ANTIBIOTIC DISCOVERY

RESISTANCE PROFILING OF MICROBIAL GENOMES TO REVEAL NOVEL ANTIBIOTIC NATURAL PRODUCTS

By CHELSEA WALKER, H. BSc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2017) Hamilton, Ontario (Biochemistry and Biomedical Sciences)

TITLE: Resistance Profiling of Microbial Genomes to Reveal Novel Antibiotic Natural Products. AUTHOR: Chelsea Walker, H. BSc. (McMaster University) SUPERVISOR: Dr. Nathan A. Magarvey. COMMITTEE MEMBERS: Dr. Eric Brown and Dr. Michael G. Surette. NUMBER OF PAGES: xvii, 168

Lay Abstract

It would be hard to imagine a world where we could no longer use the antibiotics we are routinely being prescribed for common bacterial infections. Currently, we are in an era where this thought could become a reality. Although we have been able to discover antibiotics in the past from soil dwelling microbes, this approach to discovery is being constantly challenged. At the same time, the bacteria are getting smarter in their ways to evade antibiotics, in the form of resistance, or self-protection mechanisms. As such is it essential to devise methods which can predict the potential for resistance to the antibiotics we use early in the discovery and isolation process. By using what we have learned in the past about how bacteria protect themselves for antibiotics, we can to stay one step ahead of them as we continue to search for new sources of antibiotics from bacteria.

Abstract

Microbial natural products have been an invaluable resource for providing clinically relevant therapeutics for almost a century, including most of the commonly used antibiotics that are still in medical use today. In more recent decades, the need for new biotherapeutics has begun to grow, as multi-drug resistant pathogens continue to emerge, putting into question the long-term efficacy of many drugs that we routinely depend on to combat infectious diseases. To affect this growing medical crisis, new efforts are being applied to computationally mine the genomes of microorganisms for biosynthetic gene clusters that code for molecules possessing anti-microbial activities that circumvent known resistance mechanisms. To this end, cutting-edge software platforms have been developed that can identify, with high predictive accuracy, microbial genomes that code for natural products of potential interest. However, with such analyses comes the need to thoroughly vet each predicted gene cluster, to identify those high-value candidate molecules that are not associated with known resistance mechanisms. In this work, a new strategy was developed that involved cataloguing all known 'self-resistance' mechanisms encoded by natural product producing microorganisms, which protect the producer from the highly toxic effects of their secreted anti-microbial agents. This collection of resistance data was leveraged and used to engineer an automated software-based pipeline that interrogates biosynthetic gene clusters and relates them to previously identified resistance mechanisms. Gene clusters that are revealed to be independent of known resistance mechanisms can then be flagged for further chemical and biological

study in the laboratory. Such in-depth interrogations of microbial genomes aim to help reveal the full biological repertoire of antibiotics yet to be discovered from microorganisms, and will lead to the development of the next generation of biotherapeutics to quell the growing medical crisis of antibiotic-resistance among human pathogenic organisms.

Acknowledgments

I would like to first thank my supervisor Nathan Magarvey for allowing me the opportunity to undertake graduate studies, and all his guidance throughout. His endless ideas and creative thinking made this experience of incredible value. The skills I have gained with the help of Nathan throughout this process will be incredibly useful in furthering my career development.

I would also like to acknowledge my committee members Dr. Eric Brown, and Dr. Mike Surette for their insightful comments and suggestions during committee meetings.

The individuals of the Magarvey lab have been a privilege to work with during my time at McMaster University. I would like to acknowledge Haoxin Li, and Chad Johnston for their guidance in my research projects. I would also like to thank Robyn Maclellan and Dave Capstick, for their friendship, and always being there to help me navigate the stressful times of graduate studies. A big acknowledgment also goes out the members of the Magarvey Lab who handle the computational processes and development, who without, much of my research would not have advanced to the same degree. Finally, to Agata Kieliszek, a dedicated undergraduate student and co-op student, who helped me with my research projects, and never failed to have a smile on her face.

Finally, a huge thank you goes out to my family and friends for their love and support over the course of this degree. Their constant support and encouragement was pivotal during my graduate studies.

Table of Contents

List of Tables

Chapter 2.

List of Figures

Chapter 1.

Chapter 2.

Abbreviations

ATCC- American Type Culture Collection

ATP- Antimicrobial Target Predictor

AMR-Antimicrobial resistance

BGC- Biosynthetic gene cluster

CLAMS- Computational Library for Analysis of Mass Spectral Data

CoA- Coenzyme A

DNA- Deoxyribonucleic acid

DSMZ- German Resource Centre for Biological Material

ESKAPE- *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp.

ESI- Electrospray ionization

GARLIC- Global Alignment for Natural Product Chemoinformatics

GRAPE- Generalized Retro-biosynthetic Assembly Prediction Engine

HMM- Hidden Markov Model

HPLC- High-performance liquid chromatography

HTS- High throughput screening

LC-MS/MS- Liquid chromatography-tandem mass spectrometry

MDR- Multidrug resistant

MRSA-Methicillin-resistant *Staphylococcus aureus*

NCBI- National Center for Biotechnology Information

NMR- Nuclear magnetic resonance (spectroscopy)

NP-Natural product

NRP-Nonribosomal Peptide

NRPS-Nonribosomal Peptide Synthetase

NRRL- Agricultural Research Service Culture Collection

PK- Polyketide

PKS-Polyketide Synthetase

pNP- Predicted natural product

PRISM- Predictive Informatics of Secondary Metabolomes

RNA-Ribonucleic acid

SIPE- Siderophore Identification Prediction Engine

SMILES- Simplified molecular-input line-entry system

Declaration of Academic Achievement

Chapter 2 of this thesis is prepared in paper format for publication. The specific details regarding the research contribution is described within the chapter preface.

Chapter 1. Introduction

1.1 Thesis Context

The widespread emergence and rapid increase of antibiotic resistance within clinically-relevant microorganisms has become one of the leading causes of death worldwide, resulting in at least 23, 000 deaths in the United States alone (https://www.cdc.gov/drugresistance/). The threat of multi-drug resistant infections calls for increased effort and efficiency by the scientific community to discover new sources of anti-infective agents. Bacteria create a wide array of bioactive natural products that have formed the basis for many therapeutic regimes still used in the clinic, particularly for the treatment of infectious disease. However, over-reliance on traditional discovery efforts has led to the rediscovery of known molecules, resulting in a loss of industrial interest in natural product discovery[1]. Despite this, recent genome sequencing efforts have shown that much of these valuable small molecules remain undiscovered. Traditional discovery approaches are thought to be largely exhausted in their ability to identify new chemical classes of microbial natural products. As such, new, unconventional techniques are required to fill the discovery void. Given the wealth of information available concerning biosynthetic gene clusters (BGCs) and their encoded chemistry, development of a prioritization schema is key to effectively mine for novel small molecules.

Despite the emphasis placed on associated resistance once a product reaches the clinic, little is done to infer the downstream potential for emergence of resistance during initial isolation processes. The aim of this body of work is to develop a mean to prioritize those strains based on the novelty of their BGCs and predicted natural products (pNP) that

not only display divergent chemistry, but importantly, lack known resistance genes which may infer cross resistance with the current repertoire of antimicrobial agents.

1.2 The Antibiotic Resistance Crisis

Antibiotics are one of the most influential discoveries of modern medicine, beginning with the discovery of penicillin by Alexander Flemming in 1979[2]. However, the rapid emergence of multidrug resistant (MDR) pathogens in recent years has suggested a drastic movement towards a post-antibiotic era; a concerning concept in which modern day antibiotics will be rendered ineffective[3,4]. In response to this global crisis, early in 2017, the World Health Organization released a global priority pathogens list to draw attention to the urgent need for new anti-infective agents[5]. A dangerous escalation to the antibiotic resistance crisis is the decline in the discovery rate of new antimicrobials following the discovery surge exhibited in the "golden era" of microbial natural product discovery[6]. The discovery void has further favoured the emergence of resistant pathogens, as few new chemical scaffolds of antibiotics are actively being pursued in comparison to recent decades[3,6]. As of March 2017, approximately 41 new antibiotics were in the pipeline for clinical development[7].

Tremendous effort has been dedicated towards elucidating the various mechanisms of antibiotic resistance, with focus on the ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp.), a set of antibiotic-resistant bacteria that are particularly difficult to treat[8,9]. Despite an increased understanding in the molecular mechanisms

involved in antibiotic resistance, much less is understood on how this information can be adequately translated to assist in finding the new iteration of microbial natural products.

1.3 Traditional Approaches to Natural Product Discovery

Historically, natural product discovery has relied on the ability of microorganisms, specifically soil microbes, to produce a large array of chemically distinct secondary metabolites; many of which have become crucial members of the repertoire of clinically used antibiotics[10]. During the 'golden era' of natural product discovery, researchers relied on methods which have been coined the traditional approach of natural product discovery. The traditional method is a top-down approach focusing on the repeated fractionation of the excreted metabolite profile of candidate microbes to hone in on a bioactive fraction of interest[11, 12]. This method was wildly successful throughout the 1950s, 1960s, and 1970s, as many of the purified bioactive secondary metabolites were approved for clinical use[10,13].

The traditional approach to natural products discovery has been revamped in recent years by incorporating high-throughput screening techniques (HTS). HTS provided a cost and time efficient means to highlight or eliminate candidate compounds of interest by screening large libraries of natural or synthetic compounds for an activity of interest[14]. Screens can be readily adapted for broad biological activities, or enlist more focused methods using target-based screening approaches[15, 16]. Over the years, HTS has incorporated the use of various natural product libraries, but has not been associated with the high success rates as was initially postulated[17]. However, HTS has received criticism

as of late, as there is a need to establish universal standards of use, as well as increased quality control of the compound libraries utilized[18-20].

Despite previous success, the traditional approach is associated with high re-discovery rates, bias towards high abundance molecules, and time consuming procedures[11]. Thus, it has been suggested that the traditional method may have reached its limit to identify new, and chemically distinct natural products. This has been followed by the call for new unconventional, and new approaches to natural product discovery to overcome the limitations associated with traditional discovery methods[21].

1.4 Modes of Action of Evolved Antimicrobial Natural Products

The "golden era" of natural product discovery provided a wealth of chemical entities with desirable properties as antibacterial agents. Microbial natural products have evolved over time to provide a competitive advantage within their respective environments[22]. Many of these secondary metabolites have evolved to be valuable sources of antibacterial agents, with evolution further favouring the emergence of families of secondary metabolites that share a common structural core[23]. Considering this, it is seemingly unsurprising the diversity which has been achieved in respect to microbial secondary metabolites to target almost every known bacterial target[24]. Despite targeting a variety of molecular targets, a large cohort of microbial natural products has also evolved to target the same bacterial targets such as the bacterial ribosome or cell wall[23, 24]. This concept of narrow spectrum behaviour of many derived microbial natural products, has further favoured the emergence of bacterial resistance[3]. As such, it becomes ever more

apparent of the need to develop isolation efforts directed at identifying microbial natural products with divergent mechanisms.

Identifying the mode of action of bioactive metabolites remains to be one of the major caveats of natural product discovery. Due to the chemical complexity and diversity within a single microbial extract, several compounds can be present, representing several modes of action, which makes target-based screening attempts inherently difficult[25]. An effort to address such concerns was accomplished by employing cytological profiling to decipher possible mode of actions exhibited within a single extract[26]. Despite associated efforts, current research has not met the demands in respect to identifying new sources of microbial products which display divergent modes of action.

1.5 Diversity in Resistance Mechanism

In nature, antibiotic-producing bacteria employ several resistance mechanisms to evade the effects of their own antibiotics, and the effects of the active molecules that are excreted by neighbouring species[22]. Furthermore, it has been established that environmental resistance mechanisms can act as a reservoir for the exchange of resistance genes to clinical pathogens, allowing preliminary insight into the evolutionary origins of antibiotic resistance[27]. This has been further suggested by detecting the presence of known resistance genes in environmental organisms which significantly pre-date the introduction of antibiotics as therapeutic agents[28]. The way in which bacteria avoid the effects of antibiotics can be broadly grouped into mechanisms where the target, or antibiotic is directly modified, or mechanisms in which an indirect effect occurs (i.e. translocation

pumps, duplicate targets). The ability of bacteria to resist the effect of antibiotics is diverse, and several reviews have been dedicated to summarizing the resistance mechanisms of bacteria, their ability to acquire resistance genes, and understanding the complex relationships involving the emergence of antibiotic resistance[29, 30].

1.6 Polyketides and Nonribosomal Peptides Family of Natural Products

A large cohort of identified microbial natural products are of the polyketide (PK) and non-ribosomal peptide (NRP) classes. PK and NRP natural products are derived through a series of biosynthetic enzymes known as polyketide synthetases (PKS) and non-ribosomal peptide synthetases (NRPS) respectively. Bioactive entities have been characterized from both individual assembly lines, and hybrids of the two[31-33]. NRPs are made by multimodular systems that act in a stepwise fashion to incorporate specific substrates to result in diverse peptidic natural products, known as non-ribosomal peptide synthetases[34, 35]. This structural diversity is driven by the numerous substrates that can be included, such as proteinogenic and non-proteinogenic amino acids, allowing for numerous combinations of possible products[36]. Examples of NRPs from characterized assembly line systems are vancomycin, bacitracin, and daptomycin all of which have been developed as clinical agents[37-39]. Polyketides are also assembled in a multimodular enzymatic fashion through polyketide synthetases, that enable the addition of variable monomer biosynthetic units (e.g. small organic acids) resulting in a high degree of structural diversity[40, 41]. Example polyketides with antibacterial activities are erythromycin and rifamycin[42, 43].

The comprised assembly line systems resulting in the production of PKs, and NRPs have an outstanding ability to create products with chemical diversity. Due to their modular nature, the potential for molecular diversity is significantly increased through the aid of molecular promiscuity caused by the biosynthetic enzymes[23]. Diversity within these systems can be further achieved through the addition of various tailoring enzymes affording additional chemical complexity[44]. In comparison to other classes of microbial natural products, PKs and NRPs account for most bioactive natural products, including those with antimicrobial activity[24].

1.7 Organization of Natural Product Biosynthetic Gene Clusters

Characterization of NRP and PK assembly line systems has demonstrated the clustering of genes related to the biosynthesis of natural products to certain locations within a genomic sequence, known as a biosynthetic gene cluster (BGC). Greater understanding of these assembly line systems has provided insight into finite details of the biosynthetic assembly systems, such as those involving predicting the stereochemistry of PKs, or the influence of thioesterase domains (enzymes which catalyze the release of the peptide) in generating further chemical diversity[45, 46].

Depositories have been established to store information regarding what is currently known in the context of BGCs to facilitate future endeavours involving the biosynthesis of secondary metabolites[47, 48]. Further investigation into the biosynthetic assembly lines involved in the production of secondary metabolites, also has revealed accessory genes that are involved in regulation, export, and self-protection, all relating back to the product of

the BGC. Self-protection genes play a pivotal role in protecting the host producing organism from the effects of their own repertoire of antibiotics[49].

BGCs dedicated to the production of several classes of natural products were identified in the late 1990's and further characterization and annotation of BGCs continues today. Due to the nature of resistance genes providing self-protection, several works have demonstrated the beginnings of defining associations between resistance genes, chemical scaffold, and the BGC for a given antibacterial agent [24,50,51]. Close genomic proximity between the BGCs and the resistance gene for that product have been demonstrated within many families of natural products including glycopeptides, aminoglycosides, macrolides, and beta- lactam natural products.

1.7.1 Glycopeptide Biosynthetic Gene Clusters

Glycopeptide antibiotics such as vancomycin, have a notable presence within the clinic as they are used as one of the treatment of choice agents against methicillin-resistant *Staphylococcus aureus* (MRSA)[52]. The emergence of resistance to glycopeptide antibiotics pushed the field to not only identify the mechanisms of resistance but also begin to identify the molecular genetics underlying the resistance mechanisms. Five genes necessary for providing a high level of glycopeptide resistance were identified on a transposable element within *Enterococcus faecium*[53]. Three of the five genes, VanH, VanA, and VanX, were later identified within two known producers of glycopeptide antibiotics[54]. This finding demonstrated that the source of the observed resistance genes within the clinic was the original producers of the antibiotic. This demonstrated for the first time a significant

contribution to delineating the relationships between clinically observed resistance genes, and their respective origins.

The first BGC identified for a glycopeptide antibiotic was in 1998 from *Amycolatopsis orientalis*, the producer of chloroeremomycin[55]. Since then, several other BGCs have been identified, and characteristic genes for the biosynthesis of glycopeptide antibiotics have been summarized[56]. Apart from chloroeremomycin, identified BGCs for glycopeptide natural products contained a form of self-protection, or resistance gene. Certain producers, such as *Amycolatopsis sp.*, producer of balhimycin, contain the characteristic VanHAX cassette on a separate contig[57]. Instead, the balhimycin BGC contains an additional gene within the boundaries that functions as a resistance gene. Glycopeptide antibiotics share the same mode of action, and this similarity is further reflected within the BGCs and their forms of self-protection mechanisms being reflective of their target.

1.7.2 Aminoglycoside Biosynthetic Gene Clusters

Aminoglycoside antibiotics represent an extensive family of natural products with clinical importance, particularly due to their potent antimicrobial activity against *Mycobacterium tuberculosis*. However, the effectiveness of many aminoglycosides is being challenged by emerging resistance[58]. Several aminoglycoside producing organisms have been identified and, in partnership with sequencing information, have provided a strong understanding of the involved biosynthetic machinery[59,60]. The most commonly encountered resistance mechanisms associated with aminoglycosides are the expression of

enzymes that directly modify the antibiotic through phosphorylation, adenylation, or acetylation[61,62].

The first BGC identified for an aminoglycoside antibiotic was identified for streptomycin from *Streptomyces griseus* in 1987[63]. Although only partial, the major resistance gene associated with streptomycin was found located within the BGC boundaries. Several other BGC have been identified for the aminoglycoside family of natural products, and their associated resistance genes detected within the cluster boundaries reflect the same mechanisms which were detected within the clinic. Despite there being subclasses of aminoglycosides, they share significant overlap in their associated resistance mechanisms, further representing the developing relationship between resistance genes, BGCs, and mode of action of antibiotics.

1.7.3 Macrolide Biosynthetic Genetic Clusters

Macrolide antibiotics are produced by PKSs, and represent a family of therapeutically relevant antibiotics with Gram-positive activity. The first resistance gene for a macrolide antibiotic was originally classified in 1982 from *Saccharopolyspora erythreus*, as an enzyme capable of methylating the bacterial 23S ribosomal RNA[64]. Subsequent research in years following elucidated the BGC for erythromycin, including the previously identified resistance gene[65,66]. In many instances, the determined resistance genes for macrolide antibiotics have significant sequence similarity between one another, as seen with the genes encoding ABC transporters conferring resistance to three macrolide antibiotics[67]. Overall, extensive research has been put towards elucidating the self-

protection mechanisms afforded by microbes that posses the ability to produce macrolide antibiotics in the form of target modification, alterations of cell permeability (e.g. ABC transporters), and through the aid of antibiotic modifying enzymes[68-70].

1.7.4 Beta-Lactam Biosynthetic Gene Clusters

Beta-lactam antibiotics represent one of the defining families of natural products in terms of clinical relevance. Despite the excitement surrounding the isolation of penicillin, the first beta-lactamase capable of rendering penicillin ineffective was isolated within a strain of *Escherichia coli* before the antibiotic's debut as a therapeutic agent[71].

Penicillin was the first beta-lactam antibiotic to undergo BGC characterization as several of the genes necessary for biosynthesis were identified in close proximity within the *Penicillium chrysogenum* genome[72]. Upon initial identification, no speculation was made to infer plausible self-protection genes within the proposed cluster. In years following, several BGCs were identified for the main classes of beta-lactam antibiotics. As with other classes of natural products, several self-protection mechanisms were identified within the BGCs of producing organisms (e.g. target modification, expulsion pumps, and antibiotic modifying enzymes)[73,74].

1.8 Current Considerations of Resistance Genes in Natural Product Discovery

The traditional approach to natural product discovery places little emphasis on the potential for resistance to a new antibiotic, despite the importance it plays in defining a successful therapeutic agent. A resistance guided approach was established using the known self-protection mechanisms of antibiotic producers to effectively screen, and enrich

for the isolation of both glycopeptide and ansamycin antibacterials[50]. This strategy enhanced the ability to define new members of these classes of antibiotics in comparison to previous years. Despite the attributed success of the technique, the current era of antimicrobial resistance requires as a necessity, the discovery of new chemical scaffolds with less likelihood to be effected by current resistance mechanisms.

Moving closer to such a scenario, was accomplished by the Wright lab through the development of the Antibiotic Resistance Platform in 2017[75]. The platform incorporates the use of several individual resistance elements on separate plasmids, which are used to screen extracts to decrease the identification of previously identified antibiotics, and highlight those which may possess different self-protection mechanisms[75]. Furthermore, the developed platform was also utilized to identify extracts which may act as novel inhibitors of common resistance mechanisms[75]. The developed platform begins to highlight the importance, and usefulness, of self-protection mechanisms in the modern era of antibiotic discovery. However, the platform is currently limited to those which have well defined self-protection mechanisms.

1.9 Bioinformatic Platforms for Identification of Secondary Metabolites

As previously mentioned, NRP and PK natural products are produced by assembly line systems that can result in extensive structural diversity. The surge in availability of genomic information has not only allowed for a better understanding of the genetic basis of these assembly line systems, but also the generation of bioinformatic platforms capable of inferring secondary metabolites directly from the DNA sequence (DNA-RNA-protein-

small molecules)[76-79]. To this end, a bioinformatic platform for the Predictive Informatics of Secondary Metabolomes (PRISM) has been developed by the Magarvey lab to infer the location of these assembly line systems directly from the genome of an organism of interest[76]. Furthermore, the understanding of the domain selectivity within the assembly line systems allows PRISM to predict a possible structure from the detected gene cluster[76,77]. The ability to uncover the secondary metabolite potential of organisms begins to reflect how valuable microbes can be in their ability to produce several chemically distinct entities. It has been estimated that a mere 10% of the biosynthetic potential of microbes has been currently characterized[80]. Moreover, strategies are readily being developed to compare the predicted chemistry of secondary metabolites against natural product chemistries that are already known[81]. This can be extended further to relate known natural products to their respective gene clusters, also known as de-orphaning clusters, to ensure the focus remains on those gene clusters encoding potentially novel natural products[81]. To this end, the Magarvey lab has developed several bio- and chemo-informatic tools to assist in generating more targeted, or guided approaches to natural product discovery. Of them is PRISM as described above, and two others are Generalized Retro-biosynthetic Assembly Prediction Engine (GRAPE), and the Global Alignment for Natural Product Chemoinformatics (GARLIC)[81]. GRAPE is a retrobiosynthetic algorithm that enables known natural products to be broken down into their biosynthetic units, and can be directly compared against detected BGCs[81]. Tools such as these serve as valuable resources to effectively identify new sources of microbial natural products in the genomic era of discovery.

1.10 Bioinformatic mining for Microbial Natural Products

Bioinformatic programs such as PRISM present an enormous advantage moving forward in the modern era of natural product isolation, but require pairing with additional bioinformatic tools that can identify the predicted molecules in microbial extracts. Efforts to address this need have been developed in recent years such as Informatic Search Algorithm for NAtural Products and DEREPLICATOR[82,83]. Keeping inline with the interests of the Magarvey lab to develop bioinformatic tools to accelerate natural product discovery, an additional resource known as the Computational Library for Analysis of Mass Spectral Data (CLAMS) has been developed (Internal Bioinformatic Tool from Dejong *et al.*, McMaster University). CLAMS enables the detection of plausible small molecules within a mass spectra file and identifies them as "peaks". Furthermore, CLAMS can report the respective mass to charge ratio of detected peaks, which is then used to compare against an in-house database of small molecule data to identify those which may relate to previously known microbial products. Further aiding this resource as a valuable bioinformatic tool, is the ability to infer which peaks are the same, or different between collected mass spectra of interest. An extension of CLAMS has further been developed to assert those inherently unique peaks that may be explicated related to a single strain through comparison against a database containing the mass spectral information gathered within the lab.

Continued development of bioinformatic tools are inherently necessary to bridge the gap between detection of BGCs within a genome and identifying those encoded products in microbial extracts. Having devised methods to aid in this task, brings forth a

unique opportunity to identify those potential peaks of interest that may have been missed during the traditional era of antibiotic discovery.

1.11 Resistance Genes and the Genomic Era of Natural Product Discovery

The steady growth of genomic information has not only allowed for a greater understanding of the genetic basis of secondary metabolite biosynthesis, but also an increased ability to assess changes in the resistance landscape. Significant research has been dedicated to surveillance and monitoring of clinically relevant resistance genes. Databases such as the Comprehensive Antibiotic Resistance Database, have been compiled to highlight what is known about clinically-relevant antibiotic resistance, and its molecular basis[84]. Resources such as these have also been accompanied by several studies dedicated to providing surveillance measures in a clinical context[85,86].

As microbes produce antibiotics, they are required to also co-produce resistance genes to allow the host to avoid lethality from the effects of the small molecules they produce[49]. As described above, this notion has resulted in the identification of several self-protection, or resistance genes within the confines of BGCs, that exhibit relationships to the specific class of molecules they encode. This impending relationship between resistance genes and chemical scaffolds has been used to enrich for isolation efforts of molecules with similar chemical scaffolds, highlighting the underlying relationship that exists between the two[50]. As such, further studies involving the resistance genes present within the BGCs have demonstrated a relationship to the classes of small molecules, and corresponding molecular target, such as the ribosomal methylation resistance genes of macrolide antibiotics[65]. The

repeated defining of these relationships, especially in respect the main classes of natural products, suggests that these antimicrobial resistance genes may serve as a characteristic, or defining feature of these chemical class of natural products.

Further investigation into these relationships involved with antimicrobial resistance genes and target prediction, may allow one to envision methods that assert the mode of action without extensive structural knowledge. By pairing with the advancements of the genomic era brings forth a unique opportunity to reveal resistance genes associated with BGCs. By further focusing on the relationships between antimicrobial resistance genes and their respective targets, may allow for us to further delineate the complex relationships occurring between the two. If a sufficient method to address this concept can be generated in a systematic fashion, would allow for significant advancement in the ability to postulate the potential molecular target of predicted natural products from the genomic information. By creating such a process, may help better navigate the wealth of predicted BGCs, to ensure focus remains on those which may diverge mechanistically.

1.12 Challenges of Natural Product Discovery in a Genomic Era

The genomic era brings forth an abundance of sequencing information that goes beyond what is generally able to be deciphered manually. At the same time, technology is rapidly advancing in ability to develop appropriate bioinformatic platforms to manoeuver the wealth of sequencing information. This would allow one to envision a generation of natural product isolation that diverts away from the classical means of bioactivity guided fractionation and move forward with data driven approaches. Despite the promising

potential of genomic-driven natural product discovery, challenges with such methods exist[87]. Often, BGCs are silent, or not expressed under normal laboratory conditions[88,89]. Overcoming such a hurdle can often require laborious efforts to activate or upregulate a given cluster or involve heterologous expression which is accompanied with its own set of inherent challenges[90,91]. Furthermore, direct identification of structure predicted secondary metabolites within a complicated microbial extract can vary greatly due to enzymatic promiscuity and the inherent potential for post-translational modifications[23].

It is also well appreciated that not all microbial BGCs encode secondary metabolites with antibacterial properties[92,93]. For example, this would include molecules such a siderophores that are important for a microbe's ability to sequester iron[94]. Furthermore, siderophore compounds are encoded by the same NRPS, and PKS machinery as other microbial natural products[95]. The presence of these types of BGCs, among others, would hinder or convolute the search for BGCs that encode for new antibacterial small molecules. Therefore, it is critical that genomic-guided discovery methods employ a logic to discount both BGCs associated to known molecules, as well as BGCs that encode for secondary metabolites with undesirable activities or characteristics.

1.13 Microbial Strains with Potential to be Sources of Novel Antimicrobials

Actinobacteria in the past have been a major focus of traditional discovery methods, with many clinically relevant antibiotics being isolated from them[96]. Extending research further into microbes outside of the of the Actinomycetes is hoped to bring forth new sources of chemically distinct natural products. Delving into organisms which were not

extensively investigated during the golden era of discovery is accompanied by its own series of potential complications as many organisms now, are being isolated from unique environments, leading to possible difficulty in cultivations under normal laboratory conditions[97].

The potential to identify microbes who might be "genetically primed" to produce secondary metabolites would bring forth an additional level of confidence in tackling the growth conditions of these underexplored producers. This can essentially allow for guided isolation efforts into various aquatic and soil dwelling bacteria such as *Aquimarina muelleri,* and *Lysobacter gummosus*. Shifting focus to the genome of microbes would allow insight to the secondary metabolite capacity of these microbes, but also show the depth of potential that may remain in those producers that may have also been mined in the past.

1.14 *Flexibacter* sp. as a Potential Producer of Bioactive Natural Products

Flexibacter sp. is a Gram-negative environmental bacterium, which is most commonly known for it association as a fish pathogen[98]. Different *Flexibacter* species have been isolated from various environmental locations, including areas of rotting swamp grass, marine environments, and general soil locations[99-101]. In the past, two *Flexibacter* sp. ATCC 35208, and *Flexibacter* sp. ATCC 35103, have been interrogated and found to produce an intriguing family of secondary metabolites, the monobactams[99,100]. Monobactams are monocyclic beta-lactam containing molecules with weak antibacterial activities[102]. Monobactams have been of clinical interest with the development of Aztreonam, and their stability in the presence of beta-lactamases[102]. Despite the finding of

one of the industry's most intriguing families of natural products, *Flexibacter* sp., ATCC 35208 has not been extensively mined for the potential of other bioactive secondary metabolites. Recent genomic sequencing of *Flexibacter* sp., has revealed that its secondary metabolite potential goes beyond monobactams in form of other possible NRPs, suggesting *Flexibacter* sp., may be a valuable candidate for further investigation. Pairing this knowledge with the appreciation for the biosynthetic capacity already achieved by this organism suggests there may be potential to extend even further to other uncharted molecular targets.

1.15 Thesis Overview

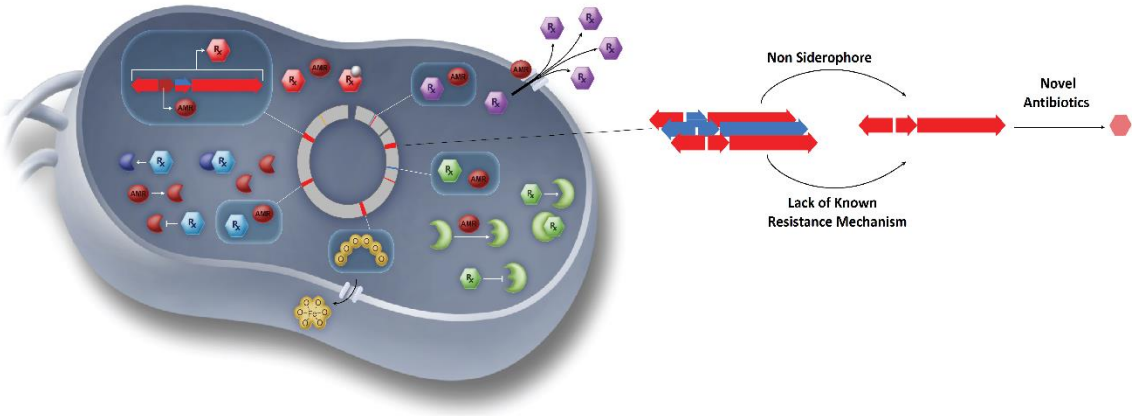


Figure 1.1 Thesis overview. Using genome mining and resistance profiling to highlight natural products of interest that may serve as valuable sources of new antibiotics.

The impending crisis involving antibiotic resistance further reiterates the need for new platforms for antibiotic discovery. Natural products in the past of proved to be a valuable resource of bioactive metabolites, and history urges us to revisit them again in

efforts to address the antibiotic resistance crisis. This need for new discovery platforms that consider antibiotic resistance genes is the overarching objective of this thesis. By defining such a method using resistance profiling, and genomic mining will reveal new sources of microbial natural products with antibacterial properties, the hypothesis to my work.

