MAPPING BACTERIAL METABOLISM GUIDES NATURAL PRODUCT INQUIRY

COMPREHENSIVE MAPPING OF BACTERIAL METABOLISM GUIDES INQUIRY INTO SPECIALIZED METABOLITES

By Victor Blaga, B.H.Sc.

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

McMaster University

Master of Science (2022)

Hamilton, Ontario (Biochemistry and Biomedical Sciences)

TITLE: Comprehensive Mapping of Bacterial Metabolism Guides Inquiry into Specialized Metabolites

AUTHOR: Victor Blaga, B.H.Sc. (McMaster University)

SUPERVISOR: Dr. Nathan Magarvey

COMMITTEE MEMBERS: Dr. Paul Ayers, Dr. Brian Golding, Dr. Jakob Magolan

NUMBER OF PAGES: xiv, 90

Lay Abstract

A substantial portion of the antimicrobial drugs located within our very own medicine cabinets are produced and secreted by bacteria. Scientists have long been drawn towards the isolation of these molecules, called specialized metabolites, for their therapeutic potential. Today, however, emerging diseases and the antibiotic resistance have rendered many of our conventional medicines ineffective. Now more than ever, there is a need for updated methods in discovering new specialized metabolites; we cannot afford to continue using outdated reductionist approaches. To this end, I have developed a set of software tools which can universally predict the structures of specialized metabolites from the genomes of bacteria, and even connect this information to anticipated biological activity. These tools combine the breadth of knowledge published by experts over decades of research, with state-of-the-art computer algorithms, in order to guide drug discovery in the era of big data and artificial intelligence.

Abstract

Bacterial specialized metabolites (SMs) have long interested scientists due to their diverse chemistries which can harbour antimicrobial activity. The discovery of these molecules experienced a period of exponential success in the mid-1900s and today is guided by next-generation sequencing and liquid chromatograph-mass spectrometry technology. Despite these technological advances, however, they remain under-leveraged. Now, merging disease and antibiotic resistance threaten the security provided by existing antimicrobial medicines. There is an urgent need for a more targeted approach in the discovery of novel SMs, which leverages the tremendous efforts of the past alongside modern big data analytics.

To this end, I set out to develop a foundation that could guide SM discovery in the future. I began by bridging all available bacterial metabolism data into a common latent space. Using known metabolic pathways, I generated a library of biosynthetic units used in a novel program to encode metabolites. Each unit was connected to its requisite genes, which were leveraged to reverse engineer this platform by predicting the chemical structures of SMs directly from their encoding genes. Finally, deep learning models were used to annotate some of these chemical-genomic connections as being associated with a particular activity, in this case siderophore activity.

This suite of software tools offers promising opportunities to pursue downstream applications, including the connection of unknown genes to metabolites and the identification of novel chemical-genomic connections. In order for these prospects to be realized, however, this intricately-connected network of data necessitates more sophisticated interpretation. Indeed, interconnected chemical, genomic and activity data

lends itself particularly well to analysis by graph neural networks. The foundational work described in this communication builds the basis for this comprehensive analysis, which may uncover new insights into the bacterial metabolism we thought we knew.

Acknowledgements

First, I would like to sincerely thank my supervisor Nathan Magarvey for his continuous support and guidance throughout my graduate studies. Nathan, you have been a stellar role model who has pushed me to think big and communicate myself clearly and purposefully. I have no doubt that these skills will follow me throughout life and impact me in ways I am yet to even know.

I would also like to thank my committee members Paul Ayers, Brian Golding and Jakob Magolan for their insightful comments and suggestions during committee meetings. Your patience and enthusiasm for my ambitious interdisciplinary work was greatly appreciated.

The constant support and collaboration within the Magarvey Lab sits at the core of the work that we do. I cannot express enough gratitude to my colleagues, who have been an absolute pleasure to work with. Mathusan, Norman and Keshav, your unwavering optimism and scientific curiosity is both inspiring and energizing. Mathusan, you in particular have helped me immeasurably throughout my time with the lab as a role model and mentor, and for that, I am deeply appreciative. Xiaxia, Tonya and Noah, I am grateful for the work that you do which enables us to try our crazy computational ideas, and I appreciate your patience when they don’t always pan out.

Finally, I would like to thank my friends and family for their unconditional love and support throughout this experience. None of my successes are entirely my own, and I am fortunate beyond measure to have you all in my life.

Table of Contents

List of Figures

List of Tables

Abbreviations

A — *adenylation*

Specific genetic domain used by Nonribosomal Peptide Synthetases to load amino acids.

AGGH — *activity-guided gene hook*

Combinations of genes associated with a particular activity.

BGC — *biosynthetic gene cluster*

Proximally-located groups of genes responsible for the biosynthesis of a natural product.

BLAST — *basic local alignment search tool*

Program for comparing the similarity of sequence data.

C — *condensation*

Specific genetic domain used by Nonribosomal Peptide Synthetases to catalyze peptide bond formation.

EC — *enzyme commission*

Numbers used to classify biological enzymes based on their function.

HMM — *hidden Markov model*

Statistical method used to model biological sequence data.

KEGG — *Kyoto encyclopedia of genes and genomes*

Repository for metabolic pathway data.

LCMS — *liquid chromatography—mass spectrometry*

Analytical technique combining liquid chromatography with mass spectrometry.

NGS — *next-generation sequencing*

Broad term describing state-of-the-art sequencing technologies.

NIS — *nonribosomal peptide synthetase-indepdendant siderophore*

Class of chemical compounds comprised of siderophores which are not non ribosomal peptide synthetases.

NRPS — *nonribosomal peptide synthetase*

Class of chemical compounds comprised mainly of amino acids and biosynthesized in a modular assembly line manner.

SM — *specialized metabolite*

A molecule produced and secreted by a living organism (in this case bacteria) which is not necessary to life but serves a specialized function instead.

SMARTS — *SMILES arbitrary target specification*

Chemical language which represents molecules as text. Can capture more information than SMILES strings.

SMILES — *simplified molecular input line entry system*

Chemical language which represents molecules as text.

Type1PK — *type 1 polyketide synthase*

Class of chemical compounds comprised mainly of discrete polyketide units and biosynthesized in a modular assembly line manner.

Type2PK — *type 2 polyketide synthase*

Class of chemical compounds which feature an often large core scaffold with tailoring groups attached. Biosynthesized through a series of iterative reactions that act on top of each other.

Declaration of Authorship

I, Victor Blaga, declare that this thesis titled, “Comprehensive Mapping of Bacterial Metabolism Guides Inquiry into Specialized Metabolites” and its constituent works contained herein are my own. Each chapter within this thesis corresponds to a software tool or program that was developed to address one of my three thesis aims. Specific details regarding research contributions are outlined in the preface of each chapter.

Chapter 1

Introduction

1.1 Thesis Context

Microbes are the dominant life forms on Earth. Their presence is so pervasive, in fact, that much of the human genome draws itself from these organisms[1]. Even the mitochondria is hypothesized to originate from bacteria of the phylum *Alphaproteobacteria*[2]. The presence of microbes within humans is not limited to these traces, however, as entire microbiomes of organisms exist in nearly every corner of the human body. In fact, the organisms of our microbiome outnumber our human cells ten-to-one[3]. These constituent microbes within the human body only begin to scratch the surface, as these organisms modulate human life in more ways than we can imagine. Bacteria in particular have served specific purposes which have been foundational throughout human history.

For example, bacterial antibiotics can be traced back to an ancient population of Sudanese Nubia from as early as 550 BCE[4]. The molecule tetracycline, produced by *antinomycetes* soil bacteria, was identified in bone samples of that era using fluorescence microscopy[4,5]. It is hypothesized that this natural antibiotic unknowingly found itself in the diet of these individuals, centuries before the modern concept of antibiotics was even established.

From an agricultural perspective, humans' relationship with bacteria may have begun even earlier[6]. Milk products were incorporated into human diet as early as 10,000-5,000 BCE, but their susceptibility to spoiling rendered them impractical. It was not until a group of Middle Eastern herdsman stumbled upon the fermentation of milk that its consumption became more feasible. It was discovered that their use of animal intestines to store milk exposed it to lacto-fermenting bacteria. Not longer after, milk-derived products such as yogurt became a staple of the human diet, as yet again bacteria affected the lives of humans.

Classical Inquiry Into Bacterial Metabolites

Whether it is ancient Sudanse antibiotic use or dairy fermentation in the Middle East, these bacterial functions were only possible due to the diverse chemical entities produced by these organisms. These so-called metabolites have been scarcely understood, even as recently as the early 1900s. At this point in time, laboratory techniques focused on the isolation and culture of entire bacterial organisms, rather than their secreted products[7]. Sporadic discoveries of individual molecules appeared in the form of *gramicidin* and others, but it was not until Sir Alexander Fleming's chance discovery of *penicillin* that a global inquest into bacterial metabolites was sparked. Today, we classify bacterial metabolites into two broad classes: *primary* (or *centralized*) and *secondary* (or *specialized*).

Primary metabolites are responsible for maintaining the basic metabolic functions necessary to life. Sugars, amino acids, nucleic acids and fatty acids are the basic chemical building blocks of life, which have enabled bacteria to live and evolve for millions of years. Given that these metabolites are so fundamental to life, their expression is ubiquitous among bacterial species[8]. By extension, the genetic encodings of these metabolites is also relatively conserved. For example, the genes *pfkA* and *pfkA2* are related homologs of the taxonomically-distant bacteria *Escherichia coli* and *Streptomyces coelicolor*, respectively[9,10]. Both genes encode *phosphofructokinase* enzymes essential to the glycolytic pathways inherent in each organism's primary metabolism. Findings such as these have historically relied on human annotation and curation, but technological advances from the 1990s onwards have expedited the field.

