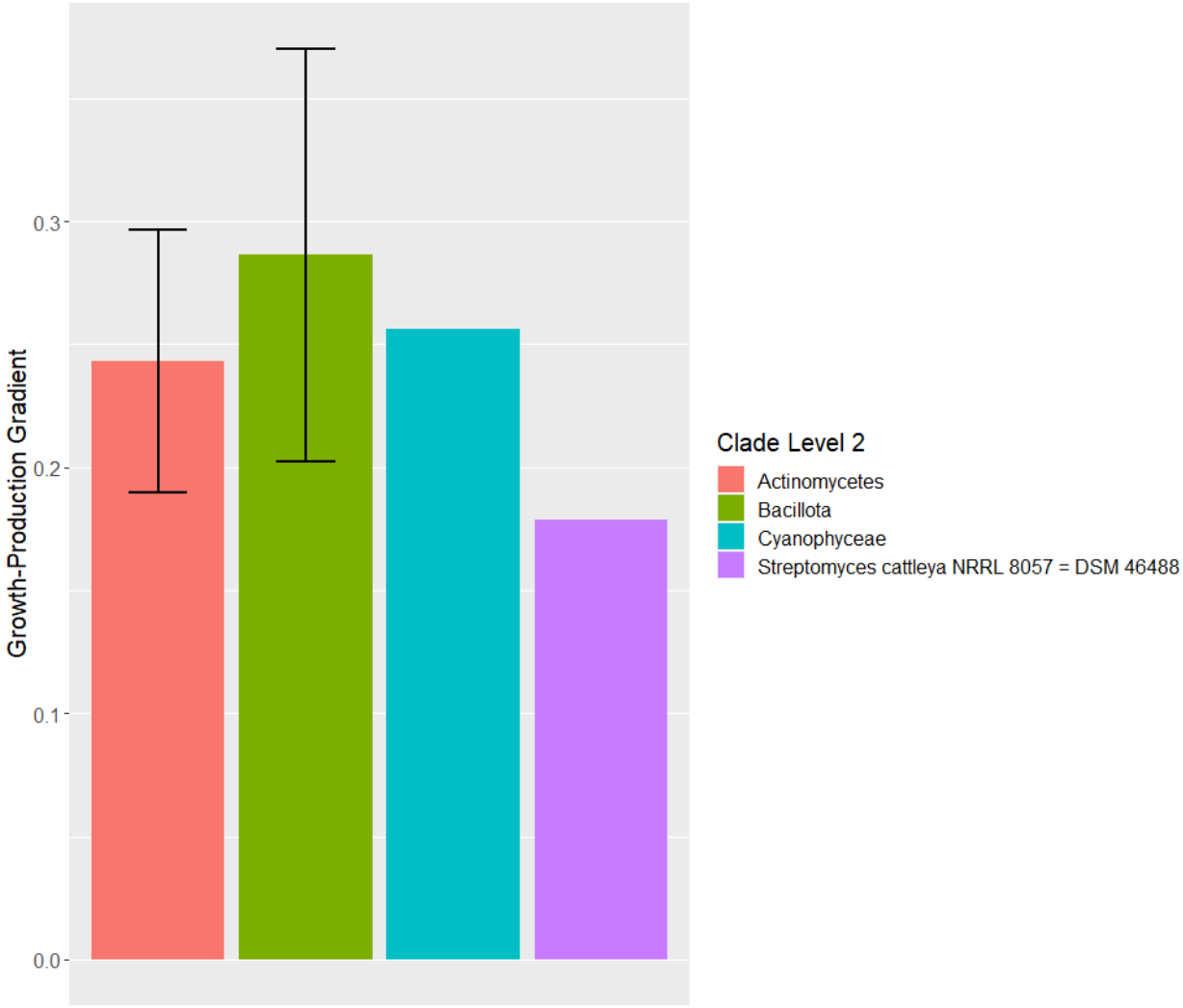


Figure 4.3.9: Bar chart of average growth-production gradient values for phylogenetic groups at the second clade level of the tree. The error bars represent the standard deviant value, and groups without error bars consist of only one member.

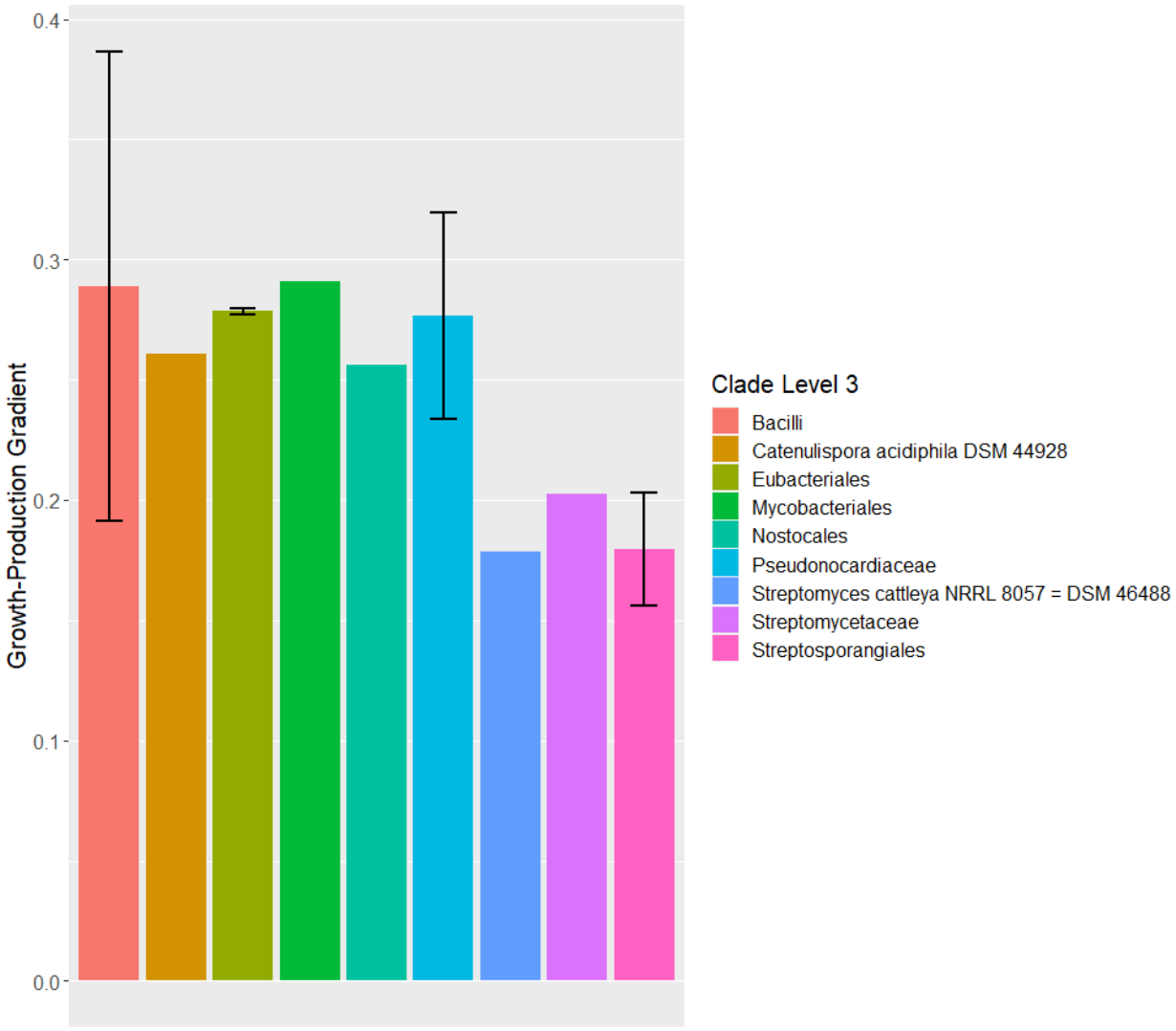


Figure 4.3.10: Bar chart of average growth-production gradient values for phylogenetic groups at the third clade level of the tree. The error bars represent the standard deviant value, and groups without error bars consist of only one member.

Comparison of Reference and CarveMe Reconstructed GEMs

The comparison of the pathway-extended reference GEMs and CarveMe GEMs was done by matching the results from the reference GEM dataset with the correct BGC ID and core in the CarveMe GEM dataset. This meant that the amount of data for comparison was limited to the 30 reconstructed pathways of the extended CarveMe GEMs. For each reconstructed pathway the growth-production gradient values of the reference GEM and CarveMe GEM were compared (Figure 4.3.11). Most of the pathways had higher gradient values in the reconstructed GEM of their source organism than in the reference GEM, indicating a lower metabolic burden for the production of the RiPP compounds. The mean gradient value for the native host GEMs was 0.27 ± 0.077 and for the reference GEMs it was 0.22 ± 0.057. A t-test was performed between the two sets of gradient values, which gave a p-value of 0.0084, indicating with high statistical confidence that there is a difference between the groups of values.

However, there are two notable exceptions which we see above the identity line in Figure 4.3.11; one is a blue cross and the other a purple dot. By investigating the legend we see that the blue cross is a single *Streptomyces cattleya* species, while the purple dot indicates that it belongs to the third clade level in the common tree called Streptomycetaceae. These results are interesting given that *Streptomyces cattleya* belongs to the same genus, *Streptomyces*, as *Streptomyces coelicolor*, while the other pathway was found to be sourced from the species *Streptomyces humidus*, also of the same genus. Their MIBiG IDs are BGC0001539 and BGC0002307 respectively. Considering that both the Streptomycetaceae group and *S. cattleya* were shown to have much lower gradients than most of the other phylogenetic groups in the phylogenetic analysis, it may not be surprising that their RiPP pathways performed better in another host. However the Streptosporangiales group also had similarly low gradient values in the phylogenetic analysis, but their associated pathways both lie under the identity line. The *S. cattleya* RiPP compound was found to be a class II lasso peptide with a precursor sequence length of 44, while the *S. humidus* RiPP compound was a class I lasso peptide with a length of 42. The RiPP compounds in the dataset had an average precursor length of 50 ± 11 and a median of 45. In Figure 4.3.12 we can note that there are class II and III lasso peptides present below the identity line as well.