The central research project, prepared for submission for publication, delves into a devised platform for the discovery of microbial natural products. It is aimed to reveal which organisms have a strong potential to produce bioactive molecules that evade common resistance mechanisms, based on a selective approach to avoid re-discovery of known molecules and identify those which differ mechanistically. These findings will enlist a multidimensional approach leading towards a method to reveal new antibiotic agents, showcasing how valuable microbes continue to be in efforts to identify the next iteration of antibacterial agents.

Chapter 2. Defining Antibiotic Molecules Refractory to Known Resistance Mechanisms from Microbial Genomes via Machine Learning

2.1 Chapter Preface

The genomic era of discovery brings forward an unique environment for defining the next iteration of antimicrobial natural products. As with the challenges associated with the traditional era of discovery, the genomic era brings forth its own set of challenges to move forward in defining new isolation methods. As bioinformatic, and computational technology advances, we can leverage this information to build new resources to aid in the search for new natural products. As antimicrobial resistance continues to be a major health concern, it is ever more apparent of the need to define new antimicrobials that lack noted resistance genes. As such, it is essential that new methods of antibiotic discovery are focused on assessing the possibility for resistance to emerge, at the beginning of isolation efforts.

The following chapter is formatted as a manuscript that is prepared for submission for publication, in which I am the lead author. I conducted the experimental design, and performed all experiments (except for those stated below), curated ATP datasets, interpreted results, and wrote the manuscript. Dr. Haoxin Li prepared the work related to the siderophore prediction engine, curated siderophore datasets, isolated siderophore compounds, and contributed to the manuscript. Dr. Maclean Edwards also contributed to the experimental design, generated SIPE and ATP algorithms, and validated statistical methods. Dr. Jabed Tomal carried out the random forest experiments, and validated

statistical methods. Nishanth Merwin performed the global analysis of siderophore, and ATP predicted products, and provided guidance on the manuscript. Chris Dejong assisted in curation of datasets. Michael Skinnider programmed the additions of hidden Markov models to PRISM, and contributed to the manuscript. Dr. Nathan Magarvey contributed to the experimental design, and contributed to the manuscript.

2.2 Abstract

Microorganisms have historically been the main source of antibiotic agents. The central approach to realize these antibiotics is random screening and cultivation of microbes, collecting organic extracts and conducting screening of them using bioactivity guided fractionation. Accumulated information of existing natural product structural classes, and known antibiotic resistance is often not influential of this discovery process, and lacks modern data-driven methodology. Genome sequencing has revealed many potential antibiotic pathways, yet methods to decipher this information, and devise predictive algorithms to define relationships to known antibacterials is lacking. Moreover, it remains unclear how to select biosynthetic gene clusters and the natural products they encode, that are refractory to current observed antibiotic resistance. In this work a strategy is described to change how microbes are explored for antibiotic agents using a series of allied algorithms that creates a platform to delineate biosynthetic clusters that create novel antibiotics lacking appreciated resistance profiles. We identify features within the known chemical space to permit classification of compounds, or encountered biosynthetic gene clusters as siderophores with a high degree of accuracy. Using a devised compendium of antimicrobial resistance genes in the form of hidden Markov

models built into PRISM, and the associated known chemistry of natural product space, we have enabled a method capable of replicating for known antimicrobial resistance mechanisms in the form of the Antimicrobial Target Predictor. We also survey the global landscape of predicted natural products that lack defining features related to siderophores, or other known natural products to highlight the numerous entities yet to be discovered that may serve as agents in defining the next iteration of antibacterial agents.

2.3 Introduction

The emergence of multidrug resistant human pathogens refractory to the current clinical antibiotics demands new strategies to define the next generation antibiotics[1]. Defining such novel antibacterial chemistry from synthetic compounds has proved challenging, and existing screening methods of microbial natural product extracts has as well[2,3]. Origins of these challenges differ, but both display a general inefficiency and a bias towards existing antibiotic scaffolds. Natural product screenings implementing cellular bioactivity tracking is challenged with rediscovery of knowns, whereas synthetic compound leads realized through target-based screening are often inactive on whole cells[4]. Of the known targets that lead to killing, natural antibiotics hit all the known described 58 targets[5]. Mapping of these naturally-derived antibacterials, particularly of the polyketide (PK) and nonribosomal peptide (NRP) classes, shows that their chemical diversity parallels target diversification[5]. Using retrobiosynthetic clustering also demonstrated that ~40 per cent of the time agents with chemically distinct scaffolds act through distinct biological targets/mechanisms[5].

Genomics of microorganisms have revealed large collectives of natural product biosynthetic pathways[6]. Products of the vast majority of these pathways can not be readily ascribed to known compounds, evidenced by the ~1000 cluster-to-compound matches known to date[7,8]. Of the biosynthetic pathways identified, PK and NRP pathways are in particular abundance, and of interest, as they may encode novel entities. The molecular diversity afforded by the modularity of the PK and NRP biosynthetic logic, when considering the combinatorics and permutations would certainly cover vast chemical space[9]. If one envisions the possible combinatorics from genetic code, the natural product diversity using the known ~500 PK and NRP monomers and associated capacity to create scaffold diversity is greater than 500 factorial. Those detectable in genomes, in theory, encode agents that have been selected for, and the failed chemical combinations with no favorable action would have been lost over evolutionary time[10]. Though to date, the known number of NRP, PK and hybrids thereof only represent less than 10% of the encodable space, and suggests an intriguing notion that those newfound clusters could encode novel antibiotic chemistry[11]. A key challenge is to define the predicted natural products (pNPs) and compare to the known chemistry. The concept of divergent molecules on its own, is not sufficient to suggest the actions of these pathways will produce antibiotics with other functions. Siderophore molecules, agents with iron-binding/retrieval functions lack antibacterial action, and are also produced via PK and NRP assembly logic[12]. As such, decoding this information to define the unknowns and infer functions using machine processing is of the essence to classify the molecular actions (siderophore, known antibiotics, and possibly novel antibiotics).

New algorithms are required to compare pNPs to natural products (NPs) that can account for, in some cases, poor prediction accuracy (degeneracy from DNA-RNA-protein-small molecules), yet define differentiated novel chemicals that lack susceptibility to known resistance. Most tools have focused on biosynthetic gene cluster (BGC) detection, and to varying degrees' structural prediction alone[13-16]. Other confounding issues are the need to sort pNPs that are not antibiotics (e.g. siderophores), reveal novel agents, and selectively target their isolation. Connectivity of BGCs to those pNPs and physical isolation of these unknowns is also still in its relative infancy, and must also be addressed to conduct systematic mining to impact antibiotic discovery[17]. Also, many known antibiotics and siderophores have not yet been connected to their cognate BGCs[18]. Construction of an integrated framework that also takes into consideration antibiotic resistance is required to create intelligent selection of the next generation of antibiotic leads. These methods would also need to be efficient enough to deal with the large deluge of biosynthetic data. Machine learning and algorithms to define the molecular functions of predicted structures has not yet been presented with a scale to prioritize molecule isolation. Here we present a machine-driven process to create a new method to identify evolved antibiotic molecules that are chemically and mechanistically differentiated, and not immediately susceptible to observed resistance. At the core of this new strategy is the formative training data of the known chemistry, resistance, and principles of machine learning/algorithm bio-chemoinformatic design. Application of this new approach is set forward on a large collective of sequenced BGCs defining new molecules with varying

spectrums of functions (siderophores and antibiotics), and activities on drug resistant pathogens.

2.4 Results and Discussion

2.4.1 A Compendium of Resistance Genes Connected to Molecular Targets

Antibiotics may be described as "selfish molecules" analogous to the concept of a "selfish gene". Within the operons encoding these selfish-molecules are likewise counter-strategies to ensure it is protected from the inhibitory actions of its product[19, 20]. Self-protection mechanisms parallel the four main antibiotic resistance strategies found in clinical resistance, including: target inactivation, target decoy, antibiotic modification, and antibiotic expulsion (Fig. 2.1). Initial cataloging of resistance determinants has focused on defining their presence in clinically relevant pathogens but given the origins, one may conceive their utility in sorting and clustering antibiotic producing pathways[21,22]. Several resources have been dedicated to the tracking and providing surveillance of clinically associated resistance genes but not as extensive for producer encoded resistance[21, 23-25]. Despite the acknowledgment of importance of resistance genes in governing the success of antibiotics, a smaller emphasis has been placed on defining the intertwined relationship between resistance genes and the mechanism of action of antibiotics.

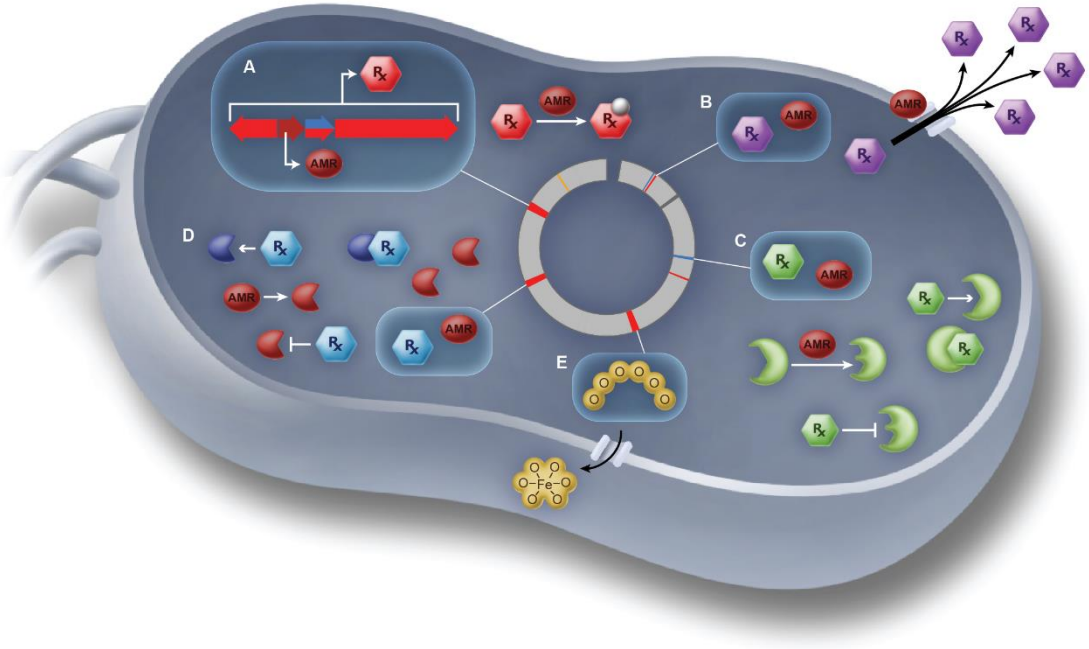


Figure 2.1. Microbial secondary metabolites are encoded by biosynthetic gene clusters that contain a gene associated with serving a self-protection function. Common resistance mechanisms encoded by resistance genes in biosynthetic gene clusters are (A) target modifications, (B) transporters, (C) target modifications, and (D) decoy mechanisms. Other non-antibacterial products encoded by the same assembly line systems include (E) iron-binding molecules (siderophores). Image courtesy of Sheena Gingerich.

To survey the known antibiotic self-protection genes, an extensive literature search of published data was compiled of the antimicrobial resistance (AMR) genes encountered within environmental and clinical microbial organisms. Each instance of a BGC associated resistance gene was flagged, and a hidden Markov model (HMM) was created for their subsequent detection. Cut-off scores were empirically determined, culminating in 301 HMMs that also included a set of clinical resistance genes that were combined into a single searchable set (Supplementary Table 2.1)[23]. For the noted clinical

observed resistance genes, we refined for several of the available HMMs to ensure specificity. Having set clear cut-off scores that would lead to few instances of false reporting, we integrated these within the Prediction Informatics of Secondary Metabolomes (PRISM) platform[15]. Within the PRISM framework, these HMMs can be used to query input BGCs for the presence of candidate resistance genes. Taking these as a complete set, we moved forward with creating a linkage of these resistance determinants to the molecular targets of the antibiotics that they resist. This map of correlations between the curated resistance genes and molecular targets revealed associations to 25 antibacterial targets (Fig. 2.2, Supplementary Table 2.2). Comprised in Fig. 2.2, is a series of depicted modifications that are representative of individual AMR genes, which have been developed into HMMs. Furthermore, respective modifications have been segregated into those that are specific to the cell wall/membrane, or cellular targets. Segregating the compiled AMR genes one step further, aimed to highlight those determined modifications within each category that are reflective of direct target modifications, or those which are cytosolic modifications (e.g. antibiotic modifications, or expulsion pumps). Further breakdown of curated resistance mechanisms revealed 37% to be associated with direct antibiotic modifications, 32% as transporters, 15% representing modifications to antibacterial targets, 5% as target decoys, and 11% were found to be of less common resistance functions (e.g. immunity proteins, or clinical genes modulating resistance). The creation of the compendium of environmental and clinically relevant resistance genes, and drawing associations to known antibacterial targets was a critical first step to enable the use of the information in an integrated fashion.

Of the accumulated antibiotic associated resistance genes, the majority (78%) could be specifically tied to a known antibacterial target, whereas nonspecific, or multidrug resistance genes were less frequent (21%), and finally a small percentage were identified to an unknown molecular target (1%). Of the surveyed resistance genes, the most frequently observed target relationship was those resistance genes related to inhibition of the bacterial ribosome with 90 instances (38%). This follows a similar pattern as exhibited in the context of microbial natural products, with most evolved chemical entities also targeting the bacterial ribosome[5].The second most commonly observed relationship between resistance genes and molecular targets was to penicillin-binding proteins with 52 instances (22%), followed by D-Ala-D-Ala chelators with 20 instances (9%). Following the top three molecular targets, most remaining targets exhibit small percentages, or fewer than 10 instances of a direct relationship between resistance genes and molecular targets. The entire list of molecular targets and their associated resistance genes can be seen in Supplementary Table 2.3.

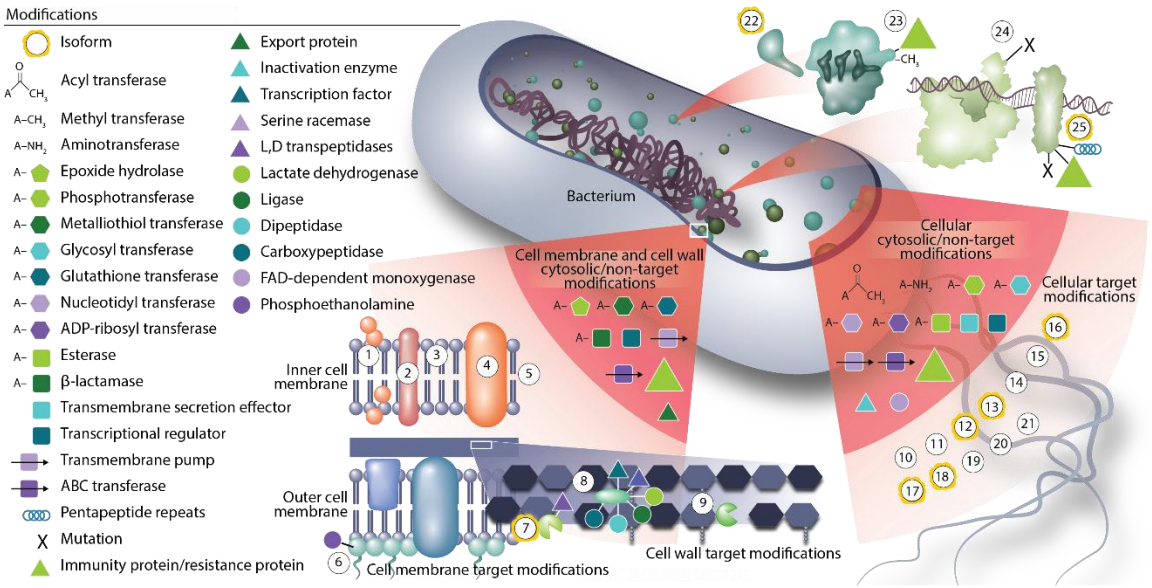


Figure 2.2. A compendium of antibiotic resistance determinants connected to their associated molecular target. A comprehensive survey of the known resistance mechanisms exhibited by bacteria identified 25 targets of the inner cell wall membrane (1-5), outer cell membrane (6), cell wall associated enzymes (7-9), amino acid metabolism (10-21), and individual enzymes (22-25).Modifications can be broadly grouped by those which are cell membrane and cell wall cytosolic modifications, direct modifications to the cell membrane and cell wall, as well as cellular cytosolic or non-target modifications, and cellular target direct modifications. Full target legend is listed in Supplementary Table 2.2. Image courtesy of Sam Holmes.

2.4.2 Devising a Training Set of Predicted Natural Products from Known Biosynthetic Gene Clusters

The decrease in costs associated with next generation sequencing has spurred a generation of characterizing the biosynthetic genes responsible for various classes of microbial natural products. To this end, a set of microbial natural product BGCs was

collected and comprised of 227 antibacterial, 31 siderophore, and 104 antifungal BGCs (Supplementary Table 2.4). This curated set of known BGCs provided the basis for creating pNPs with known molecular functions to generate the training and test sets for the custom optimization algorithms. Each of the curated BGCs were analyzed for their predicted biosynthetic substrates and resistance genes (from section above) using PRISM, to be used as a feature set. The results of this surveying, were also used to manually annotate to relate known BGCs, detected AMR HMMs and molecular targets of antibacterial BGCs (Supplementary Table 2.5). Each of these known natural products were also broken into chemical subunit pieces, using a Generalized Retro-biosynthetic Assembly Prediction Engine (GRAPE)[18]. GRAPE breaks down each represented natural product into its respective substrates, and monomer composition. It was reasoned that upon breaking down microbial natural products into their biosynthetic origins, would reveal distinct features that would allow the differentiation between antibiotics, and siderophore natural products. GRAPE generated subunits and the PRISM predicted substrates are relatable using the Global Alignment for Natural Product Chemoinformatics (GARLIC)[18]. This resource enables the direct comparison between PRISM predicted substrates, and the subunits of previously known microbial products to allow comparisons to be established between pNPs and the structural features of known natural products. Collecting the data of the known GRAPE subunits and the PRISM subunits from these curated sets alongside the resistant profiling, would create a feature set for downstream predictive algorithms to predict the molecular function. Furthermore,

these features will be used as a foundation for devised algorithms to infer a prediction when a new BGC is encountered.

2.4.3 Machine Learning of Known and Predicted NP Siderophore Chemistry

A key challenge in mining microbial genomes for natural products is selecting what BGCs may encode in respect to their function, such as those with antibacterial properties. PK and NRP biosynthesis paradigms lead to diversified molecules with a range of functions, and siderophore chemistry is one of such functions[12, 26-28]. As previously mentioned, siderophore molecules function as iron scavengers for microbes, and are unlikely to possess the necessary features to be developed as clinically useful antibiotics. To define what features of siderophore chemistry may be differentiating, the training set of knowns described above was used. In addition to the BGC data above, we also undertook a comprehensive literature review, covering review papers in addition to primary literature, to collect the chemical structures of all known siderophore compounds. A total of 384 siderophores were collected with information including their structures, producers, and membrane receptors (Supplementary Fig. 2.2)[12, 29-37]. Manual annotation revealed that 204 of these siderophores, from 29 genera, were associated with one of 14 known membrane receptors (Supplementary Table 2.6). Among the 384 siderophores, 43% are peptidic siderophores, whereas the other 57% are non-peptidic siderophores.

To model the structural characteristics of siderophores, a control set of non-siderophores was additionally required. As we decided to use 284 siderophore compounds in our training set, we randomly selected a control set of twice the size of the training set,

selecting 568 non-siderophores from our in-house microbial natural products database. After curation of all compounds, we used GRAPE to reveal the monomers within each chemical structure. We first considered the distribution of substrate monomers between siderophores and non-siderophores, by simply looking for those monomers with the highest occurrence within each set, as determined by GRAPE. There was a clear difference between the distributions, wherein six of the top ten most common monomers from the siderophore set have the characteristics to facilitate iron binding; hydroxyornithine, lysine, 5,6-dihydropyoverdine chromophore, 2,3-dihydroxybenzoic acid, beta-hydroxyaspartic acid, and benzoic acid. The entire list of siderophore monomers and the top monomers from non-siderophores can be seen in Supplementary Table 2.7.

Following the initial insight into the notion of there being distinct differences between siderophore and non-siderophore structures, we developed a random forest model to classify compounds as siderophores. A random forest model was trained for binary siderophore classification using the 131 structural features identified by GRAPE. From the 384 siderophore compounds, 74% were randomly selected as the training set, and 26% were retained as the test set. In regards to non-siderophore compounds, 91% were randomly selected as the training set, and 9% were retained as the test set. Our model showed a sensitivity (siderophores correctly predicted as siderophores) of 97% and specificity (non-siderophores correctly predicted as non-siderophores) of 98%, with a false negative rate (siderophores incorrectly predicted as non-siderophores) of 3% and a

false positive rate (non-siderophore incorrectly predicted as siderophores) of 2%. The overall prediction accuracy was 95%.

Having shown that a random forest model permitted classification of siderophores from structural information, we next considered siderophore classification of pNPs from BGCs. Our first attempt was to use the same random forest model built from structural information to predict siderophore pNPs obtained from PRISM. However, because the accuracy of PRISM in substrate prediction is less than 100%, the performance of the model was not adequate. We therefore sought to develop a second random forest model based on known siderophore and non-siderophore pNPs. We collected all 31 known siderophore BGCs and 406 non-siderophore BGCs. The second random forest model was trained using 20 siderophore and 254 non-siderophore pNPs with all the information that PRISM detects, including amino acid and polyketide substrates, tailoring genes, sugar genes, and resistance genes (see Methods). Performance on the remaining pNPs showed a sensitivity of 91% and specificity of 93%, with a false negative rate of 9%, and a 7% false positive rate. The overall prediction accuracy was 84%.

Random forest models are effective at revealing variable importance and their relationship with prediction error. Within this model, the most important variables for structural and cluster predictions are shown in Fig. 2.3B. Common important variables for both structural and pNP classification included 2,3-dihydroxybenzoic acid, malonate, methyl malonate, sugar presence, lysine, valine, cyclic tailoring reaction, hydroxyornithine, and serine.

Within the training dataset, there were significantly more non-siderophore pNPs leading to misclassifications favouring non-siderophores. Biased weighting and sampling techniques favouring the underrepresented siderophore pNPs could alleviate this misclassification error[38]. To tune the random forests, we plotted the out-of-bag prediction errors against number of trees grown for the compound and pNP data (Supplementary Fig. 2.3). For the compound data, the out-of-bag error is stabilized at around 400 trees grown. For the pNP data, the out-of-bag error is stabilized at around 300 trees. Hence, in both the data sets, we have achieved our tuned random forest models by growing 500 trees. However, when we investigated classes of pNPs that were often incorrectly predicted, we found that lipopeptides, such as daptomycin, polymyxin and teixobactin, were always predicted incorrectly (Supplementary Table 2.8). This issue using the existing machine learning strategies lead us to consider how we may create an alternate algorithm that would create the differentiating power to separate lipopeptides (often antibacterial) from siderophores. Lipopeptides shares some structural similarities with NRPS-derived siderophores: for example, both classes have several amino acids with charged side chains. The limited number of annotated siderophore BGCs complicated the ability to accurately capture the differences between siderophore and lipopeptide pNPs. In lieu of these challenges, we decided to use a second approach, linear regression, for siderophore pNP classification.

2.4.4 Broyden-Fletcher-Goldfarb-Shanno-optimized Linear Coefficients for Siderophore and Antibiotic Prediction

A custom optimization strategy was used to specifically address the issue of poor predictive power for siderophore pNPs from lipopeptide BGCs. Here, the training strategy does not aim to achieve the best total separation, but to minimize the total pNPs that are misclassified. This idea was translated into a function as described in Methods, and referred to as the siderophore identification prediction engine (SIPE). Of note, this function only uses the errors in misclassified observations to iterate and improve. This aims to create a better classification by specifically improving misclassification instead of boosting total separation. Using the Broyden-Fletcher-Goldfarb-Shanno minimization algorithm to minimize this defined function, a set of optimal linear coefficients are produced. When applied to a given pNP, an unbounded score is output, which can be used to represent the likelihood that a given pNP is a siderophore. The performance of this method is defined by the following stats: an overall accuracy of 97%, a false positive rate of 1%, and a false negative rate of 40%. 55 of 160 pNPs received an unknown classification when compared with random forest model using the same training and testing sets. However, this model correctly classified products from lipopeptides BGCs as non-siderophore producing BGCs, in addition to other non-siderophore BGCs in the testing set (Supplementary Table 2.8). To compare the performance of the two models, we plotted the receiver operating characteristic curves. The overall performance of SIPE and the random forest model were comparable but a key differentiating function of SIPE is the reduction of misclassification such that lipopeptide antibiotics were not defined as

siderophores (Fig. 2.3). We therefore selected the linear regression model as our default for potential siderophore pNP classification, as we only wish to eliminate pNPs that are siderophores, and keep those with potential antibacterial properties to be assessed in the next segment of the pipeline.

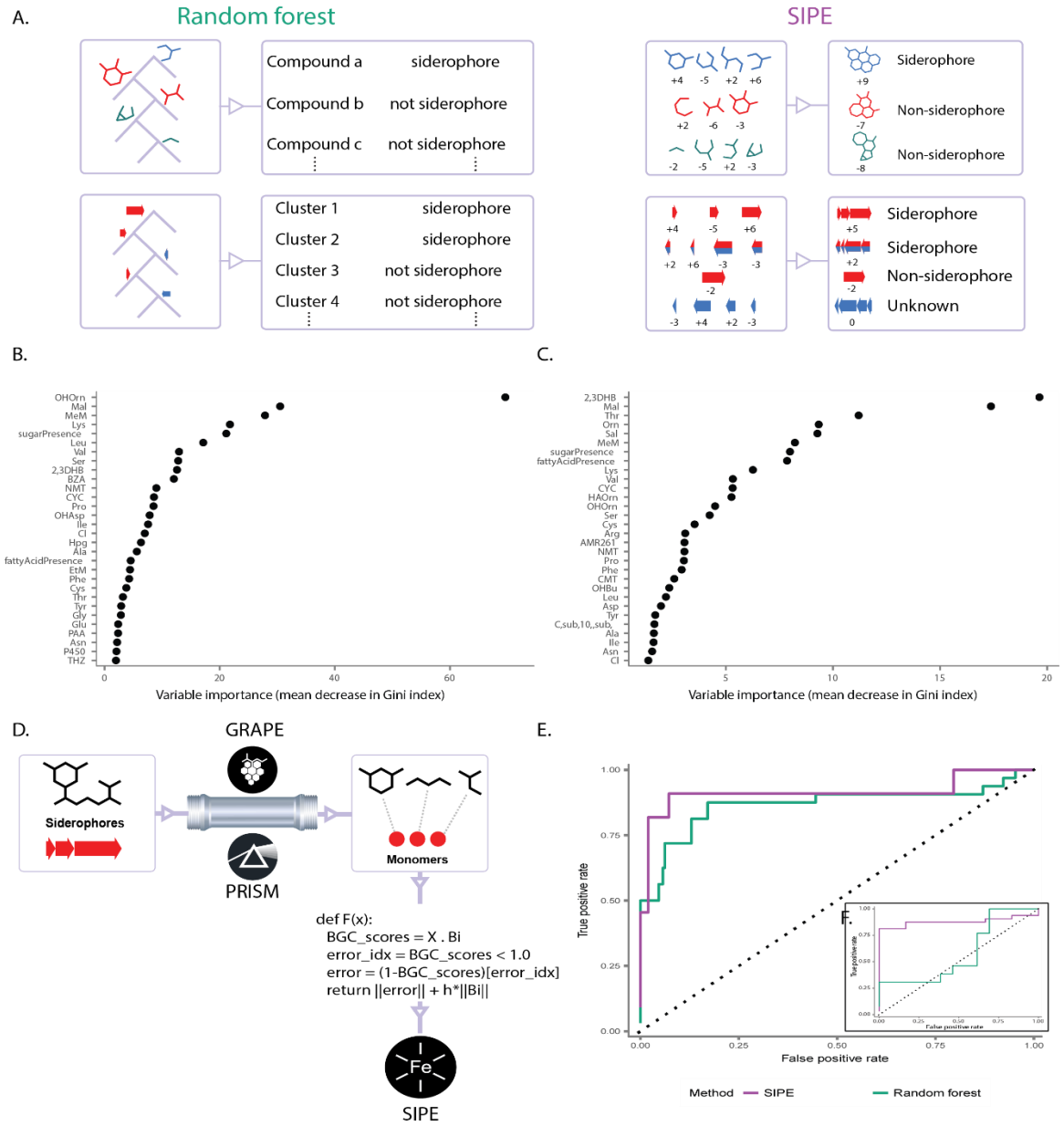


Figure 2.3 Random forest and SIPE classification of NPs as siderophore compounds. A. Work flow representation of random forest and SIPE models. The important variable plot of random forest model shows the top 30 most important variables for classifying compounds (B) or predicted natural products from biosynthetic gene clusters (C) as siderophores. D. Information gathered via PRISM and GRAPE through structural and genetic data is used in the generation of SIPE. E. Receiver operating characteristic curve comparison between random forest and SIPE models for PRISM predicted siderophores using the whole test set. F. Reciever operating characteristic curves of random forest and SIPE using only lipopeptides from the devised test set.

2.4.5 Optimization Strategy for the Generation of the Antimicrobial Target Predictor (ATP)

As we have now generated a method to classify pNPs as siderophores using SIPE, the next task was to address the likelihood of a pNP diverging mechanistically through its predicted features. A custom optimization strategy was developed to assist in predicting which pNPs will both diverge chemically, but also do not display immediate inference of a currently known resistance mechanism. The generation of the ATP pipeline can be seen in Fig. 2.4. The first arm of the ATP pipeline demonstrates the use of the AMR genes in inferring a potential target, based on the frequency, and precision of the AMR HMMs within the devised BGC set, as described in Methods (Supplementary Table 2.9). Of the 301 antibacterial specific resistance genes curated, 101 were present within the devised BGC training set. Of the AMR HMMs present, 88 were classified as single target resistance genes, meaning that they correlate with an exclusive target, based on a calculated precision above 70%. The limited number of annotated BGCs limits the ability to capture all AMR genes in the representative set to allow for full prediction power. Furthermore, a subset of the compendium is comprised of clinically associated AMR genes, which are likely to have evolved in locations within bacterial genomes outside of

dedicated BGCs. As described in methods, the ATP only asserts a prediction based on the detected AMR HMMs present within the identified BGC.

Within the devised training set of BGCs, 168 (149 antibacterial, 14 antifungal, and 5 siderophore) contained detectable known resistance genes. This suggests that within our devised training set, 149 of 227 antibacterial BGCs have previously characterized, or known self-protection genes. Using this subset as a notion to test the simple prediction engine using only AMR genes, a target prediction was generated 65% of the time, with an accuracy of 99%. As in, the simple prediction engine could infer a target 65% of the time within the devised training set due to the presence of a noted single-target resistance gene. In cases of the presence of a noted multi-target resistance gene, no target is predicted, and rather all encountered targets associated with that resistance gene are listed. Three individual testing measures were applied to assert the ability to identify the target of a pNP that is outside of the devised training set of known BGCs. Using a leave-one-out cross validation, a prediction was made 39% of the time, with an accuracy of 80%. The second validation used was Monte Carlo cross-validation, which resulted in the ability to infer a target prediction 47% of the time, with a mean accuracy of 76%.

Although the performance was adequate in predicting the antimicrobial target when a resistance gene was present, it was essential to devise a method that would also infer the potential target of a pNP based on predicted chemical features. In this sense, the simple prediction optimization approach was combined with GARLIC, a previously published method to compare the known chemistry between known NPs, and pNPs from a detected BGC[18]. This enabled an approach to account for both chemical and resistance

features associated with a pNP. Using the devised BGC set, those labelled with antimicrobial activity (227 BGCs) were used to assess the ability to properly predict an antimicrobial target, using the devised combined approach, or ATP. When ATP is applied to a given pNP, a predicted target and unbounded confidence score is outputted. This represents the likelihood of the pNP to assert its mode of action via a previously characterized target. Confidence scores aside, a correct target prediction was made 40% of the time, and was unknown 10% of the time. Although, when confidence scores are considered, a correct prediction is made 86% of the time with a confidence score above 0.5, and 90% of the time when above a confidence score of 1.0 (Supplementary Fig. 2.4). Although incorrect predictions are made, the assigning of a confidence score shows the potential of the engine to correctly identify those with both strong chemical and resistance features to assert those which may diverge in mode of action. The predicted targets, and overall confidence scores of the known antibacterial BGCs can be seen in Supplementary Table 2.10.