Research surrounding these primary metabolites has benefited from software such as the Basic Local Alignment Search Tool (BLAST), which compares sequence data and calculates its similarity[11]. In the case of *pfkA* and *pfkA2*, BLAST can readily identify a high similarity score between these two genes in order to determine they are homologs. Conversely, BLAST-searching a gene across a library of known genes can identify if the gene is unrelated to anything previously observed. This strategy is dependant on having a large, curated library of known genes, which has not always been available. Significant work was required to build these libraries in early stages, and even today BLAST is limited in that it cannot identify a new gene in a bacterial genome if no similar sequence exists in the reference library. Nonetheless, as a result of these early curation efforts, and owing to the chemical and genetic ubiquity of primary metabolism, the scientific community has been able to largely characterize primary metabolism within bacteria.

As a result, primary metabolism is now well-documented in accessible, manually-reviewed databases. The BRENDA database is a repository of enzymatic reactions which describe metabolism broadly across all domains of life[12]. Over 16,000 reactions in the BRENDA database are manually classified with enzyme commission (EC) numbers. The Kyoto Encyclopedia of Genes and Genomes (KEGG) organizes many of these reactions into metabolic pathways[13]. Overall, KEGG organizes 18,918 molecules and 11,774 reactions into 551 pathways. The MetaCyc database aims to break down the large metabolic pathways contained in KEGG into smaller, more manageable sub-pathways[14]. This manually-curated database contains 17,780 reactions organized into 3,006 metabolic pathways. Although exhaustive efforts have been undertaken to curate and organize the information within these databases, significant but incomplete overlap exists between sources. The MetAMDB database was introduced in 2021 in an effort to consolidate the information contained within BRENDA, KEGG and MetaCyc. Currently, MetAMDB contains 135,064 metabolites and 64,795 reactions organized into 3,237 metabolic pathways[15]. Moreover, MetAMDB contains atom mapping information which tracks atoms between the reactants and products of each of its reactions. Throughout this work, MetAMDB is used as the gold standard dataset of comprehensively mapped primary metabolism, and serves as the template for curation of similar specialized metabolism data.

Contrary to primary metabolism, specialized metabolism is not required for basic bacterial life. Instead, specialized metabolites are produced and secreted by organisms in order to confer a selective advantage in a given environment[16]. Sometimes, these specialized metabolites function to kill other microbes in the surrounding environment, and can be leveraged by humans to create new antibiotics and pesticides. Given that specialized metabolism is context-specific, it is often unique to a bacterial species or subset of species and its diversity tends to outnumber that of primary metabolism. Specialized metabolism, therefore, is especially labour intensive to research and not as comprehensively-documented as primary metabolism.

The challenges presented with specialized metabolism research, however, were not enough to suppress the growing interest following the discovery of penicillin. The golden-age of specialized metabolism discovery of the mid-1900s saw the identification of thousands of novel metabolites in a short amount of time[7]. During this period, bacteria were isolated and cultured at will, and their secreted metabolites investigated. Success rates were understandably high, given the thousands of bacterial species in existence and that relatively few attempts had previously been made to identify their specialized metabolites. By the time these widespread curation efforts began to lose momentum in the 1990s/2000s, over 23,000 specialized metabolites had been identified — the majority sourced from bacteria[7].

The conventional methods used to discovery specialized metabolites in the 1900s, however, were pre-disposed to uncovering the same data. As certain bacteria are more commonly found in the environment, the probability of exhausting the readily-available catalogue of bacterial species is high. In other words, as the scientific community discovered countless specialized metabolites from bacteria, their chances of discovering new metabolites decreased in favour of rediscovering known molecules.

Adding to this challenge is the fact that many specialized metabolites are only produced and secreted under very specific conditions. As a result, the genes that produce these metabolites are often “cryptic” and not induced under typical laboratory conditions[17]. In these cases, new specialized metabolites which may be fundamental to medicine, agriculture or industry may be hiding in plain sight. As a result of these challenges and others, many large pharmaceutical endeavours into specialized metabolites were driven to a halt by the early 2000s.

Inquiry Into Bacterial Metabolites in the Genomic Era

The rise of next-generation sequencing (NGS) has reignited interest in novel specialized metabolites, by reducing the time and resources required for their discovery[18]. Several factors contribute to the efficiency provided by NGS. Firstly, sequencing a bacterial genome can occur with much less material than culturing and isolation methods. As little as a single bacterial cell may be needed to sequence the organism’s genome[19]. Moreover, sequencing a single bacterial genome may take as little as 30 minutes, as opposed to days of growing bacterial cultures. Finally, the cost of NGS has declined exponentially in recent years, and may continue to do so in the future. For these reasons, NGS has become well-positioned to quickly and cost-effectively generate libraries of bacterial genome sequencing data. The challenge, then, is to use this data to identify novel metabolites and rule out those which have likely been explored already.

It turns out that the innate nature of bacterial specialized metabolism renders it amenable to analysis at the genomic level. More specifically, the genes responsible for the biosynthesis of specialized metabolites are often clustered together in close proximity, in what is referred to as biosynthetic gene clusters (BGCs)[20]. As a result, each specialized metabolite is connected to a BGC which can be detected using NGS. Curated BGCs are publicly available in the MIBiG repository, which contains annotated BGCs for 1,927 specialized metabolites[21]. Computational programs such as PRODIGAL are able to identify new BGCs within a broader genome sequence[22]. Using these BGCs, BLAST can be used to compare the similarity of their sequences. BLAST comparisons, however, are not sufficient to discern if a given BGC will produce a novel metabolite or not. Subsequent technologies have been developed which are able to annotate BGCs and even predict their encoded structures.

The antibiotics and secondary metabolite prediction shell (antiSMASH) is a software program designed to annotate BGCs and predict their encoded structures[23]. AntiSMASH is built around a series of hidden Markov models (HMMs) which predict the likelihood that a gene sequence matches a family of sequences (unlike BLAST which compares against a single sequence at a time). Using these HMMs, antiSMASH can annotate query BGCs with the presence of specific biosynthetic genes used in bacterial specialized metabolism. Additionally, antiSMASH can annotate specific biosynthetic domains corresponding to modular-type specialized metabolites. These metabolites, which encompass the non-ribosomal peptide synthetases (NRPSs) and type 1 polyketide synthetases (Type1PKs), are composed of discrete chemical monomer units which are appended together in an assembly-line manner[24]. Owing to this formulaic and largely-predictable biosynthetic process, antiSMASH is able to predict the encoded chemical structures of these modular-type BGCs. Furthermore, antiSMASH is integrated into the MIBiG repository to automatically update all newly-added BGCs.

The PRediction Informatics for Secondary Metabolites (PRISM) engine takes antiSMASH's structure prediction feature one step further to include nonmodular-type specialized metabolites[25,26]. Unlike modular SMs, these metabolites are biosynthesized through a series of enzyme-catalyzed reactions which act iteratively on top of each other. These metabolites are not readily representable by chemical monomer units and, as a result, are more difficult to predict the structures of. Nonetheless, PRISM uses a library of biosynthetic genes connected to chemical reactions to predict the structures of nonmodular classes of metabolites, such as aminoglycosides and betalactams. However, due to the need to manually program each biosynthetic reaction within the PRISM program, it is laborious to add new reactions and PRISM lacks the ability to predict certain nonmodular classes. Moreover, if the annotations for any genes in a BGC are missing, there is a chance that the associated reaction will not be registered and any downstream reactions will be missed.

In order to address the rigidity and resource demand of rule-based approaches such as PRISM, more generalizable methodologies have emerged in recent years. DeepBGC is a software tool which leverages state-of-the-art deep learning models to predict and classify different types of BGCs from sequence data[27]. Unlike PRISM's predefined rule set, DeepBGC uses a natural language processing model to enable computer-derived insights into the data at hand. Although DeepBGC does not make structural predictions the way antiSMASH and PRISM do, it does challenge the use of rigid, rule-based methodologies in the analysis of such complex data.

Holistic Inquiry Into Bacterial Metabolites in the Future

Many of the existing bioinformatics tools leveraged in bacterial metabolism research aim to interrogate discrete types of data; BLAST analyzes individual sequences, DeepBGC investigates entire BGCs, and PRISM takes it a step further to predict chemical structures. All of these programs treat bacterial metabolism data on the basis of discrete, unrelated entries. Given the pervasiveness and complexity of bacteria, however, this approach may be limited in its ability to fully identify patterns in the associated data. It is already known that there is significant interconnectedness among the metabolism of multiple bacterial organisms. In the case of organ transplants, for example, the composition of the host's microbiome plays a critical role in the success of the operation[28,29]. Different collections of organisms within a microbiome give rise to different collections of secreted metabolites, leading to variable downstream effects on hosts. With this in mind, then, it may be better-suited to investigate bacterial metabolism as a whole, rather than discrete components of it.