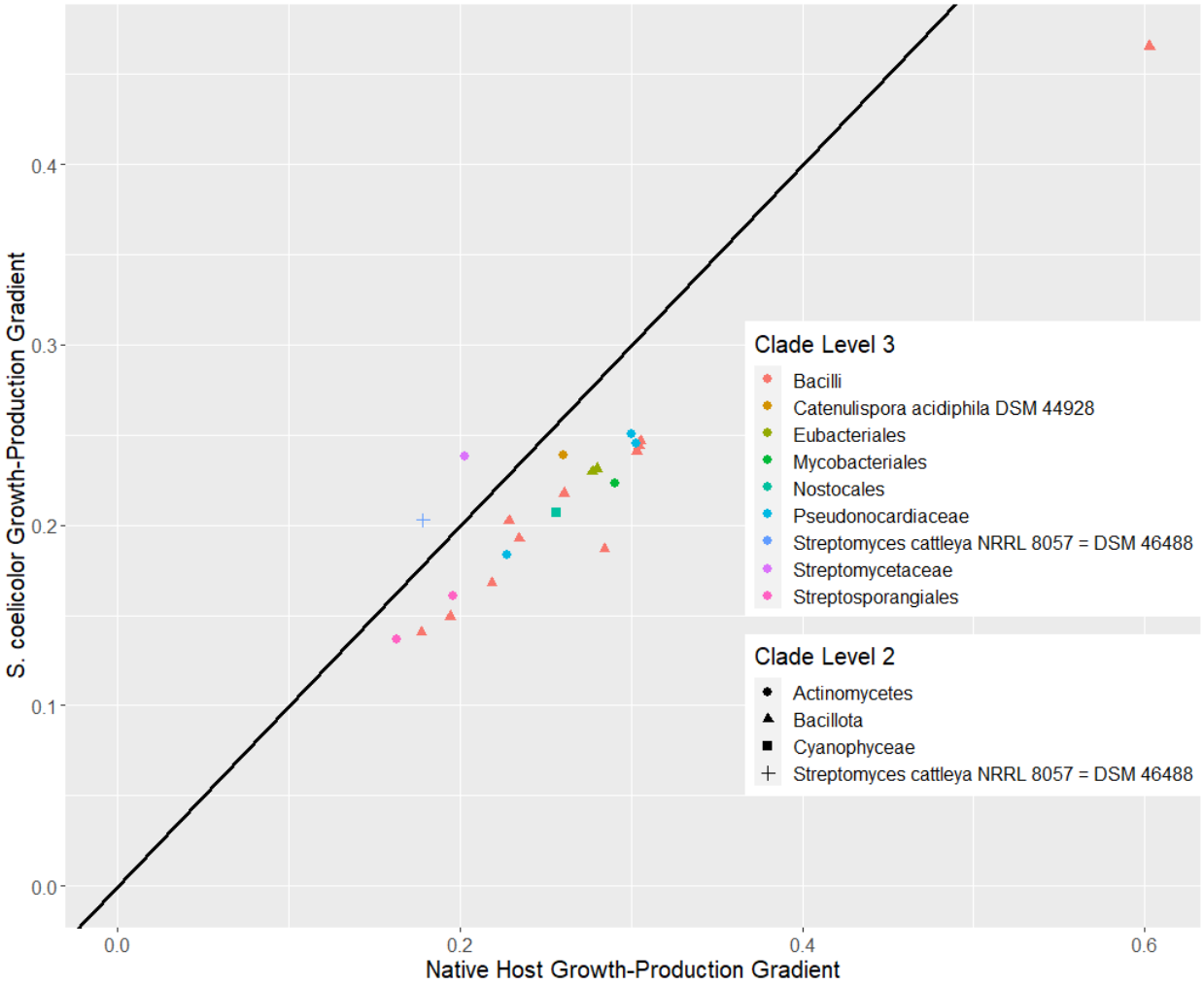


Figure 4.3.11: Scatter plot of growth-production gradient values of pathway-extended CarveMe reconstructed GEMs against gradient values of pathway-extended Sco-GEM reference GEMs. Each point represents the gradient value of a single reconstructed BGC pathway. Points above the identity line indicate that the pathway has a lower gradient in the reference GEM than in its reconstructed source organism GEM, while points below indicate the opposite. The shape of the dots indicate its identity in the second clade level for the common phylogenetic tree between the organisms, while color indicates its third clade level. The red color indicates no association at that clade level.

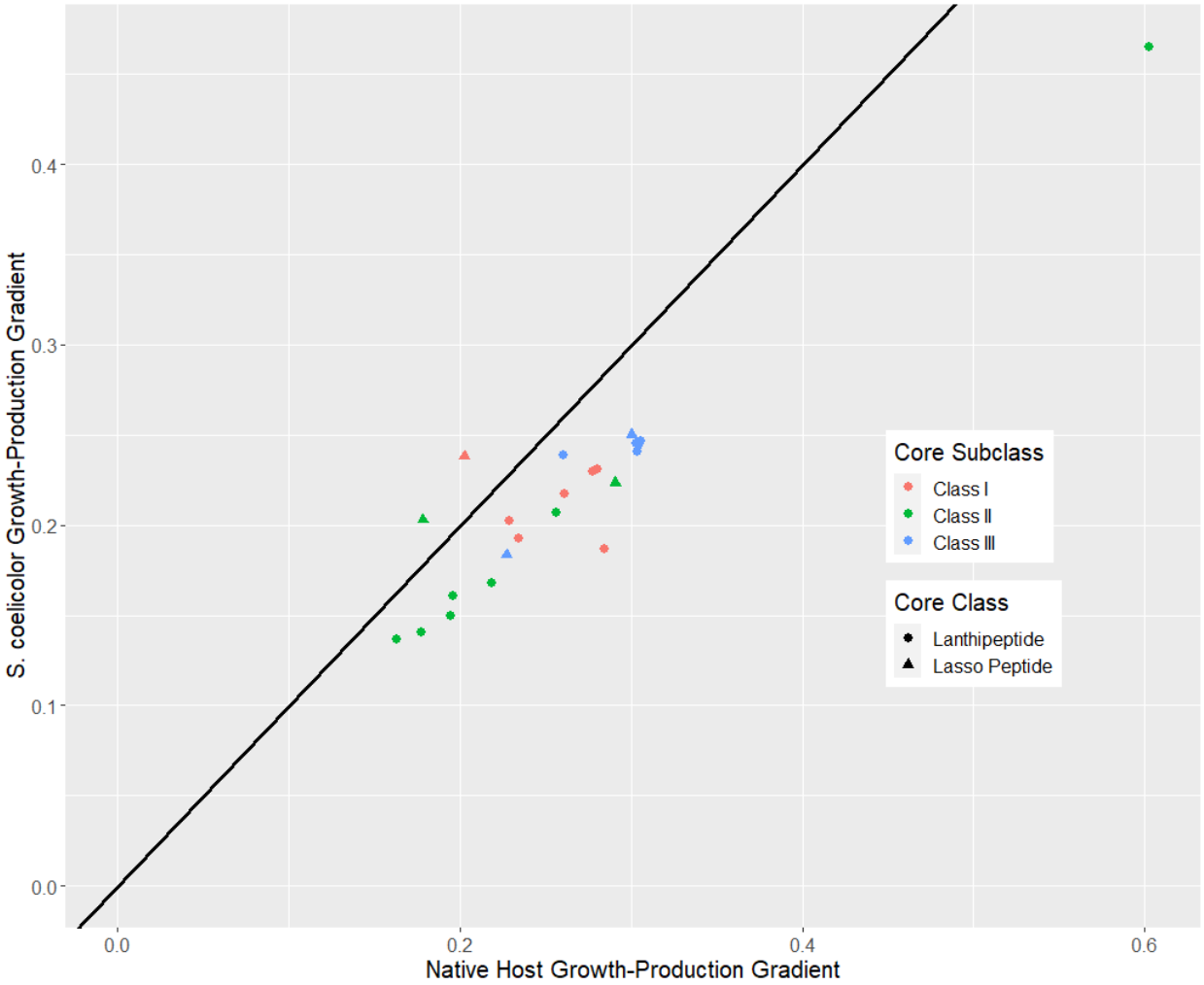


Figure 4.3.12: Scatter plot of growth-production gradient values of pathway-extended CarveMe reconstructed GEMs against gradient values of pathway-extended Sco-GEM reference GEMs. Each point represents the gradient value of a single reconstructed BGC pathway. Points above the identity line indicate that the pathway has a lower gradient in the reference GEM than in its reconstructed source organism GEM, while points below indicate the opposite. The shape of the dots indicate its RiPP class, while color indicates its RiPP subclass.

Metabolic Burden in Heterologous Hosts

We might expect from what was seen in the comparison between the extended reference and CarveMe GEMs that there would be a general skew towards higher gradient/lower metabolic burden for the native GEMs of the RiPP pathways, with possible exceptions for hosts of the same phylogenetic group. This was investigated for the set of pathway-extended GEMs with heterologous hosts by comparing the growth-production gradient values for the RiPP pathways in heterologous hosts with their respective native host gradient values. Observing Figure 4.3.13, we would expect to generally see more data points below the identity line, and that more of the data points above the line would have the same phylogenetic clade between native and heterologous host.