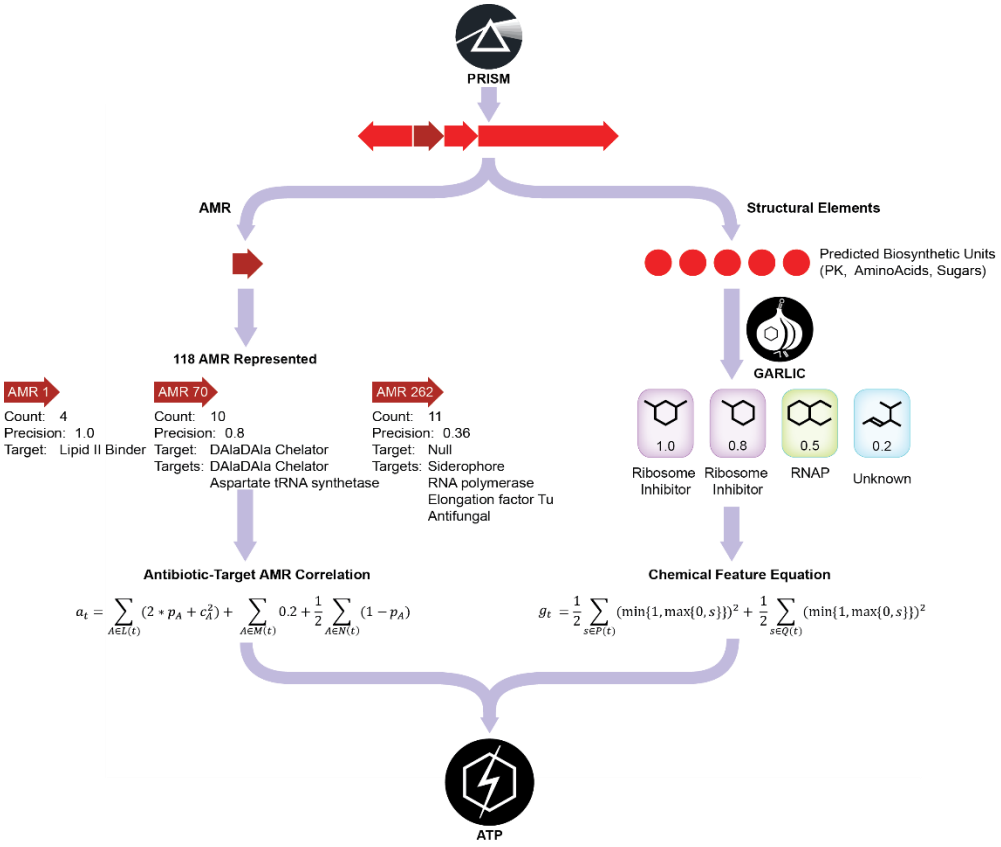


Figure 2.4. Workflow of the generation of ATP using antibiotic-target AMR correlations, and chemical features. AMR correlations are made by surveying the set of known BGCs and corresponding AMR HMM hits. Chemical feature comparisons of predicted natural products are made through GARLIC and consider associated molecular target annotations. Red represents NRPS, brown indicates antimicrobial resistance genes.

2.4.6 Analysis of pNPs Arising from the ATP/SIPE Pipeline

To further assert where focus should be placed in respect to identifying pNPs with desired activities, we sought to analyze the pNPs of all public and in-house available genomic sequences. The multi-stage process is depicted in Fig. 2.5 to identify those pNPs with the potential diverge in mode of action, and do not encode pNPs with chemotypes indicative of siderophores. We used PRISM to analyze all prokaryotic genomes (65,423 as of March 2016) from NCBI, as well as an internal library of bacterial genomes (339

genomes), resulting in the identification of 293,712 BGCs from which pNPs were generated[39]. The first stage of the pipelined process, was to limit the analysis to those pNPs with more than 3 predicted domains, and were derived from NRPS or PKS producing pathways. This eliminated 106,022 pNPs from non NRPS or PKS pathways, and 143,142 which contained fewer than 3 predicted domains. With the remaining 44,556 pNPs, we eliminated, or grouped those pNPs with high similarity scores as determined by GARLIC, to account for multiple producers of the same pNP. This was accomplished by grouping those pNPs with similarity scores ≥0.98 as determined by GARLIC, as a single entity. Upon doing so, 32,493 pNPs were determined to be duplicates, leaving 12,063 pNPs for further analysis. The first stage of the pipelined process, involved the use of SIPE to eliminate 383 pNPs as encoding possible siderophores. Throughout the ATP analysis, 744 pNPs were eliminated based on their AMR gene confidence scores alone, whereas 2705 were eliminated based on associated GARLIC scores, or a combination of GARLIC and AMR genes. This left 8231 pNPs for further investigation.

For interest purposes, the diversity of all pNPs identified (before removing duplicates) as siderophores were plotted against their similarity to known siderophore NPs. This revealed many unknown siderophore pNPs bearing different structural features, and from several different microbial producers (Supplementary Figure 2.5). This demonstrated the pNPs from genomically-detected NRPS dependent siderophore pathways contained a large degree of structural diversity from those currently known (see above section on).

To verify predictions made by our models with compounds and BGCs that have not previously been identified as siderophores, we investigated three BGCs and their products: acidobactins[40], vacidobactins[40] and potensibactin[41]. SIPE predicted that acidobactin A and B, vacidobactin A and B, and potensibactin were siderophore compounds. To confirm the prediction, crude extracts of *Acidovorax citrulli* DSM 17060, *Variovorax paradoxus* S110, and *Nocardiopsis potens* DSM 45234 were analyzed by LC/MS. In each of these instances, we could detect the masses of their iron-bound species when iron was added that increased over time to the detriment of their corresponding apo-peaks of acidobactins (*A. citrulli* DSM 17060), vacidobactins (*V. paradoxus* S110) and potensibactin (*N. potens* DSM 45234) (Supplementary Fig. 2.6, Supplementary Table 2.11). Thus, through testing these three pNPs from their associated BGCs determined by SIPE analysis, shows they are in fact siderophore molecules.

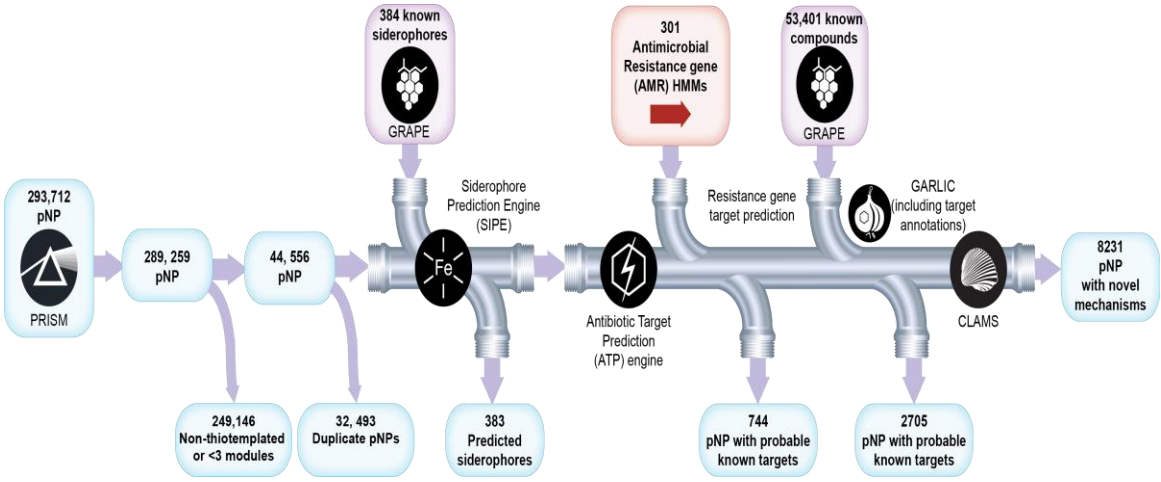


Figure 2.5. ATP workflow to identify pNPs by eliminating predicted siderophores, and identifying pNPs without common resistance mechanisms, and possess unique chemistry. PRISM predicted natural products from detected biosynthetic gene clusters are analyzed by SIPE to identify probably siderophores, followed by ATP to identify those with probable targets. The predicted natural products of determined BGCs can be identified through the aid of CLAMS by analyzing collected mass spectral data.

2.4.7 Non-siderophore Antibiotic pNPs

As predicted siderophores were eliminated in the first step of the process, the next task was to identify those with a potential to encode potentially novel antibacterials. A main theme throughout the ATP pipeline is to identify those BGCs and corresponding pNPs relating to known antibiotic pathways, or those which may exhibit cross resistance with previously identified natural products. At the first exit point, 744 pNPs were eliminated from the pipeline based solely on their associated resistance gene scores, as determined by ATP. The BGCs at this respective point were likely to contain a noted single target resistance gene, as determined by ATP, resulting in a high confidence score target prediction to be made. For example, contained within the pNPs predicted from

NCBI, would be *Saccharopolyspora erythraea*, a known producer of erythromycin[41]. A detailed ATP analysis of the erythromycin BGC, demonstrates the detection of a resistance gene with a high precision score, which would infer a target prediction to be made. Although, ATP still reports GARLIC analysis to highlight the chemical similarities to previously identified NPs (Supplementary Table 2.12). Several other noted examples of previously characterized BGCs, and associated products can be examined in a similar fashion as depicted in Supplementary Tables 2.13-2.15. This particular exit point highlights the importance of considering AMR genes, when structural predictions may not be suffcient due to inaccurate predictions, resulting in low GARLIC scores to known compounds, as displayed with the andrimid BGC (Supplementary Table 2.13).

The true potential of the devised method results from the analysis of those pNPs which were not contained within the original test set. In this respect, the recently published BGC of aldgamycin[42], and bananamide[43] were analyzed (Supplementary Table 2.16 and Supplementary Table 2.17). Aldgamycin is macrolide antibacterial agent, with members isolated from *S. avidinii*, and *Streptomyces* sp. KMA-001[42, 44, 45]. Detailed ATP analysis reveals a strong target prediction towards the bacterial ribosome, through identification of a resistance gene previously found to be associated with another known macrolide natural product chalcomycin. Interestingly, chalcomycin has been found to be co-produced with aldgamycin in *Streptomyces* sp. KMA-001[42]. The prediction is further increased through high GARLIC scores towards other macrolide antibiotics with activity annotations also suggesting the bacterial ribosome. Despite aldgamycin not undergoing mode of action characterization in the past, its strong similarities to other macrolide

antibiotics has likely suggested it would indeed target the bacterial ribosome, and this has been further suggested by ATP analysis. A second example analyzed at this portion of the ATP pipeline is bananamide from *Pseudomonas fluorescens* BW11P2[43]. Detailed ATP analysis of the BGC, revealed detection of an AMR gene encoding a MatE efflux transporter, which in the devised training set, was routinely related to other known natural products known to cause membrane disruption. Despite exhibiting low GARLIC scores towards known NPs with annotated molecular targets, the emphasis placed towards the presence of AMR genes, allows for a high confidence prediction to be made inferring the likely mode of action of bananamide to be through membrane disruption.

Moving to the second exit point along the ATP pipeline, acts to eliminate those BGCs that did not possess resistance genes with high precision scores, or lacked detection of a known AMR gene. This point focuses more on the chemical features of the pNPs alone, or in the presence of noted multi-target resistance genes. Furthermore, this may allow for inferring, and identifying potential BGCs for previously known NPs that have yet to be connected to their associated BGC. In this sense, we have identified what we believe to be the BGC responsible for LL-19020, and LL-AO341. LL-19020 was initially characterized from *S. lydicus* subspecies *tanzanius* NRRL 18036, as an antibacterial agent, and is reported to be related to the family of elfamycin antibiotics[46]. This is further reflected in the detailed ATP analysis, demonstrating despite the absence of a resistance gene inferring the mode of action, a series of GARLIC scores with associated annotations towards elongation factor Tu, allows for a high confidence prediction to be made (Supplementary Table 2.18). This further correlates with what has long been appreciated

of elfamycin type antibiotics exerting their mode of action through targeting bacterial elongation factor Tu[47].

The second example involved the identification of the BGC for LL-AO341 from its known producer, *S. candidus* NRRL 3147[48]. Detailed ATP analysis of the determined cluster revealed a target prediction to be made inferring the mode of action to involve targeting cardiolipin within the bacterial cell membrane (Supplementary Table 2.19). The prediction was made based on the chemical similarities between the pNP to the known NP telomycin, and associated molecular target annotations. Despite some structural differences in the final structure of LL-AO341, the similarities in the core backbone of each compound, allows for a target prediction through the pNP of the determined cluster. To further assert that a correct prediction was made, a crude extract of *S. candidus* NRRL 3147 was tested against previously generated telomycin-resistant strains of *Staphylococcus aureus*[5]. In comparison to wild-type strains, the crude extract was less effective at causing inhibition towards telomycin-resistant strains (Supplementary Figure 2.7). The mutation rendering the strains resistant to telomycin, had been previously mapped back to a series of inactivating mutations in the cardiolipin synthase gene of *S. aureus*[5]. This notion of cross-resistance between telomycin and LL-AO341 has been demonstrated in the past, as well as the suggested mode of action towards a component of the cytoplasmic membrane[49]. This similarity in observed activity suggests LL-AO341 is acting through the same mode of action as telomycin, through a direct interaction with cardiolipin within the bacterial membrane. Interestingly, further investigation into the BGC responsible for LL-AO341, suggests the presence of a gene encoding phospholipase

A2, that may serve as a self-protection enzyme. Cardiolipin has been found to serve as a substrate for phospholipase A2, and subsequently leads to the hydrolysis of cardiolipin[50]. Furthermore, telomycin has previously been demonstrated to possess valuable activity *in vivo*[51]. Thus, suggesting further studies involving telomycin and LL-AO341 could be of value in pursuing either as clinically relevant agents. Overall, these examples showcase the ability of the pipeline to de-replicate against pNPs that may exhibit cross resistance with known antibacterial agents.

2.4.8 Directed Isolation and Testing of pNP Antibiotics

The end pNPs of the ATP pipeline hold the strongest potential to identify those with divergent modes of action, and further highlight the chemical diversity exhibited within the pNPs from a range of microbial producers. By plotting the diversity observed within all the pNPs that pass ATP (confidence scores below 0.5), and before removing duplicate pNPs (29,191 pNPs), can reveal those that may represent new chemical classes of natural products that have yet to be identified (Supplementary Fig. 2.7). As we see by the overlapping points, GARLIC identified several pNPs with high structural similarity. This supports the notion that several microbes can produce the same pNPs. This could also occur because of sequencing bias within the NCBI database. This diversity can be further analyzed by revealing the total number of pNPs identified by ATP in different bacterial species (Supplementary Figure 2.9A). As was similar in the siderophore global analysis, most identified pNPs by ATP were from *Pseudomonas*, *Burkholideria*, and *Streptomyces* species. To account for potential bias within the NCBI database towards clinically relevant pathogens, we also analyzed the number of pNPs identified by ATP

within a certain genus, in comparison to the total number of PRISM identified BGCs (Supplementary Figure 2.9B). This revealed a distinctly different bacterial profile than had previously been identified when looking at total pNPs identified by ATP only. Although, a degree of bias exists, as some of these organisms may only be represented by a few genomic sequences. Regardless, both analysis highlight the untapped potential of microbes to produce pNPs that meet the ATP criteria to infer potential pNPs that diverge mechanistically, and good candidates for further analysis.

The large set of pNPs remaining at the end of the ATP pipeline can be further segregated for analysis. As shown in Fig. 2.5, by removing duplicates for those pNPs with high structural similarity scores, we could condense the original pool of 29,165 pNPs to 8231 pNPs. By analyzing BGCs associated with these pNPs, we can assess those remaining for their presence or absence of a previously known AMR genes. Upon doing so, 26% of those remaining at the end had an associated resistance gene, leaving 74% with no previously known AMR gene. Also important to note, those BGCs remaining at the end of the devised pipeline, are not always guaranteed to be associated with antibacterial activity. As previously mentioned, it is well appreciated that these assembly line systems produce NPs other than those associated with antibacterial activity, such as antifungal compounds, or new classes of siderophore molecules. These pNPs are likely to be present within the end pool, including those without an associated resistance gene. Although the pNPs without an associated known resistance gene highlight those with a strong potential to produce a product with a new mode of action, there is also a potential

for that pNP to be associated to a non-bacterial target, which would not require an associated self-protection gene to be present.

As several pNPs were identified for further analysis and to increase the likelihood of identifying a pNP with antibacterial activity, it was essential to add an additional layer to prioritize for further analysis. The results of the ATP pipeline, were combined with in-house bioactivity screening data performed against *P. aeruginosa* PAO1, and methicillin-resistant *S. aureus* USA 300. Upon doing so, we sorted for strains with multiple pNPs from the ATP analysis, and those that exhibited Gram positive, Gram negative, or broad spectrum activity. Several strains were identified that possessed the necessary characteristics to act as a starting point to verify the ATP platform. To this end, we focused on four strains interest; *Flexibacter* sp. ATCC 35208, *Aquimarina muelleri* DSM 19832, *Lysobacter gummosus* DSM 6980, and *Aquimarina* sp. DSM 19860 (Fig. 2.6). Overlapping with accumulated bioactivity data allows a first glance into the probability of the excreted metabolites within these organisms to produce natural products which may diverge mechanistically, which is an essential feature to identify the pNPs of the ATP pipeline with desriable antibacterial activity. Due to the impending threat of pathogens such as the ESKAPE pathogens, finding new sources of therapeutic agents that circumvent these screening strains is of utmost importance in defining the next iteration of antibiotics[52].

Further identifying the pNPs from ATP analysis in the microbial extracts requires bioinformatic tools to infer possible locations within complex microbial extracts. Using an in-house software program, Computational Library for Analysis of Mass Spectral Data

(CLAMS) (Internal Bioinformatic Tool from Dejong *et al.,* McMaster University), we can readily decipher complex microbial extracts to reveal those respective peaks within the LC-MS/MS chromatogram that may be correlated to the pNP identified by the ATP pipeline. Using CLAMS, peaks within the chromatogram can be compared on a large scale, against in-house LC-MS/MS data to identify those peaks with low observed occurrences in other microbial extracts. The resulting analysis of *Flexibacter* sp. ATCC 35208, *A. muelleri* DSM 19832, *L. gummosus* DSM 6980, and *Aquimarina* sp. DSM 19860 can be seen in Supplementary Fig. 2.10-2.13 respectively. Although the relationship between unique peaks directly correlating to the pNPs is not concrete, it provides an opportunity to explore this concept, and assess the relationship between unique peaks of microbial extracts, and BGCs within organisms.

Flexibacter sp. ATCC 35208 represents an interesting organism to further elucidate the pNPs identified from the ATP pipeline with observed Gram positive activity against methicillin-resistant *S. aureus* USA 300. As seen in Supplementary Fig. 2.10, there are several possible locations that could relate to the pNPs determined from the ATP pipeline. As depicted in Fig. 2.6, *Flexibacter* sp. ATCC 35208 contains two pNPs that possess the necessary predicted features to be NPs with divergent activity. The other pNPs present within *Flexibacter* sp. ATCC 35208, did not encode any with siderophore features, but did contain a pNP for a monobactam, SQ 28, 332, in which the BGC has recently been identified, and this is further supported by GARLIC analysis[53,54]. Monobactams are of notable interest due to their modes of action, and desirable activities that have resulted in the development of therapeutic agents[55]. Findings

such as this, highlight the metabolic potential of *Flexibacter* sp., and suggest it may be able to produce other interesting NPs. By allowing the unique peaks identified to guide isolation efforts, candidate peaks have been narrowed down to five peaks (observed m/z 614.273, m/z 574.238, m/z 988.434, m/z 963.403, and 1032.425), that we believe could be related to the pNPs of the ATP pipeline, and account for the exhibited antibacterial activity. NMR studies are currently underway to elucidate the structure of the natural product, and will be followed up with BGC analysis to confirm the relationship to one of the pNPs of the ATP pipeline.

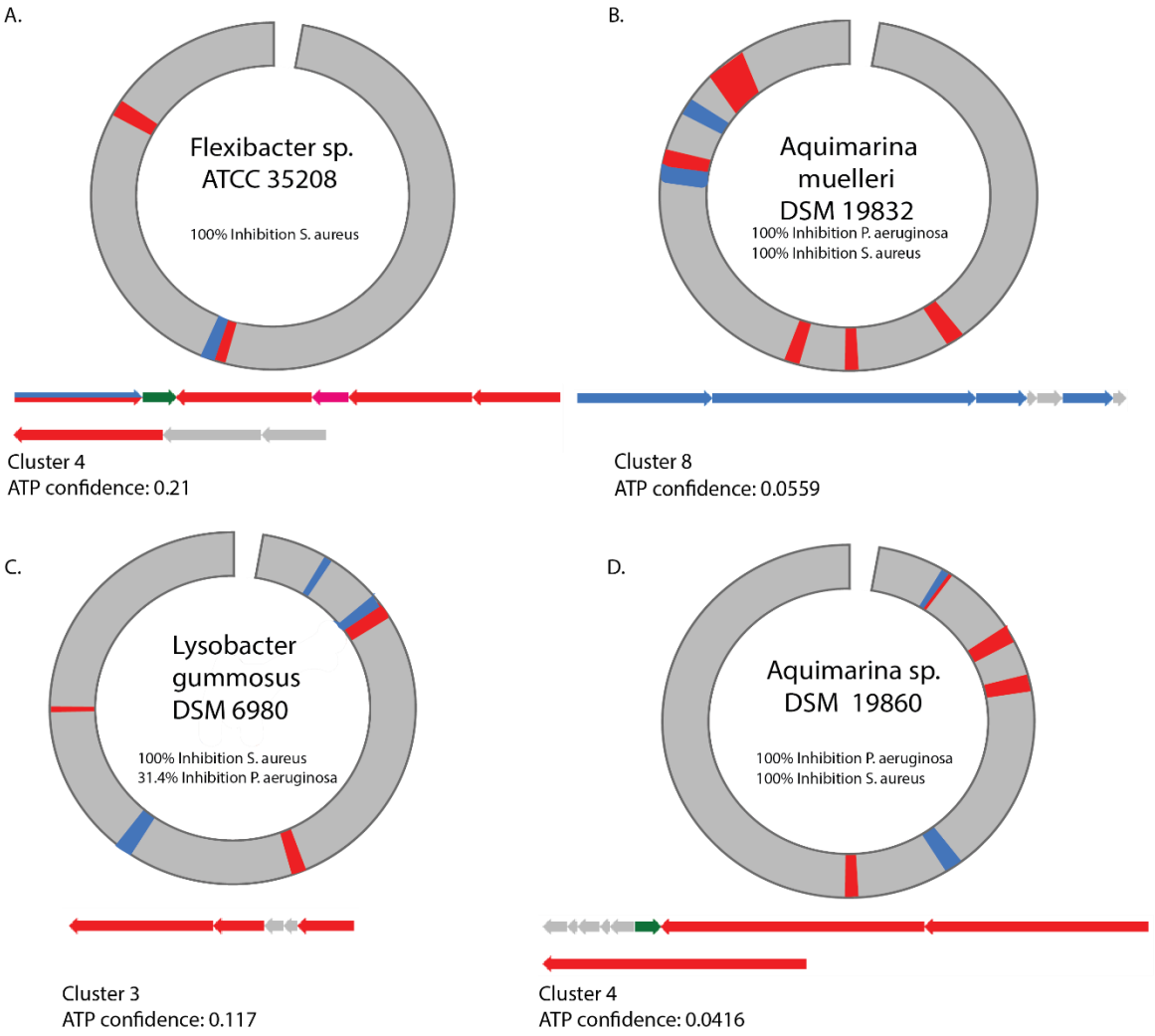


Figure 2.6. ATP analysis reveals four organisms with pNPs that lack direct association to known resistance mechanisms, and exhibit divergent predicted chemistry. Analysis focuses on (A) *Flexibacter* sp. ATCC 35208, (B) *Lysobacter gummosus* DSM 6980, (C) *Aquimarina muelleri* DSM 19832, (D) *Aquimarina* sp. DSM 19860. Represented genome circles depict number of predicted natural products from the genomic information which meet ATP requirements. Red indicated NRPS gene clusters, blue indicates PKS gene clusters, green indicates tailoring enzymes.

2.5 Materials and Methods

2.5.1 Generation of the Antibiotic Resistance Determinant Compendium

To identify resistance determinants present within biosynthetic gene clusters, a comprehensive analysis was undertaken. This set of devised HMMs was developed in continuation of the previous set of hidden Markov models previously published within the PRISM platform[5]. The previous resistance determinant library of 257 hidden Markov models was further extended by 44 hidden Markov models associated with antimicrobial resistance. 166 hidden Markov models were obtained from the Resfams antibiotic resistance hidden Markov model database[23]. Antibiotic resistance sequences were manually collected based on homology to literature defined antimicrobial resistance sequences to generate alignments via MUSCLE[56], and subsequently trimmed using trimA1[57]. From resulting alignments, hidden Markov models were generated using hmmbuild program, version 3.1b1 via the HMMer3 software package[58]. For each devised hidden Markov model bitscore cutoffs were determined via manual analysis of the search results acquired through the UniProtKB database[59], using the HMMER web server[58]. The generated list of antimicrobial specific hidden Markov models is presented in Supplementary Table 2.1.

To provide a further expansion of the compendium of AMR genes, each developed hidden Markov model was manually annotated for the broader mechanism of the AMR gene (e.g. antibiotic modification, target modification), specific modification (e.g. glycosyl transferase, ABC Transporter), as well as mode of action of the antibiotic it

is associated to. The accumulation of this information, would provide the foundation for the generation of devised optimization approaches.

2.5.2 Chemical Structures Used for Training and Testing Sets

To first assert whether a compound is a siderophore, as well as identify those which may be previously characterised compounds, all chemical structures were required to be converted to a standardized format to allow for adequate comparison. To this end, GRAPE[18], was employed to break down each chemical structure into individual monomer units. In respect to SIPE, 374 Simplified molecular-input line-entry system (SMILES) of siderophore compounds were curated based on extensive literature review. This was combined with 568 SMILES of non-siderophore compounds randomly selected from an in-house bacterial secondary metabolite database.

In respect to ATP, the GARLIC portion of the pipeline incorporates all microbial natural products compiled from extensive literature search and represented within an in-house bacterial secondary metabolite database (53,401). Each present within the in-house database was analyzed by GRAPE. Each microbial natural product that is known to be associated with antibacterial activity is also correlated with a target if known. Siderophore, and anti fungals are represented by as their own respective target for simplicity in the generation of ATP.

2.5.3 Biosynthetic Gene Clusters Used in Training and Test Sets

The BGCs used for analysis in regards to both SIPE and ATP were accumulated through extensive literature review to reveal those BGC of known microbial compounds.

In respect to SIPE, 31 fully annotated siderophore BGCs were included. BGCs used in SIPE training and testing also included 456 non-siderophore BGCs. The BGCs curated for ATP analysis was accomplished in a similar manner and represented by 227 antimicrobial, 31 siderophore, and 104 antifungal BGCs. The DNA sequence of each respective BGC was analyzed by PRISM version 2.1.2 to gather the predicted amino acid or polyketide monomers, associated tailoring genes, and access for the presence of resistance genes. The output of the PRISM analysis was used in the building and testing of both the SIPE and ATP models.

2.5.4 Features

We used 131 structural features from GRAPE to identify siderophore and non-siderophore compounds and 822 genetic features from PRISM to identity siderophore and non-siderophore biosynthetic clusters. The full lists of features were previously described in references 5 and 16.

2.5.5 Random Forests

Random forests[62] is an ensemble of classification trees which each tree is grown to its maximal depth using a bootstrap sample[63] of the training data and at each node of the tree the best split is chosen from a random sample of variables instead of all feature variables. One example of decision tree is shown in Supplementary Figure 2.1. The aggregation of classification trees in an ensemble is done by majority vote and ties are handled by a random mechanism. In this study, we applied random forests with under sampling and oversampling of the majority class and minority class, respectively. To

construct a random forests ensemble, we grew 500 trees with tuning parameter using the R package *randomForest*. Default parameters in *randomForest* were used for other settings in this model.

2.5.6 Siderophore Identification Prediction Engine (SIPE)

Linear coefficients were optimized using the Broyden-Fletcher-Goldfarb-Shanno algorithm to minimize the following function:

```
def F(x):
 BGC_scores = X . Bi
 error_idx = BGC_scores < 1.0
 error = (1-BGC_scores)[error_idx]
 return ||error|| + h*||Bi||
```

where *X* represents a 2D matrix comprised of multiple BGCs and their features, and *Bi* represents a vector containing a set of coefficients for each feature. This function returns a single value identifying the total error with the given set of coefficients. The parameter *h* (set to 0.5), is used to limit the total coefficients. An initial set of coefficients of 0 were used to initialize the optimization function. When applying this model, a linear combination of features multiplied by their trained weights is used to generate a score.

2.5.7 Antimicrobial Target Predictor

Using the devised set of biosynthetic gene clusters, and the AMR HMMs which hit within BGCs, the precision of the most frequently seen target is calculated as the relative frequency of that target for each occurrence of that AMR gene. The relative confidence represents the amount of data that supports the prediction, and is calculated by the target count of the most common target for that AMR gene divided by the highest such count among all AMR genes. All other targets for that gene are also stored. The

precision is the probability of correct target prediction (on the dataset itself), by choosing the most frequently seen target for the given AMR gene, and is given by equation (1).

$$p_A = P(C|A) = \frac{ma}{na} \quad (1)$$

Where p_a is the precision, C is the state of a correct prediction, A, represents an AMR gene, m_a is the count for the most common target seen for the AMR gene A, and n_a is the total number of times that AMR gene occurred in the data set.

Relative confidence is the normalized probability that of a given AMR gene being responsible for a correct prediction within the dataset given by equation (2).

$$c_A = \frac{P(C|A)}{max_{B \in D} P(B|C)} \quad (2)$$

$$= \frac{P(C|A) * P(A)}{max_{B \in D} P(C|B) * P(B)}$$

$$= \frac{m_a}{max_{B \in D} m_B}$$

Where C , A, m_a are as above, and D is the dataset of AMR genes. As some resistance genes can be associated with multiple targets, for prediction purposes, a single target resistance gene is a resistance gene with a determined precision above 0.7. Resistance genes with lower precisions are classified as multi-target resistance genes.

Also, included in the prediction process is the information gathered from GARLIC, a software comparison for that biosynthetic gene cluster against all known fragments (GRAPE) of all known compounds in our database, and their associated target,

if any. To make an overall prediction for the target of a given cluster is given by equation (3).

$$s_{t=a_t+g_t} \tag{3}$$

Where s_t is the overall target prediction, a_t is the AMR-target based score, and g_t is the chemical feature based equation.

The AMR-based target score is defined by the following (4).

$$a_t = \sum_{A \in L(t)} (2 * p_A + c_A^2) + \sum_{A \in M(t)} 0.2 + \frac{1}{2} \sum_{A \in N(t)} (1 - p_A) \tag{4}$$

Where t is the target, p_a is the precision for the AMR gene A, c_A is the relative confidence for that gene, $L(t)$ is the set of AMR genes in the cluster that are single target and predict the target, $M(t)$ is the set of multi-target AMR genes in the cluster that have been associated with that target, and $N(t)$ is the set of single-target AMR genes for which the target has been seen, but is not the main predicted target.
The chemical feature based equation is defined by the following equation (5).

$$g_t = \frac{1}{2} \sum_{s \in P(t)} (\min\{1, \max\{0, s\}\})^2 + \frac{1}{2} \sum_{s \in Q(t)} (\min\{1, \max\{0, s\}\})^2 \tag{5}$$

Where t is the target, $P(t)$ is the set of relative garlic scores that are annotated by that target among the top five annotated hits. The min and max functions are used so that the score is treated as one when it is above one and as zero when below zero.