There currently exists no framework which enables this collective analysis of all of bacterial metabolism data. The development of such a resource would allow all existing bacterial metabolism data to be integrated into a unified medium, such that broad computer-guided analyses could be performed to derive new patterns within the data (Figure 1). It turns out that modular SMs may offer a source of inspiration towards creating this unified medium. As polymers of chemical monomer units, modular SMs are inherently representable as series of biosynthetic units. Moreover, each of these units can be directly tied to its genetic origin. For example, the incorporation of a valine within an NRPS will always require a thiolation (T) domain, a condensation (C) domain, and an adenylation (A) domain whose binding pocket is specific to valine[24]. The genes associated with these domains are readily-searchable within a gene sequence, using such tools as HMMs, BLAST or state-of-the-art deep learning models. This direct translation between chemistry and genomics enables one to readily predict the genes associated with a SM structure, and conversely the structures encoded for by a set of genes. Expanding these properties beyond just modular metabolites would present an opportunity to comprehensively connect the encoded chemistries of metabolism to their genomic origins.

This interconnected network of data could be further expanded upon to include additional information, such as taxonomical associations of genes and chemical units. Of particular interest would be the ability to associate certain combinations of genes and chemical units to biological activities. This may better facilitate analyses where the collective makeup of bacterial metabolism directly affects results, such as in the case of organ transplants. For example, if the encoded chemistries of a patient’s microbiome were annotated with associated biological activities, a computer model would be better able to predict their modulatory effects on each other and the host’s immune system.

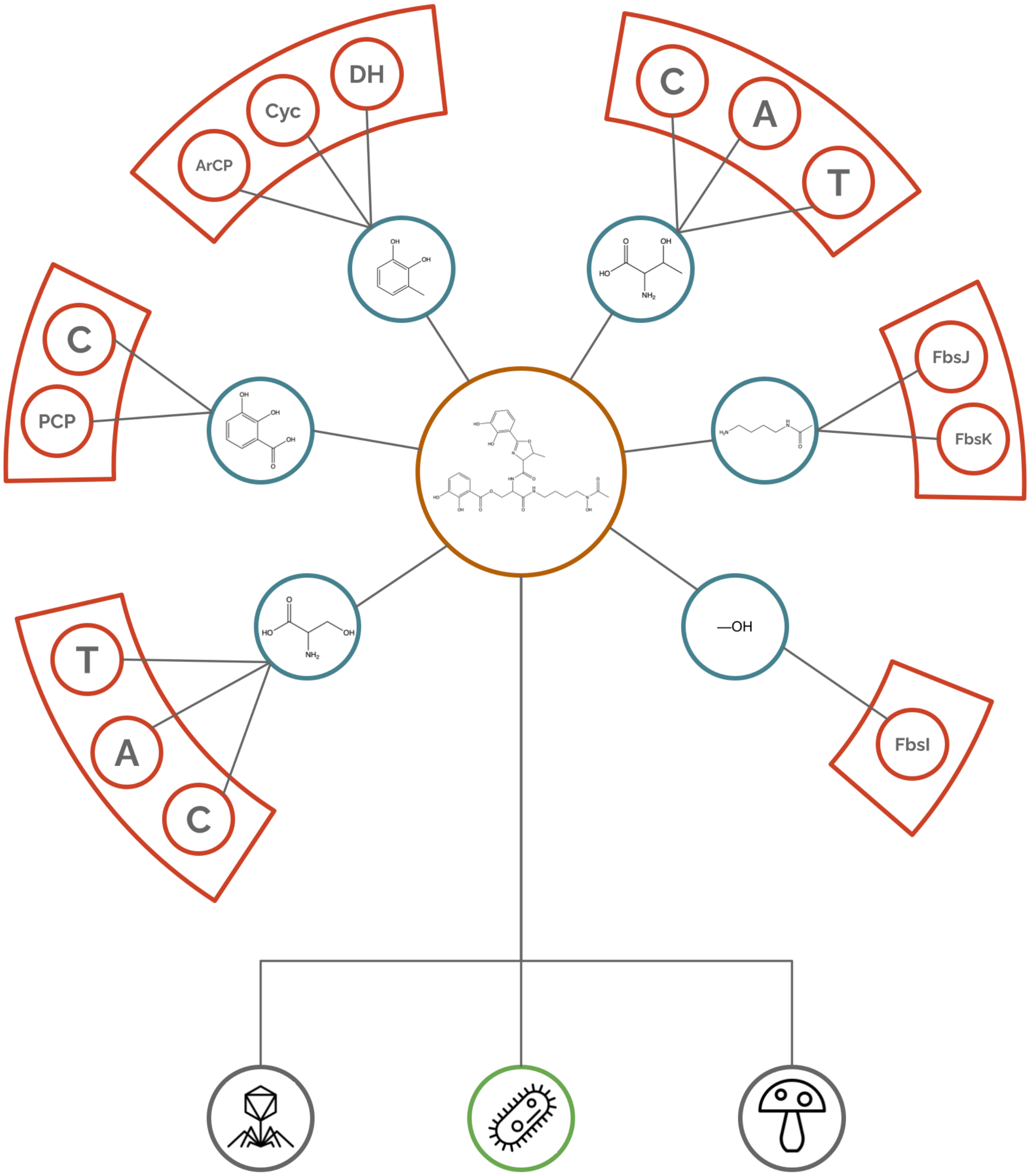


FIGURE 1.1 | **Hypothetical network graph of interconnected bacterial data pertaining to *Fimsbactin*.** Depicted in red is the set of requisite biosynthetic genes, which are connected to their corresponding biosynthetic units (blue). All of these data nodes lead to the same metabolite, *Fimsbactin* (yellow). Associated activities of this metabolite are depicted in green. Note, only the data associated with Fimsbactin is depicted for clarity purposes. In a comprehensive network covering all of bacterial metabolism, any of these data nodes may be connected to other nodes derived from different metabolites. For example, a given biosynthetic unit of Fimsbactin may be found in another metabolite with a different biological activity, and therefore assigned an activity not typically associated with Fimsbactin itself.

Much of the work involved in characterizing metabolism has already been completed, but remains to be fully leveraged. The MetaAMDB database provides comprehensive mapping of bacterial primary metabolic pathways, and even some specialized metabolic pathways. Hundreds of additional specialized metabolic pathways have already been described in individual literature papers, which may be manually curated to augment the data contained within MetAMDB. Moreover, curated in-house datasets and large public repositories such as PubChem contain activity annotations of bacterial specialized metabolites, which could be harnessed to further annotate the integrated bacterial metabolism data. In fact, I believe that the breadth of existing data is comprehensive enough that it can be utilized to design a system which represents all metabolites, whether primary, modular or nonmodular, as a series of annotated, genomically-encoded biosynthetic units.

Put simply, I believe that the way in which we think about bacteria in the future will be drastically different from today. Rather than see them as individual organisms with discrete functions, we will appreciate them for the systems-wide interconnected entities which they are. In order to analyze bacteria as such, we will need new technological frameworks upon which we can build the necessary analytical platforms. It is this guiding principle that has led me to the work contained herein, whose purpose is the following:

> *To develop a framework which integrates known bacterial metabolism into an interconnected network of data, such that it may be used to collectively investigate bacterial metabolism in the future.*

1.2 Purpose of This Work

The work presented herein involves the collective efforts undertaken during my graduate studies to develop tools which integrate and expand upon the existing knowledge of bacterial metabolism. In order to achieve this, my work was broken down into the following 3 aims, each corresponding to a chapter in this work:

1. Develop a chemically-informed biosynthetic representation of SMs which can be connected to genomic information
2. Develop a system for reverse-engineering genomic annotations into the newly-formed biosynthetic representation of SMs
3. Explore the potential to expand this new biosynthetic representation of SMs to contain information regarding biological activity

I begin in Chapter 2 by presenting an approach which generates biosynthetic representations of metabolites. This work rests upon a foundational library of biosynthetic units manually curated by my peers and I. Chapter 2 also discusses my approach to deriving unconventional nonmodular biosynthetic units from metabolic pathways, in order to round out the unit library. Using these units, I present a jointly-developed tool, Bear, which is used to generate the biosynthetic representation of metabolites. The Bear tool is tested on its ability to generate successful annotations of nearly 20,000 bacterial SMs from our in-house database. Finally, Bear is used alongside the curated metabolic pathways to generate so-called gene hooks — combinations fo genes which, when observed, can be expected to give rise to a particular chemical unit or reaction.

Chapter 3 builds upon Bear's gene hook library to reverse engineer the newly-developed biosynthetic representations of SMs. An extension to Bear, *BearClaws*, is developed which predicts the encoded structure of SMs from a set of genomic annotations. BearClaws is run on 95 known SMs and the results are compared to the true structures of the SMs and those predicted by a state-of-the-art competitor program. Finally, selected examples are explored in greater depth to establish the strength and weaknesses of BearClaws, as well as recommend future directions for its improvement.

Chapter 4 concludes my work by exploring the potential to expand upon Bear's biosynthetic unit library to connect genomic information directly to biological activity information. I begin by presenting a series of deep learning models which can classify the activity of SMs. A particular model specific to siderophore activity is used as a test case for the annotation of biosynthetic units with activity. The model's ability to capture the importance of particular units to siderophore activity is investigated. Finally, the corresponding gene hooks for each siderophore-related unit are pooled and labelled to be *activity-guided gene hooks* — combinations of genes associated with a particular activity, in this case siderophore activity.