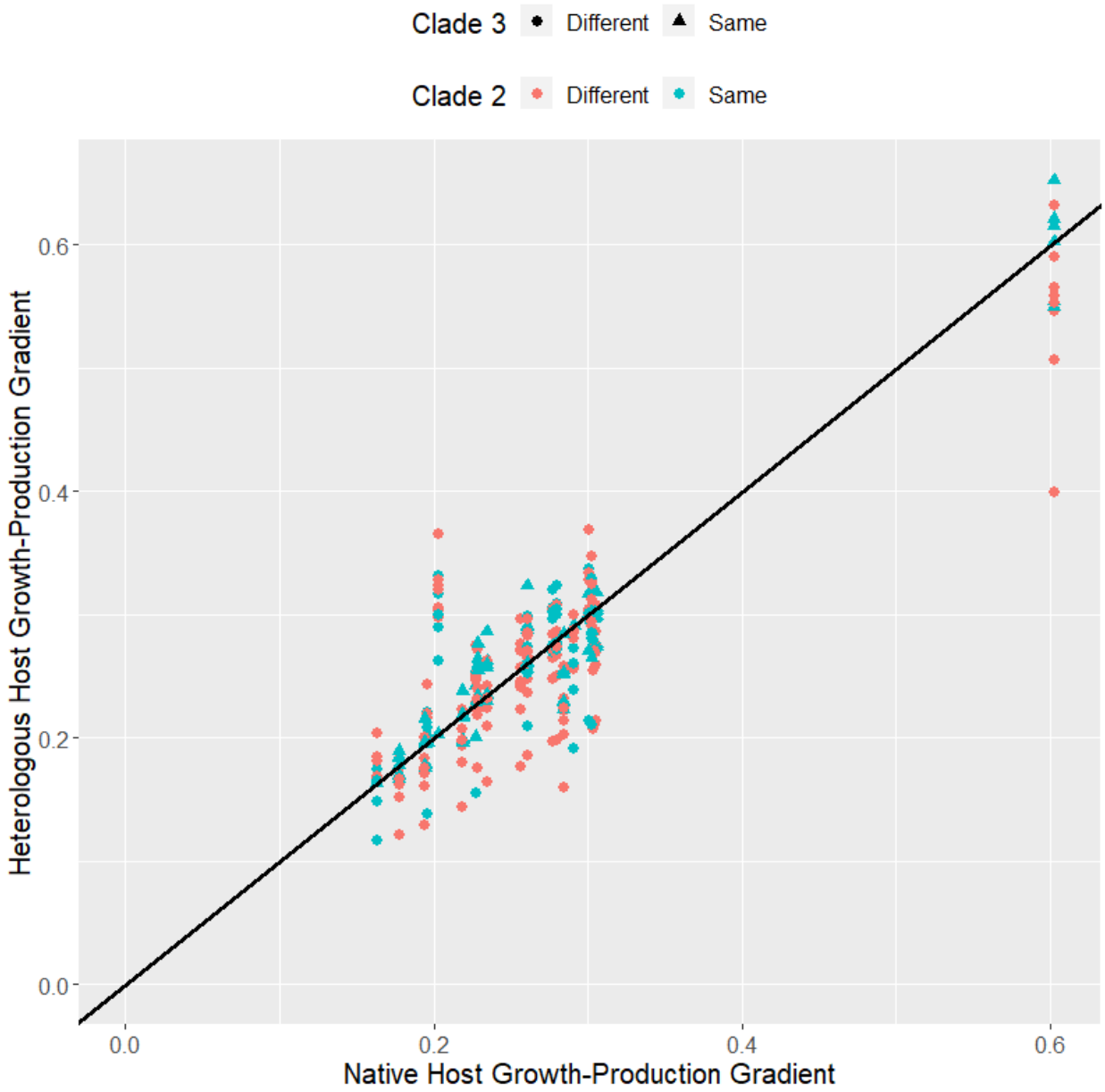


Figure 4.3.13: Scatter plot of growth-production gradient values for pathways in multiple heterologous host GEMs plotted against the growth-production gradient in their source organism GEM. Dots above the identity line indicate higher gradient values in the host GEM, while dots below indicate higher gradient values in the source organism GEM. Dot color indicates whether the source and host organism belong to the same second clade of the common tree, while dot shape indicates whether they belong to the same third clade. Each column of dots indicates the same pathway in different hosts.

Since there does not appear to be an observable pattern in the scatter plot that follows the expectation, the gradient value means for the data points grouped by similar and different clade levels were instead considered (Table 4.3.1). For all of the clade similarity groups, a t-test was

performed between the native host and heterologous host gradient values. We see that there is generally a slightly lower gradient value mean for the heterologous hosts when compared to the native hosts for all of the groups, indicating as expected that the metabolic burden is generally higher for heterologous hosts. However, the gradient means all have high variances, particularly the native GEM gradients of the same third clade group, which together with the high p-values of the t-tests gives us less statistical confidence to say that the heterologous and native host gradient value means are different. We could however expect the difference between gradient values to be the greatest for the group of data points with both different second and third clade levels (4th row of Table 4.3.1), and we see here that the p-value is much closer to 0.05 than the other groups.

If we consider that more phylogenetically similar hosts would have less metabolic burden than dissimilar ones, we might expect there to be differences between the heterologous host gradient values of the group with the same third clade level and the heterologous gradient values of the group with different third clade level (3rd and 4th row of Table 4.3.1). Performing a t-test for these two sets of gradient values we get a p-value of 0.05. Furthermore, we would expect there to be less of a difference between the native gradient values of these same two groups, and performing a t-test for the sets of native gradient values gives us a p-value of 0.24. It was observed that the native gradient values for the group of pathways with hosts of the same third clade had a larger variance than the other sets of gradient values. Although the statistical confidence is not great, in general these results seem to fit with the expectation that the RiPP pathways in heterologous host GEMs with similar phylogeny to their native host GEM have less of a metabolic burden.

Table 4.3.1: Table of mean gradient values of BGCs in their source GEMs and in heterologous host GEMs, grouped by similar and different clades. The first row of values are the gradient value means for the entire dataset. The significance value (p) from the t-test between native and heterologous gradient values of the same clade group (row) indicates if they are statistically different when $p < 0.05$.

Native GEM gradient	**Heterologous GEM gradient**	**Clade 2**	**Clade 3**	**Significance value (p)**
0.2733 ± 0.0755	0.2662 ± 0.0738	Same & Different	Same & Different	0.16
0.2727 ± 0.0763	0.2722 ± 0.0763	Same	Same & Different	0.94
0.2850 ± 0.0902	0.2792 ± 0.0887	Same	Same	0.61
0.2739 ± 0.0751	0.2608 ± 0.0712	Different	Different	0.06

It was observed that one of the RiPP pathways had a higher gradient value in every heterologous host (Figure 4.3.13). Investigating the data set this was found to be the class I lasso peptide pathway from *S. humidus* which we have seen in earlier chapters. Treating this as an outlier, the analysis of the gradient values by phylogenetic groups was redone (Table 4.3.2). We can observe that the p-values for the native and heterologous host gradient values of the phylogenetic similarity groups are generally lower. For the group containing all the pathways except the outlier (1st row of Table 4.3.2) the p-value is now significant, indicating with greater statistical confidence that there is a difference between the gradient values. The same applies for the group of pathways with hosts of the same third clade. There was no change to the p-value for the pathways with similar hosts at the third clade level, however the p-value for the pathways with similar hosts at the second clade level was now much lower, and the same as the previous group. As before, t-tests were performed for the gradient values of the pathways in hosts of the same third clade and different third clade (3rd and 4th row in Table 4.3.2). For the heterologous gradient values the p-value was 0.03, which is lower than before, and for the native values the p-value was 0.32, which is higher than before. Generally, by removing the outlier we see that the results more significantly fit the expectation that the RiPP pathways perform better in their native hosts, and that they perform better in heterologous hosts of the same phylogenetic group as its native host.

Table 4.3.2: Table of mean gradient values of BGCs in their source GEMs and in heterologous host GEMs without the *S. humidus* outlier, grouped by similar and different clades. The first row of values are the gradient value means for the dataset without the outlier. The significance value (p) from the t-test between native and heterologous gradient values of the same clade group (row) indicates if they are statistically different when $p < 0.05$.

Native GEM gradient	**Heterologous GEM gradient**	**Clade 2**	**Clade 3**	**Significance value (p)**
0.2759 ± 0.0757	0.2649 ± 0.0745	Same & Different	Same & Different	0.03
0.2756 ± 0.0765	0.2716 ± 0.0774	Same	Same & Different	0.61
0.2857 ± 0.0902	0.2798 ± 0.0888	Same	Same	0.61
0.2761 ± 0.0751	0.2590 ± 0.0714	Different	Different	0.01

5 Discussion

In light of the results, we will consider the degree to which the methods of this thesis manage to achieve its aim, which is to reconstruct the metabolic pathways of RiPP BGCs and showcase the method in the context of microbial ecology to explore if bacteria has optimized its metabolism to reduce the metabolic burden of producing these costly secondary metabolites. First we will take into account the results of the RiPP compound structure prediction, which was used to check the accuracy of the pathway reconstruction to arrive at the correct products during the development of the methods. Then we will consider the pathway reconstruction as a whole and discuss the observed results as well as the limitations of the different tools used to achieve it. Lastly we will look at what the results of the metabolic burden analysis appear to indicate, and how they may have been influenced or skewed by different factors.

5.1 Structure Prediction

We chose a Tanimoto score of 0.85 as a boundary for whether to consider a high structural similarity between the predicted RiPP compounds and their PubChem structures. Using this boundary we observed that 42 of the 57 chosen RiPP compounds had ARMRiPP predicted structures over 0.85 Tanimoto score when compared with their PubChem structures, indicating high structural similarity. When comparing with the structural similarity results of the unmodified core peptide structures, ARMRiPP showed a mean improvement of 38% for its predicted structures.