Finally, the predicted target is the target that has the highest score for that cluster, and the confidence score is given as the difference between the highest and second highest scores defined by (6).

$$confidence = \hat{s} - \max\{s_t : s_t \in S \setminus \{\hat{s}\}\}, \tag{6}$$

Where S is the set of all target scores for that cluster and

$$\hat{s} = \max\{s_t : s_t \in S\} \tag{7}$$

2.5.8 Global Analysis of Predicted Siderophores and Other pNPs from ATP analysis

65,423 microbial genomes, and 339 in-house genomes were obtained from the NCBI Genome database (downloaded March 2016)[39], and processed through PRISM version 2.1.2.. pNPs with GARLIC scores ≥0.98, were grouped, and considered to be the same pNP based on structural similarity. Distances between siderophore pNPs, and other pNPs were generated via the Manhattan distance of biosynthetic features (all PRISM features excluding resistance genes). Of the pNPs revealed through the ATP pipeline, a pairwise similarity matrix was generated per the relative score identified by GARLIC. A 2D projection of these relationships was then generated using t-SNE as implemented within Scikit-learn using the Barnes-Hut approximation[63].

2.5.9 General Chemical Procedures

High-resolution MS spectra were collected on AB Sciex 5600+ TripleToF mass spectrometer (AB Sciex LLC, USA), equipped with an electrospray ionization source (ESI), coupled to a Shimadzu Nexera XR HPLC system using a Luna C18 column (50mm x 3.0mm, Phenomenex), running acetonitrile with 0.1% formic acid, and ddH_2O with 0.1% formic acid as the mobile phase for analytical separations.

2.5.10 Microbial Strains

A. citrulli DSM 17060, *V. paradoxus* DSM 30034, *N. potens* DSM 45234, *A. muelleri* DSM 19832, *Lysobacter gummosus* DSM 6980, and *Aquimarina* sp. DSM 19860 were obtained from the German Resource Centre for Biological Material (DSMZ). Flexibacter sp. ATCC 35208 was purchased from the American Type Culture Collection (ATCC). *S. candidus* NRRL 3147 was purchased from the Agricultural Research Service

Culture Collection (NRRL). *A. citrulli* and *V. paradoxus* were maintained on Acidovorax Complex Media[56]. *N. potens,* and *S. candidus* were maintained on Bennett's media. *A. muelleri* DSM 19832 and *Aquimarina* sp. DSM 19860 were maintained on Bacto Marine Agar. *L. gummosus* and *Flexibacter* sp. were maintained on Nutrient agar. All strains were grown at 30℃. The following strains were used in susceptibility assays. *P. aeruginosa* PAO1 was maintained on Tryptic Soy agar at 37°C. *S. aureus* Newmann, and *S. aureus* USA 300 were maintained on Tryptic Soy agar at 37°C.

2.5.11 Production of Natural Products

The following was performed to isolate the predicted siderophore natural products. Fresh colonies of DSM 17060 and DSM 30034 were inoculated 50 mL of acidovorax complete media and grown for 72 hr at 30°C. Single colonies of DSM 45234 were initially inoculated into 50 mL of KE media, followed by Bennett's media for 72 hr at 30°C. After fermentation, cultures were centrifuged at 7000 rpm, supernatant was extracted by Diaion HP-20 (2%) for 2hr. Methanol eluent of the HP-20 resin was prepared for LC/MS analysis. Cultures were spun down at 8000xg for 20 minutes at 4 °C. The supernatant was extracted with 2% absorbent HP-20 resin (Diaion). Following a 2 hr incubation, resin was eluted to with excess methanol, and evaporated to dryness. The dried fraction was re-suspended in water, and followed by a liquid-liquid partition with butanol. The organic fraction was kept, and evaporated to dryness.

A fresh colony of S. candidus NRRL 3147 was used to inoculate 50mL of KE media, and grown for 72 hr at 28°C at 200 RPM. A 1% inoculation was made into LL-A0341 seed media[48], and grown for 72 hr at 28°C and 200 RPM. Fresh colonies of *A.*

muelleri and *Aquimarina* sp. were used to inoculate 50mL cultures of Bacto Marine media, and grown for 48 hr at 28°C and 200 RPM. Following growth, a 1% inoculation was made into either 50mL or 1L of the same media, followed by growth at 28°C for 120 hr and 200 rpm. A fresh colony of *L. gummosus* was used to inoculate 50mL of Tryptic Soy broth and grown at 28°C for 48 hr and 200 rpm. Following growth, a 1% inoculation was made into 50mL of 1L of the same media, and grown for 120 hr at 28°C and 200 rpm. A fresh colony of *Flexibacter* sp. was used to inoculate 50mL of Nutrient broth and grown for 72 hr at 28°C and 200 rpm. A 2% inoculation was made into either 50mL of 1L of SJ media and grown for 72 hr at 28°C and 200 rpm. All above cultures were harvested by centrifugation for 20 minutes at 8000 rpm, and 4°C. The supernatant was then extracted with 2% absorbent HP-20 resin (Diaion). Following a 2 hr incubation, resins were eluted with excess methanol and evaporated to dryness. *A. muelleri, Aquimarina* sp. And *L. gummosus* were then subjected to liquid-liquid partition with butanol, and keeping the organic fraction, and evaporating to dryness. Production of possible pNPs were analyzed via activity profiling, and LC/MS analysis.

The collected extract from Flexibacter sp. was dissolved in methanol, and prepared for semi-preparative scale LC-MS. Fractions containing determined unique peaks, and activity were kept. Semi-preparative chromatography was performed using a Luna 5µm C_{18} column (Phenomenex, 250 x 10mm) with water (0.1% formic acid) and acetonitrile (0.1% formic acid) as the mobile phase, at a flow rate of 4mL/min. After 3 minutes, acetonitrile was increased in a linear manner (curve 5) from 5% to 30% at 7 minutes, then maintained until 10 minutes, then increased to 60% at 15 minutes, followed

by a wash off 100% acetonitrile. pNPs of interest eluted at 10-11 minutes. This was followed up with a second round of semi-preparative chromatography. After 3 minutes, acetonitrile was increased in a linear manner (curve 5) from 5% to 15% by 25 minutes followed by a wash with 100% acetonitrile. pNPs of interested eluted between 15.5-16.5 minutes.

2.5.12 Determination of Antibacterial Activity

Crude microbial activity of collected fractions from each strain in the ATP analysis were determined using bioactivity testing in cation-adjusted Mueller Hinton broth. *P. aeruginosa* PAO1 and *S. aureus* USA 300 were cultured at 37 °C overnight, followed by 1:50 dilution into fresh media and grown until an O.D. of 0.6 was reached. Once reached, strains were further diluted to 10^{-3} before being used in activity assays. *S. aureus* Newmann was treated in the same manner for their respective assays. Crude extracts collected were tested at a final concentration of 200µg/mL and 400µg/mL after incubation for 16 hr and 37°C. Absorbance readings to determine percent inhibition were measured at 600nm.

2.5.13 Genome Sequencing and Analysis

Gram positive strains within the in-house library of strains were extracted by inoculating a single colony into 50mL of appropriate media, and grown for generally 72 hr at 28°C. Incubation times may be altered depending on strain. 500-1000µL of culture was centrifuged at 12xg for 5 minutes and suspended in 500µL SET buffer (75 mM NaCl, 25 mM EDTA pH 8.0, 20 mM Tris HCl pH 7.5, 2 mg/mL lysozyme), and incubated for 2 hr at 37°C. Following incubation, a final concentration of 0.5mg/mL Proteinase K and

1% SDS were added. Mixture was incubated at 55°C for 2 hr, following adjustment of NaCl to 1.25M. This was followed by extracting twice with phenol-chloroform. Genomic DNA was precipitated by the addition of isopropanol, and washed twice with ethanol. DNA was suspended in sterile water before being sent for sequencing. Gram negative organisms were inoculated into 50mL of appropriate media, and generally incubated at 28°C, unless strain calls for otherwise. Genomic DNA was isolated per the protocol from GenElute Genomic DNA Extraction kit (Sigma Alrich). Library preparation and Illumina sequencing was performed at the Farncombe Metagenomics Facility at McMaster University, by an Illumina HiSeq DNA sequencer. Genomes are assembled by 3 assembly software systems (Velvet, IDBA, SPAdes), and then followed by SPPACE to scaffold assembled contigs[64-66]. The best assembly is selected using summary statistics reported by QUAST quality assessment reports[67].

2.6 Conclusion

In this work, we have created a methodology to engage the genomic capacity of microbes using a method that is data-driven and represents a departure in how microbes have historically been interrogated for antibiotic molecules. From identified BGCs, ATP can define and classify pNPs as probable antibacterials or other chemotypes such as siderophores. SIPE, represents a linear regression model for the classification of pNPs as siderophore chemotypes with a high degree of accuracy. Furthermore, ATP employs a custom optimization based approach developed from a compendium of AMR genes, and structural elements of previously characterized natural products to identify those pNPs which diverge in both respects. The accuracy of the developed method aims to ensure

laborious efforts involving genome mining are not centered upon predicted chemotypes associated with siderophores, but rather those with a high probability of producing antibacterial agents. The wealth of genomic information, and transition towards computational based approaches has significantly revived efforts in respect to defining new methods of natural product discovery that move away from the traditional discovery methods of the golden era. Despite efforts involving genome mining inferring pNPs, identifying those with a strong potential to be clinically relevant as antibacterial agents remained untouched prior to development of ATP. By providing a resource to classify possible bioactivity of pNPs based on similarities to other natural products, and resistance genes, allows focus and efforts to be directed towards isolation of pNPs with diverse bioactivities. It is hoped that the devised method will act as a resource to diminish the impact of the associated challenges of the traditional era of discovery, and result in positive contributions to meet the demand for new antibacterials to circumvent current antibiotic resistance mechanisms.

2.7 Supplementary Tables

Supplementary Table 2.1 Curated hidden Markov Models used by PRISM to detect antibiotic resistance genes within BGCs

Resistance Legend	Antibiotic	Gene	Gene Function	HMM L	Cut off
AMR 1	Bacitracin	bcrB,bcrA	ABC Transporters	bacitracin_ ABC_transp orters.hmm	150
AMR 2	Bacitracin	bacA, bcrC	Phosphatase	bacitracin-phosphatase s.hmm	230
AMR 3	Mersacidin	mrsG, mrsE	ABC Transporters	mersacidin-ABC_transp orter.hmm	310
AMR 4	nukacinISK-1	NukE, NukH	ABC Transporters	nukacinISK-1-ABC_Trans porter.hmm	360
AMR 5	Subtillin	SpaI	Immunity Protein	subtillin-immunity_p rotein	250
AMR 6	Subtillin	spaE, spaG	ABC transporters	SubtilinAB CTransporte rs.hmm	170
AMR 7	Fruilimycin	expA	ABC transporters	fruilimycin-ABC_transp orter.hmm	400
AMR 8	Cinnamycin	cinT, cinH	ABC Transporters	cinnamycin-ABC_transp orters.hmm	350
AMR 9	A500359	orf21	Phosphotran sferase	A500359_p hosphotrans ferase.hmm	300

AMR 10	Caprazamycin	cpz22	ABC Transporters	caprazamycin-ABC_transporter.hmm	720
AMR 11	Caprazamycin	cpz12, cpz27	Acyl Transferase	caprazamycin-acetyl_transferase.hmm	110
AMR 12	A54145	IptM, IptN	ABC Transporters	A54145_ABC_transporter.hmm	300
AMR 13	Daptomycin	DptM, DptN, DptP	ABC Transporters	daptomycin-ABC_transporter.hmm	270
AMR 14	Daptomycin	mprF	Target Modification	daptomycin-Phosphatidylglycerol_lysyltransferase.hmm	630
AMR 15	calcium-dependent antibiotic	hasP	Phosphotransferase	CDA_phosphotransferase.hmm	600
AMR 16	Cytolysin	cylI	Immunity Protein	cytolysin-immunityprotein.hmm	190
AMR 17	Epidermin	epiG, epiG	ABC Transporters	epidermin_ABC_transporters.hmm	255
AMR 18	Epidermin	epiH	Membrane Protein	epidermin_membrane_protein.hmm	300
AMR 19	Nisin	NisE, nisG	ABC Transporters	nisin-ABCtransporter.hmm	260

AMR 20	Nisin	nisI	Immunity Protein	Nisin Immunity Protein	390
AMR 21	Pep5	pepI	Immunity Protein	pep5-immunity_protein.hmm	145
AMR 22	Polymyxin	pmxD, pmxC	ABC Transporters	polymyxin-ABC_transporters.hmm	815
AMR 23	Beta Lactam	BCII	Class B metallo beta lactamase	Class_B_metallo_betalactamases.hmm	115
AMR 24	Cephamycin	pbp1, pbp1a	penicillin binding protein isoform	penicillian_binding_protein-isoform.hmm	410
AMR 25	Cephamycin	pbp1a	penicillin binding protein	cephamycin-PBP.hmm	770
AMR 26	Beta Lactam	Beta_Lactam1	Beta Lactamase II	outfile66.hmm	240
AMR 27	Beta Lactam	Beta_Lactam2	Beta Lactamase	outfile07.hmm	240
AMR 28	Beta Lactam	Beta_Lactam3	Beta lactamase II	outfile03.hmm	200
AMR 29	Beta Lactam	BJP	Beta lactamase subclass B3 metallo-beta lactamase	outfile67.hmm	179

AMR 30	Beta Lactam	BlaB	Beta Lactamase B	outfile68.hmm	231
AMR 31	Beta Lactam	blaI	Gene modulating beta lactam resistance, regulates BlaZ	outfile69.hmm	270
AMR 32	Beta Lactam	blaR1	Gene modulating beta lactam resistance, regulates BlaZ	outfile70.hmm	1300
AMR 33	Beta Lactam	CARB-PSE	Beta lactamase class A	outfile71.hmm	320
AMR 34	Cephalosporin	cblA	Beta Lactamase	outfile72.hmm	280
AMR 35	Cephalosporin	CepA	Beta Lactamase	outfile73.hmm	290
AMR 36	Cephalosporin	cfxA	Beta Lactamase	outfile75.hmm	216
AMR 37	Beta Lactam	ClassA	Class A beta lactamase	outfile79.hmm	415
AMR 38	Beta Lactam	ClassB	Class B beta lactamase	outfile80.hmm	275
AMR 39	Beta Lactam	AmpC	Class C Beta lactamase	outfile81.hmm	430
AMR 40	Beta Lactam	Class D	Class D beta lactamase	outfile82.hmm	220

AMR 41	Beta Lactam	CMY_LAT_MOX_ACT_MIR_FOX	Class C Beta lactamase	outfile83.hmm	535
AMR 42	Beta Lactam	ctxM	Class A Beta lactamase	outfile84.hmm	550
AMR 43	Beta Lactam	DHA	Class C Beta Lactamase	outfile85.hmm	485
AMR 44	Beta Lactam	DIM_GIM_SIM	B1 metallo-beta-lactamases	outfile86.hmm	300
AMR 45	Beta Lactam	Exo	Class A beta lactamase	outfile94.hmm	240
AMR 46	Beta Lactam	GES	Class A beta lactamase	outfile96.hmm	250
AMR 47	Beta Lactam	GOB	Subclass B3 metallo-beta lactamases	outfile97.hmm	288
AMR 48	Beta Lactam	IMP	Plasmid mediated IMP-type carbapenemases (subclass B1 (metallo-) beta-lactamase)	outfile98.hmm	343
AMR 49	Beta Lactam	IND	IND beta-lactamases (subclass B1 (metallo-) beta-lactamase)	outfile99.hmm	242

AMR 50	Beta Lactam	KHM	KHM beta-lactamases (subclass B1 (metallo-) beta-lactamase)	outfile100.hmm	340
AMR 51	Beta Lactam	KPC	Klebsiella pneumoniae carbapenem resistant (KPC) beta-lactamases (class a)	outfile101.hmm	371
AMR 52	Beta Lactam	L1	L1 beta-lactamase (subclass B3 (metallo-) beta-lactamase)	outfile102.hmm	585
AMR 53	Beta Lactam	Lactamase B	Beta-lactamase superfamily domain	outfile43.hmm	240
AMR 54	Beta Lactam	LRA	LRA beta-lactamase (subclass B3 (metallo-) beta-lactamase)	outfile103.hmm	270
AMR 55	Beta Lactam	mecR1	mecR1: gene modulating beta-lactam resistance	outfile108.hmm	600
AMR 56	Beta Lactam	moxA	MoxA beta-lactamase (class a)	outfile116.hmm	220

AMR 57	Beta Lactam	PC1	PC1: blaZ beta-lactamase (class a)	outfile121.hmm	250
AMR 58	Beta Lactam	NDM_ccrA	NDM-CcrA", "A grouping of related NDM and CcrA beta-lactamases	outfile119.hmm	256
AMR 59	Beta Lactam	sfh	sfh beta-lactamases (subclass B2 (metallo-) beta-lactamase)	outfile128.hmm	520
AMR 60	Beta Lactam	SHV_LEN	A grouping of the related SHV and LEN beta-lactamases (class a)	outfile129.hmm	435
AMR 61	Beta Lactam	SME	SME beta-lactamase (class a)	outfile130.hmm	490
AMR 62	Beta Lactam	SPM	Sao Paulo metallo-beta-lactamase (SPM-1) (subclass B1 (metallo-) beta-lactamase)	outfile131.hmm	120

AMR 63	Beta Lactam	SubclassB1	Subclass B1 (metallo-) beta-lactamase	outfile133.hmm	215
AMR 64	Beta Lactam	SubclassB2	Subclass B2 (metallo-) beta-lactamase	outfile134.hmm	440
AMR 65	Beta Lactam	Subclass B3	Subclass B3 (metallo-) beta-lactamase hydrolize penicillins	outfile135.hmm	315
AMR 66	Beta Lactam	TEM	TEM beta-lactamase (class a)	outfile136.hmm	437
AMR 67	Beta Lactam	Transpeptidase	Penicillin binding protein transpeptidase domain", "Target Redundancy /Overexpression	outfile05.hmm	390
AMR 68	Beta Lactam	VEB_PER	VEB and PER beta-lactamases (class a)	outfile161.hmm	290
AMR 69	Beta Lactam	VIM	Verone integron-encoded (VIM) metallo-beta-lactamase	outfile162.hmm	183

			(subclass B1 (metallo-) beta-lactamase)		
AMR 70	A40926, balhimycin	VanY, VanyB,	D-Ala-D-Ala Carboxypeptidases	carboxypeptidases.hmm	200
AMR 71	Vancomycin	VanE, VanG, VanSc	D-Ala-D-Serine ligase	vancomycin-d-ala-d-serine_ligases.hmm	405
AMR 72	Balhimycin, A47934, vancomycin	VanA, VanXST, VanX	D-Ala-D-Ala Dipeptidases	dipeptidases.hmm	200
AMR 73	Teicoplanin, Vancomycin	VanH	D-lactate dehydrogenase	d-lactate_dehydrogenases.hmm	540
AMR 74	Teicoplanin, Vancomycin	Van A	D-Ala-D-lactate Ligase	glycopeptides-D-ala-D-ala_Ligases.hmm	630
AMR 75	Vancomycin	Ligase_1	D-Ala-D-Ala-Ligase C	outfile22.hmm	100
AMR 76	Vancomycin	Ligase_2	D-Ala-D-ala-Ligase N	outfile23.hmm	100
AMR 77	Vancomycin	DalaDala	D-alanine--D-alanine ligase	outfile21.hmm	245.65
AMR 78	Vancomycin	VanA	VanA: D-Ala-D-Ala	outfile149.hmm	800

			ligase that can synthesize D-Ala-D-Lac		
AMR 79	Vancomycin	VanB	VanB: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Lac	outfile150.hmm	800
AMR 80	Vancomycin	VanC	anC: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Ser	outfile151.hmm	700
AMR 81	Vancomycin	VanD	VanD: D-Ala-D-Ala ligase that can synthesize D-Ala-D-Lac	outfile152.hmm	680
AMR 82	Vancomycin	VanH	VanH: D-specific alpha-ketoacid dehydrogenase that synthesizes D-lactate	outfile153.hmm	575
AMR 83	Vancomycin	vanR	VanR: transcriptional activator regulating VanA,	outfile154.hmm	271

			VanH and VanX		
AMR 84	Vancomycin	vanS	VanS: trasncriptional regulator of van glycopeptide resistance genes	outfile155.hmm	230
AMR 85	Vancomycin	vanT	VanT: membrane bound serine racemase, converting L-serine to D-serine	outfile156.hmm	650
AMR 86	Vancomycin	vanW	VanW: glycopeptide resistance gene	outfile157.hmm	460
AMR 87	Vancomycin	vanX	VanX: glycopeptide resistance gene	outfile158.hmm	270
AMR 88	Vancomycin	VanY,	VanY: glycopeptide resistance gene", "Gylcopeptide Resistance	outfile159.hmm	130
AMR 89	Vancomycin	vanZ	VanZ: glycopeptide resistance gene	outfile160.hmm	330

AMR 90	fosfomycin	FomA	Phosphotransferase	fosfomycin-fomA-Phosphotransferase.hmm	240
AMR 91	fosfomycin	FosB	Phosphotransferase	fosfomycin-fomB-phosphotransferase.hmm	450
AMR 92	Dapadiamides	DdaI	Transmembrane Pump	dapadiamides-transmembrane_protein.hmm	500
AMR 93	Fosmidomycin	fsr	MSF Transporter	Fosmidomycin-MFS.hmm	556
AMR 94	FR 90098	dxrB	DOXP Isoform	FR90098-DOXP_reductoisomerases.hmm	750
AMR 95	Platencin, plantensimycin	PtmP3, PtnP3, FabF	beta-ketoacyl-acyl-carrier-protein synthase II Isoform	beta-ketoacyl-acyl-carrier-protein_synthase_II_Isoforms.hmm	800
AMR 96	Andrimid	admT	Acyl CoA Carboxylase Isoform	AdmT-acyl CoA carboxylase isoform.hmm	700
AMR 97	Pantocin A	paaC	Transmembrane Transporter	pantocinA-Transmembrane_transporter.hmm	335

AMR 98	Mupirocin	mupA	isoleucyl-tRNA synthetase isoform	mupirocin-isoleucyl-tRNA_synthetase_isoform.hmm	625
AMR 99	Chuangxmycin, Indolmycin	TrpRS1, sgr3809	Tryptophan-tRNA synthetase isoform	tryptophanyl_tRNA_synthtase_isoforms.hmm	720
AMR 100	Borrelidin	THR	threonyl-tRNA synthetase	borreldin-threonyl tRNA synthetase	1415
AMR 101	Albomycin	ambK	seryl-tRNA synthetase	albomycin_seryltRNAsynthetase.hmm	800
AMR 102	Phosphothricin	rimL	phosphothricin-N-acetyltransferase	phosphothricin-acetyltransferase.hmm	310
AMR 103	Chlorohthricin	chlG	MFS Transporter	chlorothricin-MFS.hmm	500
AMR 106	Microcin C7	mccE	Immunity Protein	microcin_C7-Acyl_transferase.hmm	895
AMR 109	Factumycin	FacT	MFS Transporter	factumycin-ABC_transporter.hmm	900
AMR 110	GE2270A, kirromycin	tuf, tufB1	Ef-Tu Isoform	kirromycin-Ef_Tu_isoforms.hmm	840

AMR 111	Pikromycin, Thiostrepton, Clindamycin, erythryomycin , Gentamicin, hygromycin, kanamycin	PikR1, PikR2, TsnR, LmrB, erm, ermE, gtmJ, hyg6, kmr	rRNA methyltransferase	rRNA_methyltransferasesI20150513-211754-0720-8042522-pg.hmm	210
AMR 112	Clindamycin	LmrA, LmrC	ABC Transporters	clindamycin-ABC_transporters.hmm	810
AMR 113	Althiomycin	almE	MSF Transporter	althiomycin-MFS_transporter.hmm	300
AMR 114	fortimycin	ForP	Phosphotransferase	fortimycin-phosphotransferase.hmm	400
AMR 115	puromycin	pac	Acetyl Transferase	puromycin-acetyltransferase.hmm	220
AMR 116	puromycin	pur8	MFS Transporter	39, 40	196
AMR 117	fusidic acid	fusB	Detoxification protein	fusB_fusardic_acid_resistance.hmm	400
AMR 118	Chloramphenicol	mdtL, cml-e	MSF Transporter	chloramphenicol_MFS_transporters.hmm	440
AMR 119	Chloramphenicol	Cata10, cata12, cata2	Acetyl Transferase	chloramphenicol_acetyltransferases.hmm	180

AMR 120	Chloramphenicol	Cat	Acetyl Transferase	outfile16.hmm	80
AMR 121	Chloramphenicol	CAT	Acetyl Transferase	outfile75.hmm	216
AMR 122	Chloramphenicol	CAT	Acetyl Transferase	outfile76.hmm	75
AMR 123	Chloramphenicol	C_MSF	Efflux Pump	outfile77.hmm	250
AMR 124	Chloramphenicol	CPT	Phosphotransferase	outfile78.hmm	250
AMR 125	Chloramphenicol	CPT	Phosphotransferase	outfile14.hmm	80
AMR 126	Gentamicin	gtmJ	Phosphotransferase	gentamicin-phosphotransferase.hmm	275
AMR 127	Hygromycin A	hyg28	ABC Transporters	hygromycin-ABC_transporter.hmm	1000
AMR 128	Hygromycin A	hyg19	MFS Transporter	hygromycin-MFS.hmm	300
AMR 129	Hygromycin B	Hyg, hyg21	Phopshotransferase	hygromycin-phosphotransferase.hmm	700
AMR 130	Florfenicol	florR	MSF Transporter	florfenicol-MFS.hmm	775
AMR 131	Istamycin	istP	Phosphotransferase	istamycin-phosphotransferase.hmm	200
AMR 132	Kanamycin	KanM	Acetyl Transferase	kanamycin-acetyltransferase.hmm	225

AMR 133	Neomycin	AAC8	Acetyl Transferase	neomycin-acetyltransferase.hmm	480
AMR 134	Neomycin	Neo	Phosphotransferase	neomycin-phosphotransferase.hmm	230
AMR 135	Paromycin	PPH	Phosphotransferase	paromycin-phosphotransferase.hmm	220
AMR 136	Paromycin	AAC7	Acetyl Transferase	paromycin-acetyltrasnferase.hmm	400
AMR 137	Erythromycin	ereA, ereB, depI	Esterase	esterases.hmm	113
AMR 138	Avilamycin	aviABC1, AviBC2	ABC Transporters	availamycin-ABCtransporter.hmm	500
AMR 139	Tylosin	tlrC	ABC Transporters	TylosinABC Transporters.hmm	470
AMR 140	Streptomycin	str	Acyl Transferase	streptomycin-acetyltransferase.hmm	400
AMR 141	streptomycn	strA	Phosphotransferase	StreptomycinPhosphotransferases.hmm	300
AMR 142	Streptogramin	vgaA	ABC Transporters	streptogramin0ABC_transporter.hmm	760

AMR 143	Spiramycin	smrB	ABC Transporters	spiramycin-ABC_transporter.hmm	650
AMR 144	Spectinomycin	spcN	Phosphotransferase	spectinomycin-phosphotransferase.hmm	400
AMR 145	Lincomycin	lnuA, lnuB	Nucleotidyltransferase	lincoasmide-nucelotidyltransferases.hmm	100
AMR 146	Macrolides	mphA	Phosphotransferase	macrolide-phosphotransferases.hmm	285
AMR 147	tetracycline	Tet37	resistance protein	tetracycline-resistance_protein.hmm	700
AMR 148	Tetracyclines	TetO, TetW, otrA	Ribosomal Protection protein	tetracycline-ribosomal_protection_proteins.hmm	790
AMR 149	Tetracycline	TetX	Oxidoreductase	tetracycline-oxidoreductase.hmm	400
AMR 150	Tetracycline	TetH, tcr3, TetA, otrB	MSF	tetracycline-MFS.hmm	429
AMR 151	Tetracycline	TetA	tetA: tetracycline resistance MFS efflux pump	outfile139.hmm	447
AMR 152	Tetracycline	TetA_B	tetA(B): tetracycline	outfile137.hmm	520

			resistance MFS efflux pump		
AMR 153	Tetracycline	TetA_G	TetA-G", "tetA(G): tetracycline resistance MFS efflux pump	outfile138.hmm	550
AMR 154	Tetracycline	TetD	tetD: tetracycline resistance MFS efflux pump	outfile140.hmm	550
AMR 155	Tetracycline	TetH_TetJ	tetH and TetJ: tetracycline resistance MFS efflux pumps	outfile142.hmm	490
AMR 156	Tetracycline	TetM_TetW_TetO_TetS	TetM-TetW-TetO-TetS", "tetM, tetW, tetO, and tetS: tetracycline resistance ribosomal protection protein	outfile143.hmm	700
AMR 157	Tetracycline	MFS_Tet	Tetracycline_Resistance_MFS_Efflux_Pump", "tetracycline resistance	outfile144.hmm	186

			MFS efflux pump		
AMR 158	Tetracycline	Tet_Resistance1	tetracycline resistance ribosomal protection protein	outfile145.hmm	940
AMR 159	Tetracycline	TetX	tetX: tetracycline inactivation enzyme	outfile146.hmm	750
AMR 160	Tetracycline	TetY	tetY: tetracycline resistance MFS efflux pump	outfile147.hmm	542
AMR 161	Tetracycline	Tex_N	Tex-like protein N-terminal domain	outfile32.hmm	157
AMR 162	Tetracycline	TetE	TetE", "tetE: tetracycline resistance MFS efflux pump	outfile141.hmm	480
AMR 163	Macrolides	ermD, ermE, ermF	rRNA adenine N-6methyltransferase	macrolides-rRNA_adenine_dimethylases	275
AMR 164	Pristamycin	ptr	Membrane Protein	pristamycin-MFS.hmm	740
AMR 165	Macrolides	Macro_glycosyl	macrolide glycosyltransferase: macrolide	outfile106.hmm	685

			inactivation enzyme		
AMR 166	Aminoglycosides	ACC3	Aminoglycoside Acetyltransferase (AAC3)	outfile46.hmm	300
AMR 167	Aminoglycosides	ACC3-I	Aminoglycoside Acetyltransferase (AAC3-I)	outfile47.hmm	150
AMR 168	Aminoglycosides	ACC6-I	Aminoglycoside Acetyltransferase (AAC6-I)	outfile49.hmm	300
AMR 169	Aminoglycosides	ACC6-Ib	Aminoglycoside Acetyltransferase (AAC6-Ib)	outfile48.hmm	400
AMR 170	Aminoglycosides	ACC6-II	Aminoglycoside Acetyltransferase (AAC6-II)	outfile50.hmm	350
AMR 171	Aminoglycosides	ANT2	Aminoglycoside nucleotidyltransferase 2	outfile57.hmm	480
AMR 172	Aminoglycosides	ANT3	Aminoglycoside nucleotidyltransferase 3	outfile58.hmm	450

AMR 173	Aminoglycosides	ANT4	Aminoglycoside nucleotidyltransferase 4	outfile59.hmm	490
AMR 174	Aminoglycosides	ANT6	Aminoglycoside nucleotidyltransferase 6	outfile60.hmm	530
AMR 175	Aminoglycosides	ANT9	Aminoglycoside nucleotidyltransferase 9	outfile61.hmm	202
AMR 176	Aminoglycosides	Antibiotic_NAT	Aminoglycoside 3-N-Acetyl Transferase	outfile15.hmm	100
AMR 177	kanamycin	APH3	Aminoglycoside phosphotransferase 3	outfile62.hmm	70
AMR 178	Aminoglycosides	APH6	Aminoglycoside phosphotransferase 6	outfile63.hmm	370
AMR 179	Aminoglycosides	ANT	Aminoglycoside Nucleotidyltransferase	outfile165.hmm	200
AMR 180	Sorangicin	sorF	Glycosyl Transferase	sorangicin-glycosyltransferase.hmm	640
AMR 181	Rifamycin	rifP	MFS Transporter	rifamycin-MFS.hmm	830