The collective works presented in Chapters 2,3 and 4 demonstrate a previously-unexplored strategy for the representation of metabolites, as well as connect this encoded information to genomic and activity-related annotations. This work, however, merely sets the foundation for countless opportunities to expand upon it. Future work may seek to match novel metabolites to BGCs, or comprehensively investigate the relationship of biosynthetic units to all known molecular activities.

Chapter 2

Molecular Representations of Metabolites Using *Bear*

2.1 Preface

The work presented in this chapter describes a collaborative effort undertaken in various degrees by several lab members. Mathusan Gunabalasingam conceptualized the initial Bear program in late 2020, and I joined the effort in early 2021. Tonya Malcolm and Xiaxia Di curated the sugars, tailoring groups and amino acids contained in Bear, as well as the specialized metabolic pathways. I curated additional amino acids and derivative reactions, as well as nucleosides and terpene units. I developed the software for biosynthetic pathway mapping and unit extraction. Mathusan curated polyketide units and associated reactions, as well as fatty acid units. Mathusan developed the initial Bear program, then we worked together on it over the course of several months and several iterations. The *NPMage* program referenced in this chapter was developed by Mathusan several years ago. Dr. Nathan Magarvey provided oversight, mentorship and scientific expertise throughout this work.

2.2 Abstract

Bacterial specialized metabolites (SMs) are molecules produced and secreted by bacteria, which harbour unique chemistries and associated pharmacological activities. A key property of SMs is their ability to be reliably connected to their genomic origins by the biosynthetic gene clusters (BGCs) which produce them. Much of the foundational work to identify and characterize bacterial metabolism has already been completed, but work remains to integrate the totality of the known metabolic information in a way which aids further inquiry. Here, we present a program, *Bear*, which leverages the known bacterial metabolism information to generate chemically and genetically guided encodings of metabolites. As part of this program, we manually curated a comprehensive library of biosynthetic chemical units repeatedly observed within specialized metabolism. We further developed a system to track atoms in both primary and specialized bacterial metabolic pathways, in order to generate pathway-derived units which augment this unit library. Finally, we used the Bear program to identify the genes necessary for the production of these biosynthetic units, and suggest some of the many potential use cases for this novel metabolite-relevant molecular encoding.

2.3 Introduction

Experts have long sought to develop universal encoding methods which can represent all molecules. Chemistry is diverse, however, and subjecting all molecules to a generalized mode of representation can reduce resolution. Bacterial metabolites are a specific class of compounds which inherently contain biosynthetic information within their composition. Representing metabolites in a way which captures this biosynthetic information would provide more accurate encodings within this specific class of molecules. In doing so, the gap between chemical and genomic data would be bridged. The successful development of a biosynthetic representation of metabolites would introduce countless downstream applications, including the connection of metabolites to biosynthetic gene clusters (BGCs), and of biosynthetic data to molecular activity.

Historically, molecules have often been represented using the Standardized Molecular-Input Line-Entry System (SMILES)[30]. This text-based encoding of molecules benefits from the fact that it can contain atom-level resolution, and each unique chemical structure can be represented by a unique SMILES entry. Unfortunately, the text-based nature of SMILES renders it less directly processable by statistical computer models. Molecular fingerprints aim to mitigate this shortcoming by representing molecules as feature vectors[31]. These linear stretches of 0's and 1's indicate whether a molecular feature is present (1), or absent (0), in a given molecule. Despite their ability to be more readily interpreted by statistical models, molecular fingerprints are often sparsely-populated and can lack resolution themselves[31]. However, one of the most significant shortcomings of both SMILES and molecular fingerprints is the fact that both are limited to representing chemical data and completely lack additional layers of information. In the case of genomically-encoded bacterial metabolites, there is a need for a molecular representation which contains both chemical and genomic information.

Modular SMs, such as those produced by non-ribosomal peptide synthetases (NRPSs), are innately composed of biosynthetic units containing both chemical and genomic information. For example, amino acids are always incorporated into NRPS compounds by way of a condensation (C) domain, a thiolation (T) domain and an adenylation (A) domain[24]. BGCs can readily be queried for the presence of these domains. Other so-called modular units may include sugars, fatty acids and nucleosides, all of which are discrete chemical structures readily connected to genes. Modular SMs, therefore, could be represented by the biosynthetic units which they are composed of. Developing a methodology to expand these biosynthetic units to include nonmodular and primary metabolic compounds would present an opportunity to encode all metabolites using their biosynthetic origins. In order to do so, the biosynthetic pathways of the known nonmodular and primary metabolites can be analyzed.

The work contained within this chapter builds upon the curated biosynthetic pathways of primary metabolites contained within the open-access MetAMDB database[15]. In addition, our group's previous curation efforts augment this data with pathways specific to specialized metabolites. A methodology is presented which extracts chemical units from these biosynthetic pathways. Using the collective library of biosynthetic units, a tool which represents metabolites as a series of units, *Bear*, is presented and validated.

2.4 Methodology

2.4.1 — Unit Library

A total of 18,385 biosynthetic units are contained within the Bear library. Of these, 3,693 are manually-curated modular units, while 14,692 are pathway-derived units generated computationally. Figure 2.1 illustrates the distribution of units among different classes.

Curated Modular Units

Modular units consisting of amino acids, fatty acids, polyketide monomers, sugars, nucleosides, terpenes, tailoring units and unique miscellaneous units were manually curated by our team. Amino acids and fatty acids were cross-referenced with the *NORINE* database to ensure completeness[32]. Additionally, alpha amino acids were subjected to a series of reactions to generate the following amino acid variants: *beta amino acids, hydroxyacids, alpha-ketoacids, alkanolamines, dehydro aminoacids, cyclo aminoacids, azoles* and *hybrid peptide-polyketide units*. Polyketide units were manually selected to include all substrates (Supplementary Table A1) and known specialized starter units. Additionally, polyketide monomers were subjected to reactions in order to form pyran rings. Sugars, nucleosides and tailoring groups (ex. methyl and hydroxyl groups) were manually selected in order to represent the totality of those found in the known specialized metabolites. Simple linear terpene chains were generated from isoprene units, to a maximum size of 20 isoprene units. Additionally, terpene chains were subjected to hydroxylation at the C5 locus of the isoprene units in every non-redundant combination possible. Finally, unique miscellaneous units observed within specialized metabolic pathways, such as the decalin ring moiety found in *tetrocarins*, were added.