Although the ARMRiPP structure prediction performed well overall, the performance differed between the three tested RiPP classes. The thiopeptide structures performed the worst, with only 11 out of the 22 structures above the 0.85 cutoff and a mean Tanimoto score of 0.85. There were likely multiple causes for this, relating to both the implementation in ARMRiPP, limitations of current knowledge about their biosynthesis, and incomplete antiSMASH gene annotation. Only the central biosynthesis and secondary macrocyclization of thiopeptides was modeled in ARMRiPP, despite thiopeptide compounds being known for having additional tailoring reactions. These tailoring reactions can often be highly specific to certain thiopeptide groups or even specific compounds, and can have poorly understood chemical mechanisms and genetic origin. Perhaps as a result of this, during the investigation of the thiopeptide genbank files it was challenging to find antiSMASH annotations which reliably were associated with certain thiopeptide tailoring reactions. In the cases where such associations were found, they accounted for such a small group of thiopeptide compounds and compared to secondary macrocyclization their metabolic and structural impact was much lower. For these reasons thiopeptide tailoring reactions were prioritized lower and not implemented. Another missing implementation for

thiopeptides was the lack of the oxazole formation reaction. The thiopeptide groups that contain oxazoles are known for having them in their central nitrogen heterocycle substitution pattern, and so the lack of these greatly affects the modeling of the macrocyclization event in thiopeptides, which has large effects on the predicted end structure. The lack of oxazole implementation was due to challenges in determining which serines and threonines would react, as particularly serine plays a vital role in the later biosynthesis, and challenges in associating the antiSMASH functional annotation with the presence of enzymes which facilitate oxazole formation. Lastly, the general assumptions made for the reaction modeling of the thiopeptide reactions can affect the accuracy of the structure prediction in cases where they do not hold true, such as the assumption that there is only one pair of possible reacting motifs for the macrocyclization reaction or that the 4π component is always flanked by azole groups on each side. Despite performing the worst in general, the structure prediction of thiopeptides did show the largest improvement when compared to their unmodified core peptide structures, which is likely a result of the large amount of modification reactions that the thiopeptide core undergoes.

The predicted ARMRiPP structures for lanthipeptides performed better, with 19 out of 21 above the 0.85 threshold and a mean Tanimoto score of 0.92. Potential sources of inaccuracies for the lanthipeptide structure prediction are similar as the ones for thiopeptides. Lanthipeptides are known to have more tailoring reactions outside the two class I reactions that were implemented, but they tend to be fewer per compound than the ones in thiopeptides. Additionally, the association of certain antiSMASH annotations with tailoring reactions proved to be unreliable. A much larger source of inaccuracy is likely from inconsistencies in the assumptions made around the amount of dehydrated residues and the cyclization pattern. As it was assumed in the methods that every lanthipeptide serine/threonine gets dehydrated, and as it is known from literature that they do not (32, 103, 104), then within the reaction modeling of ARMRiPP there will be the potential of lanthipeptide products with thio-ether rings between residues which should not be possible. Furthermore, the cyclization assumption that it occurs in a C-N direction with rings above a minimum size does not hold true in every case. This assumption was made particularly due to the limited knowledge on the relationship between the LanC cyclase and ring topologies, which is known to differ both across and within lanthipeptide subclasses, and it was therefore challenging to implement a more precise heuristic to guide the ring formation in lanthipeptides.

Lasso peptides was the RiPP class which ARMRiPP performed the best for, with 12 out of 14 above the 0.85 threshold and a mean Tanimoto score of 0.96. It was however the RiPP class with the lowest improvement over the unmodified core peptide, with only 3% compared to lanthipeptide's 15% and thiopeptide's 170% improvement. When taking into account the low amount of modifications done during the lasso peptide biosynthesis, the low improvement is expected as the backbone from the unmodified core peptide remains intact to a larger degree than the other RiPP classes. Nonetheless, the accuracy is not perfect, and similarly to before this may be caused by the ARMRiPP implementation and limitations both in antiSMASH annotation and

literature. With the current research of lasso peptide biosynthesis, few reactions beyond the central modification reactions have been identified, and these are often only minor alterations to the overall structure (111). Additionally, antiSMASH had very limited lasso peptide tailoring reaction annotations, and so they were not implemented. The assumption made in the ARMRiPP methods that the macrolactam formation proceeds with the largest possible ring does not hold true if for example an acid residue is present further downstream than the actual reacting residue. Furthermore, the expectation that there is always only one pair of cysteine residues for class III and IV, and always two pairs for class I lasso peptides, may not hold true either and so it is not guaranteed that ARMRiPP predicts the correct product.

ARMRiPP was compared with PRISM, a state of the art secondary metabolite structure prediction tool. Comparing the ARMRiPP structure prediction results with PRISM's median results, we saw that they generally performed similarly when considering the amount of structures above the 0.85 threshold, 42 for ARMRiPP and 41 for PRISM. However when comparing mean Tanimoto scores we saw that ARMRiPP had a mean positive difference of 17% from the PRISM results, which was due to the large number of failed structure predictions by PRISM. When grouping the results by RiPP class we found that PRISM performed on par or better than ARMRiPP for thiopeptides, while worse for lanthipeptides and lasso peptides. When comparing with the PRISM max results we saw that it performed better than ARMRiPP overall using the 0.85 threshold, with 45 structures. However, ARMRiPP still had a mean positive difference of 13% over the PRISM max results, which again was largely affected by the group of structures which PRISM did not manage to predict.

It is important to note the distinctions between PRISM and ARMRiPP. When it comes to structure prediction, the differences between the two greatly affect the use-cases. PRISM is designed to generate all feasible compound structures which arise from the features it can identify in a BGC, and so it is for example useful in cases where you have a known chemical structure or substructure that you wish to screen against a large number of BGCs for the best fitting candidates. However, for forward prediction where the goal is to predict the structure of the end product from the annotated BGC, one would not know which of the PRISM predicted structures is the most relevant. ARMRiPP was on the other hand designed with literature-based rules, heuristics and assumptions to detail the most likely biosynthesis pathway and predict the most likely end-point structure. The results reflect these different use-cases, as the PRISM median value gives an idea of the expected accuracy from picking a structure without preference, and the PRISM max value gives the accuracy for the existence of at least one highly accurate structure in the generated set. Another distinction to note is that ARMRiPP is fully open-source and can be analyzed and used by anyone through accessing the github, while PRISM has a more protected source-code and can have limited use for individual users.

A limitation of the ARMRiPP structure prediction is that it evaluates the structures for only three out of more than 20 RiPP classes During development of the ARMRiPP script, lanthipeptides were chosen as the first implemented RiPP class as it is currently the largest identified group of RiPPs, and had extensive available literature on its biosynthetic mechanisms. The choice of thiopeptides and lasso peptides later in development was done after analyzing the RiPP classifications of the predicted cores found in the set of antiSMASH annotated BGC files from MIBiG, where they had the second and third most identified cores. The only remaining RiPP classification found with identified core sequences was sactipeptides, however there were only four of them. Due to low priority the sactipeptide RiPP class was therefore not implemented. As can be seen in the methods, it is crucial for the structure prediction of ARMRiPP that the annotated RiPP genbank file contains predicted peptide sequences for the RiPP precursor. Despite the set of genbank files from MIBiG containing BGCs of many different known RiPP classifications outside of the four mentioned, none of these had predicted precursor sequences and could therefore not be implemented. Furthermore, lanthipeptides, thiopeptides and lasso peptides together make up a majority of currently identified RiPP clusters in MIBiG, with 59% (112). Additionally, in a study of over 60 000 prokaryotic genomes where over 30 000 RiPP clusters were predicted using PRISM, it was found that 49% of them were identified as lanthipeptides, 4.8% were identified as lasso peptides and 1.6% were identified as thiopeptides (113). This does not however mean that the various smaller RiPP classes are of less interest, and with the continued research and improved annotation we believe that ARMRiPP in the future could be extended to include more RiPP subclasses.

Although the boundary of 0.85 in Tanimoto score is arbitrary, it was not picked without reason. In an influential study from 1996 where in-house sets of active compounds were compared by using the similarity value, it was found that molecules which had a Tanimoto score of 0.85 or above had a high probability of having the same activity (102). However, in later years this has come to be referred to as the “0.85 myth” (93), as later studies revealed that in reality far fewer compounds sharing this definition of similarity reliably shared the same activity (114, 115), having closer to a 30% probability in some assays (115). The larger point of this when it comes to the structure prediction results of the thesis is to clarify the potential of these results in predicting RiPP compound activity, or in doing active structure searching. One could consider that with a 30% probability of similarities in activity, functional structures of interest could be searched among a large library of generated results from ARMRiPP to find good candidates at a faster rate than at random. It is however important to note that the limitations of the Tanimoto similarity method is that it only considers arbitrary groups of similarity across the entire 2D structural composition, and so activity due to specific local or global structures will not reliably be taken into account if the value is not close to 1. Furthermore, Tanimoto similarity does not take into account stereochemistry and 3D conformations, which can have large effects on a compound’s biological activity (93). In the case of the thesis results, the structures with a value close to 1 only accounted for a small fraction of the structures tested.

Within this thesis, a subset of the structures present in the structure prediction results were used to guide the development of the code and the estimation of certain parameters such as thio-ether ring sizes. Given the use-case of the Tanimoto similarity, we would expect it to perform well in this case, as we were most interested in checking the correct presence of modifications for establishing its metabolic accuracy, and less interested in whether it fit the correct global structure. However, it is important to note that this could contribute to a skew in the observed results for the structure prediction. Not only was part of the results data used to develop ARMRiPP, but in general the RiPP compounds with known PubChem structures can be expected to be more well characterized. As such, the assumptions based on the literature for these more well characterized RiPP compounds may also be skewed so that they have higher accuracy for these cases, while not for less characterized RiPP compounds. The real accuracy in predicting RiPP compound structure could therefore be lower.