AMR 182	Microcin J	mcjD	ABC Transporters	microcinJ-ABC_transporter.hmm	1025
AMR 183	Microcin B17	mcbE, mcbF	ABC Transporters	microcinb17-ABC_transporter.hmm	115
AMR 184	Novobiocidin	GyrB	DNA gyrase B Isoform	novobiocidin-DNA_gyrase.hmm	1460
AMR 185	Albicidin	albG	Pentapeptide Repeats	albicidin_pentapeptide_repeats.hmm	350
AMR 186	Albicidin	albF	ABC Transporters	albicidin_ABC_transporters.hmm	600
AMR 187	Quinolones	Qnr	Pentapeptide Repeats	pentapeptide repeats.hmm	223
AMR 188	Fluorouinolones	FRT	Fluoroquinolone Resistant DNA Topoisomerase	outfile95.hmm	1370
AMR 189	Quinolones	Quin	quninolone resistance protein (Qnr): antibiotic target protection protein	outfile123.hmm	330
AMR 190	Tunicamycin	tmrB	resistance protein	tunicamycin-	150

				resistance_protein.hmm	
AMR 194	MultiDrug	mdtE, acrB, amrA, amrB, aprA	RND Transporters	AllRND_transporters.hmm	320
AMR 195	Multi Drug	ykkD, ykkC	small multidrug resistant	small_multidrug_resistant_protein.hmm	115
AMR 197	Trifolitoxin	tfxE	Trifolitoxin Operon Protein	TrifolitoxinResistance.hmm	550
AMR 198	Zwittermicin	ZmaR	Acetyl Transferase	ZwittermicinAcetyltransferase.hmm	600
AMR 199	MultiDrug	QacF, emrE	Cationic multidrug transporters	cationic_multidrug_transporters.hmm	135
AMR 200	MultiDrug	16s rRNA1	16S ribosomal RNA methyltransferase	outfile45.hmm	370
AMR 201	MultiDrug	ABC Trans1	ABCAntibioticEffluxPump	outfile51.hmm	245
AMR 202	MultiDrug	ABC Trans2	ABC Transporters	outfile06.hmm	285
AMR 203	MultiDrug	ABC Trans3	Abc Transporters	outfile17.hmm	240
AMR 204	MultiDrug	ABC Trans4	ABC 2 type transporter	outfile08.hmm	650

AMR 205	MultiDrug	Acetyl_Trans1	Acetyl Transferase 1	outfile09.hmm	130
AMR 206	MultiDrug	Acetyl_Trans2	Acetyl Transferase 3	outfile41.hmm	200
AMR 207	MultiDrug	Acetyl_Trans3	Acetyl Transferase 4	outfile00.hmm	150
AMR 208	MultiDrug	Acetyl_Trans4	Acetyl Transferase 7	outfile39.hmm	200
AMR 209	MultiDrug	Acetyl_Trans5	Acetyl Transferase 8	outfile40.hmm	190
AMR 210	MultiDrug	Acetyl_Trans6	Acetyl Transferase 9	outfile42.hmm	130
AMR 211	MultiDrug	Acr	AcrB/AcrD/ AcrF Family	outfile34.hmm	200
AMR 212	MultiDrug	Acetyl_Trans7	Acetyl Transferase	outfile33.hmm	125
AMR 213	Multidrug	adeI	Membrane fusion protein of multi drug efflux complex	outfile52.hmm	650
AMR 214	Multidrug	adeB	Membrane fusion protein of multi drug efflux complex	outfile53.hmm	1900

AMR 215	Multidrug	adeC-adeK-oprM	Outer membrane factor the multidrug efflux complex	outfile54.hmm	580
AMR 216	Multidrug	adeR	Regulator of AdeABC efflux system	outfile55.hmm	500
AMR 217	Multidrug	adeS	Regulator of AdeABC efflux system	outfile56.hmm	600
AMR 218	Multidrug	AminotransI_II	Aminotransferase class I and II	outfile30.hmm	372
AMR 219	Multidrug	AminotransIV	Aminotransferase class IV	outfile35.hmm	100
AMR 220	Multidrug	APH	Phosphotransferase enzyme family	outfile25.hmm	172
AMR 221	Multidrug	baeR	Subunit of gene modulating antibiotic efflux	outfile64.hmm	420
AMR 222	Multidrug	baeS	Subunit of gene modulating antibiotic efflux	outfile65.hmm	1200

AMR 223	Multidrug	cfr	23s rRNA methyltransferase	outfile74.hmm	490
AMR 224	Multidrug	emrB	subunit of efflux pump conferring antibiotic resistance	outfile87.hmm	205
AMR 225	Multidrug	emrA	smalll multirug resistance antibiotic efflux pump	outfile88.hmm	150
AMR 226	Multidrug	Erm	23s rRNA methyltransferase	outfile89.hmm	250
AMR 227	Multidrug	Erm38	23s rRNA methyltransferase	outfile90.hmm	320
AMR 228	Multidrug	ErmA	23s rRNA methyltransferase	outfile91.hmm	350
AMR 229	Multidrug	ErmB	23s rRNA methyltransferase	outfile92.hmm	370
AMR 230	Multidrug	ErmC	23s rRNA methyltransferase	outfile93.hmm	450
AMR 231	Multidrug	Fad_Binding_2	FAD Binding Domain	outfile38.hmm	500
AMR 232	Multidrug	FmrO	rRNA methyltransferase	outfile11.hmm	215

AMR 233	Multidrug	HTH_AraC	Bacterial regulatory helix-turn-helix proteins	outfile31.hmm	100
AMR 234	Multidrug	macA	macA: subunit of efflux pump conferring antibiotic resistance	outfile104.hmm	550
AMR 235	Multidrug	macB	macB: subunit of efflux pump conferring antibiotic resistance	outfile105.hmm	450
AMR 236	Multidrug	marA	marA: transcription factor induces MDR efflux pump AcrAB	outfile107.hmm	300
AMR 237	Multidrug	marR	Gene Modulating Resistance	outfile26.hmm	100
AMR 238	Multidrug	marR_2	Gene Modulating Resistance	outfile27.hmm	100
AMR 239	Multidrug	methyltrans 18	Methyltrans ferase	outfile37.hmm	120
AMR 240	Multidrug	mexA	mexA: membrane fusion protein of the MexAB-	outfile109.hmm	580

			OprM multidrug efflux complex		
AMR 241	Multidrug	mexC	mexC: membrane fusion protein of the MexCD-OprJ multidrug efflux complex	outfile110.hmm	350
AMR 242	Multidrug	mexE	mexE: membrane fusion protein of the MexEF-OprN multidrug efflux complex	outfile111.hmm	170
AMR 243	Multidrug	mexH	mexH: membrane fusion protein of the efflux complex MexGHI-OpmD", "RND Antibiotic Efflux	outfile112.hmm	100
AMR 244	Multidrug	mexW-mexI	A grouping of related mexW and mexI subunits of	outfile113.hmm	700

			efflux pumps conferring antibiotic resistance		
AMR 245	Multidrug	mexX	mexX: subunit of efflux pump conferring antibiotic resistance, RND transporter	outfile114.hmm	271
AMR 246	Multidrug	MFS1	MFS Transporter	outfile01.hmm	175
AMR 247	Multidrug	MFS3	MFS Transporter	outfle24.hmm	700
AMR 248	Multidrug	MFS	MFSAntibioticEffluxPump	outfile115.hmm	245
AMR 249	Multidrug	mprF	mprF: peptide antibiotic resistance gene	outfile117.hmm	1400
AMR 250	Multidrug	msbA	msbA: ATP-binding cassette (ABC) antibiotic efflux pump	outfile118.hmm	500
AMR 251	Multidrug	norA	norA: major facilitator superfamily (MFS)	outfile120.hmm	370

			antibiotic efflux pump		
AMR 252	Multidrug	Nuc_H_Symport	Nuc_H_symport", "PF03825.11 Nucleoside H+ symporter	outfile44.hmm	250
AMR 253	Multidrug	phoQ	phoQ: subunit of gene modulating antibiotic efflux	outfile122.hmm	750
AMR 254	Multidrug	ramA	ramA: gene modulating antibiotic efflux	outfile124.hmm	260
AMR 255	Multidrug	RND1	resistance-nodulation-cell division (RND) antibiotic efflux pump	outfile125.hmm	940
AMR 256	Multidrug	robA	obA: transcriptional activator of AcrAB antibiotic efflux pump	outfile126.hmm	620
AMR 257	Multidrug	romA	romA: trasncription factor mediating antibiotic resistance	outfle127.hmm	850

AMR 258	Multidrug	soxR	soxR: mutant efflux regulatory protein conferring antibiotic resistance	outfile131.hmm	190
AMR 260	Multidrug	Bcr_cfla	efflux Bcr CflA: drug resistance transporter	outfile04.hmm	200.9
AMR 261	Multidrug	ermB	EmrB: drug resistance MFS transporter, drug:H+ antiporter-2	outfile10.hmm	230.2
AMR 262	Multidrug	MATE_efflux	matE: MATE efflux family protein	outfile13.hmm	147.4
AMR 264	Multidrug	soxR	SoxR: redox-sensitive transcriptional activator SoxR	outfile29.hmm	116.9
AMR 266	Multidrug	tolC	tolC: subunit of efflux pump conferring antibiotic resistance	outfile148.hmm	570

AMR 267	Multidrug	Whib	Transcription factor WhiB	outfile28.hmm	100
AMR 268	Multidrug	RND_MFP	efflux transporter, RND family, MFP subunit	outfile163.hmm	240
AMR 269	Albicidin	albD	Esterase	muscle-albD-Detoxification Enzyme.clw	130
AMR 270	Enterocin A	IciA	Immunity Protein	muscle-IciA-Enterocin A Immunity Protein.clw	110
AMR 271	Multidrug	QepA	MFS	muscle-qepA-MFS transporter.clw	270
AMR 272	Tabtoxine	TbiF	Ligase	muscle-TbIF-Ligase.clw	360
AMR 273	Tabtoxine	ttr	Acetyltransferase	muscle-ttr-acetyl transferase.clw	178
AMR 275	Fosfomycin	FosC	Phosphotransferase	muscle-fosC-fosfomycin-phosphotransferase.clw	77
AMR 276	Fosfomycin	FosX	Epoxide Hydrolase	muscle-fosX-	100

				fosfomycin-epoxide hydrolase.clw	
AMR 277	Fosfomycin	FosA	Glutathione Transferase	muscle-fosA-fosfomycin-glutathione transferase.clw	120
AMR 278	Fosfomycin	FosB	Metallothiol Transferase	muscle-fosA-fosfomycin-glutathione transferase.clw	125
AMR 279	Beta Lactams	Fec1	Beta lactamase	Fec1-BetaLactamase.hmm	540
AMR 280	Nocardicin	NocD	Acetyl Transferase	nocD-acetyltransferase.hmm	316
AMR 281	Monobactam	Oxy2	Betalactamase	Oxy2_betalactamase.hmm	515
AMR 282	Monobactam	per1	beta lactamase	per2_betalactamase.hmm	400
AMR 283	Betalactams	shv2	betalactamase	shv2_Betalactamase.hmm	660
AMR 284	Polymyxin	ArnA/Pmr1	Formyl Transferase		
AMR 285	Kasuguamycin	KasKLM	ABC Transporters	kasugamycin-	160

				ABC_transporter.hmm	
AMR 286	oxazolidinones and phenicols	optrA	ABC Transporters	optrA-oxanoloid-ABCTransporter.hmm	260
AMR 287	copper	czcA	Copper resistance Protein	czcA-copper resistance protein.hmm	709
AMR 288	copper	copA	Copper Resistance Protein	copA-coppper resistance protein.hmm	300
AMR 289	Collisitin	mcr-1	phosphatidylethanolamine transferase-antibiotic modification	mcr1-collistin-phosphatidylethanolaminetransferase.hmm	1000
AMR 290	Nisin	nisG	ABC Transporter	nisin-nisG-ABCTransporter2.hmm	260
AMR 291	Chalcomycin	chrB	rRNA methyl transferase	chalcomycin-rRNAmethyltransferase.hmm	490
AMR 292	Rifampin	rph	phosphotransferase	Rifampin-Rifamyicn Phosphotransferase.hmm	1790
AMR 293	Askumycin	asuM1	MFS	AsukamycinMFSasuM1.hmm	900

AMR 294	Carnocyclin	ccII	Immunity Protein	carnocyclin-immunityprotein.hmm	500
AMR 295	Cereulide	CesC	ABC Transporters	cereulide-ABCTransporter	400
AMR 296	GE 81112	getB	ABC Transporters	Ge81112-ABC Transporter	800
AMR 297	lukacidin	lkcJ	ABC Transporters	lukacidin-ABCTransporter	975
AMR 298	Rifampin	arr	ADP Risobosyml Transferase	Rifampin-Rifamycin ADP ribosyl transferas.hmm	199.7
AMR 299	Rifampin	yjiC	Glycosyl Transferase	Rifampin-Rifamycin-GlycosylTransferase	775
AMR 300	rubradirin	rubT1	ABC Transporters	rubradirin-ABC Transporter	500
AMR 302	Terreic Acid	orf	ABC Transporters	Terreic Acid- ABC Transporter	3350
AMR 303	Tetarimycin	TamA	ABC Transporters	Tetarimycin - ABCTransporter	800
AMR 304	Tiacumucin	Tia3	ABC Transporter	Tiacumucin-ABC Transporter	780

AMR 305	Virginamycin	VarS	emrB Transporter	virginamycin-MFStransporter	570
AMR 306	Zeamine	zmn20	ABC Transporter	Zeamine-ABCTransporter	600
AMR 307	Tylosin	tlrC	rRNA methyl Transferase	Tylosin-rRNAmethyltransferase	400
AMR 308	Capreomycin	cmnU	rRNA Methyltransferase	capreomycin-cmnU-rRNAmethyltransferase.hmm	150
AMR 309	tetronasin	tnrB	ABC Transporter	tetronasin-tnrB-ABCTransporter	430
AMR 310	viomycin	vph	Phosphotransferase	Viomycin-vph-phosphotransferase	270
AMR 311	Calcimycin	calT	ABC Transporter	Calcimycin-calT-ABCTransporter	1210
AMR 312	Mycinamycin	myrA	rRNA Metyhl Transferase	Mycinamycin-myrA-rRNAmethyltransferase.hmm	350
AMR 313	Mycinamycin	myrB	rRNA Methyl transferase	Mycinamycin-myrB-rRNAmethy	335

				ltransferase.hmm	
AMR 314	Griselimycin	griR	DNA polymerase III Beta Subunit	Griselimycin-DnaNBetaSubunit.hmm	400

Supplementary Table 2.2 Extended bacterial target legend for Fig. 2.2

Target Number	**Bacterial Target**
1	Lipid II Binders
2	Bactoprenol Phosphate Binder
3	Phosphatidylethanolamine Binder
4	Translocase I
5	Membrane Destabilizers
6	LPS Binder
7	penicillin binding proteins
8	D-Ala-D-Ala Chelators
9	UDP-N-Acetylglucosamine-3-enolpryuvyl transferase
10	Glucosamine phosphate synthetase
11	Deoxyxylulose phosphate Reductoisomerase
12	FabB/F
13	Acyl CoA Carboxylase
14	l-histidinol phosphate aminotransferase
15	Isoleucine-tRNA synthetase
16	Tryptophan-tRNA synthetase
17	Threonine-tRNA synthetase

18	Serine-tRNA synthetase
19	Aspartyl tRNA synthetase
20	Glutamine synthase
21	Pyruvate carboxylase
22	Elongation Factor Tu
23	Ribosome inhibitors
24	RNAP
25	DNA gyrase

Supplementary Table 2.3 Number of resistance genes collected organized by bacterial target

Molecular Target	**Resistance Gene Instances**	**Percentage**
Acetyl-CoA carboxylase	1	0.427
Aspartate-tRNA synthetase	1	0.427
D-Ala-D-Ala chelator	20	8.547
Deoxyxylulose phosphate reductoisomerase	2	0.854
DNA gyrase	8	3.419
Elongation factor Tu	2	0.854
Fab B/F	1	0.427
Glucosamine-6-phosphate synthetase	4	1.709
Glutamine synthase	3	1.282
Isoleucine-tRNA synthetase	1	0.427
L-histidinol phosphate aminotransferase	1	0.427
Lipid II binder	9	3.846
LPS binder	3	1.282
Membrane destabilizer	9	3.846
Penicillin-binding protein	52	22.222
Phosphatidylethanolamine binder	1	0.427
Pyruvate carboxylase	1	0.427

RNA polymerase	7	2.991
Serine-tRNA synthetase	1	0.427
Threonine-tRNA synthetase	1	0.427
Translocase I	4	1.709
Tryptophan-tRNA synthetase	1	0.427
UDP-N-acetylglucosamine-3-enolpryuvyltransferase	7	2.991
Undecaprenyl phosphate binder	4	1.282
DnaN binding clamp	1	0.427

Supplementary Table 2.4 List of the devised BGC set used in generation of SIPE and ATP

Antibacterial	**Antifungal**	**Siderophore**
A102395	acetylaranotin	acidobactin
A40926	acetylaszonalenin	acinetobactin
A47934	aculeximycin	amychelin
A500359	AFtoxin	bacillibactin
A54145	ambruticin	coelibactin
abyssomicin	amphotericin	coelichelin
actagardine	apoptolidin	cupriachelin
actinorhodin	arthrofactin	delftibactin
albicidin	asperfuranone	enterobactin
albomycin	aspyridone	erythrochelin
alnumycin	aureobasidin	exochelin
Alphalipomycin	azaphilone	ferrichrome
althiomycin	bacillomycin	fimsbactin
andrimid	basiliskamide	fuscachelin
anglomycin	beauvericin	gobichelin
apramycin	bongkrekicacid	griseobactin
aranciamycin	candicidin	heterobactin

arylomycins	chivosazol	isoflavipucine
asukamycin	compactin	malleilactone
aureothin	crocacin	mirubactin
avilamycin	cycloheximide	mycobactin
azicemicin	cystothiazoleA	paenibactin
bacillaene	desmethylbassianin	pyochelin
bacitracin	echinocandin	pyoverdin
bactobolin	enniatin	rhodochelin
balhimycin	faerifungin	scabichelin
BE14106	fengycin	serobactin
bicornutin	filipin	tenellin
bogorol	FR008	vanchrobactin
borrelidin	frontalamide	vibriobactin
bottromycin	fumiquinazoline	vulnibactin
Butirosin	fumitremorgin	yersiniabactin
caerulomycinA	fumonisin	
calcimycin	fusaricidin	
calciumdependentantibiotic	fusaridioneA	
capreomycin	galbonolide	
capuramycin	gephyronicacid	
carnocyclinA	glidobactinA	
cephalosporin	griseofulvin	
cephamycinC	hassallidin	
cereulide	herbimycin	
chalcomycin	hypothemycin	

chejuenolide	jagaricin	
chelocardin	jamaicamide	
chloroeremomycin	jawsamycin	
chloramphenicol	JBIR06	
chlorothricin	JBIR34	
chlortetracycline	jerangolid	
chondrochlorens	kutznerides	
cinnamycin	lactimidomycin	
coelimycin	luminmycin	
colistin	melithiazol	
corallopyroninA	methylsalicylicacid	
cuevaene	micacocidin	
cyclomarin	microcystin	
cypemycin	microsclerodermin	
Cytolysin	ML449	
dactylocycline	monacolin	
Dapdiamides	monodictyphenone	
daptomycin	mulundocandin	
depsidomycin	mycolactone	
difficidin	mycosubtilin	
elansolid	mycotrienin	
enacyloxin	myxalamid	

enduracidin	myxothiazol	
enterocin	natamycin	
epidermin	nigericin	
erdacin	notoamide	
erythromycin	nystatin	
etnangien	octacosamicin	
Factumycin	oligomycin	
FD594	paenilarvin	
fortimicin	phoslactomycin	
Fosfomycin	pimaricin	
friulimicin	plipastatin	
fungisporin	pneumocandin	
gallidermin	pradimicin	
GE81112	puwainaphycin	
gentamicin	pyoluteorin	
goadsporin	pyrrolocin	
gramicidin	R1128	
granaticin	respirantin	
griselimycin	rhizopodin	
griseoviridin	rhizoxin	
guadinomine	rimocidin	
gulmirecin	Sch47554	

halstoctacosanolide	solanapyrone	
himastatin	soraphen	
hitachimycin	sterigmatocystin	
hormaomycin	stigmatellin	
hygrocin	tautomycin	
hygromycinA	taxillaid	
hygromycinB	terrequinone	
indanomycin	tetrahydroxynaphthalene	
Indolmycin	thanamycin	
istamycin	verlamelin	
iturin	viridicatumtoxin	
jadomycin	xenocoumacin	
kalimantacin	yanuthoneD	
kanamycin		
kasugamycin		
kijanimicin		
kirromycin		
lacticin		
lactonamycin		
laidlomycin		
lankacidin		
lankamycin		

lasalocid		
laspartomycin		
lichenysin		
Lincomycin		
lividomycin		
lobophorin		
locillomycin		
lysobactin		
lysolipin		
macrolactin		
mannopeptimycin		
marinopyrrole		
massetolide		
mersacidin		
methymycin		
Microcin C7		
microcin		
MicrocinB17		
MicrocinJ25		
midecamycin		
monensin		
mupirocin		

muraymycin		
myxopyronin		
myxovirescin		
naphthocyclinone		
naphthyridinomycin		
napsamycin		
neomycin		
niddamycin		
nisin		
nocardicinA		
nocathiacin		
nosiheptide		
Novobiocin		
Nukacin ISK1		
orfamide		
oxytetracycline		
pacidamycin		
paenibacterin		
paenilamicin		
Pantocin A		
paromomycin		
paulomycin		

pelgipeptin		
penicillin		
Pep5		
Phosphothricin		
piericidinA1		
pikromycin		
Platensimycin		
polymyxin		
potensimicin		
pristinamycin		
promysalin		
Puromycin		
putisolvin		
pyralomicin		
pyridomycin		
ramoplanin		
resistomycin		
reutericyclin		
ribostamycin		
rifamycin		
ristocetin		
rosamicin		

rubradirin		
salinomycin		
sansanmycin		
saquayamycin		
sevadicin		
simocyclinone		
sisomicin		
sorangicin		
spectinomycin		
sphaerimicin		
steffimycin		
stenothricin		
streptolydigin		
streptomycin		
streptothricin		
subtilin		
subtilosin		
surfactin		
syringafactin		
syringomycin		
tabtoxin		
taromycin		

tauramamide		
teicoplanin		
teixobactin		
telomycin		
terreicacid		
tetarimycin		
tetracycline		
tetronasin		
tetronomycin		
thienamycin		
thiomarinol		
thiomuracin		
thiostrepton		
thuggacin		
tiacumicin		
tirandamycin		
TLN05220		
tobramycin		
tolaasin		
tridecaptin		
Trifolitoxin		
Tunicamycin		

tylactone		
Tylosin		
tyrocydine		
UK68597		
valinomycin		
vancomycin		
viomycin		
virginiamycin		
zeamine		
zwittermicin		

Supplementary Table 2.5 Detected AMR HMMs by PRISM from known antibacterial BGCs, and annotations of known molecular target.

Biosynthetic Gene Cluster	**Resistance Genes**	**Target**
A102395	AMR_9	Translocase 1 inhibitor
A40926	AMR_70	DAlaDAla chelator
A47934	AMR_84, AMR_72, AMR_73, AMR_74, AMR_83	DAlaDAla chelator
A500359	AMR_9	Translocase 1 inhibitor
A54145	AMR_12	Glucosamine6phosphate synthase
Abyssomicin	AMR_103	4amino4deoxychorismate (ADC) synthase
Actinorhodin	AMR_103, AMR_116, AMR_261	Unknown
Albicidin	AMR_186, AMR_185	DNA gyrase
Albomycin	AMR_101	Serine tRNAsynthetase
Alnumycin	AMR_116	Unknown
Alphalipomycin	AMR_116	Membrane disruption
Althiomycin	AMR_286, AMR_113	Penicillin binding proteins
Andrimid	AMR_96	AcetylCoA carboxylase
Apramycin	AMR_201	Ribosome inhibitor
Aranciamycin	AMR_116	Unknown

Asukamycin	AMR_293	Unknown
Avilamycin	AMR_138	Ribosome inhibitor
Azicemicin	AMR_224	Unknown
Bacitracin	AMR_2, AMR_83	Undecaprenyl Binder
Balhimycin	AMR_70, AMR_84, AMR_83	DAlaDAla chelator
BE14106	AMR_271	Unknown
Borrelidin	AMR_100	ThreoninetRNA synthetase
Calcimycin	AMR_311	Membrane disruption
Calciumdependentantibiotic	AMR_15	Membrane disruption
Capreomycin	AMR_310	Ribosome inhibitor
Capuramycin	AMR_9	Translocase 1 inhibitor
Cephalosporin	AMR_39	Penicillin binding proteins
CephamycinC	AMR_45, AMR_116	Penicillin binding proteins
Cereulide	AMR_295	Membrane disruption
Chalcomycin	AMR_291	Ribosome inhibitor
Chelocardin	AMR_261	Unknown
Chloramphenicol	AMR_123	Ribosome inhibitor
Chlorothricin	AMR_103	Pyruvate carboxylase
Chlortetracycline	AMR_150	Ribosome inhibitor
Cinnamycin	AMR_8	Phosphatidylethanolamine Binder
Coelimycin	AMR_271	Unknown
Colistin	AMR_22	LPSBinder
Cuevaene	AMR_271, AMR_261	Unknown
Cytolysin	AMR_16	Membrane disruption
Dactylocycline	AMR_261	Ribosome inhibitor
Dapdiamides	AMR_92	Glucosamine6phosphate synthase
Daptomycin	AMR_12, AMR_201, AMR_13	Glucosamine6phosphate synthase
Depsidomycin	AMR_286, AMR_116	Unknown
Enacyloxin	AMR_262	Elongation factor Tu
Enterocin	AMR_116	Membrane disruption
Epidermin	AMR_17, AMR_18	Lipid II Binder
Erythromycin	AMR_226	Ribosome inhibitor
Etnangien	AMR_262	RNA polymerase
Factumycin	AMR_109	Elongation factor Tu
FD594	AMR_261	Unknown
Fortimicin	AMR_232, AMR_114, AMR_116	Ribosome inhibitor
Fosfomycin	AMR_90, AMR_91	UDPNacetylglucosamine3enolpyruvyl transferase
Friulimicin	AMR_7, AMR_271	Undecaprenyl Binder

Gallidermin	AMR_17, AMR_18	Lipid II Binder
GE81112	AMR_296	Ribosome inhibitor
Gentamicin	AMR_126, AMR_232, AMR_116	Ribosome inhibitor
Gramicidin	AMR_211	Membrane disruption
Granaticin	AMR_264, AMR_150	LeucinetRNA synthetase
Griseoviridin	AMR_143, AMR_305, AMR_116	Ribosome inhibitor
Hitachimycin	AMR_271	Unknown
HygromycinA	AMR_127, AMR_128	Ribosome inhibitor
HygromycinB	AMR_129	Ribosome inhibitor
Indanomycin	AMR_261	Membrane disruption
Indolmycin	AMR_99	TryptophantRNA synthetase
Istamycin	AMR_116, AMR_131	Ribosome inhibitor
Kanamycin	AMR_232, AMR_261	Ribosome inhibitor
Kasugamycin	AMR_285	Ribosome inhibitor
Kijanimicin	AMR_261	Unknown
Lacticin	AMR_1	Lipid II Binder
Lactonamycin	AMR_116	Unknown
Laidlomycin	AMR_261	Unknown
Lankacidin	AMR_297	Ribosome inhibitor
Lankamycin	AMR_143	Ribosome inhibitor
Lasalocid	AMR_224	Membrane disruption
Laspartomycin	AMR_7	Unknown
Lincomycin	AMR_112, AMR_226, AMR_116	Ribosome inhibitor
Lobophorin	AMR_261	Unknown
Locillomycin	AMR_211	Unknown
Lysobactin	AMR_268, AMR_211, AMR_201, AMR_255	Peptidoglycan transglycosylase
Macrolactin	AMR_286	FabG
Marinopyrrole	AMR_125	Unknown
Massetolide	AMR_235	Unknown
Mersacidin	AMR_1, AMR_3	Lipid II Binder
Methymycin	AMR_111, AMR_226, AMR_313	Ribosome inhibitor
Microcin C7	AMR_70	AspartatetRNA synthetase
MicrocinB17	AMR_183	DNA gyrase
MicrocinJ25	AMR_182	RNA polymerase
Midecamycin	AMR_143	Ribosome inhibitor
Monensin	AMR_261	Membrane disruption
Mupirocin	AMR_98	IletRNA synthetase
Muraymycin	AMR_11	Peptidoglycan translocase

Myxovirescin	AMR_98	Type 1 signal peptidase
Naphthyridinomycin	AMR_271	DNA polymerase
Neomycin	AMR_134, AMR_133	Ribosome inhibitor
Nisin	AMR_20, AMR_290, AMR_19	Lipid II Binder
NocardicinA	AMR_280	Penicillin binding proteins
Novobiocin	AMR_201	DNA gyrase
Nukacin ISK1	AMR_4	Lipid II Binder
Orfamide	AMR_211, AMR_235	Membrane disruption
Paenibacterin	AMR_22	LPSBinder
Pantocin A	AMR_97	lhistidinol phosphate aminotransferase
Paromomycin	AMR_134, AMR_131	Ribosome inhibitor
Paulomycin	AMR_261	Ribosome inhibitor
Pelgipeptin	AMR_22	Unknown
Pep5	AMR_21	Membrane disruption
Phosphothricin	AMR_102	Glutamine synthase
Pikromycin	AMR_111, AMR_226	Ribosome inhibitor
Platensimycin	AMR_196	FabB F
Polymyxin	AMR_22	LPSBinder
Potensimicin	AMR_226	Unknown
Pristinamycin	AMR_305, AMR_164	Ribosome inhibitor
Putisolvin	AMR_235	Membrane disruption
Pyralomicin	AMR_116	Unknown
Resistomycin	AMR_196	RNA polymerase
Reutericyclin	AMR_261	Membrane disruption
Ribostamycin	AMR_134, AMR_133	Ribosome inhibitor
Rifamycin	AMR_181	RNA polymerase
Ristocetin	AMR_72, AMR_73, AMR_74	DAlaDAla chelator
Rosamicin	AMR_143	Ribosome inhibitor
Rubradirin	AMR_300	Ribosome inhibitor
Salinomycin	AMR_301, AMR_196, AMR_116	Membrane disruption
Saquayamycin	AMR_264, AMR_261	Unknown
Simocyclinone	AMR_261	DNA gyrase
Sisomicin	AMR_126, AMR_232, AMR_116	Ribosome inhibitor
Sorangicin	AMR_262, AMR_180	RNA polymerase
Spectinomycin	AMR_144	Ribosome inhibitor
Steffimycin	AMR_116	RNA polymerase
Stenothricin	AMR_271	Unknown
Streptolydigin	AMR_116	RNA polymerase

Streptomycin	AMR_141, AMR_271	Ribosome inhibitor
Syringafactin	AMR_235	Membrane disruption
Tabtoxin	AMR_272	Glutamine synthase
Taromycin	AMR_12	Membrane disruption
Teicoplanin	AMR_84, AMR_72, AMR_73, AMR_74, AMR_83	DAlaDAla chelator
Teixobactin	AMR_211, AMR_201	Lipid II Binder
Tetarimycin	AMR_303	Unknown
Thiomarinol	AMR_98	IletRNA synthetase
Tiacumicin	AMR_304, AMR_224	RNA polymerase
Tirandamycin	AMR_116	RNA polymerase
TLN05220	AMR_116	Unknown
Tolaasin	AMR_235, AMR_215	Membrane disruption
Trifolitoxin	AMR_197	Unknown
Tylosin	AMR_307	Ribosome inhibitor
Tyrocydine	AMR_201	Membrane disruption
UK68597	AMR_70, AMR_84, AMR_73, AMR_75, AMR_83	DAlaDAla chelator
Vancomycin	AMR_70, AMR_72, AMR_73, AMR_74	DAlaDAla chelator
Viomycin	AMR_310	Ribosome inhibitor
Zeamine	AMR_306	Membrane disruption
Zwittermicin	AMR_198	Unknown

Supplementary Table 2.6 Complete legend of siderophore receptors and related compounds for Supplementary Fig. 2.2

Receptor types	**Figure 1 receptor label**	**Siderophores**	**Figure 1 compound label**
CirA/TonB	1	Vulnibactin	A
FatA/TonB	2	Anguibactin	B
FecA/TonB	3	Citrate	C
femA/TonB	4	Mycobactin	D
FEpA/TonB	5	Amphi-Enterobactin	E
Ferric Hydroxamate Receptor/TonB	6	Vicibactin	F
Ferrichrom receptor/TonB	7	Ferrichrome	G
FhuE/TonB	8	Coprogen	H
FpvA/TonB	9	Pyoverdin CHAO	I
FyuA/TonB	10	Yersiniabactin	J
iutA/TonB	11	Aerobactin	K
FhuD/ABC transpoter	12	Staphyloferrin A	L
FxuD/ABC	13	Exochelin (1,H)	M
FepB/ABC	14	Desferrioxamine A1a	N

Supplementary Table 2.7 Compound substrate counts and rank for siderophore and non-siderophore compounds. All substrates are obtained through deconstruction of compound structures through GRAPE.