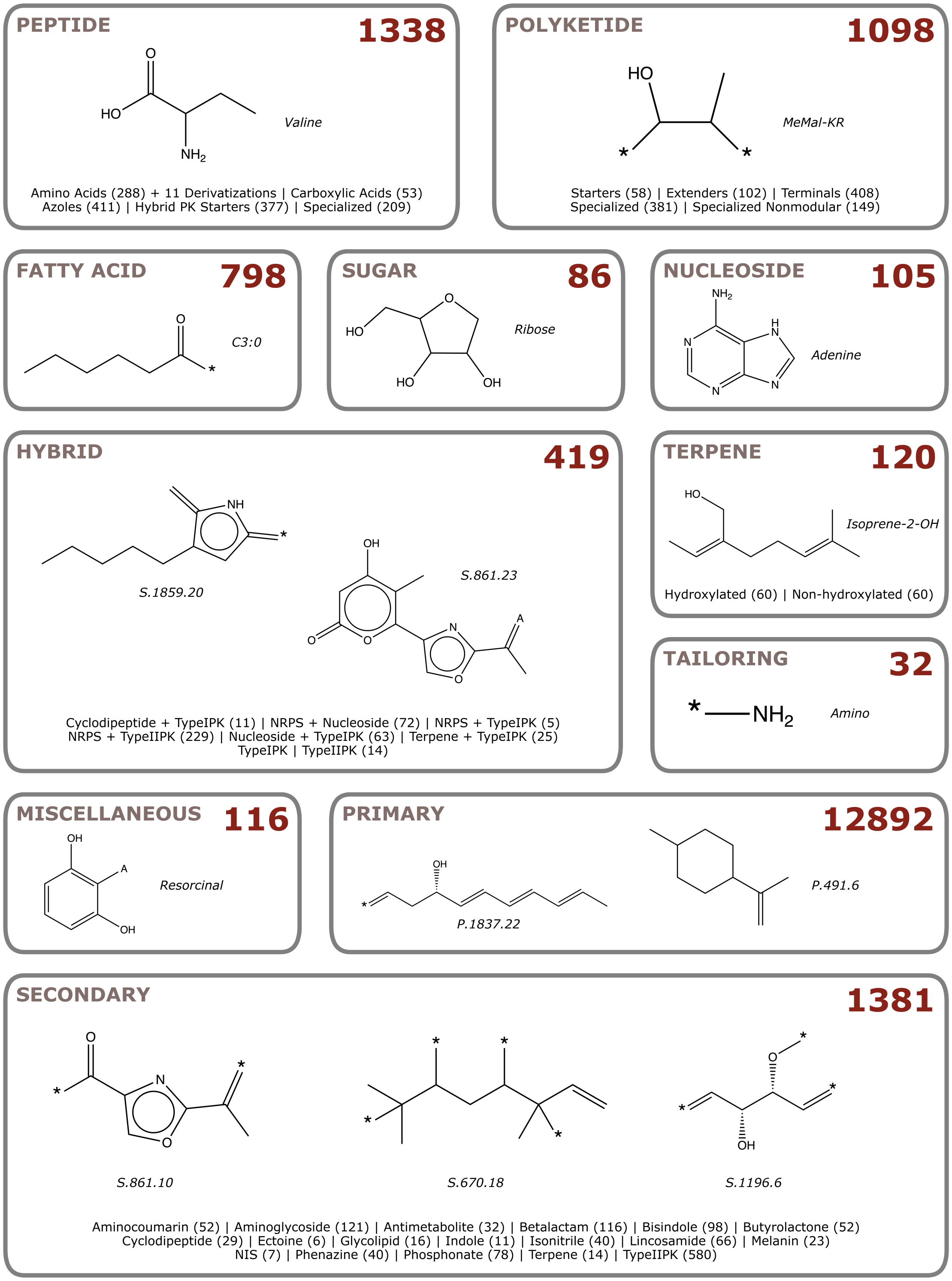

PEPTIDE
1338
O
HO
NH2
Valine
Amino Acids (288) + 11 Derivatizations | Carboxylic Acids (53)
Azoles (411) | Hybrid PK Starters (377) | Specialized (209)
POLYKETIDE
1098
HO
*
*
MeMal-KR
Starters (58) | Extenders (102) | Terminals (408)
Specialized (381) | Specialized Nonmodular (149)
FATTY ACID
798
O
*
C3:0
SUGAR
86
O
HO
HO
OH
Ribose
NUCLEOSIDE
105
NH2
N
H
N
N
N
Adenine
HYBRID
419
NH
*
S.1859.20
OH
S.861.23
O
O
N
O
A
Cyclodipeptide + TypeIPK (11) | NRPS + Nucleoside (72) | NRPS + TypeIPK (5)
NRPS + TypeIIPK (229) | Nucleoside + TypeIPK (63) | Terpene + TypeIPK (25)
TypeIPK | TypeIIPK (14)
TERPENE
120
HO
Isoprene-2-OH
Hydroxylated (60) | Non-hydroxylated (60)
TAILORING
32
*—NH2
Amino
MISCELLANEOUS
116
OH
A
OH
Resorcinal
PRIMARY
12892
OH
*
P.1837.22
P.491.6
SECONDARY
1381
O
*
N
O
*
S.861.10
*
*
*
*
S.670.18
*
O
*
*
OH
S.1196.6
Aminocoumarin (52) | Aminoglycoside (121) | Antimetabolite (32) | Betalactam (116) | Bisindole (98) | Butyrolactone (52)
Cyclodipeptide (29) | Ectoine (6) | Glycolipid (16) | Indole (11) | Isonitrile (40) | Lincosamide (66) | Melanin (23)
NIS (7) | Phenazine (40) | Phosphonate (78) | Terpene (14) | TypeIIPK (580)

FIGURE 2.1 | **Bear unit library.** Depicted are selected examples of biosynthetic units in the Bear library. The unit type is written in the top left of each box and the number of units of that type is written in the top right of each box. Certain types of units contain subtypes, which are written at the bottom of the corresponding box.

Pathway-Derived Nonmodular Units

Nonmodular biosynthetic units were derived from metabolic pathways described in literature. Pathways were processed such that each unique atom and its source (the reaction or intermediate that it came from) was tracked throughout the pathway. For each intermediate in the pathway, atoms from the same source were grouped together and “extracted” as a unit. Redundant units were removed and the resulting units were added to Bear’s final library.

Biosynthetic Pathway Data

Primary metabolic pathways were sourced from the *MetAMDB* database. A total of 3,498 primary metabolic pathways were analyzed. In-house curation efforts by staff members were successful in recording the biosynthetic pathways of 263 specialized metabolites described in literature. Each pathway was recorded as a series of unique reactions. The atoms between the reactants and products of each reaction were mapped using the *ReactionDecoder* tool[33]. Reactions which could not be mapped by ReactionDecoder were mapped manually. For each reaction within the specialized pathways, the gene sequences for the corresponding enzymes were recorded, if available.

Atom Mapping Throughout Pathways

Each metabolic pathway was processed as a set of reactions. Using the *networkx* package in Python, each set of reactions was loaded as a graph where nodes corresponded to reactions[34]. Edges between nodes were creating if the product of one reaction (node) was the reactant of another. This was checked for using the *rdkit* package in Python[35]. To begin, connections were only drawn if the reactant and product represented the exact same molecule, including stereochemistry. Due to minor inconsistencies in the reporting of stereochemistry in the primary metabolic reactions of MetAMDB, a few cases existed where reactants and products were not connected when they should have been. To account for these situations, reactant/product matches were subsequently searched for without stereochemistry requirements, only in cases where there were multiple subgraphs rather than a single unified graph (i.e. the set of pathway reactions was not fully connected). With this contingency in place, all pathways were fully connected and additional nodes were added in between reaction nodes to correspond with each chemical intermediate.

Beginning with starter reactions (those without any preceding reactions), atoms from the reactants and products were tracked and assigned unique map numbers. The unique map numbers of the products were assigned to the intermediates following the starter reactions. Subsequent reactions were iterated in order of proximity, such that atoms carried forward their unique map numbers from intermediate to intermediate. New atoms which were introduced through a particular reaction, as in the case of a hydroxyl group added by a hydroxylase enzyme, were assigned the next-highest, unoccupied map number.

After all atoms within the pathway were assigned a unique map number, the earliest point at which each map number was observed was recorded. This was taken to be the source of the atom. For example, in the case of the hydroxyl group added by a hydroxylase enzyme, the source of the map number corresponding to the hydroxyl's oxygen atom is the hydroxylation reaction itself. A dictionary mapping atoms to their sources was generated for each intermediate in each pathway (Figure 2.2). Atom sources exist in two forms: starter unit sources and reaction sources. Starter units sources correspond to intermediates in the biosynthetic pathway with no reported predecessor (ex. the biosynthesis of some metabolites begins with an amino acid). Reaction sources correspond to reactions which incorporate chemical moieties into an intermediate part way through the biosynthetic pathway. Reaction sources are directly tied to their corresponding genes, whereas starter unit sources are not.

A subset of reactions act without adding or removing atoms. As a result, the map numbers between reactants and products remain the same. For example, a hydroxyl which is reduced to a ketone describes a chemical change in which atom map numbers remain the same. These changes are not captured by the existing source mapping methodology, but contain valuable information that should be documented. To this end, a secondary dictionary pertaining to additional modification reactions of atoms was generated.

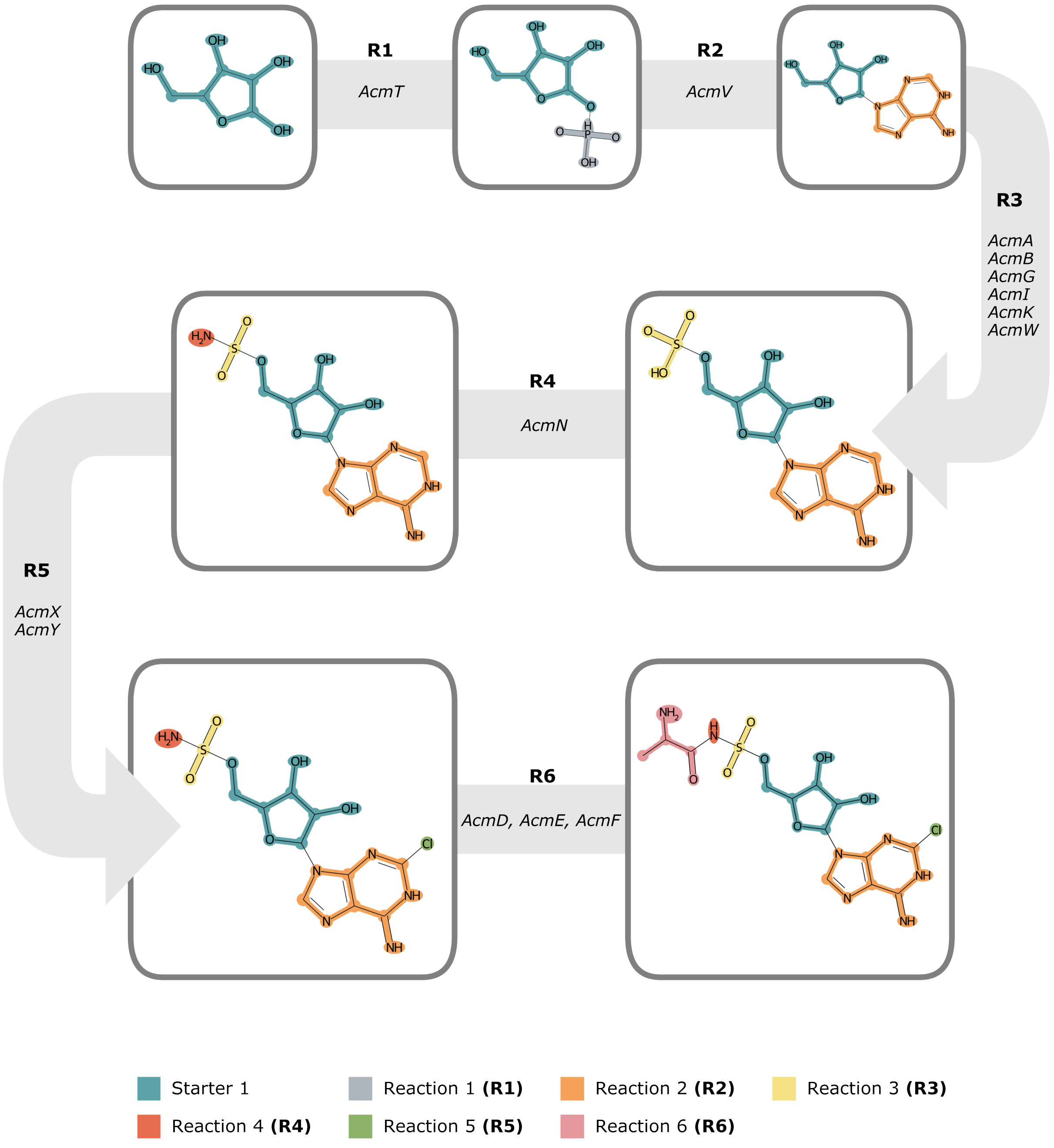


FIGURE 2.2 | **Mapping of *ascamycin* biosynthetic pathway.** Shown is the illustrated pathway mapping of *ascamycin* biosynthesis. Each unique colour corresponds to a unique reaction or starter molecule source within the pathway. The genes responsible for each reaction are shown in italics (ex. *AcmN* for reaction 4 **(R4)**). Note, the additional modification sources are not shown.