5.2 Pathway Reconstruction

The results from the metabolic pathway prediction reveals several important points about the pathway reconstruction. Firstly, we see in the Lacticin 481 pathway that the general assumption of every serine and threonine residue in the RiPP dehydration reactions being dehydrated does not always hold. We then see that the missing implementation of the thiopeptide tailoring reactions in the Thiomuracin A pathway cause the FBA results of the reconstructed pathway to be further away from the literature pathway results than only modeling the ribosomal synthesis. With the Felipeptin A1 pathway we see that the discrepancy in precursor peptide length heavily influences the production potential from the FBA and with the Felipeptin A2 pathway we see the difference a single amino acid has on the production yield.

Creating the general assumptions for the reaction modeling and pathway construction was done in part due to limitations of the current RiPP literature and in part limitations of the methods used. An example of this is the type II thiopeptide designation used by antiSMASH. From literature we know that the mechanisms for series a-c thiopeptides are poorly understood (33, 75), and additionally antiSMASH has no universal annotations for detecting series a and c thiopeptides, and so the decision for having every type II thiopeptide be the initial reaction step, series b, was made both due to literature and antiSMASH limitations. In some cases, such as the missing implementation for thiopeptide tailoring reactions, the mechanisms could be well enough known in literature to be implemented (33), but antiSMASH again had no present universal annotation for such cases. In the case of the two class I lanthipeptide tailoring reactions, antiSMASH did have annotations for each of the two reactions, but they were found to not be reliable during development. Instead the gene identifier and related SMCOGs were used to determine the presence of these reactions, but this works only for well characterized clusters

with the correct gene identifiers. For the reconstruction to include these tailoring reactions in uncharacterized clusters it would have to be made less strict.

With the exception of the Felipeptin A1 pathway, the FBA RiPP production rate error values from the predicted pathways were very low when compared to the literature pathways, and even when taking only the ribosomal synthesis reaction into account. In the Felipeptin A1 pathway we saw that the longer precursor predicted by antiSMASH in the reconstructed pathway caused a much larger error from the literature pathway than any other factor. Furthermore, in both the Thiomuracin A and Felipeptin A2 pathways there was a difference of a single amino acid from the literature precursor sequence, and both of these had a substantially larger error than the Lacticin 481 pathway which had the correct precursor length. This indicates that the production potential of the RiPP compounds is heavily influenced by the RiPP precursor peptide length.

As mentioned earlier, the ARMRiPP reconstruction script was developed for mechanisms known from well characterized RiPP pathways, and it is therefore not as likely to get the same accuracy for the reconstruction of uncharacterized RiPP pathways. The results of the pathway reconstruction accuracy could also be skewed by the fact that the literature pathways ARMRiPP was compared with necessarily must be well characterized. Furthermore, the assumptions underlying the stoichiometry in the pathways reconstructed by ARMRiPP were chosen from effects observed for a smaller number of well characterized RiPPs, which may not be the case for RiPPs that have yet to have their structures and biosynthetic mechanisms elucidated.

5.3 Metabolic Burden of RiPPs

The ability to accurately reconstruct RiPP (and other BGC) pathways opens new opportunities across different fields. In this work, we chose to explore ARMRiPP's value in the context of microbial ecology, asking the question: Have species adapted their metabolism to reduce the cost of producing their secondary metabolites? If this is true, it indicates that the cost of producing the RiPP end product has been a significant factor relating with the evolution of its host.

Differences in the metabolic burden for the RiPP pathways could be found both when considering the different RiPP classes in the Sco-GEM as well as the different phylogenetic groups for the CarveMe reconstructed native host GEMs. Outliers were found in the thiopeptide class gradient values, which all had in common that they incorporated the thiopeptide secondary macrocyclization pathway. It was not definitely determined why this caused such a large increase to the gradient value, but it may have been due to the limited literature on the reaction mechanisms of the addition reactant pathways and the reaction steps not being properly mass balanced in the reconstructed pathways (33, 75). Outside of the thiopeptide outliers, we saw that the lasso peptide classes appeared to have generally higher gradient values. The class II

lanthipeptides appeared to have a much larger range of possible gradient values than the other subclasses, while class I and III lanthipeptides had less variance. A possible explanation for this lies in the observed trend of the precursor length affecting the gradient value. Here we saw that particularly class II lanthipeptides had a large range of possible precursor peptide lengths, and a large range of gradient values as a result of the trend. Furthermore, we also saw that lasso peptides generally have smaller precursor lengths than lanthi- and thiopeptides, which may explain why their gradient values appeared higher.

When considering the calculated lower and upper bounds for the metabolic burden of ribosomal synthesis, determined by varying peptide lengths of glycine and methionine, it was found that most RiPP pathways lie within this range, with a skew towards the methionine bound. However, two pathways were found outside this range, indicating that not all variance is due to differences in the precursor peptide amino acid composition. Due to the fact that several RiPP modification enzymes are known to act on most or all of specific amino acids in the core domain of the precursor, there is an added element to the metabolic cost from the core domain composition. For example, a high amount of cysteines in a thiopeptide core will have a large effect due to the metabolic cost of the thiazole forming reaction. Additionally, even though serine is considered a metabolically cheap amino acid (110), in both lanthipeptides and thiopeptides it has the added cost which comes from their dehydration reactions.

In the analysis of the metabolic burden for the different phylogenetic groups it was found that at certain clade levels there were apparent differences between the groups. As all of the considered CarveMe GEMs were of organisms from the same taxon, Terrabacterium, the distinct groups were found from the second clade level of the phylogenetic tree. At the second clade level it was observed that Bacillota had a higher average gradient than Actinomycetes, but with much more variance. At the third clade level it was clear that the source of the high variance in Bacillota was likely from the Bacilli class. This group was the largest at the third clade level, which may explain why the variance was so much higher than the other groups. In larger scale studies of RiPP native organisms Bacilli also appear overrepresented, and so it is reasonable that they would be so here as well (113). At the third clade level it could also be observed that particularly Streptomycetaceae (including *S. cattleya*) and Streptosporangiales had lower average gradients than the other groups.

A large limitation of the phylogenetic group results was the low number of CarveMe GEMs that had growth, which caused many of the groups especially at the third clade level to only consist of 1-2 members. Given how CarveMe functions, it can only do so much in cases where the genome annotation may be limited and it cannot be presumed that the organism will contain specific metabolically crucial reactions. Furthermore, it may be the case that certain organisms in reality would not be able to grow with the nutrients used in the FBA simulated media. The same media as Sco-GEM was used, a complex media with minimal glucose as the only carbon-source, and so

organisms which are dependent on other nutrients or carbon sources would not be expected to grow.

When comparing the CarveMe native host reconstructed GEMs with the Sco-GEM reference models it was found that the metabolic burden was on average lower for pathways in their native host GEMs, with two notable exceptions. These exceptions were found to both be part of the same Streptomycetaceae group. This group, as we saw earlier, had comparatively low gradient values indicating high metabolic burden for the RiPP pathway in their native host GEM. It could therefore be reasoned that it performed better in the Sco-GEM due to it being a more robust producer, and not due to their shared phylogeny. On the other hand, the Streptosporangiales group also had a similar gradient value as the Streptomycetaceae, but the pathways of this group were lower in the Sco-GEM. Furthermore, when considering the RiPP class of the two exceptional pathways they were found to both be lasso peptides, but several lasso peptide pathways were also found with higher gradient values in their native host. A factor which could possibly skew the data in favor of higher gradient values for the CarveMe GEMs is the fact that Sco-GEM is a much more well curated model, and can therefore have a more realistic solution space than the CarveMe GEMs. Since the CarveMe GEMs have not been experimentally verified and may possess reactions from cryptic genes, their solution spaces may be less restricted than a well curated model like Sco-GEM.

This possible effect of RiPP pathways having a lower metabolic burden in hosts with similar phylogeny to their native host was investigated further using the set of heterologous host CarveMe GEMs. With statistical confidence, a general trend could be seen that appeared to indicate that the metabolic burden was higher in heterologous hosts when compared to native hosts, and more so if the heterologous host was of a different third clade level than the native host. When comparing the gradient values for pathways put into hosts of the same third clade with the gradient values for pathways put into hosts of different third clades, a statistically significant difference was found between the heterologous gradient values, and not for the native values. However, there are more confounding variables to this result compared with the previous, as the pathways compared are not the same. Also, a potential skew to this data is the fact that half of the pathways belong to Bacilli as seen earlier, which could weigh the average results for pathways of the same third clade toward the Bacilli gradient values. Bacilli was the phylogenetic group with the highest amount of variance in gradient values. This may be the cause behind the higher variance observed for the native host gradient values of the group of pathways with hosts of the same third clade.

Nevertheless, the results follow from the question asked at the beginning of this chapter; that certain groups of species seem to have adapted their metabolism to more efficiently produce certain RiPP compounds over others. This observation coincides with studies that show that there is a positive association for the distance between BGCs and the phylogenetic distance of BGC

hosts (22), indicating that these hosts have evolved with these BGCs over larger stretches of time and not acquired them recently through horizontal gene transfer. Although horizontal gene transfer is known to play a role in the diversification of BGCs (116, 117), a recent study found that for a specific RiPP class, bacteriocins, certain genetic and metabolic preconditions had to be met for effective horizontal transfer of the BGC (118). If this effect is more universal then it would strongly favor the dispersal of BGCs among species with more similar genetics and metabolism. Furthermore, a study of over 100 strains in the *Salinispora* genus found that vertical inheritance of BGCs over evolutionary time plays a large role in the interspecies diversification of BGCs (116). With all of this in mind, the observation that native hosts and heterologous hosts of similar phylogeny would be able to produce RiPP compounds with a lower metabolic burden than heterologous hosts of different phylogeny fits with the observations of the aforementioned studies (22, 116, 118).