Siderophore substrate	**Counts**	**Rank**	**Non-siderophore substrate**	**Counts**	**Rank**
Ser	386	1	Val	79	1
OHOrn	371	2	Leu	77	2
Lys	192	3	Ala	74	3
Gly	129	4	Ser	73	4
Thr	122	5	Thr	69	5
ChrD	88	6	Gly	54	6
2,3DHB	79	7	Hpg	52	7
Ala	78	8	PAA	50	8
OHAsp	77	9	Pro	49	9
BZA	54	10	Cys	42	10
Asp	30	11	Dab	40	11

Supplementary Table 2.8 Accuracy comparison between random forest and SIPE. The test set is separated by chemical families.

Chemical Family	**Random forest accuracy**	**SIPE accuracy**	**SIPE Predicted**	**Total**
Aminoglycoside	1.00	0.00	0.00	1
Lipopeptide	0.67	1.00	0.72	18
Cyclic nrp	1.00	1.00	1.00	2
Hybrid pk/nrp	1.00	1.00	0.94	16
Depsipeptide	0.85	0.90	0.77	13
Siderophore	0.91	0.60	0.45	11
Beta lactam	1.00	0.00	0.00	1
Modified nrp	1.00	1.00	1.00	1
Tetramic acids	1.00	0.00	0.00	1
CDPs	1.00	0.00	0.00	1
RIPP	1.00	1.00	0.22	9
Dipeptide	1.00	1.00	0.50	2
Nucleotide antibiotic	1.00	0.00	0.00	3
Ergopeptine	1.00	0.00	0.00	1
Cyclic peptide	0.88	1.00	0.88	8
Polyketides	1.00	1.00	0.69	64
Indole	1.00	1.00	0.75	4
Linear nrp	0.75	1.00	0.50	4
Glycopeptide	1.00	1.00	1.00	3

Supplementary Table 2.9 Resistance gene precision and frequency results from the devised BGC set. The determined values were used within the generation of the ATP.

Name	Count	Correct Count	Precision	Relative confidence	Frequency	Target	Target(s)
AMR_84	4	4	1	0.8	0.017544	DAlaDAla chelator	DAlaDAla chelator
AMR_232	4	4	1	0.8	0.017544	Ribosome inhibitor	Ribosome inhibitor
AMR_235	5	4	0.8	0.8	0.02193	Membrane disruption	Unknown, Membrane disruption
AMR_83	5	4	0.8	0.8	0.02193	DAlaDAla chelator	DAlaDAla chelator, Undecaprenyl Binder
AMR_8	1	1	1	0.2	0.004386	Phosphatidylethanolamine Binder	Phosphatidylethanolamine Binder
AMR_9	3	3	1	0.6	0.013158	Translocase 1 inhibitor	Translocase 1 inhibitor
AMR_138	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_4	1	1	1	0.2	0.004386	Lipid II Binder	Lipid II Binder
AMR_134	3	3	1	0.6	0.013158	Ribosome inhibitor	Ribosome inhibitor
AMR_7	2	1	0.5	0.2	0.008772		Unknown, Undecaprenyl Binder
AMR_133	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor
AMR_2	1	1	1	0.2	0.004386	Undecaprenyl Binder	Undecaprenyl Binder
AMR_3	1	1	1	0.2	0.004386	Lipid II Binder	Lipid II Binder

AMR_70	5	4	0.8	0.8	0.02193	DAlaDAla chelator	Aspartatet RNA synthetase, DAlaDAla chelator
AMR_72	4	4	1	0.8	0.017544	DAlaDAla chelator	DAlaDAla chelator
AMR_73	5	5	1	1	0.02193	DAlaDAla chelator	DAlaDAla chelator
AMR_74	4	4	1	0.8	0.017544	DAlaDAla chelator	DAlaDAla chelator
AMR_75	1	1	1	0.2	0.004386	DAlaDAla chelator	DAlaDAla chelator
AMR_96	1	1	1	0.2	0.004386	AcetylCoA carboxylase	AcetylCoA carboxylase
AMR_97	1	1	1	0.2	0.004386	lhistidinol phosphate aminotransferase	lhistidinol phosphate aminotransferase
AMR_92	1	1	1	0.2	0.004386	Glucosamine6phosphate synthase	Glucosamine6phosphate synthase
AMR_90	1	1	1	0.2	0.004386	UDPNacetylglucosamine3enolpyruvyl transferase	UDPNacetylglucosamine3enolpyruvyl transferase
AMR_91	1	1	1	0.2	0.004386	UDPNacetylglucosamine3enolpyruvyl transferase	UDPNacetylglucosamine3enolpyruvyl transferase
AMR_201	6	1	0.166667	0.2	0.026316		Lipid II Binder, Ribosome inhibitor, Membrane

							disruption, Glucosamine6phosphate synthase, DNA gyrase, Peptidoglycan transglycosylase
AMR_98	3	2	0.666667	0.4	0.013158		Type 1 signal peptidase, IletRNA synthetase
AMR_99	1	1	1	0.2	0.004386	TryptophantRNA synthetase	TryptophantRNA synthetase
AMR_285	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_286	3	1	0.333333	0.2	0.013158		Unknown, FabG, Penicillin binding proteins
AMR_280	1	1	1	0.2	0.004386	Penicillin binding proteins	Penicillin binding proteins
AMR_128	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_125	1	1	1	0.2	0.004386	Unknown	Unknown
AMR_127	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_126	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor
AMR_123	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_131	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor

AMR_211	5	2	0.4	0.4	0.02193		Membrane disruption, Lipid II Binder, Peptidoglycan transglycosylase, Unknown
AMR_215	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_293	1	1	1	0.2	0.004386	Unknown	Unknown
AMR_291	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_290	1	1	1	0.2	0.004386	Lipid II Binder	Lipid II Binder
AMR_297	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_296	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_295	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_111	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor
AMR_112	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_113	1	1	1	0.2	0.004386	Penicillin binding proteins	Penicillin binding proteins
AMR_114	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_116	20	7	0.35	1.4	0.087719		Unknown, Penicillin binding proteins, RNA polymerase,

							Ribosome inhibitor, Membrane disruption
AMR_196	3	1	0.333333	0.2	0.013158		Membrane disruption, FabB F, RNA polymerase
AMR_197	1	1	1	0.2	0.004386	Unknown	Unknown
AMR_198	1	1	1	0.2	0.004386	Unknown	Unknown
AMR_1	2	2	1	0.4	0.008772	Lipid II Binder	Lipid II Binder
AMR_18	2	2	1	0.4	0.008772	Lipid II Binder	Lipid II Binder
AMR_19	1	1	1	0.2	0.004386	Lipid II Binder	Lipid II Binder
AMR_16	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_17	2	2	1	0.4	0.008772	Lipid II Binder	Lipid II Binder
AMR_15	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_12	3	2	0.666667	0.4	0.013158		Membrane disruption, Glucosamine6phosphate synthase
AMR_13	1	1	1	0.2	0.004386	Glucosamine6phosphate synthase	Glucosamine6phosphate synthase
AMR_11	1	1	1	0.2	0.004386	Peptidoglycan	Peptidoglycan

						translocase	translocase
AMR_103	3	1	0.333333	0.2	0.013158		Unknown, 4amino4deoxychorismate (ADC) synthase, Pyruvate carboxylase
AMR_102	1	1	1	0.2	0.004386	Glutamine synthase	Glutamine synthase
AMR_101	1	1	1	0.2	0.004386	Serine tRNAsynthetase	Serine tRNAsynthetase
AMR_100	1	1	1	0.2	0.004386	Threonine tRNA synthetase	Threonine tRNA synthetase
AMR_109	1	1	1	0.2	0.004386	Elongation factor Tu	Elongation factor Tu
AMR_268	1	1	1	0.2	0.004386	Peptidoglycan transglycosylase	Peptidoglycan transglycosylase
AMR_264	2	1	0.5	0.2	0.008772		Unknown, Leucine tRNA synthetase
AMR_262	3	2	0.666667	0.4	0.013158		Elongation factor Tu, RNA polymerase
AMR_261	15	8	0.533333	1.6	0.065789		DNA gyrase, Unknown, Ribosome inhibitor, Membrane disruption

AMR_186	1	1	1	0.2	0.004386	DNA gyrase	DNA gyrase
AMR_185	1	1	1	0.2	0.004386	DNA gyrase	DNA gyrase
AMR_183	1	1	1	0.2	0.004386	DNA gyrase	DNA gyrase
AMR_182	1	1	1	0.2	0.004386	RNA polymerase	RNA polymerase
AMR_181	1	1	1	0.2	0.004386	RNA polymerase	RNA polymerase
AMR_180	1	1	1	0.2	0.004386	RNA polymerase	RNA polymerase
AMR_22	4	3	0.75	0.6	0.017544	LPSBinder	Unknown, LPSBinder
AMR_21	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_20	1	1	1	0.2	0.004386	Lipid II Binder	Lipid II Binder
AMR_271	8	5	0.625	1	0.035088		DNA polymerase, Unknown, Ribosome inhibitor, Undecaprenyl Binder
AMR_272	1	1	1	0.2	0.004386	Glutamine synthase	Glutamine synthase
AMR_39	1	1	1	0.2	0.004386	Penicillin binding proteins	Penicillin binding proteins
AMR_164	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_313	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_310	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor

AMR_311	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_45	1	1	1	0.2	0.004386	Penicillin binding proteins	Penicillin binding proteins
AMR_255	1	1	1	0.2	0.004386	Peptidoglycan transglycosylase	Peptidoglycan transglycosylase
AMR_150	2	1	0.5	0.2	0.008772		LeucinetRNA synthetase, Ribosome inhibitor
AMR_305	2	2	1	0.4	0.008772	Ribosome inhibitor	Ribosome inhibitor
AMR_304	1	1	1	0.2	0.004386	RNA polymerase	RNA polymerase
AMR_307	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_306	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_301	1	1	1	0.2	0.004386	Membrane disruption	Membrane disruption
AMR_300	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_303	1	1	1	0.2	0.004386	Unknown	Unknown
AMR_226	5	4	0.8	0.8	0.02193	Ribosome inhibitor	Unknown, Ribosome inhibitor
AMR_224	3	1	0.333333	0.2	0.013158		Unknown, RNA polymerase, Membran

							e disruption
AMR_143	4	4	1	0.8	0.017544	Ribosome inhibitor	Ribosome inhibitor
AMR_141	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_144	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor
AMR_129	1	1	1	0.2	0.004386	Ribosome inhibitor	Ribosome inhibitor

Supplementary Table 2.10 ATP analysis of the devised BGC set and their respective target predictions, and confidence scores.

Antibacterial Biosynthetic Gene Cluster	Predicted Target	Confidence Score
Polymyxin	LPS Binder	7.44
Ribostamycin	Ribosome Inhibitor	0.61
Nocardicin A	Penicillin Binding Protein	1.78
Viomycin	Ribosome Inhibitor	0.28
Etnangien	Unknown	0
Gulmirecin	RNA polymerase	0.49
Nukacin ISK-1	Lipid II Binder	3.7
Nosiheptide	Tryptophan tRNA synthetase	0.5
Laspartomycin	Unknown	0
Tridecaptin	LPS Binder	0.24
Thuggacin	RNA polymerase	0.022
Tylactone	Ribosome Inhibitor	2.03
Tirandamycin	Unknown	0
Tetarimycin	Unknown	2.04
Streptothricin	Unknown	0
BE-14106	Elongation factor Tu	0.33
Tetracycline	Tryptophan tRNA synthetase	0
Paenilamicin	Ribosome Inhibitor	0.014
Bacitracin	DAlaDAla Chelator	0.1
Iturin	DNA Gyrase	0.013
Pristinamycin	Ribosome Inhibitor	4.44
Bactobolin	Peptidoglycan translocase	0.084
Subtilin	Tryptophan tRNA synthetase	0.5
Chalcomycin	Ribosome Inhibitor	3.48
Myxovirescin	Type 1 Signal Peptidase	0.95
Resistomycin	Tryptophan tRNA synthetase	0.3
Nocathiacin	Ribosome Inhibitor	0.020
Fungisporin	Peptidoglycan translocase	0.24
Coelimycin	Unknown	0
Zwittermicin	Unknown	2.04
Pyralomicin	Ribosome Inhibitor	0.62
Pacidamycin	Penicillin Binding Protein	0.0023
Tyrocydine	Membrane Destabilizer	0.91
Calcium dependent antibiotic	Membrane Destabilizer	2.37
Colistin	LPS Binder	7.50
Zeamine	Membrane Destabilizer	2.04

Tabtoxin	Glutamine synthase	1.54
Epidermin	Tryptophan Synthetase	0.5
Actinorhodin	Peptidoglycan transglycosylase	0.060
Enduracidin	DAlaDAla Chelator	0.036
Nisin	Lipid II Binder	5.62
Putisolvin	Membrane Destabilizer	2.01
Taromycin	Glucoasmine-6-phosphate synthase	0.12
Macrolactin	Unknown	0
Tolaasin	Membrane Destabilizer	4.15
Syringomycin	LPS Binder	0.53
Balhimycin	DAlaDAla Chelator	4.91
Telomycin	Cardiolipin	0.27
Ristocetin	DAlaDAla Chelator	5.97
Cephamycin C	Penicillin Binding Protein	0.13
Erdacin	Tryptophan-tRNA synthetase	0.5
Muraymycin	Peptidoglycan translocase	2.08
Chlortetracycline	Ribosome Inhibitor	0.16
Friulimicin	Unknown	0
Factumycin	Elongation Factor Tu	2.04
Thiomuracin	Tryptophan-tRNA synthetase	0
Chondrochlorens	RNA polymerase	0.19
Asukamycin	Unknown	0
Thiostrepton	Tryptophan-tRNA synthetase	0
Teixobactin	Siderophore	0.088
Microcin B17	Tryptophan-tRNA synthetase	0.5
Rifamycin	RNA polymerase	0.48
FD-594	Unknown	0
Terreic Acid	Transcription termination factor Rho	0.049
Simocyclinone	Ribosome Inhibitor	0.34
Aureothin	Ribosome Inhibitor	0.18
Lankacidin	Ribosome Inhibitor	1.95
Mycinamycin	Ribosome Inhibitor	2.67
Chlorothricin	Elongation Factor Tu	0.28
Daptomycin	Glucosamine-6-phosphat synthase	1.92
Lacticin	Lipid II Binder	1.66
TLN-05220	Ribosome Inhibitor	0.017
Subtilosin	Tryptophan-tRNA synthetase	0.5
Lichenysin	Membrane Destabilizer	0.44

Bicornutin	Peptidoglycan translocase	0.020
Azicemicin	Unknown	0
Alnumycin	Tryptophan-tRNA synthetase	0.3
Pelgipeptin	LPS Binder	3.81
Massetolide	Membrane Destabilizer	2.06
Sansanmycin	Penicillin-Binding Protein	0.0023
Phosphothricin	Glutamine Synthase	1.81
Laidlomycin	Ribosome Inhibitor	0.059
Bacillaene	Peptide deformylase, FabG	0.25
Enacyloxin	RNA Polymerase	0.046
Tetronomycin	RNA Polymerase	1.54
Sorangicin	RNA polymerase	2.12
Dactylocycline	Ribosome Inhibitor	1.19
Avilamycin	Ribosome Inhibitor	3.76
Aranciamycin	Unknown	0
Difficidin	Ile-tRNA synthetase	0.15
Steffimycin	Unknown	0
Aureusimine	Glutamine Synthase	0.086
Thienamycin	Tryptophan-tRNA synthetase	0
Caerulomycin A	Penicillin-Binding Protein	0.060
Surfactin	Membrane Destabilizer	0.41
Hormaomycin	DAlaDAla Chelator	0.20
Granaticin	Ribosome Inhibitor	1.38
Albomycin	Serine tRNA Synthetase	2.04
Tetronasin	RNA Polymerase	1.21
Mupirocin	Ile-tRNA Synthetase	0.70
Desmethylbassianin	Unknown	0
Gramicidin	Unknown	0
Griselimycin	DnaN DNA polymerase sliding clamp	0.016
Auricin	Tryptophan-tRNA synthetase	0
Althiomycin	Penicillin Binding Protein	2.04
Lankamycin	Ribosome Inhibitor	5.60
Locillomycin	Unknown	0
Capreomycin	Ribosome Inhibitor	0.29
Bottromycin	Tryptophan-tRNA synthetase	0
Lysolipin	Ribosome Inhibitor	0.44
Andrimid	Acetyl CoA Carboxylase	1.72
Goadsporin	Tryptophan-tRNA synthetase	0
Virginamycin	RNA Polymerase	0.22
Vancomycin	DAlaDAla Chelator	5.16

Paenibacterin	LPS Binder	3.47
A-500359	Translocase I Inhibitor	2.32
Tiacumicin	RNA Polymerase	3.45
Rubradirin	Ribosome Inhibitor	0.23
Marinopyrrole	Unknown	1.2
Arylomycins	Type 2 Signal Peptidase	1.59
Jadomycin	Ribosome Inhibitor	1.11
Pikromycin	Ribosome Inhibitor	5.17
Lysobactin	Peptidoglycan Transglycosylase	6.41
Chejuenolide	Penicillin-Binding Protein	0.027
Teicoplanin	DAlaDAla Chelator	10.43
Kalimantacin	Peptide Deformylase, FabG	0.013
Corallopyronin A	Na-dependent NADH-quinone reductase	0.18
Microcin J25	RNA Polymerase	1.54
Napthyridomycin	Ribosome Inhibitor	0.022
Cytolysin	Membrane Destabilizer	1.54
Niddamycin	Ribosome Inhibitor	1.50
Anglomycin	Ribosome Inhibitor	2.02
Reutericyclin	Siderophore	0.29
Lobophorin	Membrane Destabilizer	0.16
Syringafactin	Membrane Destabilizer	2.09
Elansolid	Peptide Deformylase, FabG	0.64
Coelibactin	Siderophore	0
Borrelidin	Threonine-tRNA Synthetase	1.50
Erythromycin	Ribosome Inhibitor	4.29
Cephalosporin	Penicillin-Binding Protein	0.57
Methymycin	Ribosome Inhibitor	4.49
Cyclomarin	Elongation Factor G	0.036
Hygromycin A	Ribosome Inhibitor	3.58
Pyridomycin	Peptidoglycan Translocase	0.12
Cypemycin	Tryptophan-tRNA synthetase	0
Griseoviridin	Ribosome Inhibitor	8.03
Hygrocin	Ribosome Inhibitor	0.53
Lactonamycin	Ribosome Inhibitor	1.12
Rosamicin	Ribosome Inhibitor	7.88
Enterocin	Siderophore	0.35
Napthocyclinone	Tryptophan-tRNA synthetase	0.5
Tauramamide	Peptidoglycan translocase	0.5
Cereulide	Membrane Destabilizer	1.80
Sevadicin	Peptidoglycan Translocase	0.60

Halstoctacosanolide	RNA Polymerase	0.14
A54145	Glucosamine-6-phosphate synthase	1.65
Potensimicin	Ribosome Inhibitor	3.72
Hitachimycin	Elongation Factor Tu	0.57
Trifolitoxin	Unknown	1.54
Mannopeptimycin	Lipid II Binder	0.051
Alpha Lipomycin	Elongation Factor Tu	0.20
Ramoplanin	DAlaDAla Chelator	0.094
Saquayamycin	Ribosome Inhibitor	1.16
Sphaerimicin	Siderophore	0.077
Penicillin	Penicillin-Binding Protein	2.61
A-102395	Translocase 1 Inhibitor	2.36
Capuramycin	Translocase 1 Inhibitor	2.32
Orfamide	Membrane Destabilizer	2.14
Gallidermin	Lipid II Binder	5.98
Depsidomycin	Membrane Destabilizer	0.090
A40926	DAlaDAla Chelator	0.59
Chelocardin	Unknown	0
UK-68597	DAlaDAla Chelator	10.64
Oxytetracycline	Tryptophan-tRNA synthetase	0
Abyssomicin	RNA Polymerase	0.053
Cuevaene	Unknown	0
A47934	DAlaDAla Chelator	3.15
Valinomycin	Ribosome Inhibitor	0.314
Thiomarinol	Ile-tRNA Synthetase	0.52
Albicidin	DNA Gyrase	4.03
Kirromycin	Elongation Factor Tu	1.23
Kijanimicin	RNA Polymerase	0.022
Calcimycin	RNA Polymerase	0.031
Midecamycin	Ribosome Inhibitor	7.34
Stenothricin	Ribosome Inhibitor	0.14
Myxopyronin	RNA Polymerase	0.054
GE81112	Ribosome Inhibitor	1.90
Napsamycin	Penicillin-Binding Protein	0.0023
Streptolydigin	Ribosome Inhibitor	0.86
Chloroeremomycin	DAlaDAla Chelator	2.36

Supplementary Table 2.11 High resolution mass measurements for SIPE identified compounds: acidobactins, vacidobactins and potensibactin.

Compound	**Calculated m/z ($M+H^+$; iron binding form)**	**Observed m/z ($M+H^+$; iron binding form)**	**Delta ppm**	**Calculated m/z ($M+H^+$; apo-form)**	**Observed m/z ($M+H^+$; apo-form)**	**Delta ppm**
Acidobactin A	791.2267	791.2294	3.4	738.3152	738.3183	4.2
Acidobactin B	775.2317	775.2347	3.9	722.3203	722.3233	4.1
Vacidobactin A	805.2423	805.2458	4.3	752.3309	752.3338	3.9
Vacidobactin B	789.2473	789.2501	3.5	736.3359	736.3387	3.8
Potensibactin	780.2013	780.2017	0.5	727.2899	727.2903	0.5

Supplementary Table 2.12 Detailed ATP analysis of the erythromycin BGC, a known inhibitor of the ribosome.

Resistance		**Target**		**Precision**		**Other Target(s)**	
AMR 226		Ribosome Inhibitor		0.83		Unknown	
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target (s)**	**Name**	**Garlic Score**	**Activity**	**Target (s)**	**Name**
0.76			Pseudo-erythromycin A-6,9-hemiketal	0.69	Antibacterial	Ribosome Inhibitor	N-Demethyl-erythromycin A
0.69			Erythromycin A N-oxide	0.69	Antibacterial	Ribosome Inhibitor	Erythromycin B
0.69			Erythromycin G	0.69	Antibacterial	Ribosome Inhibitor	Clarithromycin
0.69			Erythromycin analogue	0.69	Antibacterial	Ribosome Inhibitor	6-Deoxy-15-Norerythromycin B
0.69			6 DEOXY 15 NORERYTHROMYCIN A	0.69	Antibacterial	Ribosome Inhibitor	Kujimycin C
Best Target Prediction		Ribosome Inhibitor		**Confidence**		4.29	

Supplementary Table 2.13 Detailed ATP analysis of the andrimid BGC, a known inhibitor of acetyl CoA carboxylase

Resistance		**Target**		**Precision**		**Other Target(s)**	
AMR 96		Acetyl CoA Carboxylase		1			
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target (s)**	**Name**	**Garlic Score**	**Activity**	**Target(s)**	**Name**
0.63			PF-1022-D	0.30	Antibacterial		Ilamycin-C1
0.59			Syringolin-D	0.22	Siderophore		Acinetoferrin
0.59			Syringolin A	0.22	Antibacterial	Peptidoglycan translocase	Pacidamycin 1
0.59			Asporchracin	0.22	Antibacterial	Peptidoglycan translocase	Pacidamycin 2
0.50			Eurystatin-C	0.22	Antibacterial	Peptidoglycan translocase	Pacidamycin 3
Best Target Prediction		Acetyl CoA Carboxylase		**Confidence**		1.72	

Supplementary Table 2.14 Detailed ATP analysis of the teicoplanin BGC, a known inhibitor of D-Ala-D-Ala chelator

Resistance	**Target**	**Precision**	**Other Target(s)**
AMR 73	DAlaDAla Chelator	1	
AMR 74	DAlaDAla Chelator	1	
AMR 72	DAlaDAla Chelator	1	
AMR 84	DAlaDAla Chelator	1	
AMR 83	DAlaDAla Chelator	0.8	Undecaprenyl Binder

Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target(s)**	**Name**	**Garlic Score**	**Activity**	**Target(s)**	**Name**
0.92			Teicoplanin A3-1	0.92	Antibacterial	DAlaDAla Chelator	Teichomycin-A1
0.92	Antibacterial		A 84575 A	0.91	Antibacterial	DAlaDAla Chelator	Parvodicin C4
0.92			A-41030-G	0.91	Antibacterial	DAlaDAla Chelator	A-40926-MDC1
0.92			Teicoplanin A2-5	0.91	Antibacterial	DAlaDAla Chelator	Parvodicin B1
0.92	Antibacterial	DAlaDAla Chelator	Teicoplanin-RS-2	0.91	Antibacterial	DAlaDAla Chelator	Parvodicin C2
Best Prediction		DAlaDAla Chelator		**Confidence**		10.43	

Supplementary Table 2.15 Detailed ATP analysis of the rifamycin BGC, a known inhibitor of RNA polymerase

Resistance		**Target**		**Precision**		**Other Target(s)**	
AMR 305		Ribosome Inhibitor		1			
AMR 181		RNA Polymerase		1			
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	Activity	Target (s)	Name	Garlic Score	Activity	Target(s)	Name
0.54			Ammocidin C	0.3	Antibacterial	RNA polymerase	Lipiarmycin B4
0.53			Ammocidin A	0.3	Antibacterial	RNA polymerase	Lipiarmycin A4
0.52			Ammocidin B	0.3	Antibacterial	RNA polymerase	Clostomycin-A
0.48			Ammocidin	0.25	Antibacterial	RNA polymerase	Proansamycin-X
0.44			Ammocidin D	0.25	Antibacterial	RNA polymerase	Rifamycin-W
Best Target Prediction		RNA polymerase		**Confidence**		0.48	

Supplementary Table 2.16 Detailed ATP analysis of the BGC of the macrolide antibiotic aldgamycin

Resistance	**Target**	**Precision**	**Other Target(s)**
AMR 291	Ribosome Inhibitor	1	

Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target (s)**	**Name**	**Garlic Score**	**Activity**	**Target (s)**	**Name**
0.62	Antibacterial		Chalcomycin B	0.57	Antibacterial	Ribosome Inhibitor	Deacetyl-15-Deoxy-15-oxolankamycin
0.61	Antibacterial		Chalcomycin A	0.52	Antibacterial	Ribosome Inhibitor	Kujimycin C
0.57	Antibacterial	Ribosome Inhibitor	Deacetyl-15-Deoxy-15-oxolankamycin	0.49	Antibacterial	Ribosome Inhibitor	15-Deoxylankamycin
0.56	Antibacterial		15-OAlpha-4-L-O-Acetylarcanosyl-lankamycin	0.49	Antibacterial	Ribosome Inhibitor	Kujimycin D
Best Target Prediction		Ribosome Inhibitor		**Confidence**		4.68	

Supplementary Table 2.17 Detailed ATP analysis of bananamide from *P. fluorescens* strain BW11P2, predicting a mode of action as a membrane destabilizer

Resistance		**Target**		**Precision**		**Other Target(s)**	
AMR 235		Membrane Disruption		0.8		Unknown	
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target (s)**	**Name**	**Garlic Score**	**Activity**	**Target(s)**	**Name**
0.81			MDN-006	0.24	Antibacterial	Ribosome Inhibitor	Fujimycin C
0.75			Lokisin	0.13	Antibacterial	Ribosome Inhibitor	Viridogrisein 1
0.75			Tensin	0.10	Antibacterial	Ribosome Inhibitor	Grividomycin I
0.75			Amphisin	0.072	Antibacterial	Membrane Disruption	[Ala4]-Surfactin
Best Target Prediction		Membrane Disruption		**Confidence**		1.84	

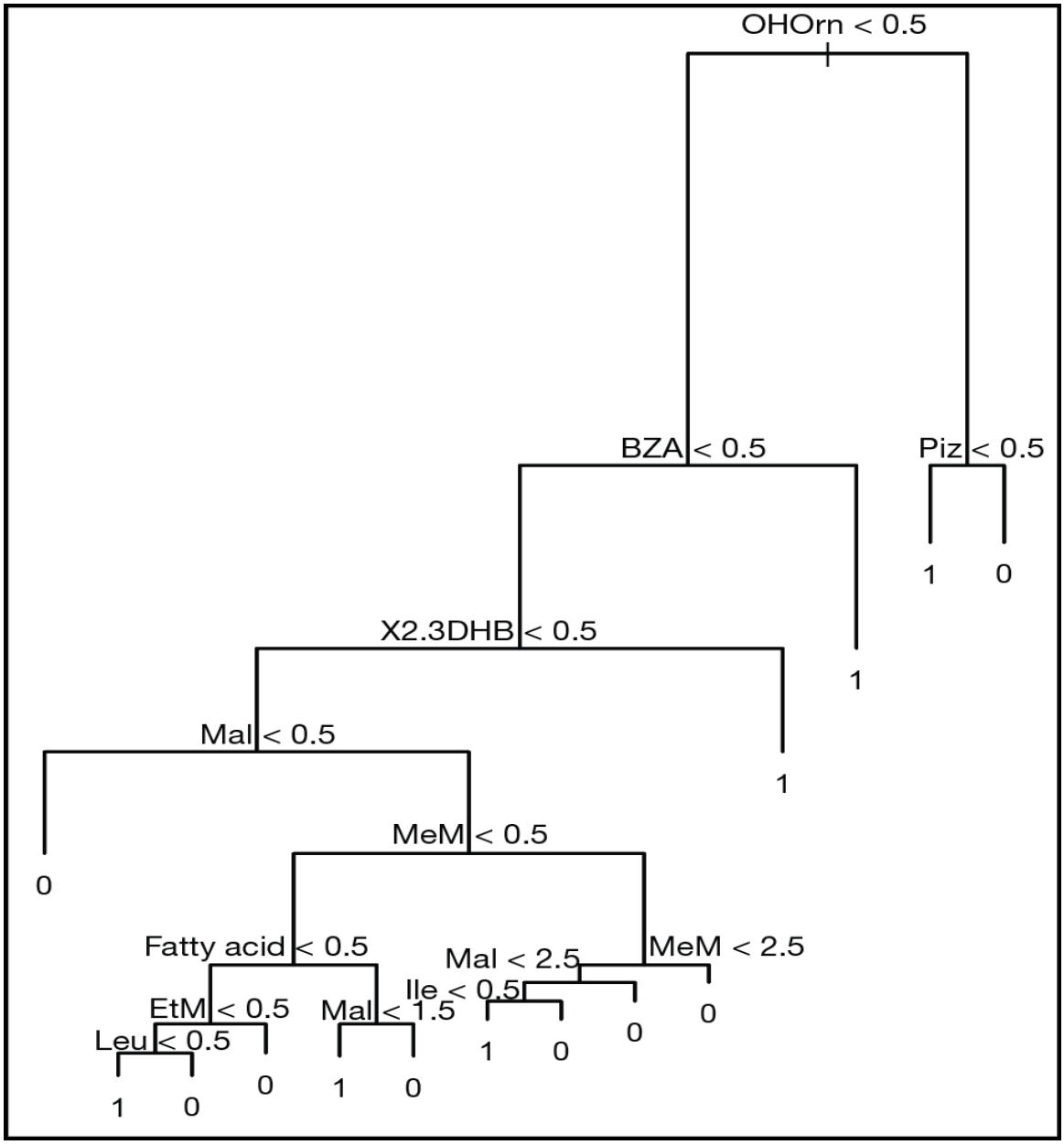
Supplementary Table 2.18 Detailed ATP analysis of the identified cluster for LL-19020 from *S. lydicus tanzanius* NRRL 18036, a previously known antibiotic with a mode of action targeting elongation factor Tu