For each pair of adjacent intermediates in a pathway, atoms were analyzed for chemical changes using the *rdkit* package. Chemical changes were defined as those situations in which the neighbouring atoms, bonds, valence, charge or hybridization state of an atom changed. Atoms whose map number remained the same, but exhibited a chemical change, were added to the additional modification reactions dictionary. The pathway reaction which gave rise to this chemical change was taken to be the additional modification source of the change. Not all atoms within intermediates were connected to additional modification reactions.

Unit Extraction

The intermediates and final compounds of each pathway were analyzed individually. Adjacent atoms from the same source were grouped together. Molecules were fragmented at the bonds corresponding to the boundaries between sources, and the resulting fragments were taken to be biosynthetic units. Bonds were only considered for fragmentation if they were single bonds and the generated fragments were representable by a valid SMILES string.

After nonmodular units were extracted from each metabolic pathway, duplicates were removed. Each unit was represented by a standardized, canonical SMILES string devoid of map numbers, and a final set of non-redundant units was kept. Each unit was represented with the notation *S.X.Y*, where *X* represents an internal pathway ID number and *Y* represents the unit number from that pathway. The same unit found in multiple pathways was represented by the notation using the lowest pathway ID number. Data tracing units back to all pathways in which they were found was saved. The dereplicated set of pathway-derived units was added to the library of chemical units contained in Bear.

Chemotype Association

Each pathway-derived unit was traced back to the pathways which it is found in. Each of these pathways corresponds to the biosynthesis of a specialized metabolite whose chemotype has been annotated using an in-house program called *NPMage*. The chemotypes of the associated pathways were pooled for each unit, and the most frequent chemotype was assigned to be the chemotype label for the unit itself.

SMARTS Representation

In order to successfully achieve the final goal of Bear — to map query structures as series of biosynthetic units — it was imperative to minimize the number of false positive unit annotations. Using native SMILES strings would limit the specificity of each unit, as SMILES strings only capture information regarding the configuration of atoms bound to each other. As a result, the false positive rate of biosynthetic units mapping to query structures would be high, and computational time would increase.

SMILES Arbitrary Target Specification (SMARTS) strings, on the other hand, are a more complex language for representing chemical structures[36]. Unlike SMILES, SMARTS strings can contain information regarding the valence state, degree, neighbours and hybridization state of atoms. During the extraction of pathway-derived units, a SMARTS string with strict chemical queries was generated for each unit, alongside the SMILES string. Chemical reactions contained within Bear were also represented by SMARTS strings in order to increase specificity and limit false positives. It is these SMARTS strings which are utilized during Bear's annotation of chemical structures.

Flexible Units

Although the use of strict chemical queries in SMARTS strings greatly reduces false positive annotations, it also limits the applicability of Bear units to structures outside of the labelled data used. For example, if a Bear unit is only ever observed within the known data as having a chlorine group attached to a particular carbon, the strict queries of the SMARTS string will ensure that the unit cannot map to derivative structures lacking the chlorine group. It is reasonable to infer, however, that a variation of that structure without the chlorine likely exists within the totality of metabolites in existence. In order for Bear to be able to map these structures as well, so-called flexible units whose carbons are attached to tailoring groups (such as a chlorine) had their chemical queries removed from the SMARTS strings. The chemical queries for all other atoms remained unchanged. This way, the SMARTS strings corresponding to these units are specific enough to avoid false positive mapping, but contain enough flexibility to be applicable to new structures not previously observed. These flexible units are used as a secondary round of annotations within the Bear tool, for query structures which are not fully mapped using the default library.

2.4.2 — Biosynthetic Representation

The Bear tool was designed for the purpose of mapping metabolites as series of biosynthetic units. The tool leverages its library of curated and pathway-derived units to annotate query structures. Redundant annotations are dereplicated and the combination of non-overlapping units which annotates the greatest number of atoms on the query structure is presented to the user (Figure 2.3).

Unit Mapping

To begin, the Bear tool iterates through the library of units to check which ones are found within the query structure. In order to do this, each unit is represented by its SMARTS string, and the *rdkit* package is used to check if the unit is contained within the query structure. All units and the atoms they overlap with in the query structure are recorded.

Inevitably there will often be more unit annotations per query structure than are required to fully map the structure. Using the total set of unit annotations, the optimal combination of non-overlapping units needs to be computed. In order to expedite this computation, a series of dereplication steps removes redundant annotations.

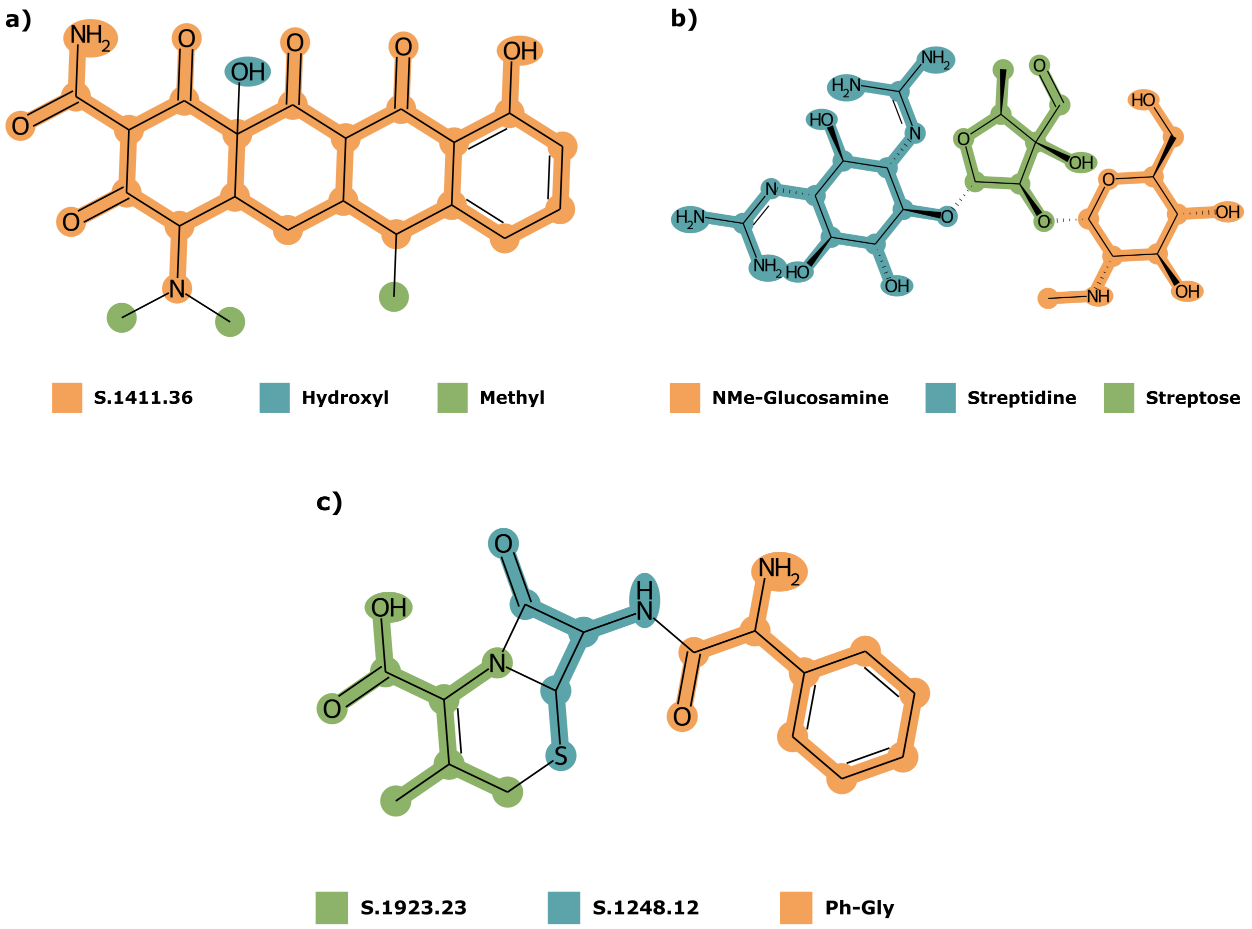


FIGURE 2.3 | **Selected examples of Bear biosynthetic representations.** Shown are: 6-deoxytetracycline (a), Streptomycin (b), Cephalexin (c), Erythromycin (d) and Daptomycin (e).