6 Conclusion and Outlook

In summary, this thesis presented a novel method for automatically reconstructing the metabolic pathways of BGCs from annotated gene data, with a focus on ribosomally synthesized and post-translationally modified peptides (RiPPs). The developed software was shown to have a high accuracy in predicting correct structures and correct pathways, with an average Tanimoto score of 0.9 for structure prediction and an average error of 7% for prediction of RiPP production yields when compared to literature. The accuracy of the structural prediction was considered in light of the performance of a similar software, PRISM. ARMRiPP was found to have better performance for the use-case of predicting a single most likely end-point structure, while PRISM had better performance in cases where one wishes to search a known structure among all potential structures. However, the accuracy of ARMRiPP in predicting pathways and structures was found to be limited by factors such as missing implementations for various tailoring reactions, limited mechanistic knowledge of certain RiPP pathways, limitations in antiSMASH annotations, and potential sources of data skewness.

To investigate a possible use-case of ARMRiPP for studying effects of BGC metabolism in the same and in multiple organisms, the reconstructed pathways were used to estimate the metabolic burden of RiPP compound production in multiple genome-scale metabolic models (GEMs). Slight differences in metabolic burden were observed between the RiPP classes, with larger differences observed between different phylogenetic groups. Furthermore, a general trend was observed that metabolic burden would be lower for RiPP pathways when put into GEMs of their native host compared to the metabolic burden in another heterologous host. Statistically significant differences in the metabolic burden of pathways were observed between native and heterologous hosts, and the difference was more significant for less closely related organisms. The observations from the metabolic burden results were discussed from the perspective of microbial ecology, and how this potentially relates to the evolution of BGCs in microbial communities.

Overall, the method presented in this thesis has the potential to be a valuable tool for pathway reconstruction and metabolic modeling of RiPP biosynthesis in various microorganisms. However, further research is needed to address the limitations of the software and improve its accuracy in predicting metabolic pathways and structures, as well as to further investigate the ecological and evolutionary implications of RiPP clusters.

Some points of interest for the development of future projects like this are the continued development of tools such as antiSMASH, and the further elucidation of RiPP modification mechanisms and classes. During the writing of this thesis, version 7 of antiSMASH was released, promising improved detection rules and expanded pHMMs (119). Furthermore, after

corresponding with MIBiG, they explained that they will be releasing a new batch of BGC genbank files with antiSMASH 7 annotations in the near future. Considering how antiSMASH plays a role in the accuracy of the ARMRiPP method, the improvement of its annotations would consequently improve both this and future methods. The continued research of RiPP biosynthetic mechanisms can also improve upon the method by facilitating the formulation of more precise rules and assumptions for guiding the pathway reconstruction, and potentially allow for the implementation of new RiPP classes. Some more immediate improvements could be made to the ARMRiPP method by modeling more precise mass balancing, such as taking into account what happens to the leader peptide after it is cleaved off. Additionally, the method could be tried on a larger set of BGCs including uncharacterized ones, to investigate if the observations of this thesis holds true in a larger dataset. It would also be of interest to investigate and verify the results of ARMRiPP using molecular genetic laboratory methods, for example with the candidates found to have lowered metabolic burden in *S. coelicolor*.

Bibliography

1. Beutler JA. Natural Products as a Foundation for Drug Discovery. Curr Protoc Pharmacol Editor Board SJ Enna Ed--Chief Al. 2009 Sep 1;46:9.11.1-9.11.21.

2. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020 Mar 27;83(3):770–803.

3. Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol. 2022 May;7(5):726–35.

4. Monfil VO, Casas-Flores S. Chapter 32 - Molecular Mechanisms of Biocontrol in Trichoderma spp. and Their Applications in Agriculture. In: Gupta VK, Schmoll M, Herrera-Estrella A, Upadhyay RS, Druzhinina I, Tuohy MG, editors. Biotechnology and Biology of Trichoderma [Internet]. Amsterdam: Elsevier; 2014. p. 429–53. Available from: https://www.sciencedirect.com/science/article/pii/B9780444595768000321

5. Sharrar Allison M., Crits-Christoph Alexander, Méheust Raphaël, Diamond Spencer, Starr Evan P., Banfield Jillian F. Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. mBio. 2020 Jun 16;11(3):e00416-20.

6. Awad H, EL–Shahed K, Aziz R, Sarmidi M, El Enshasy H. Antibiotics as Microbial Secondary Metabolites: Production and Application. J Tekologi. 2012 Sep 15;59:101–11.

7. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021 Mar;20(3):200–16.

8. van Duin D, Paterson D. Multidrug Resistant Bacteria in the Community: Trends and Lessons Learned. Infect Dis Clin North Am. 2016 Jun;30(2):377–90.

9. Chen R, Wong HL, Kindler GS, MacLeod FI, Benaud N, Ferrari BC, et al. Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. Front Microbiol [Internet]. 2020;11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01950

10. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc Natl Acad Sci U S A. 2014 Jun 24;111(25):9259–64.

11. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011 Jul 1;39(suppl_2):W339–46.

12. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 2021 Jul 2;49(W1):W29–35.

13. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 2020 Nov 27;11(1):6058.

14. Liu Z, Zhao Y, Huang C, Luo Y. Recent Advances in Silent Gene Cluster Activation in Streptomyces. Front Bioeng Biotechnol [Internet]. 2021 [cited 2023 May 11];9. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2021.632230

15. Rowe SM, Spring DR. The role of chemical synthesis in developing RiPP antibiotics. Chem Soc Rev. 2021 Apr 13;50(7):4245–58.

16. Sulheim S, Fossheim FA, Wentzel A, Almaas E. Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters. BMC Bioinformatics. 2021 Feb 23;22(1):81.

17. Fossheim FA. Constructing Metabolic Pathways from Identified Biosynthetic Gene Clusters [Internet] [Master thesis]. NTNU; 2020 [cited 2023 May 13]. Available from: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2782574

18. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013 Jan;30(1):108–60.

19. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015 Sep;11(9):625–31.

20. Chavali AK, Rhee SY. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. Brief Bioinform. 2017 Apr 7;19(5):1022–34.

21. Martinet Loïc, Naômé Aymeric, Deflandre Benoit, Maciejewska Marta, Tellatin Déborah, Tenconi Elodie, et al. A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. mBio. 2019 Aug 13;10(4):e01230-19.

22. Xia L, Miao Y, Cao A, Liu Y, Liu Z, Sun X, et al. Biosynthetic gene cluster profiling predicts the positive association between antagonism and phylogeny in Bacillus. Nat Commun. 2022 Feb 23;13(1):1023.

23. Kwon MJ, Steiniger C, Cairns TC, Wisecaver JH, Lind AL, Pohl C, et al. Beyond the Biosynthetic Gene Cluster Paradigm: Genome-Wide Coexpression Networks Connect Clustered and Unclustered Transcription Factors to Secondary Metabolic Pathways. Microbiol Spectr. 9(2):e00898-21.

24. Eddy SR. Profile hidden Markov models. Bioinforma Oxf Engl. 1998;14(9):755–63.

25. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep 1;10:402–15.

26. Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, et al. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV

genomes. BMC Bioinformatics. 2006 Dec;7(1):1–15.

27. Blin K, Kazempour D, Wohlleben W, Weber T. Improved Lanthipeptide Detection and Prediction for antiSMASH. PLOS ONE. 2014 Feb 20;9(2):e89420.

28. Covington BC, Xu F, Seyedsayamdost MR. A Natural Product Chemist's Guide to Unlocking Silent Biosynthetic Gene Clusters. Annu Rev Biochem. 2021 Jun 20;90:763–88.

29. Wang S, Lin S, Fang Q, Gyampoh R, Lu Z, Gao Y, et al. A ribosomally synthesised and post-translationally modified peptide containing a β-enamino acid and a macrocyclic motif. Nat Commun. 2022 Aug 26;13(1):5044.

30. Walker JA, Hamlish N, Tytla A, Brauer DD, Francis MB, Schepartz A. Redirecting RiPP Biosynthetic Enzymes to Proteins and Backbone-Modified Substrates. ACS Cent Sci. 2022 Apr 27;8(4):473–82.

31. Burkhart BJ, Hudson GA, Dunbar KL, Mitchell DA. A Prevalent Peptide-Binding Domain Guides Ribosomal Natural Product Biosynthesis. Nat Chem Biol. 2015 Aug;11(8):564–70.

32. Repka LM, Chekan JR, Nair SK, van der Donk WA. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. Chem Rev. 2017 Apr 26;117(8):5457–520.

33. Vinogradov AA, Suga H. Introduction to Thiopeptides: Biological Activity, Biosynthesis, and Strategies for Functional Reprogramming. Spec Issue Chem Tools Biol Discov. 2020 Aug 20;27(8):1032–51.

34. Moffat A, Santos-Aberturas J, Chandra G, Truman A. A User Guide for the Identification of New RiPP Biosynthetic Gene Clusters Using a RiPPER-Based Workflow. In: Methods in molecular biology (Clifton, NJ). 2021. p. 227–47.

35. Liu T, Ma X, Yu J, Yang W, Wang G, Wang Z, et al. Rational generation of lasso peptides based on biosynthetic gene mutations and site-selective chemical modifications. Chem Sci. 2021;12(37):12353–64.

36. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013 Jul 1;41(W1):W204–12.

37. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015 Jul 1;43(W1):W237–43.

38. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017 Jul 3;45(W1):W36–41.

39. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019 Jul 2;47(W1):W81–7.

40. Schnell N, Entian KD, Schneider U, Götz F, Zähner H, Kellner R, et al. Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. Nature. 1988 May 1;333(6170):276–8.

41. McAuliffe O, Ross RP, Hill C. Lantibiotics: structure, biosynthesis and mode of action. FEMS Microbiol Rev. 2001 May 1;25(3):285–308.

42. Brötz H, Josten M, Wiedemann I, Schneider U, Götz F, Bierbaum G, et al. Role of lipid-bound peptidoglycan precursors in the formation of pores by nisin, epidermin and other lantibiotics. Mol Microbiol. 1998 Oct 1;30(2):317–27.