Resistance		**Target**		**Precision**		**Other Target(s)**	
N/A							
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target (s)**	**Name**	**Garlic Score**	**Activity**	**Target(s)**	**Name**
0.55			3-Furanyl-Avermectin-B1	0.44	Antibacterial	Elongation Factor Tu	GE-21604-A
0.55			3-Furanyl-Avermectin-A2	0.42	Antibacterial	Elongation Factor Tu	Unphenelfamycin
0.52			Cyclobutyl-Avermectin-B1	0.39	Antibacterial	Elongation Factor Tu	LL-E19020 zeta
0.52			Cyclohexyl-Avermectin-A1	0.39	Antibacterial	Ribosome Inhibitor	15-Deoxylankamycin
Best Target Prediction		Elongation Factor Tu		**Confidence**		1.25	

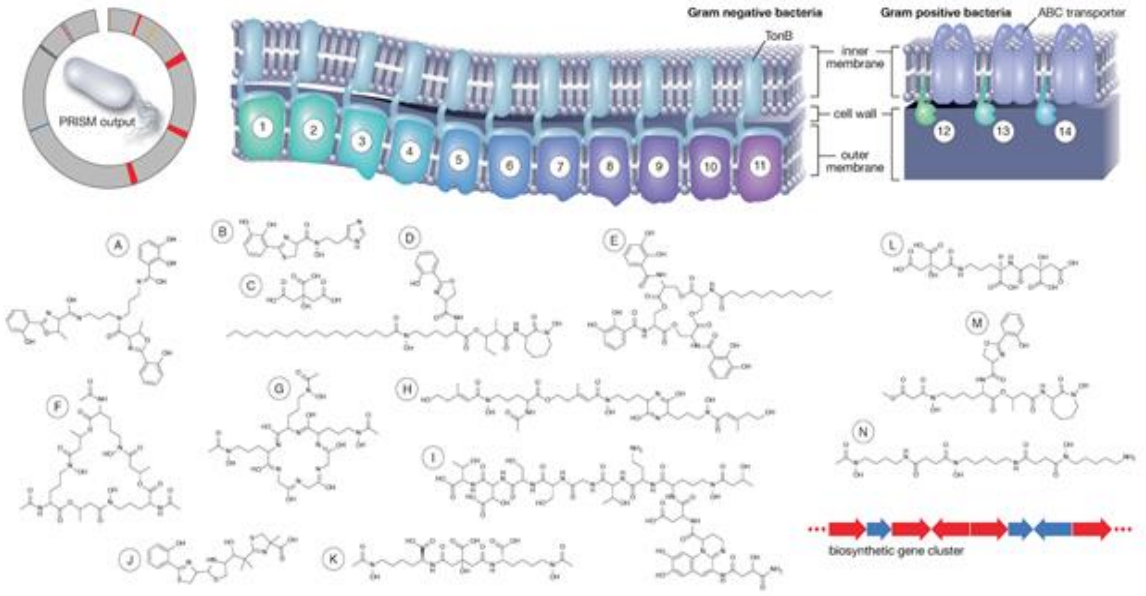
Supplementary Table 2.19 Detailed ATP results of the LL-AO341 BGC from *S. candidus* NRRL 3147, predicting a molecular target of cardiolipin

Resistance		**Target**		**Precision**		**Other Target(s)**	
N/A							
Top Overall Garlic Scores				**Top Garlic Scores with Activity/Target Annotations**			
Garlic Score	**Activity**	**Target(s)**	**Name**	**Garlic Score**	**Activity**	**Target(s)**	**Name**
0.6	Antibacterial	Cardiolipin	Telomycin	0.6	Antibacterial	Cardiolipin	Telomycin
0.55			LL-A0341B	0.12			Malonichrome
0.55			LL-A0341A	0.06			Monamycin-I
0.52			Neotelomycin	0.06			Pyoverdin
0.52			A-128-HYP	0			Heterobactin
Best Target Prediction		Cardiolipin		**Confidence**		0.36	

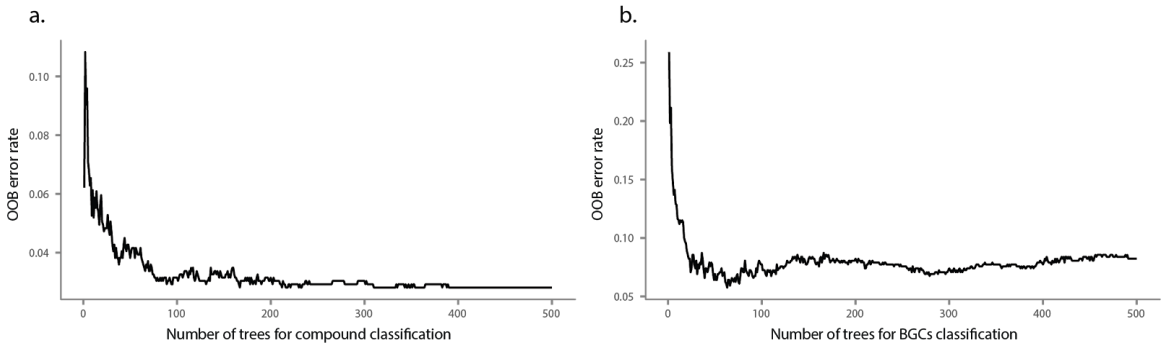
2.8 Supplementary Figures



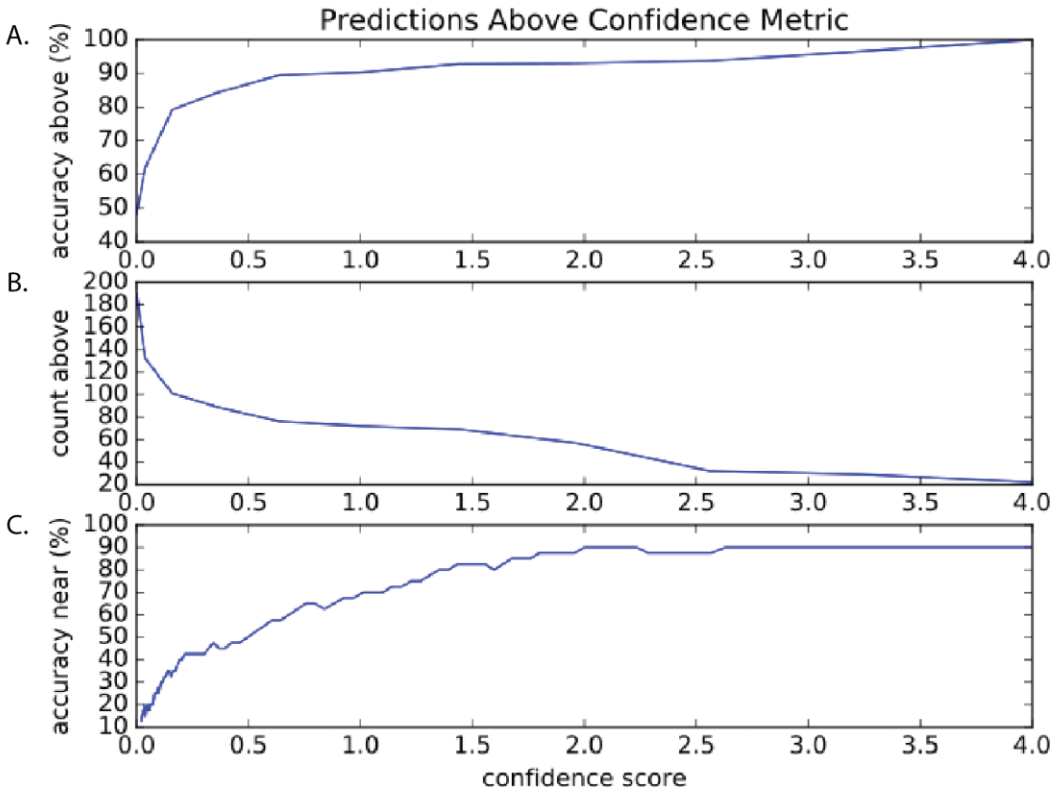
Supplementary Figure 2.1 An example of a classification tree in random forest mode for siderophore prediction.



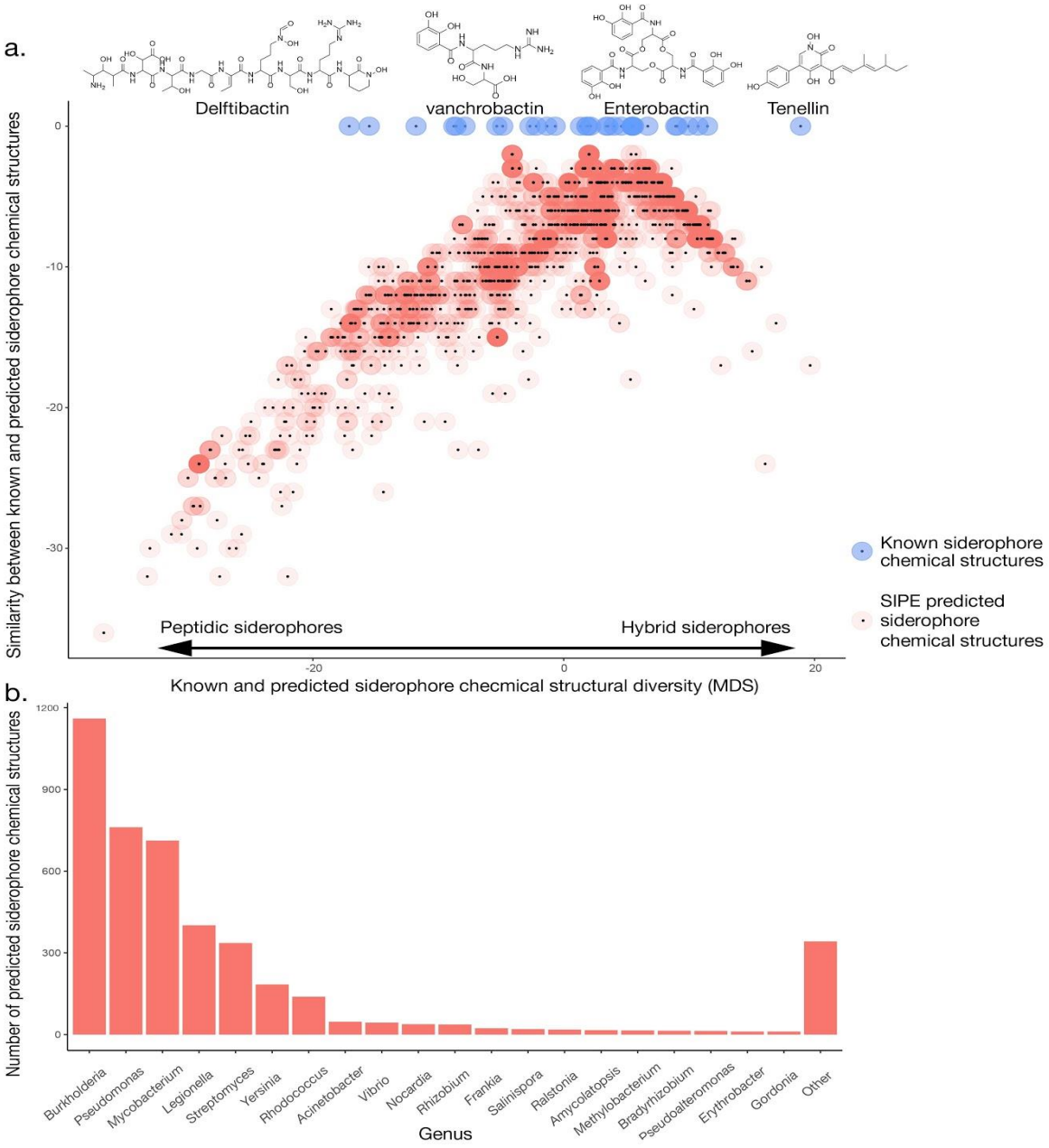
Supplementary Figure 2.2 Summary of microbial siderophore compounds and associated membrane receptors. Fourteen siderophores and their receptors are paired and labeled with letters and numbers, respectively. Receptors 1–11 are from Gram negative bacteria, and 12–14 are from Gram positive bacteria. The full legend can be found in Supplementary Table 2.6.



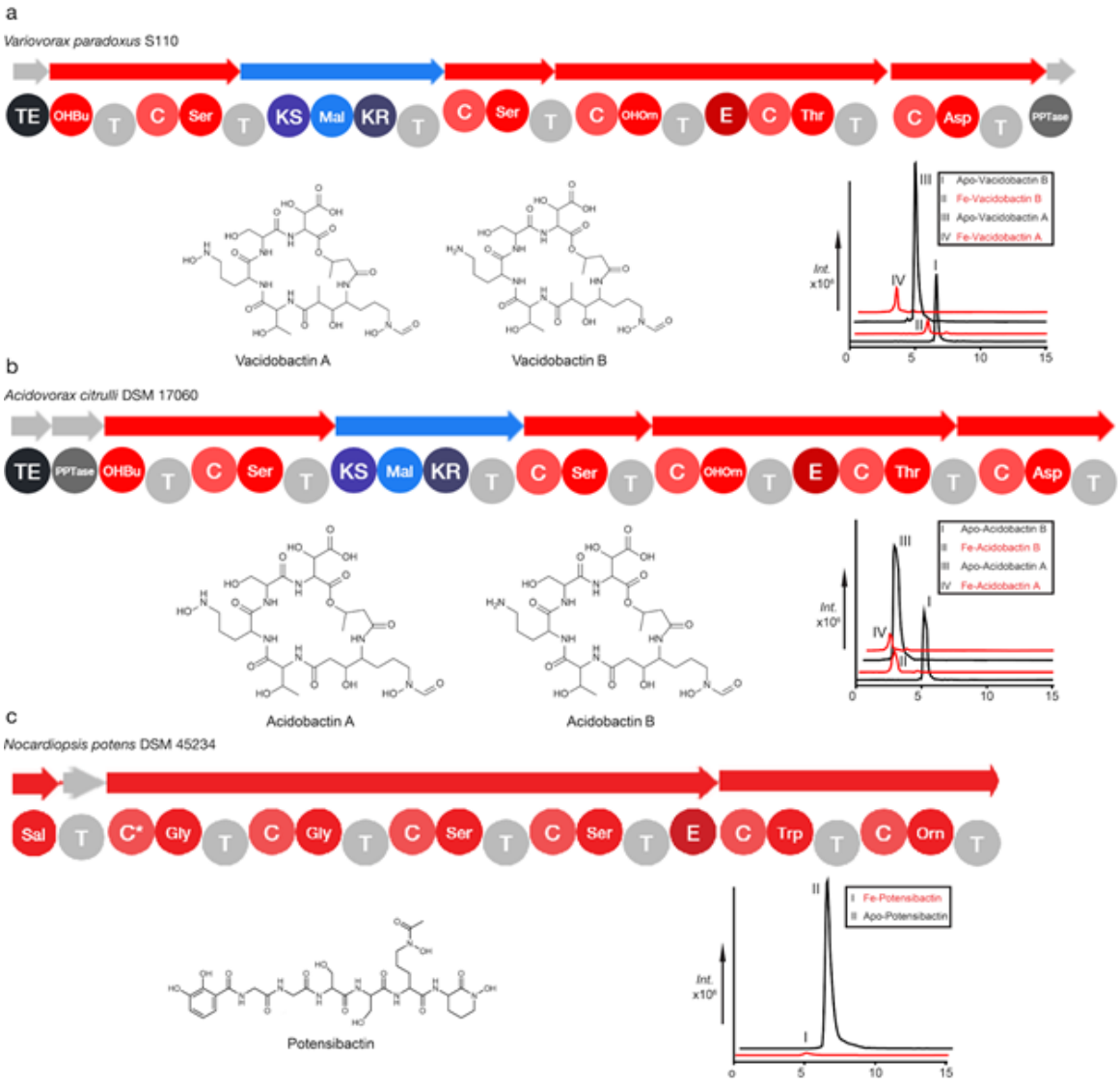
Supplementary Figure 2.3 Out-of-bag error plots for siderophore compounds and natural product biosynthetic gene clusters. The out-of-bag error shows the number of trees required for stable error in the random forest models for compounds (a) or biosynthetic gene clusters (b).



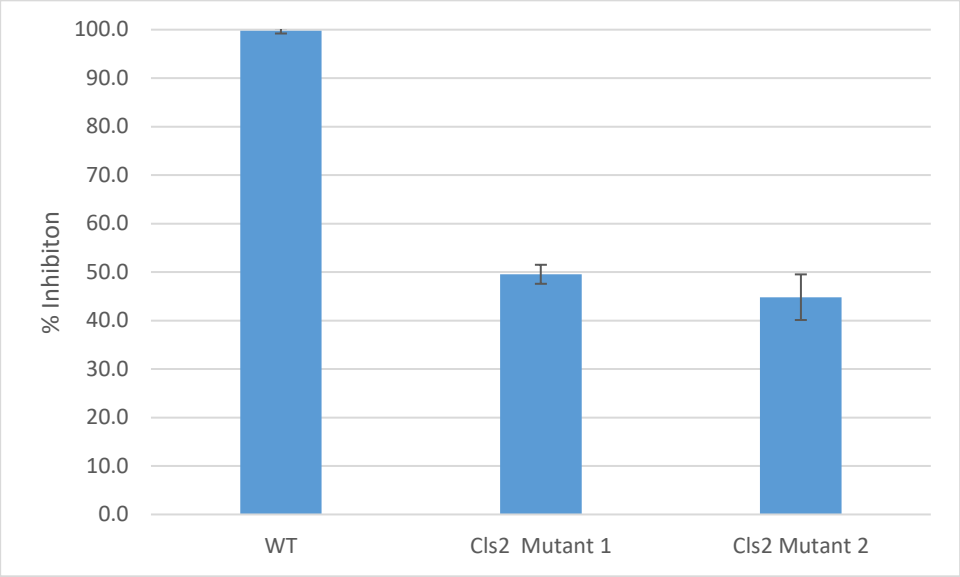
Supplementary Figure 2.4 Relationships between determined confidence score, accuracy, and counts as determined by the ATP pipelines on known antimicrobial BGCs. The devised plots depict (A) the overall accuracy of ATP above a given cut off, (B) the number of BGCs present within the devised set above a given cut off and (C) the likelihood of a correct target prediction being made at a given confidence score.



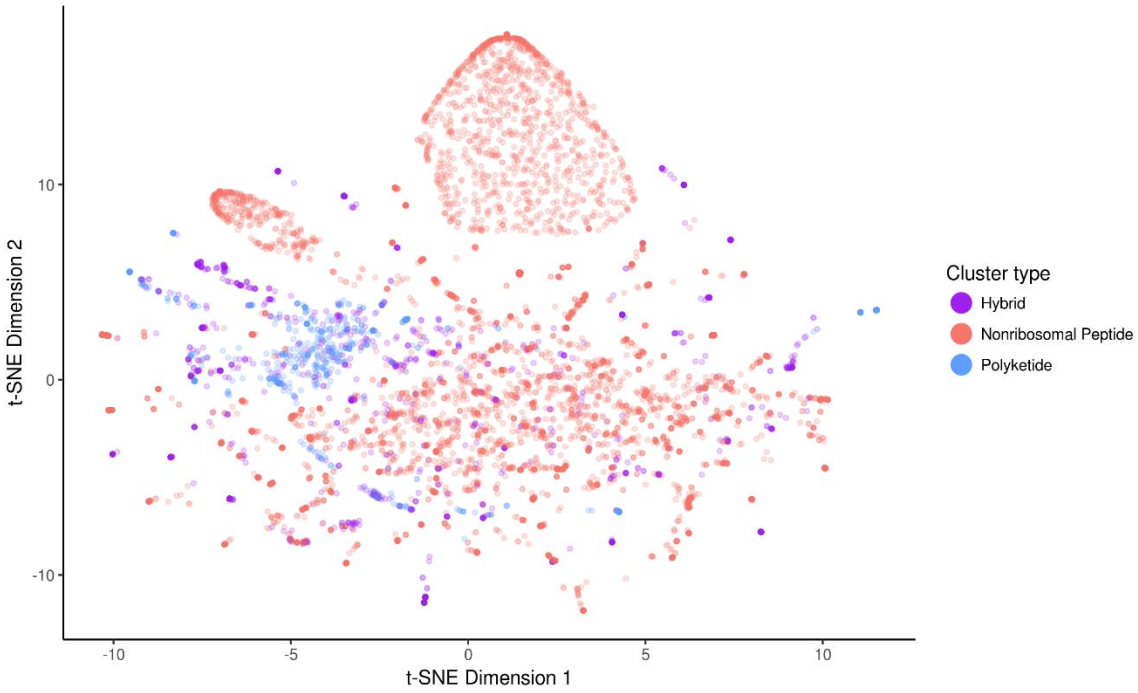
Supplementary Figure 2.5 Global mapping of genomically predicted siderophore chemistries with SIPE. A. Diversity of predicted siderophore chemistries, as measured by the Manhattan distance (rectilinear distance) using 31 known and 4,474 PRISM predicted siderophore chemical structures. The horizontal axis represents the structural diversity of known and predicted siderophore structures, whereas the vertical axis represents the similarity between known and predicted siderophore structures. B. Taxonomical distribution of predicted siderophore BGCs at the genus level.



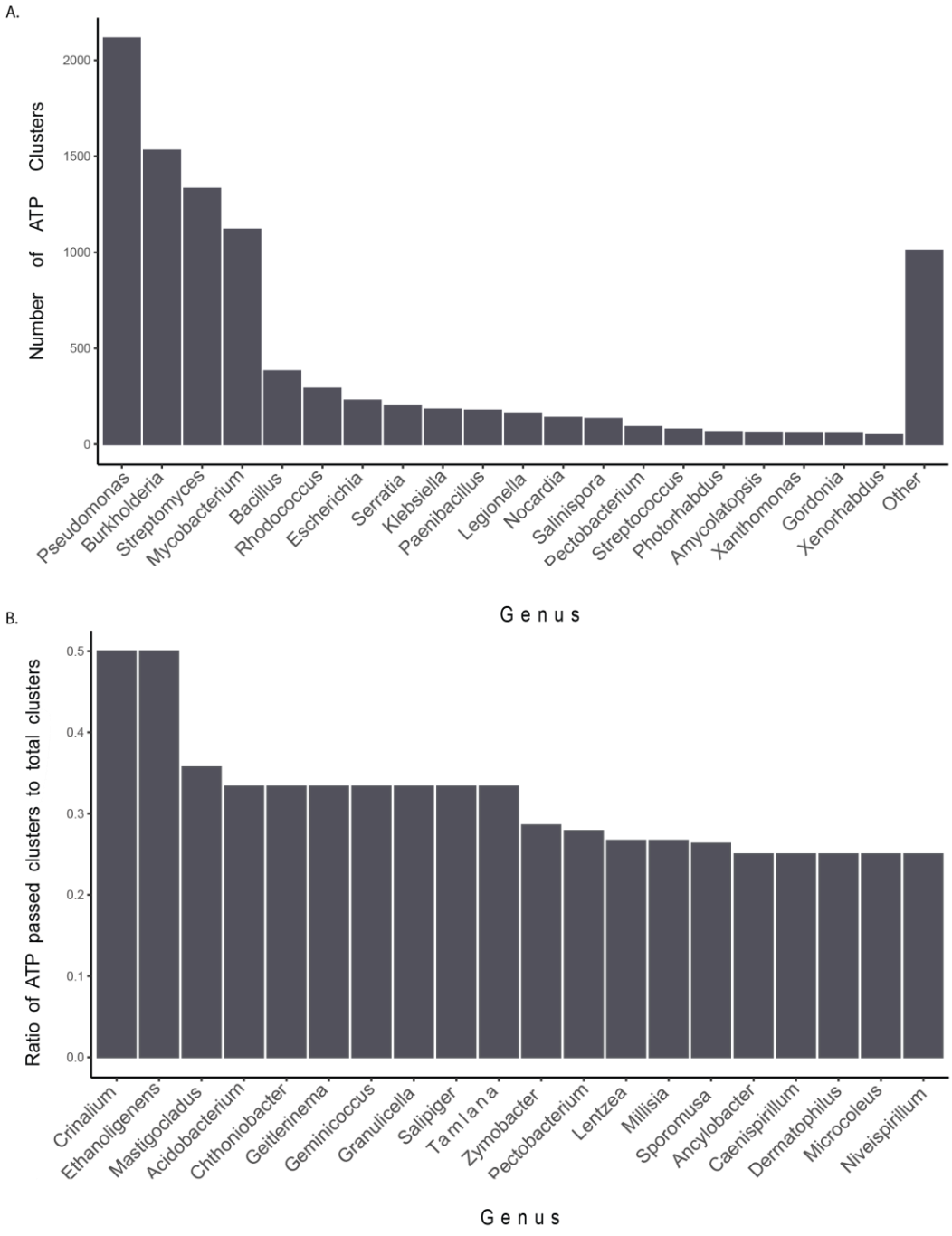
Supplementary Figure 2.6 Confirmation of predicted BGC as siderophores. A. PRISM output of vacidobactins with their structures and LC/MS identification of both Apo and iron binding forms. B. PRISM output of acidobactins with their structures and LC/MS identification of both Apo and iron binding forms. C. PRISM output of potensibactin with its structure and LC/MS identification of both Apo and iron binding forms.



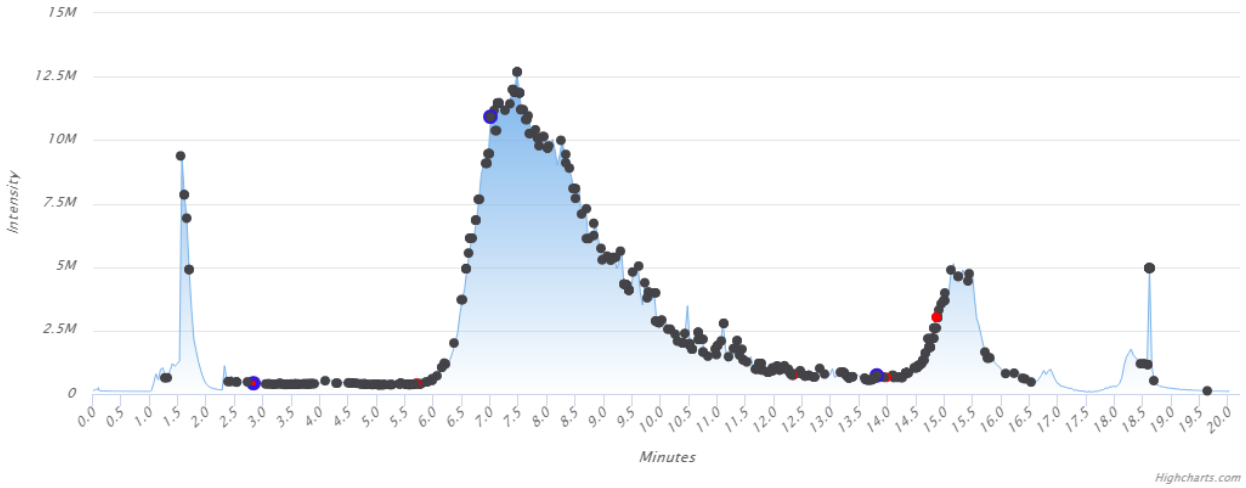
Supplementary Figure 2.7 Crude extracts of *S. candidus* NRRL 3147, producer of LL-AO341, exhibits activity against wild-type *S. aureus*, and a lesser degree to *S. aureus* with mutations in cardiolipin synthase. Wild-type and previously generated spontaneously resistant strains of *S. aureus* to telomycin were exposed to a crude extract of *S. candidus* NRRL 3147 for 16hr. Results are shown as mean percent inhibition ± s.d; n=3. Data is representative of triplicate experiments.



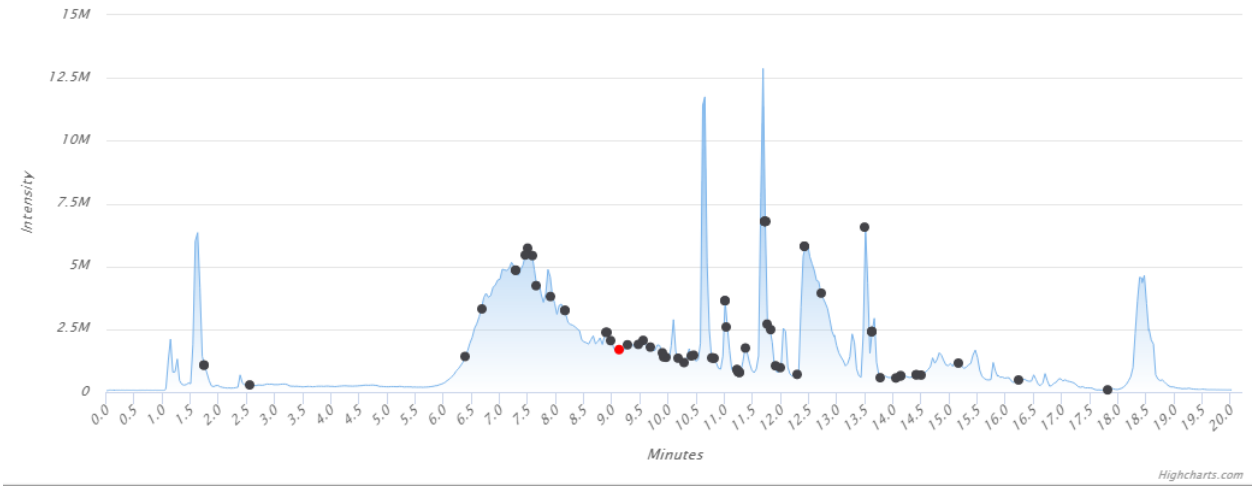
Supplementary Figure 2.8 Global mapping of genomically predicted natural products identified by ATP to potentially diverge mechanistically. Diversity of pNPs identified by ATP as measured by pairwise similarity via GARLIC. Vertical and Horizontal axis are represented by the t-SNE dimensions resulting in a 2D projection plot. Purple dots are indicative of hybrid pNPs, red as NRP pNPs, and blue as polyketide pNPs. Overlapping points indicate pNPs with high structural similarity according to GARLIC.



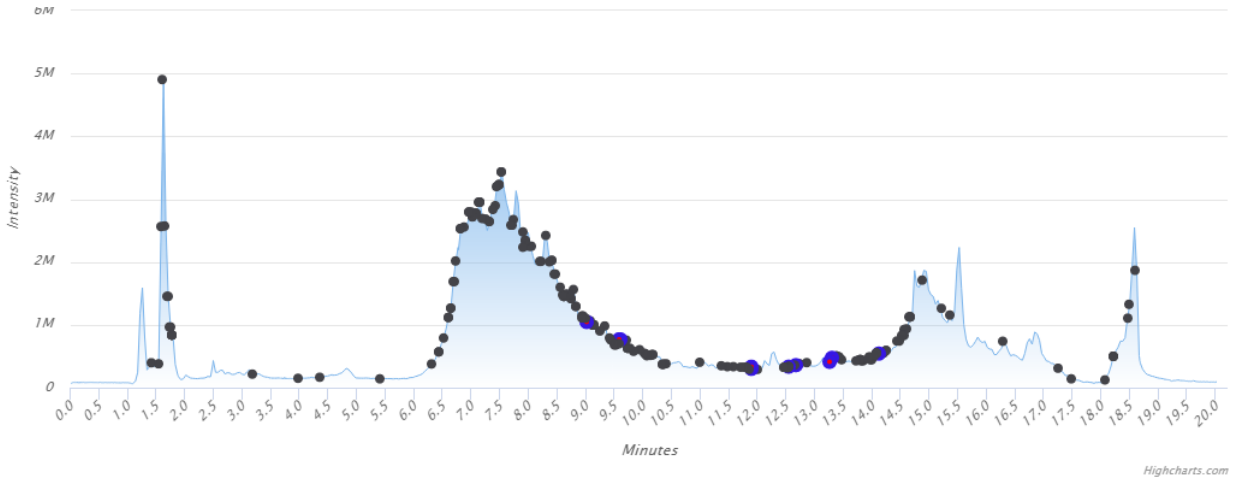
Supplementary Figure 2.9 Taxonomical distribution of pNPs producers with potential for divergent mechanisms as identified by ATP. A. Total counts of pNPs identified by ATP from potential producers at the genus level. B. Ratio of pNPs identified by ATP with potential to diverge mechanistically in comparison to total number of PRISM pNPs at the genus level.