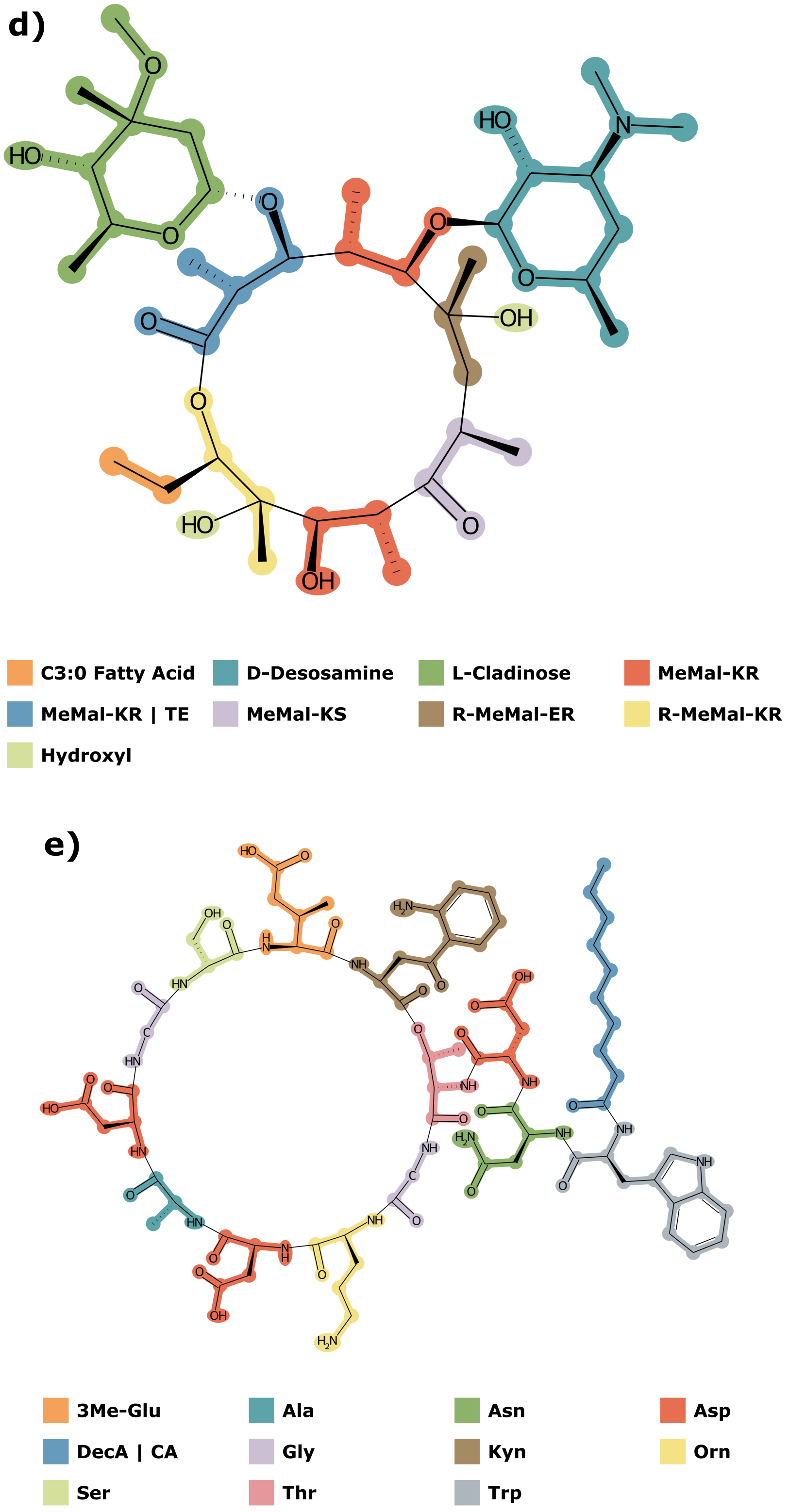


FIGURE 2.3 (continued) | **Selected examples of Bear biosynthetic representations.** Shown are: 6-deoxytetracycline (a), Streptomycin (b), Cephalexin (c), Erythromycin (d) and Daptomycin (e).

Firstly, annotations whose corresponding atoms are complete subsets of other annotations are removed. Secondly, unit combinations which perfectly overlap with stretches of 3 or more connected peptide or polyketide units are removed. Thirdly, flexible unit annotations which overlap with non-flexible unit annotations are removed. Finally, annotations which partially or fully overlap so-called confident annotations are removed. Confident annotations describe certain units whose structure is so specific that it is highly unlikely to belong to another unit. Sugars, nucleosides and certain large miscellaneous units are designated as confident units.

Finding The Optimal Solution

After the successful dereplication of units, the remaining annotations are use to compute the optimal subset of annotations which represent the query structure. The optimal solution is computed based on a series of rules which prioritize certain characteristics of the unit set (Supplementary Table A2). Most importantly, the optimal solution must annotate the greatest number of atoms in the query structure. Even after applying these prioritization rules, there may be multiple solutions to the query structure. In these cases, Bear displays the first solution available, but the remaining solutions are saved and readily-accessible by the user.

Unknown Units

Bear representations which contain unmapped regions hold unique chemical information. It is possible that these unmapped regions correspond to chemical units not yet contained within the scientific community's body of knowledge. Owing to the fact that all Bear units are representable by SMILES strings, it is possible to extract these unmapped regions as their own chemical moieties for future analyses. For example, the totality of unknown chemical units may be compared to known BGCs in order to derive new biosynthetic units and their genetic connections. To this end, Bear reports all unmapped regions as discrete units with their corresponding SMILES.

2.4.3 — Gene Connections

Units

With the successful completion of the Bear tool, the next step was to annotate its units with relevant genomic information. The natural inclination was to use the corresponding genes from a metabolite's source map to annotate the derived units. However, the biosynthetic representations presented by Bear do not always perfectly overlap with a given metabolite's source map (Figure 2.4). This is due to the fact that Bear leverages the entire library of units, and there may be units which span across multiple source regions (Figure 2.4). To account for this, the Bear solution for a given structure can be overlapped with its respective source map, and the biosynthetic units annotated with the genes corresponding to their atoms. For example, the grey biosynthetic unit in Figure 2.4b, *S.1101*, overlaps with *Reaction 4* and *Reaction 6* in the source map of Figure 2.4a. Therefore, its gene connections are registered as *AcmN* (corresponding with Reaction 4), as well as *AcmD, AcmE* and *AcmF* (corresponding with Reaction 6).

Each intermediate and final compound from both primary and specialized nonmodular biosynthetic pathways was mapped by Bear and its biosynthetic representation compared to its source map. The genes corresponding to the atoms of each unit were pooled to create so-called gene hooks — a combination of genes which gives rise to a given chemical unit. Given that it is possible for the same chemical unit to map to multiple compounds from different pathways, it is possible for the same unit to have multiple gene hooks. Therefore, each gene hook for a given unit was recorded.

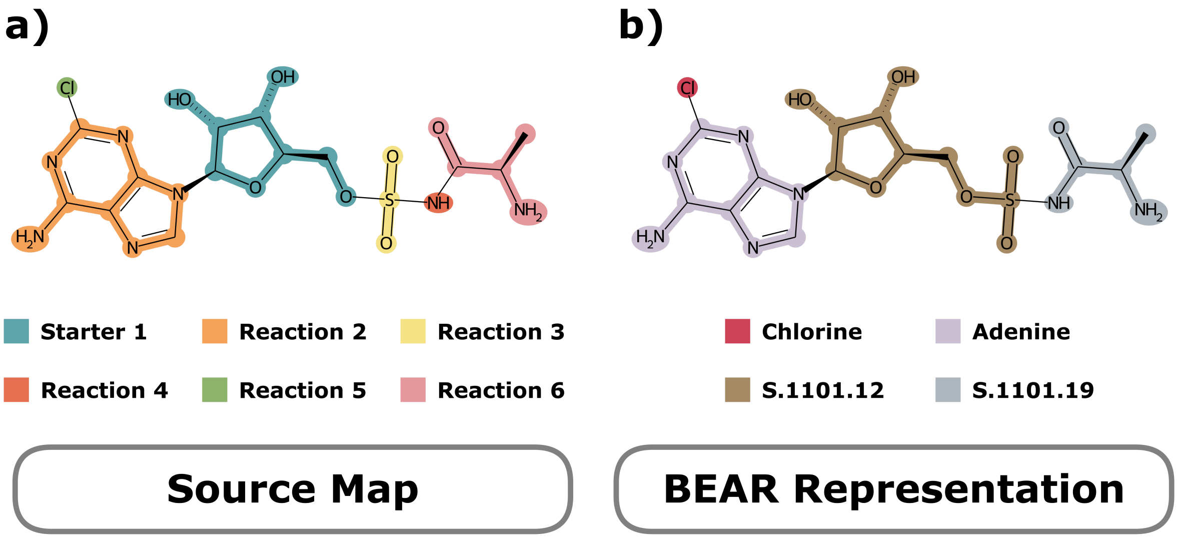


FIGURE 2.4 | **Source map to Bear representation comparison of Ascamycin.** (a) The final compound in the ascamycin biosynthetic pathway is illustrated with atoms highlighted according to its source map. (b) The Bear breakdown of ascamycin is illustrated with atoms highlighted according to biosynthetic units. To account for the discrepancy, the sources of all atoms corresponding to a biosynthetic unit are mapped to that unit. For example, the sources of the grey biosynthetic unit, *S.1101.19,* are both *Reaction 4* and *Reaction 6*.