43. Breukink E, Wiedemann I, Kraaij C van, Kuipers OP, Sahl HG, de Kruijff B. Use of the Cell Wall Precursor Lipid II by a Pore-Forming Peptide Antibiotic. Science. 1999 Dec 17;286(5448):2361–4.

44. Hasper HE, Kramer NE, Smith JL, Hillman JD, Zachariah C, Kuipers OP, et al. An Alternative Bactericidal Mechanism of Action for Lantibiotic Peptides That Target Lipid II. Science. 2006 Sep 15;313(5793):1636–7.

45. Wakamatsu K, Choung SY, Kobayashi T, Inoue K, Higashijima T, Miyazawa T. Complex formation of peptide antibiotic Ro09-0198 with lysophosphatidylethanolamine: proton NMR analyses in dimethyl sulfoxide solution. Biochemistry. 1990;29(1):113–8.

46. Märki F, Hänni E, Fredenhagen A, van Oostrum J. Mode of action of the lanthionine-containing peptide antibiotics duramycin, duramycin B and C, and cinnamycin as indirect inhibitors of phospholipase A2. Biochem Pharmacol. 1991 Oct 24;42(10):2027–35.

47. Toogood PL. Model studies of lantibiotic biogenesis. Int J Rapid Publ Prelim. 1993 Dec 3;34(49):7833–6.

48. Burrage S, Raynham T, Williams G, Essex JW, Allen C, Cardno M, et al. Biomimetic synthesis of lantibiotics. Chem Weinh Bergstr Ger. 2000 Apr 14;6(8):1455–66.

49. Okeley NM, Zhu Y, van der Donk WA. Facile Chemoselective Synthesis of Dehydroalanine-Containing Peptides. Org Lett. 2000 Nov 1;2(23):3603–6.

50. Zhu Y, Gieselman MD, Zhou H, Averin O, van der Donk WA. Biomimetic studies on the mechanism of stereoselective lanthionine formation. Org Biomol Chem. 2003 Oct 7;1(19):3304–15.

51. Zhou H, van der Donk WA. Biomimetic Stereoselective Formation of Methyllanthionine. Org Lett. 2002 Apr 1;4(8):1335–8.

52. Kiesau P, Eikmanns U, Gutowski-Eckel Z, Weber S, Hammelmann M, Entian K D. Evidence for a multimeric subtilin synthetase complex. J Bacteriol. 1997 Mar 1;179(5):1475–81.

53. Siegers K, Heinzmann S, Entian KD. Biosynthesis of Lantibiotic Nisin: Posttranslational Modification of Its Prepeptide Occurs at a Multimeric Membrane-associated Lanthionine

Synthetase Complex. J Biol Chem. 1996 May 24;271(21):12294–301.

54. Garg N, Salazar-Ocampo LMA, van der Donk WA. In vitro activity of the nisin dehydratase NisB. Proc Natl Acad Sci. 2013 Apr 30;110(18):7258–63.

55. Ortega MA, Hao Y, Zhang Q, Walker MC, van der Donk WA, Nair SK. Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. Nature. 2015 Jan 1;517(7535):509–12.

56. Kuipers Anneke, Wierenga Jenny, Rink Rick, Kluskens Leon D., Driessen Arnold J. M., Kuipers Oscar P., et al. Sec-Mediated Transport of Posttranslationally Dehydrated Peptides in Lactococcus lactis. Appl Environ Microbiol. 2006 Dec 1;72(12):7626–33.

57. Zhang Q, Ortega M, Shi Y, Wang H, Melby JO, Tang W, et al. Structural investigation of ribosomally synthesized natural products by hypothetical structure enumeration and evaluation using tandem MS. Proc Natl Acad Sci. 2014 Aug 19;111(33):12031–6.

58. Velásquez JE, Zhang X, van der Donk W. Biosynthesis of the Antimicrobial Peptide Epilancin 15X and Its N-terminal Lactate. Chem Biol. 2011 Jul 29;18(7):857–67.

59. Chatterjee C, Paul M, Xie L, van der Donk WA. Biosynthesis and Mode of Action of Lantibiotics. Chem Rev. 2005 Feb 1;105(2):633–84.

60. Thibodeaux CJ, Ha T, van der Donk WA. A Price To Pay for Relaxed Substrate Specificity: A Comparative Kinetic Analysis of the Class II Lanthipeptide Synthetases ProcM and HalM2. J Am Chem Soc. 2014 Dec 17;136(50):17513–29.

61. Mukherjee S, van der Donk WA. Mechanistic Studies on the Substrate-Tolerant Lanthipeptide Synthetase ProcM. J Am Chem Soc. 2014 Jul 23;136(29):10450–9.

62. Bagley MC, Dale JW, Merritt EA, Xiong X. Thiopeptide Antibiotics. Chem Rev. 2005 Feb 1;105(2):685–714.

63. Thompson J, Cundliffe E, Stark MJR. The Mode of Action of Berninamycin and the Mechanism of Resistance in the Producing Organism, Streptomyces bernensis. Vol. 128, Microbiology. Microbiology Society; 1982. p. 875–84.

64. CUNDLIFFE E, THOMPSON J. Concerning the Mode of Action of Micrococcin upon Bacterial Protein Synthesis. Eur J Biochem. 1981 Aug 1;118(1):47–52.

65. Cundliffe E, Thompson J. The Mode of Action of Nosiheptide (Multhiomycin) and the Mechanism of Resistance in the Producing Organism. Vol. 126, Microbiology. Microbiology Society; 1981. p. 185–92.

66. Harms JM, Wilson DN, Schluenzen F, Connell SR, Stachelhaus T, Zaborowska Z, et al. Translational Regulation via L11: Molecular Switches on the Ribosome Turned On and Off by Thiostrepton and Micrococcin. Mol Cell. 2008 Apr 11;30(1):26–38.

67. Lentzen G, Klinck R, Matassova N, Aboul-ela F, Murchie AIH. Structural Basis for Contrasting Activities of Ribosome Binding Thiazole Antibiotics. Chem Biol. 2003 Aug

1;10(8):769–78.

68. Rosendahl G, Douthwaite S. The antibiotics micrococcin and thiostrepton interact directly with 23S rRNA nucleotides 1067A and 1095A. Nucleic Acids Res. 1994 Feb 11;22(3):357–63.

69. Clementi N, Polacek N. Ribosome-associated GTPases: The role of RNA for GTPase activation. RNA Biol. 2010 Sep 1;7(5):521–7.

70. Polikanov YS, Aleksashin NA, Beckert B, Wilson DN. Front Mol Biosci. 2018;5:1–21.

71. Murumkar PR, Ghuge RB. Chapter 9 - Vicinal Diaryl Oxadiazoles, Oxazoles, and Isoxazoles. In: Yadav MR, Murumkar PR, Ghuge RB, editors. Vicinal Diaryl Substituted Heterocycles [Internet]. Elsevier; 2018. p. 277–303. Available from: https://www.sciencedirect.com/science/article/pii/B9780081022375000092

72. Hughes RA, Thompson SP, Alcaraz L, Moody CJ. Total Synthesis of the Thiopeptide Antibiotic Amythiamicin D. J Am Chem Soc. 2005 Nov 1;127(44):15644–51.

73. Burkhart BJ, Schwalen CJ, Mann G, Naismith JH, Mitchell DA. YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function. Chem Rev. 2017 Apr 26;117(8):5389–456.

74. Moutiez M, Belin P, Gondry M. Aminoacyl-tRNA-Utilizing Enzymes in Natural Product Biosynthesis. Chem Rev. 2017 Apr 26;117(8):5578–618.

75. Rice AJ, Pelton JM, Kramer NJ, Catlin DS, Nair SK, Pogorelov TV, et al. Enzymatic Pyridine Aromatization during Thiopeptide Biosynthesis. J Am Chem Soc. 2022 Nov 23;144(46):21116–24.

76. Yu Y, Duan L, Zhang Q, Liao R, Ding Y, Pan H, et al. Nosiheptide Biosynthesis Featuring a Unique Indole Side Ring Formation on the Characteristic Thiopeptide Framework. ACS Chem Biol. 2009 Oct 16;4(10):855–64.

77. Kelly WL, Pan L, Li C. Thiostrepton Biosynthesis: Prototype for a New Family of Bacteriocins. J Am Chem Soc. 2009 Apr 1;131(12):4327–34.

78. Cheng C, Hua ZC. Lasso Peptides: Heterologous Production and Potential Medical Application. Front Bioeng Biotechnol [Internet]. 2020;8. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2020.571165

79. Capelli R, Marchetti F, Tiana G, Colombo G. SAGE: A Fast Computational Tool for Linear Epitope Grafting onto a Foreign Protein Scaffold. J Chem Inf Model. 2017 Jan 23;57(1):6–10.

80. Zhu S, Fage CD, Hegemann JD, Mielcarek A, Yan D, Linne U, et al. The B1 Protein Guides the Biosynthesis of a Lasso Peptide. Sci Rep. 2016 Oct 18;6(1):35604.

81. Duan Y, Niu W, Pang L, Bian X, Zhang Y, Zhong G. Unusual Post-Translational Modifications in the Biosynthesis of Lasso Peptides. Int J Mol Sci. 2022;23(13).

82. Passi A, Tibocha-Bonilla JD, Kumar M, Tec-Campos D, Zengler K, Zuniga C. Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data. Metabolites. 2021 Dec 24;12(1):14.

83. Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biol. 2021 Feb 18;22(1):64.

84. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018 Sep 6;46(15):7542–53.

85. Sulheim S, Kumelj T, van Dissel D, Salehzadeh-Yazdi A, Du C, van Wezel GP, et al. Enzyme-Constrained Models and Omics Analysis of Streptomyces coelicolor Reveal Metabolic Changes that Enhance Heterologous Production. iScience. 2020 Sep 25;23(9):101525.

86. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016 Jan 4;44(D1):D515–22.

87. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010 Mar 1;28(3):245–8.

88. Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci. 2002 Nov 12;99(23):15112–7.

89. Sulheim S. Assembly and application of genome-scale metabolic models to study Streptomyces coelicolor and Prochlorococcus [Internet] [Doctoral thesis]. NTNU; 2021 [cited 2023 May 12]. Available from: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2734302

90. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988 Feb 1;28(1):31–6.

91. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns [Internet]. [cited 2023 May 12]. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

92. RDKit [Internet]. [cited 2023 May 12]. Available from: https://www.rdkit.org/

93. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular Similarity in Medicinal Chemistry. J Med Chem. 2014 Apr 24;57(8):3186–204.

94. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminformatics. 2015 May 20;7(1):20.

95. Creative Content by carterart [Internet]. Vecteezy. [cited 2023 May 13]. Available from: https://www.vecteezy.com/members/carterart

96. Landrum G, Tosco P, Kelley B, Ric, sriniker, Cosgrove D, et al. rdkit/rdkit: 2023_03_1 (Q1

2023) Release [Internet]. Zenodo; 2023 [cited 2023 May 12]. Available from: https://zenodo.org/record/7880616

97. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013 Aug 8;7(1):74.

98. Pelley JW. 17 - Protein Synthesis and Degradation. In: Pelley JW, editor. Elsevier's Integrated Biochemistry [Internet]. Philadelphia: Mosby; 2007. p. 147–58. Available from: https://www.sciencedirect.com/science/article/pii/B9780323034104500237

99. Li Y, Ducasse R, Zirah S, Blond A, Goulard C, Lescop E, et al. Characterization of Sviceucin from Streptomyces Provides Insight into Enzyme Exchangeability and Disulfide Bond Formation in Lasso Peptides. ACS Chem Biol. 2015 Nov 20;10(11):2641–9.

100. Chakravarthi S, Jessop CE, Bulleid NJ. The role of glutathione in disulphide bond formation and endoplasmic-reticulum-generated oxidative stress. EMBO Rep. 2006 Mar;7(3):271–5.

101. Common Taxonomy Tree [Internet]. [cited 2023 May 12]. Available from: https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi

102. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. J Med Chem. 1996 Jan 1;39(16):3049–59.

103. Uguen P, Hindré T, Didelot S, Marty C, Haras D, Pennec JP, et al. Maturation by LctT Is Required for Biosynthesis of Full-Length Lantibiotic Lacticin 481. Appl Environ Microbiol. 2005 Feb 1;71:562–5.

104. Thibodeaux CJ, Wagoner J, Yu Y, van der Donk WA. Leader Peptide Establishes Dehydration Order, Promotes Efficiency, and Ensures Fidelity During Lacticin 481 Biosynthesis. J Am Chem Soc. 2016 May 25;138(20):6436–44.

105. Zhang Z, Hudson GA, Mahanta N, Tietz JI, van der Donk WA, Mitchell DA. Biosynthetic Timing and Substrate Specificity for the Thiopeptide Thiomuracin. J Am Chem Soc. 2016 Dec 7;138(48):15511–4.

106. Nam W. 8.12 - Cytochrome P450. In: McCleverty JA, Meyer TJ, editors. Comprehensive Coordination Chemistry II [Internet]. Oxford: Pergamon; 2003. p. 281–307. Available from: https://www.sciencedirect.com/science/article/pii/B0080437486081457

107. Morris RP, Leeds JA, Naegeli HU, Oberer L, Memmert K, Weber E, et al. Ribosomally Synthesized Thiopeptide Antibiotics Targeting Elongation Factor Tu. J Am Chem Soc. 2009 Apr 29;131(16):5946–55.

108. Guerrero-Garzón JF, Madland E, Zehl M, Singh M, Rezaei S, Aachmann FL, et al. Class IV Lasso Peptides Synergistically Induce Proliferation of Cancer Cells and Sensitize Them to Doxorubicin. iScience. 2020 Nov 10;23(12):101785.

109. Cao L, Do T, Link AJ. Mechanisms of action of ribosomally synthesized and

posttranslationally modified peptides (RiPPs). J Ind Microbiol Biotechnol. 2021 Apr 1;48(3–4):kuab005.

110. Kaleta C, Schäuble S, Rinas U, Schuster S. Metabolic costs of amino acid and protein production in Escherichia coli. Biotechnol J. 2013 Sep 1;8(9):1105–14.

111. Wang M, Fage CD, He Y, Mi J, Yang Y, Li F, et al. Recent Advances and Perspectives on Expanding the Chemical Diversity of Lasso Peptides. Front Bioeng Biotechnol [Internet]. 2021;9. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2021.741364

112. Zhong Z, He B, Li J, Li YX. Challenges and advances in genome mining of ribosomally synthesized and post-translationally modified peptides (RiPPs). Synth Syst Biotechnol. 2020 Sep 1;5(3):155–72.

113. Skinnider MA, Johnston CW, Edgar RE, Dejong CA, Merwin NJ, Rees PN, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. Proc Natl Acad Sci. 2016 Oct 18;113(42):E6343–51.

114. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today. 2007 Mar 1;12(5):225–33.

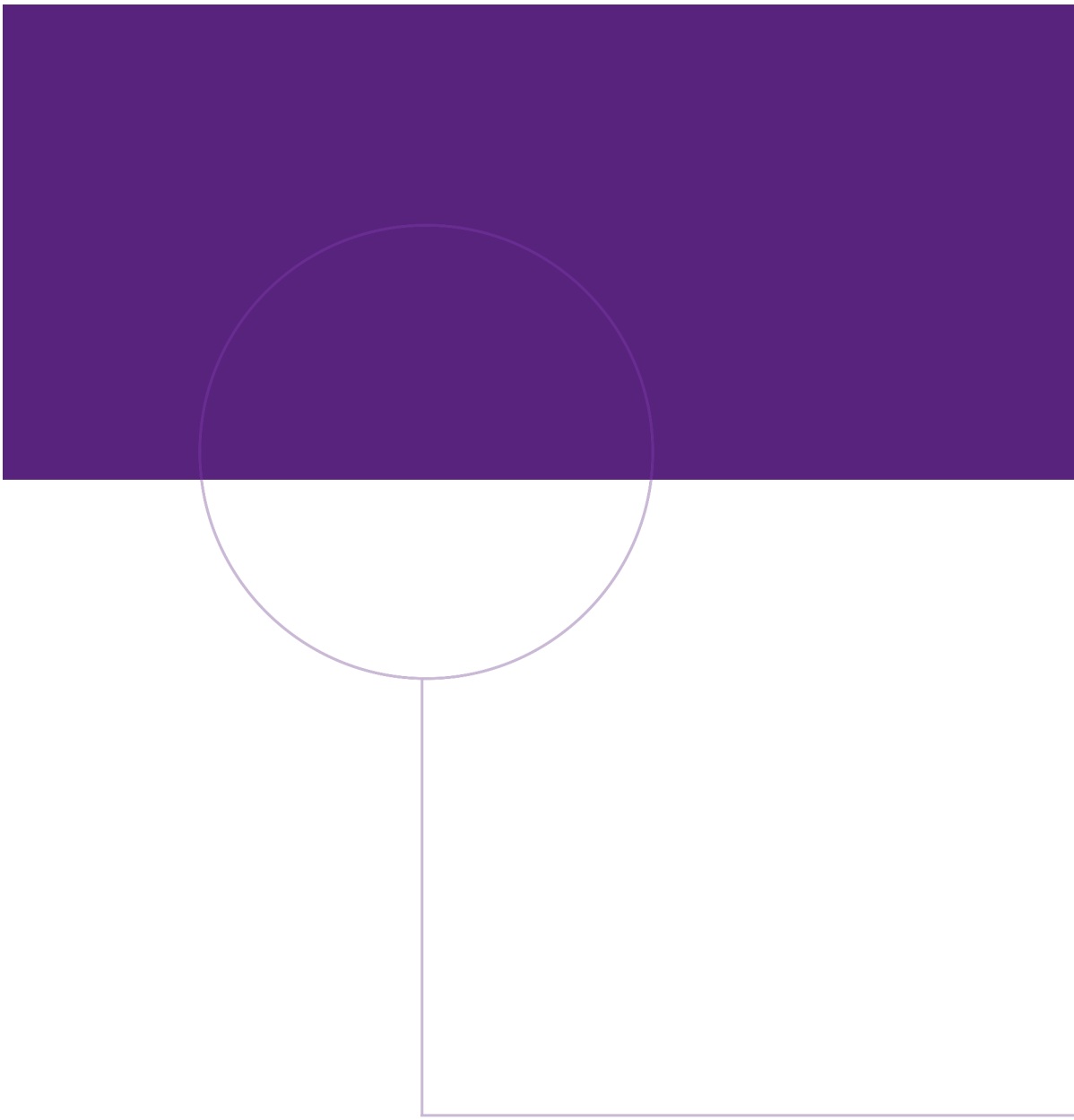
115. Martin YC, Kofron JL, Traphagen LM. Do Structurally Similar Molecules Have Similar Biological Activity? J Med Chem. 2002 Sep 1;45(19):4350–8.

116. Chase AB, Sweeney D, Muskat MN, Guillén-Matus DG, Jensen PR. Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. mBio. 2021 Nov 23;12(6):e02700-21.

117. Wu D, Jiang B, Ye CY, Timko MP, Fan L. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. Spec Issue Rice Funct Genomics. 2022 May 9;3(3):100320.

118. Krauss S, Harbig TA, Rapp J, Schaefle T, Franz-Wachtel M, Reetz L, et al. Horizontal Transfer of Bacteriocin Biosynthesis Genes Requires Metabolic Adaptation To Improve Compound Production and Cellular Fitness. Microbiol Spectr. 11(1):e03176-22.

119. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023 May 4;gkad344.



NTNU

Norwegian University of
Science and Technology