Supplementary Figure 2.10 CLAMS analysis of the microbial extracts produced by *Flexibacter* sp. ATCC 35208. Black dots represent peaks unique to the strain, red dots represent unique peaks with associated known small molecules, red/blue dots represent known small molecules with characterized activity.



Supplementary Figure 2.11 CLAMS analysis representing unique peaks present within the accumulated extracts of *A. muelleri*. Black represents a detected peak with no associated small molecule, red represents a known small molecule with no associated known activity.



Supplementary Figure 2.12 Clams analysis representing unique peaks present within the accumulated extracts of *L. gummosus*. Black dots represent a detected unique peak with no associated small molecule, red/blue dots indicated a known small molecule with characterized activity.

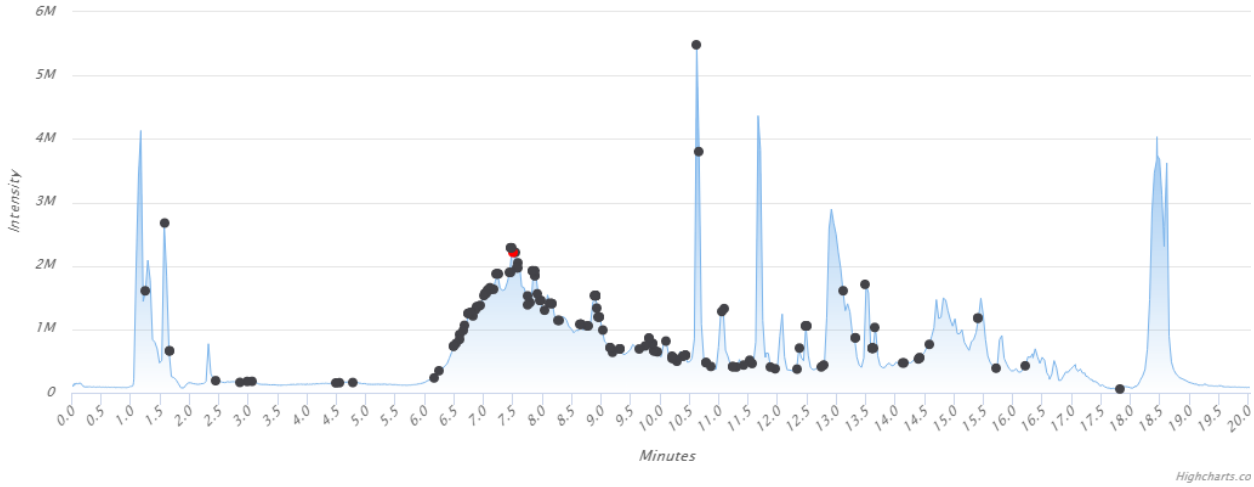


Figure 2.13 CLAMS analysis representing the unique peaks present within the acquired extracts from *Aquimarina* sp. Black dots indicate a detected unique peak, red dots indicate a detected unique peak with an associated small molecule with no activity associations.

2.7 References

1. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529,** 336–343 (2016).
2. Payne, D.J. et al. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov*. **6,** 29-40 (2007).
3. Silver, L. L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* **24,** 71–109 (2011).
4. Harvey, A. L. Natural products as a screening resource. C*urr. Opin. Chem. Biol.* **11,** 480–484 (2007).
5. Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.* **12,** 233–239 (2016).
6. Milshteyn, A., Schneider, J. S. & Brady, S. F. Mining the metabiome: identifying novel natural products from microbial communities. *Chem. Biol.* **21**, 1211–1223 (2014).
7. Medema, M. H. et al. Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11,** 625–631 (2015).
8. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45,** D555–D559 (2017).
9. Fischbach, M.A. & Clardy, J. One pathway, many products. *Nat. Chem. Biol.* **3,** 353- 355 (2007)
10. Fischbach, M. A., Walsh, C. T. & Clardy, J. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4601–4608 (2008).
11. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26,** 1362–1384 (2009).
12. Hider, R. C. & Kong, X. Chemistry and biology of siderophores. *Nat. Prod. Rep*. **27,** 637 (2010).
13. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* (2017).
14. Skinnider, M. A. et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc. Natl. Acad. Sci.* **113**, E6343–E6351 (2016).
15. Skinnider, M. A. et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res*. **43,** 9645–9662 (2015).
16. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45,** D555–D559 (2017).
17. Walsh, C. T. & Fischbach, M. A. Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.* **132,** 2469–2493 (2010).
18. Dejong, C. A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12,** 1007–1014 (2016).

19. D'Costa, V. M., McGrann, K. M., Hughes, D. W. & Wright, G. D. Sampling the Antibiotic Resistome. *Science.* **311,** 374–377 (2006).
20. Cundliffe, E. & Demain, A. L. Avoidance of suicide in antibiotic-producing microbes. J. *Ind. Microbiol. Biotechnol.* **37,** 643–672 (2010).
21. McArthur, A. G. et al. The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57,** 3348–3357 (2013).
22. Sintchenko, V., Iredell, J. R. & Gilbert, G. L. Pathogen profiling for disease management and surveillance. *Nat. Rev. Microbiol.* **5,** 464–470 (2007).
23. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 207–216 (2015).
24. Liu, B. & Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37,** D443-D447 (2009).
25. Wallace, J. C., Port, J. A., Smith, M. N. & Faustman, E. M. FARME DB: a functional antibiotic resistance element database. *Database (Oxford).* (2017).
26. Gerwick, W. H. et al. Structure of Curacin A, a Novel Antimitotic, Antiproliferative and Brine Shrimp Toxic Natural Product from the Marine Cyanobacterium Lyngbya majuscula. *J. Org. Chem.* **59,** 1243–1245 (1994).
27. Ford, P. W. et al. Papuamides A−D, HIV-Inhibitory and Cytotoxic Depsipeptides from the Sponges Theonella mirabilis and Theonella swinhoei Collected in Papua New Guinea. *J. Am. Chem. Soc.* **121,** 5899–5909 (1999).
28. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).
29. Kem, M. P. & Butler, A. Acyl peptidic siderophores: structures, biosyntheses and post-assembly modifications. *BioMetals* **28,** 445–459 (2015).
30. Figueroa, L. O. S., Schwarz, B. & Richards, A. M. Structural characterization of amphiphilic siderophores produced by a soda lake isolate, Halomonas sp. SL01, reveals cysteine-, phenylalanine- and proline-containing head groups. *Extremophiles* **19,** 1183–1192 (2015).
31. Giessen, T. W. et al. Isolation, structure elucidation, and biosynthesis of an unusual hydroxamic acid ester-containing siderophore from Actinosynnema mirum. *J. Nat. Prod.* **75,** 905–914 (2012).
32. Bosello, M., Robbel, L., Linne, U., Xie, X. & Marahiel, M. A. Biosynthesis of the siderophore rhodochelin requires the coordinated expression of three independent gene clusters in Rhodococcus jostii RHA1. *J. Am. Chem. Soc.* **133,** 4587–4595 (2011).
33. Kodani, S. et al. Structure and biosynthesis of scabichelin, a novel tris-hydroxamate siderophore produced by the plant pathogen Streptomyces scabies 87.22. *Org. Biomol. Chem.* **11,** 4686–4894 (2013).
34. Kreutzer, M. F., Kage, H. & Nett, M. Structure and biosynthetic assembly of cupriachelin, a photoreactive siderophore from the bioplastic producer Cupriavidus necator H16. *J. Am. Chem. Soc.* **134,** 5415–5422 (2012).

35. Kodani, S., Komaki, H., Suzuki, M., Hemmi, H. & Ohnishi-Kameyama, M. Isolation and structure determination of new siderophore albachelin from Amycolatopsis alba. *Biometals* **28,** 381–389 (2015).
36. Kodani, S., Komaki, H., Suzuki, M., Kobayakawa, F. & Hemmi, H. Structure determination of a siderophore peucechelin from Streptomyces peucetius. *Biometals* **28,** 791–801 (2015).
37. Wang, W., Qiu, Z., Tan, H. & Cao, L. Siderophore production by actinobacteria. *BioMetals* **27,** 623–631 (2014).
38. Chen, C., Liaw, A. & Com, A. L. *Using Random Forest to Learn Imbalanced Data.* University of California, Berkeley, California, USA, 2004.
39. Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology; 2004-March 2016. Available from: https://www.ncbi.nlm.nih.gov/genome/
40. Johnston, C. W. et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6,** 8421 (2015)
41. Oliynyk, M. et al. Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. *Nat. Biotechnol.* **25,** 447–453 (2007).
42. Park, J.-S., Yang, H. O. & Kwon, H. C. Aldgamycin I, an antibacterial 16-membered macrolide from the abandoned mine bacterium, Streptomyces sp. KMA-001. *J. Antibiot. (Tokyo).* **62,** 171–175 (2009).
43. Nguyen, D. D. et al. Erratum: Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2,** 17010 (2017).
44. Kunstmann, M.P., Mitscher, L.A. & Patterson, E.L. Aldgamycin E, a new neutral macrolide antibiotic. *Antimicrob. Agents Chemother.* **10,** 87–90 (1964).
45. Mizobuchi, S., Mochizuki, J., Soga, H., Tanba, H. & Inoue, H. Aldgamycin G, a new macrolide antibiotic. *J. Antibiot. (Tokyo).* **39,** 1776–1178 (1986).
46. Carter, G. T., Williams, D. R. & Korshalla, J. D. Antibiotic LL-E19020 Zeta and LL-E 19029 Eta. European Patent 0531642 B1 filed 29 June 1992, and issued 21 May 1997.
47. Hall, C. C., Watkins, J. D. & Georgopapadakou, N. H. Effects of elfamycins on elongation factor Tu from Escherichia coli and Staphylococcus aureus. *Antimicrob. Agents Chemother.* **33,** 322–5 (1989).
48. Arnold, W. H., Leonard, P. E., Karl, H. W. & Norman, P. J. Antibiotic ao-341 and production thereof. US Patent 3377244 A filed 13 Jan. 1965, and issued 9 Apr. 1968.
49. Oliva, B., Maiese, W. M., Greenstein, M., Borders, D. B. & Chopra, I. Mode of action of the cyclic depsipeptide antibiotic LL-AO341 beta 1 and partial characterization of a Staphylococcus aureus mutant resistant to the antibiotic. *J. Antimicrob. Chemother.* **32,** 817–830 (1993).
50. Hsu, Y.-H., Dumlao, D. S., Cao, J. & Dennis, E. A. Assessing Phospholipase A2 Activity toward Cardiolipin by Mass Spectrometry. *PLoS One* **8,** e59267 (2013).

51. Tisch, D.E., Huftalen, J.B. & Dickison, H.L. Pharmacological studies with telomycin. *Antibiot. Annu.* **5,** 863–868 (1957-1958).
52. Boucher, H. W. et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48,** 1–12 (2009).
53. Singh, P. D. et al. SQ 28,332, a new monobactam produced by a Flexibacter sp. Taxonomy, fermentation, isolation, structure determination and biological properties. *J. Antibiot. (Tokyo).* **36,** 1245–1251 (1983).
54. Webster, A.L.H., Walker, C., Li, H., Skinnider, M. & Magarvey, N. A. [Flexibacter] sp. ATCC 35208 monobactam biosynthesis gene cluster. Accession no. KY452018.1. *GenBank* (2017).
55. Sykes, R. B. & Bonner, D. P. Aztreonam: The first monobactam. *Am. J. Med.* **78,** 2–10 (1985).
56. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
57. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1273 (2009).
58. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43,** W30–W38 (2015).
59. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (oxford),* bar009. (2011).
60. Johnston, C. W. et al. Gold biomineralization by a metallophore from a gold-associated microbe. *Nat. Chem. Biol.* **9,** 241–243 (2013).
61. Loh, W. *Classification and Regression Trees*, Wadsworth International Group, Belmont,California, USA, 1984.
62. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.
63. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *JMLR* **12**, 2825-2830 (2011).
64. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829 2008.
65. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).
66. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinforma. Orig. Pap.* **28**, 1420–142810 (2012).
67. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

Chapter 3: Significance and Future Directions

Bioactive metabolites from microbes have been major influences in the medical field, especially due to their significant role in the treatment of infectious diseases. However, a general decline in discovery rates of new chemical scaffolds, and complexity of microbial extracts have hindered our success with natural product isolation efforts in present day. Despite the associated setbacks stemming from the traditional approaches of natural product discovery, the advent of next generation sequencing has provoked renewed interest in environmental microorganisms for their bioactive secondary metabolite potential[12]. While the encoded capabilities of these organisms to produce numerous undiscovered entities is undisputed, the prioritization and localization of such compounds within complex microbial extracts remains a major hurdle in moving forward with defining the next iteration of natural products. Nonetheless, novel research confronting such limitations is on the rise, with the increased development of unconventional methods for natural product isolation[21]. The combination of genomic information as well as the ability to chemically de-replicate against known products from those predicted within the genome, allows for further development of computational programs to aid in defining the next generation of microbial natural products with desired activities, whilst avoiding potential chemotypes that are unwanted.

The primary aim of my thesis was to develop a new platform for antimicrobial natural product discovery by using antimicrobial resistance genes, and the known chemistry of previously identified natural products, to sort and prioritize genetically "primed" microbes for their potential to produce natural products with divergent actions. This would be followed up by downstream fermentations, to uncover the possible bioactive predicted natural products. Chapter 2 of my thesis is the central body of work regarding this concept. The act of predicting natural products from genomic information is widely acknowledged, but the ability to readily define those with a potential to be of clinical importance had yet to be established. Through the development of the Antimicrobial Target Predictor, as described in the chapter, we sought to address this

concept by using a series of aligned algorithms to predict with a high degree of accuracy those nonribosomal peptide, and polyketide producing pathways that are likely to encode biosynthetic products of interest. The investigation into the predicted natural products identified within this chapter revealed the ability of the Siderophore Identification Prediction Engine to accurately predict the action of three newly identified siderophore compounds; acidobactin, vacidobactin, and potensibactin.

Moreover, we were able to further demonstrate the accuracy of the pipeline in eliminating previously known antibiotics such as erythromycin, and to gain further insight into the mode of action of known compounds such as aldgamycin and bananamide through the pipeline with their recently published BGCs. Additionally, we were able to showcase the pipeline's ability to infer relationships between previously characterized natural products and potential BGCs within the producing organisms. This not only allows for a greater understanding of the mode of action of these compounds, especially in respect to the known antibiotic LL-AO341, but also attracts emphasis to the pipeline's ability to eliminate BGCs of natural products that have yet to be correlated to their known BGCs through the use of GARLIC. Lastly, this chapter draws attention to four producers of pNPs whose associated chemistry, and lack of notable resistance patterns possess the features necessary to be of potential therapeutic interest in defining new antimicrobials.

Future directions of this project remain focused on the four identified producers of pNPs and isolation thereof. In the immediate future, we continue to move forward with the aid of bioinformatic tool, CLAMS, to identify, connect, and characterize the pNPs identified via ATP from *Aquimarina muelleri*, *Aquimarina* sp., and *Lysobacter gummosus*. In regards to *Flexibacter* sp., we are also actively commencing final isolation procedures and NMR experiments to identify the pNPs identified through ATP. Other future goals of this project are to extend further into the exploration of the other pNPs identified by the ATP. Through the aid of this novel platform it is expected that significant progress will be made in the ability to identify chemically distinct entities from various microbes.

The described research project showcases the value of creating allied algorithms to reveal new microbial natural products in a systematic fashion. Devising modern, data-driven methods is essential to leverage the immense amount of information gained by both traditional methods and genomic pursuits to create an all-encompassing platform for discovery. Profiling the features of pNPs of microbes is possible using the ATP platform. The identified pNPs support the central hypothesis of this work to combine genome mining and resistance profiling to reveal new sources of microbial natural products. We expect that further studies into these identified predicted natural products will reveal interesting chemical scaffolds with potential therapeutic value, to meet the demand for new antibiotics in the era of antibiotic-resistance.

References

1. Wright, G. D. Something old, something new: revisiting natural products in antibiotic drug discovery. *Can. J. Microbiol.* **60,** 147–154 (2014).
2. Fleming, A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *Br. J. Exp. Pathol.* **10,** 226-236.
3. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529,** 336–343 (2016).
4. Taubes, G. The Bacteria Fight Back. *Science* **321,** 356–361 (2008).
5. World Health Organization. *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics.* (World Health Organization, 2017). http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/
6. Bérdy, J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J. Antibiot. (Tokyo).* **65,** 385–395 (2012).
7. Pew Research Center. *Antibiotics Currently in Clinical Development.* (Pew Research Center, Washington, D.C., 2016). http://www.pewtrusts.org/en/multimedia/data-visualizations/2014/antibiotics-currently-in-clinical-development
8. Rice, L. B. Progress and Challenges in Implementing the Research on ESKAPE Pathogens. *Infect. Control Hosp. Epidemiol.* **31,** S7–S10 (2010).
9. Pendleton, J. N., Gorman, S. P. & Gilmore, B. F. Clinical relevance of the ESKAPE pathogens. *Expert Rev. Anti. Infect. Ther.* **11,** 297–308 (2013).
10. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79,** 629–661 (2016).
11. Weller, M. G. A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques. *Sensors (Basel).* **12,** 9181–9209 (2012).
12. Luo, Y., Cobb, R. E. & Zhao, H. Recent advances in natural product discovery. *Curr. Opin. Biotechnol.* **30,** 230–237 (2014).
13. Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* **1830,** 3670–3695 (2013).
14. Mishra, K. P., Ganju, L., Sairam, M., Banerjee, P. K. & Sawhney, R. C. A review of high throughput technology for the screening of natural products. *Biomed. Pharmacother.* **62,** 94–98 (2008).
15. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **4,** 206–220 (2005).

16. Singh, S. B., Young, K. & Miesel, L. Screening strategies for discovery of antibacterial natural products. *Expert Rev. Anti. Infect. Ther.* **9,** 589–613 (2011).
17. Barnes, E. C. *et al.* The use of isolated natural products as scaffolds for the generation of chemically diverse screening libraries for drug discovery. *Nat. Prod. Rep.* **33,** 372–381 (2016).
18. Boehm, M., Zhang, L., Bodycombe, N., Maciejewski, M. & Wassermann, A. M. in *Frontiers in Molecular Design and Chemical Information Science - Herman Skolnik Award Symposium 2015: Jürgen Bajorath* **1222,** 16–345 (American Chemical Society, 2016).
19. Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J. & Shoichet, B. K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **5,** 479–83 (2009).
20. Screening we can believe in. *Nat. Chem. Biol.* **5,** 127–127 (2009).
21. Farha, M. A. *et al.* Strategies for target identification of antimicrobial natural products. *Nat. Prod. Rep.* **33,** 668–680 (2016).
22. Demain, A. L. & Fang, A. The natural functions of secondary metabolites. *Adv. Biochem. Eng. Biotechnol.* **69,** 1–39 (2000).
23. Fischbach, M. A. & Clardy, J. One pathway, many products. *Nat. Chem. Biol.* **3,** 353–355 (2007).
24. Johnston, C. W. *et al.* Assembly and clustering of natural antibiotics guides target identification. *Nat. Chem. Biol.* **12,** 233–239 (2016).
25. Harvey, A. L. Natural products as a screening resource. *Curr. Opin. Chem. Biol.* **11,** 480–484 (2007).
26. Schulze, C. J. *et al.* "Function-first" lead discovery: mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20,** 285–95 (2013).
27. Forsberg, K. J. *et al.* The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens. *Science* **337,** 1107–1111 (2012).
28. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461 (2011).
29. Holmes, A. H. *et al.* Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* **387,** 176–187 (2016).
30. Crofts, T. S., Gasparrini, A. J. & Dantas, G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat Rev Micro* **advance on,** (2017).
31. Miao, V. *et al.* The lipopeptide antibiotic A54145 biosynthetic gene cluster from Streptomyces fradiae. *J. Ind. Microbiol. Biotechnol.* **33,** 129–140 (2006).
32. Anzai, Y. *et al.* Organization of the biosynthetic gene cluster for the polyketide macrolide mycinamicin in *Micromonospora griseorubida*. *FEMS Microbiol. Lett.* **218,** 135–141 (2003).

33. Royer, M. *et al.* Albicidin Pathotoxin Produced by *Xanthomonas albilineans* Is Encoded by Three Large PKS and NRPS Genes Present in a Gene Cluster Also Containing Several Putative Modifying, Regulatory, and Resistance Genes. *Mol. Plant-Microbe Interact.* **17,** 414–427 (2004).
34. Strieker, M., Tanović, A. & Marahiel, M. A. Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* **20,** 234–240 (2010).
35. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).
36. Walsh, C. T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* **33,** 127–35 (2016).
37. Jeong, H. *et al.* Genome Sequence of the Vancomycin-Producing Amycolatopsis orientalis subsp. orientalis Strain KCTC 9412T. *Genome Announc.* **1,** e00408-13-e00408-13 (2013).
38. Konz, D., Klens, A., Schörgendorfer, K. & Marahiel, M. A. The bacitracin biosynthesis operon of Bacillus licheniformis ATCC 10716: molecular characterization of three multi-modular peptide synthetases. *Chem. Biol.* **4,** 927–37 (1997).
39. Miao, V. *et al.* Daptomycin biosynthesis in Streptomyces roseosporus: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151,** 1507–1523 (2005).
40. Fischer, R. Promiscuity Breeds Diversity: The Role of Polyketide Synthase in Natural Product Biosynthesis. *Chem. Biol.* **21,** 701–702 (2014).
41. Hopwood, D. A., Lee, T., Khosla, C., Hopwood, D. & Khosla, C. Cracking the Polyketide Code. *PLoS Biol.* **2,** e35 (2004).
42. Haydock, S. F. *et al.* Cloning and sequence analysis of genes involved in erythromycin biosynthesis in Saccharopolyspora erythraea: sequence similarities between EryG and a family of S-adenosylmethionine-dependent methyltransferases. *Mol. Gen. Genet.* **230,** 120–8 (1991).
43. August, P. R. *et al.* Biosynthesis of the ansamycin antibiotic rifamycin: deductions from the molecular analysis of the rif biosynthetic gene cluster of Amycolatopsis mediterranei S699. *Chem. Biol.* **5,** 69–79 (1998).
44. Walsh, C. T. Polyketide and Nonribosomal Peptide Antibiotics: Modularity and Versatility. *Science* **303,** (2004).
45. Keating-Clay, A. T. Stereocontrol within polyketide assembly lines. *Nat. Prod. Rep.* **33,** 141–9 (2016).
46. Horsman, M. E., Hari, T. P. A. & Boddy, C. N. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? *Nat. Prod. Rep.* **33,** 183–202 (2015).

47. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45,** D555–D559 (2017).
48. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11,** 625–631 (2015).
49. Cundliffe, E. & Demain, A. L. Avoidance of suicide in antibiotic-producing microbes. *J. Ind. Microbiol. Biotechnol.* **37,** 643–672 (2010).
50. Thaker, M. N. *et al.* Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31,** 922–927 (2013).
51. Tang, X. *et al.* Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem. Biol.* **10,** 2841–9 (2015).
52. Butler, M. S., Hansford, K. A., Blaskovich, M. A. T., Halai, R. & Cooper, M. A. Glycopeptide antibiotics: Back to the future. *J. Antibiot. (Tokyo).* **67,** 631–644 (2014).
53. Arthur, M., Molinas, C., Depardieu, F. & Courvalin, P. Characterization of Tn1546, a Tn3-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. *J. Bacteriol.* **175,** 117–27 (1993).
54. Marshall, C. G., Lessard, I. A., Park, I. & Wright, G. D. Glycopeptide antibiotic resistance genes in glycopeptide-producing organisms. *Antimicrob. Agents Chemother.* **42,** 2215–20 (1998).
55. van Wageningen, A. A. M. *et al.* Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem. Biol.* **5,** 155–162 (1998).
56. Yim, G., Thaker, M. N., Koteva, K. & Wright, G. Glycopeptide antibiotic biosynthesis. *J. Antibiot. (Tokyo).* **67,** 31–41 (2014).
57. Schäberle, T. F. *et al.* Self-resistance and cell wall composition in the glycopeptide producer Amycolatopsis balhimycina. *Antimicrob. Agents Chemother.* **55,** 4283–9 (2011).
58. Kurz, S. G., Furin, J. J. & Bark, C. M. Drug-Resistant Tuberculosis: Challenges and Progress. *Infect. Dis. Clin. North Am.* **30,** 509–522 (2016).
59. Park, J. W., Ban, Y. H., Nam, S.-J., Cha, S.-S. & Yoon, Y. J. Biosynthetic pathways of aminoglycosides and their engineering. *Curr. Opin. Biotechnol.* **48,** 33–41 (2017).
60. Kudo, F. & Eguchi, T. Biosynthetic genes for aminoglycoside antibiotics. *J. Antibiot. (Tokyo).* **62,** 471–481 (2009).
61. Ramirez, M. S. & Tolmasky, M. E. Aminoglycoside modifying enzymes. *Drug Resist. Updat.* **13,** 151–171 (2010).

62. Wright, G. D. Aminoglycoside-modifying enzymes. *Curr. Opin. Microbiol.* **2,** 499–503 (1999).
63. Distler, J., Braun, C., Ebert, A. & Piepersberg, W. Gene cluster for streptomycin biosynthesis in Streptomyces griseus: analysis of a central region including the major resistance gene. *Mol. Gen. Genet.* **208,** 204–10 (1987).
64. Thompson, C. J., Kieser, T., Ward, J. M. & Hopwood, D. A. Physical analysis of antibiotic-resistance genes from Streptomyces and their use in vector construction. *Gene* **20,** 51–62 (1982).
65. Weber, J. M. *et al.* Organization of a cluster of erythromycin genes in Saccharopolyspora erythraea. *J. Bacteriol.* **172,** 2372–83 (1990).
66. Dhillon, N., Hale, R. S., Cortes, J. & Leadlay, P. F. Molecular characterization of a gene from Saccharopolyspora erythraea (Streptomyces erythraeus) which is involved in erythromycin biosynthesis. *Mol. Microbiol.* **3,** 1405–14 (1989).
67. Schoner, B. *et al.* Sequence similarity between macrolide-resistance determinants and ATP-binding transport proteins. *Gene* **115,** 93–6 (1992).
68. Liu, M., Kirpekar, F., van Wezel, G. P. & Douthwaite, S. The tylosin resistance gene *tlrB* of *Streptomyces fradiae* encodes a methyltransferase that targets G748 in 23S rRNA. *Mol. Microbiol.* **37,** 811–820 (2000).
69. Rodríguez, A. M., Olano, C., Vilches, C., Méndez, C. & Salas, J. A. Streptomyces antibioticus contains at least three oleandomycin-resistance determinants, one of which shows similarity with proteins of the ABC-transporter superfamily. *Mol. Microbiol.* **8,** 571–82 (1993).
70. Zhao, L., Beyer, N. J., Borisova, S. A. & Liu, H. β-Glucosylation as a Part of Self-Resistance Mechanism in Methymycin/Pikromycin Producing Strain *Streptomyces venezuelae* †. *Biochemistry* **42,** 14794–14804 (2003).
71. Abraham, E.P. & Chain, E. An Enzyme from Bacteria able to Destroy Penicillin. *Nature* **146,** 837–837 (1940).
72. Smith, D. J., Burnham, M. K., Edwards, J., Earl, A. J. & Turner, G. Cloning and heterologous expression of the penicillin biosynthetic gene cluster from penicillum chrysogenum. *Biotechnology. (N. Y).* **8,** 39–41 (1990).
73. Kinscherf, T. G. & Willis, D. K. The Biosynthetic Gene Cluster for the β-Lactam Antibiotic Tabtoxin in Pseudomonas syringae. *J. Antibiot. (Tokyo).* **58,** 817–821 (2005).
74. Paradkar, A. S., Aidoo, K. A., Wong, A. & Jensen, S. E. Molecular analysis of a beta-lactam resistance gene encoded within the cephamycin gene cluster of Streptomyces clavuligerus. *J. Bacteriol.* **178,** 6266–74 (1996).
75. Cox, G. *et al.* A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell Chem. Biol.* **24,** 98–109 (2017).
76. Skinnider, M. A. *et al.* Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* **43,** 9645–62 (2015).

77. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* (2017).
78. Skinnider, M. A. *et al.* Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc. Natl. Acad. Sci.* **113,** E6343–E6351 (2016).
79. Blin, K. *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* (2017).
80. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26,** 1362–84 (2009).
81. Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12,** 1007–1014 (2016).
82. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13,** 30–37 (2016).
83. Johnston, C. W. *et al.* An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6,** 8421 (2015).
84. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57,** 3348–3357 (2013).
85. Haas, W., Pillar, C. M., Torres, M., Morris, T. W. & Sahm, D. F. Monitoring antibiotic resistance in ocular microorganisms: results from the Antibiotic Resistance Monitoring in Ocular micRorganisms (ARMOR) 2009 surveillance study. *Am. J. Ophthalmol.* **152,** 567–574.e3 (2011).
86. Wray, C. & Gnanou, J. C. Antibiotic resistance monitoring in bacteria of animal origin: analysis of national monitoring programmes. *Int. J. Antimicrob. Agents* **14,** 291–4 (2000).
87. Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41,** 203–209 (2014).
88. Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13,** 509–523 (2015).
89. Ochi, K. & Hosaka, T. New strategies for drug discovery: activation of silent or weakly expressed microbial gene clusters. *Appl. Microbiol. Biotechnol.* **97,** 87–98 (2013).
90. Zarins-Tutt, J. S. *et al.* Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products. *Nat. Prod. Rep.* **33,** 54–72 (2015).

91. Luo, Y., Enghiad, B. & Zhao, H. New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters. *Nat. Prod. Rep.* **33,** 174–82 (2016).
92. Tan, G. T., Pezzuto, J. M., Kinghorn, A. D. & Hughes, S. H. Evaluation of natural products as inhibitors of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase. *J. Nat. Prod.* **54,** 143–154 (1991).
93. Gerwick, W. H. *et al.* Structure of Curacin A, a Novel Antimitotic, Antiproliferative and Brine Shrimp Toxic Natural Product from the Marine Cyanobacterium Lyngbya majuscula. *J. Org. Chem.* **59,** 1243–1245 (1994).
94. Wang, W., Qiu, Z., Tan, H. & Cao, L. Siderophore production by actinobacteria. *BioMetals* **27,** 623–631 (2014).
95. Hider, R. C. & Kong, X. Chemistry and biology of siderophores. *Nat. Prod. Rep.* **27,** 637 (2010).
96. Mahajan, G. B. & Balachandran, L. Antibacterial agents from actinomycetes - a review. *Front. Biosci. (Elite Ed).* **4,** 240–253 (2012).
97. Stewart, E. J. Growing unculturable bacteria. *J. Bacteriol.* **194,** 4151–4160 (2012).
98. Decostere, A., Haesebrouck, F. & Devriese, L. A. Characterization of four Flavobacterium columnare (Flexibacter columnaris) strains isolated from tropical fish. *Vet. Microbiol.* **62,** 35–45 (1998).
99. Singh, P. D. *et al.* SQ 28,332, a new monobactam produced by a Flexibacter sp. Taxonomy, fermentation, isolation, structure determination and biological properties. *J. Antibiot. (Tokyo).* **36,** 1245–1251 (1983).
100. Cooper, R. *et al.* Two new monobactam antibiotics produced by a Flexibacter sp. I. Taxonomy, fermentation, isolation and biological properties. *J. Antibiot. (Tokyo).* **36,** 1252–1257 (1983).
101. Chen, M. E., Henry-Ford, D. & Groff, J. M. Isolation and Characterization of *Flexibacter maritimus* from Marine Fishes of California. *J. Aquat. Anim. Health* **7,** 318–326 (1995).
102. Sykes, R. B. & Bonner, D. P. Discovery and development of the monobactams. *Rev. Infect. Dis.* **7,** S579-S593 (1985).