Linker Reactions

In addition to chemical units, source map-to-Bear breakdown comparisons can also be leveraged to derive information regarding the bonds *between* units. Chemical reactions which join two Bear units can be annotated with the sources of these bonds. These chemical reactions, herein referred to as *linker reactions*, can be represented by a reaction SMILES in which the individual Bear units are the reactants, and the connected dimer is the product.

In order to derive the gene connections for these linker reactions, adjacent units within Bear solutions were analyzed. The sources of the atom from each unit attached to the linker bond were compared. In cases where both atoms shared an additional modification reaction source, this was taken to be the definitive source of the linker reaction. This is because a linker reaction would modify the chemical nature of the two atoms involved in the bond, so the additional modification reaction would register for both atoms. In certain cases, the solution presented by Bear deviated from the source mapping of the query structure. In these situations, no overlapping additional modification reactions existed, and the entire set of source reactions was taken to be the linker reaction's source. Using these linker reactions, a second dataset of gene hooks pertaining to linker reactions was generated.

2.5 Results & Discussion

Assessing Pathway-Derived Units

Firstly, the validity of the pathway-derived unit extraction protocol was assessed with an atom coverage test. The atom coverage test refers to the percentage of atoms in a query structure which are annotated by Bear's biosynthetic breakdown. All intermediates and final compounds from biosynthetic pathways were subjected to an atom coverage test, and it was confirmed that 100% of intermediates and compounds could all be fully annotated using Bear's unit library. This is to be expected, however, as the units in Bear's library are sourced from these very pathways themselves.

Holdout Atom Coverage Test

Next, the robustness of the pathway-derived unit extraction logic was tested using a holdout dataset. Similar to those used in machine learning applications, a holdout dataset uses a portion of the entire data to test the model's performance on the remaining, never-before-seen data[37]. In the case of Bear, the units from 80% of the biosynthetic pathways were used to map the intermediates and compounds from the remaining 20% of the pathways. The units from the 20% holdout dataset were excluded from the library, in order to test the extent to which the units from a portion of the pathways could explain the remaining pathway intermediates and compounds. Figure 2.5 illustrates the results of the holdout atom coverage test per chemotype. Note, NRPS and Type1PK chemotypes did not undergo a train/test split, as their units are not pathway-derived.

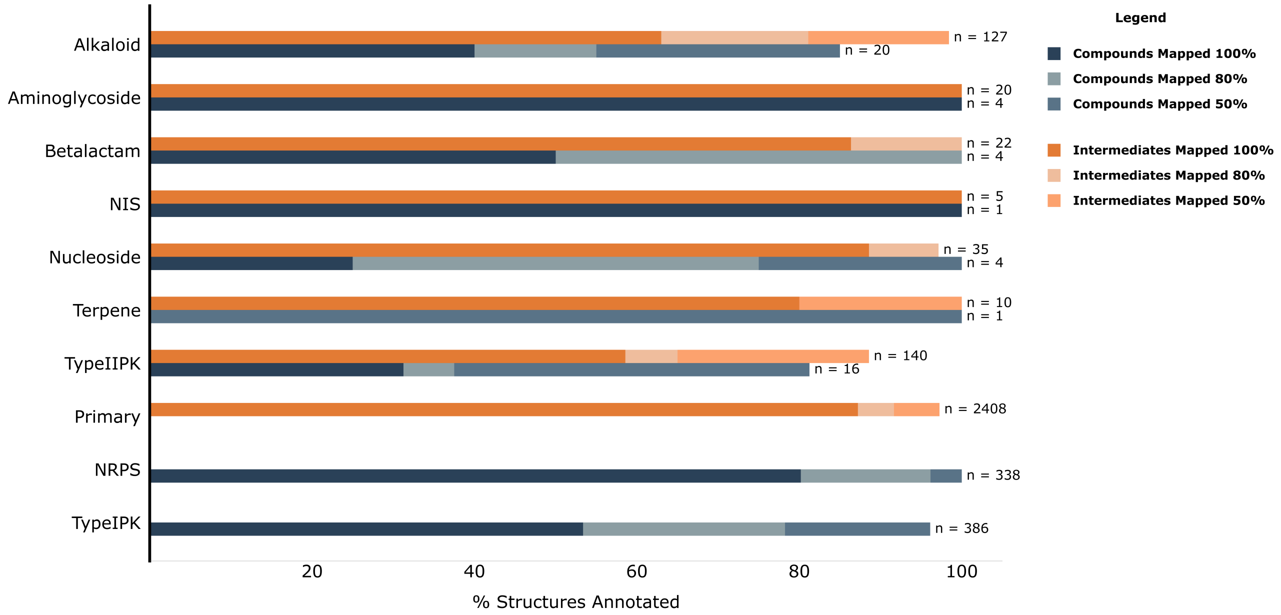


FIGURE 2.5 | **Results of holdout atom coverage test.** The percentage of compounds mapped at 50%, 80% and 100% is shown in blue. The percentage of intermediates mapped at 50%, 80% and 100% is shown in orange. NRPS and Type1PK results correspond only with known compounds from those chemotypes, as their units are not pathway-derived.

The results of the holdout atom coverage test indicate that sensitivity to random removal of units from the Bear library is dependant on chemotype. Type 2 polyketides (Type2PKs), for example, are particularly sensitive to unit removal, given they had the lowest atom coverage results for intermediates and compounds combined. This observation aligns with theoretical expectations as Type2PKs inherently contain specific scaffold backbones[38]. Often these scaffolds cannot be broken down and are represented as one large biosynthetic unit. If a given scaffold is removed from the Bear unit library, its corresponding pathway intermediate and final compound will no longer be fully annotated. Figure 2.6 illustrates this situation using *6-deoxytetracycline*, and the change in atom coverage when units derived from the *6-deoxytetracycline* biosynthetic gene cluster are removed from Bear.

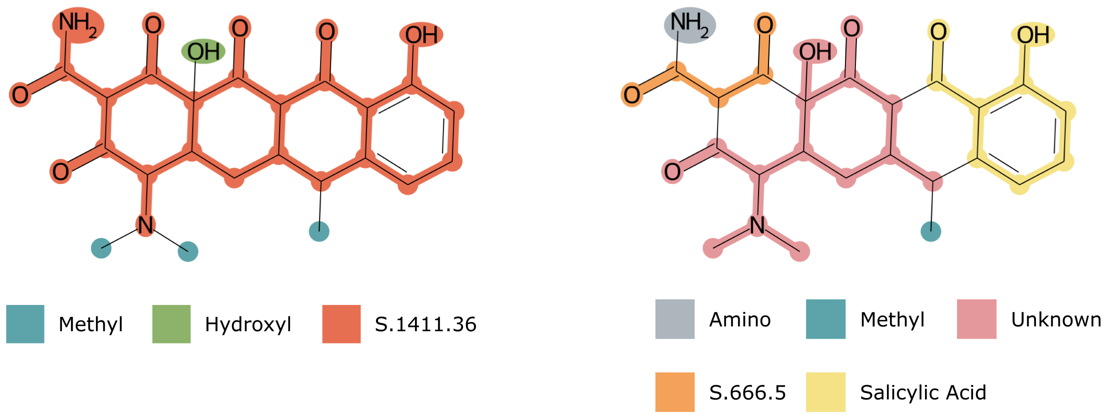


FIGURE 2.6 | **Sensitivity to Bear unit library restrictions in *6-deoxytetracycline*.** Shown on the left is the correct Bear annotation of *6-deoxytetracycline*, using the full Bear unit library. Shown on the right is the incorrect Bear annotation of *6-deoxytetracycline* when units from its biosynthetic pathway were removed from the Bear unit library. This is an example of how restrictions to the Bear unit library can remove important Type2PK scaffolds which cause certain metabolites to go unmapped.

On the other hand, chemotypes for which metabolites are largely comprised of repetitive motifs were more resistant to decreases in atom coverage performance. The aminoglycoside, betalactam and NIS chemotypes were examples of such cases. Interestingly, the nucleoside chemotype, which often includes repetitive chemical motifs, was not as robust as expected. This suggests that future data curation efforts should perhaps prioritize biosynthetic pathways belonging to nucleosides.

It is important to note, however, that the results discussed thus far are with respect to 100% atom coverage. Even where a compound is not fully annotated at 100%, there is valuable information contained in the existing annotations. For example, when querying BGCs for potential matches to a metabolite, an incomplete set of Bear annotations may still be sufficient to identify a match, or at least narrow down the number of potential candidates. With this in mind, Bear was able to annotate most chemotypes nearly perfectly at the 50% threshold. At the 80% threshold level, Bear continued to experience trouble with Type2PKs, but demonstrated a marked improvement among all other chemotypes.

Full Database Atom Coverage Test

Despite promising results with the holdout atom coverage test, we wanted to explore Bear's ability to annotate SMs outside of the known metabolites used in the pathway extraction protocol. To this end, a second atom coverage test was performed using a test set of the SMs contained within our in-house database. The test set was limited to bacterial compounds which had chemotype annotations and were able to be analyzed by Bear in 1 minute or less. A total of 19,485 compounds were analyzed. Figure 2.7 illustrates the results of this analysis across all chemotypes.

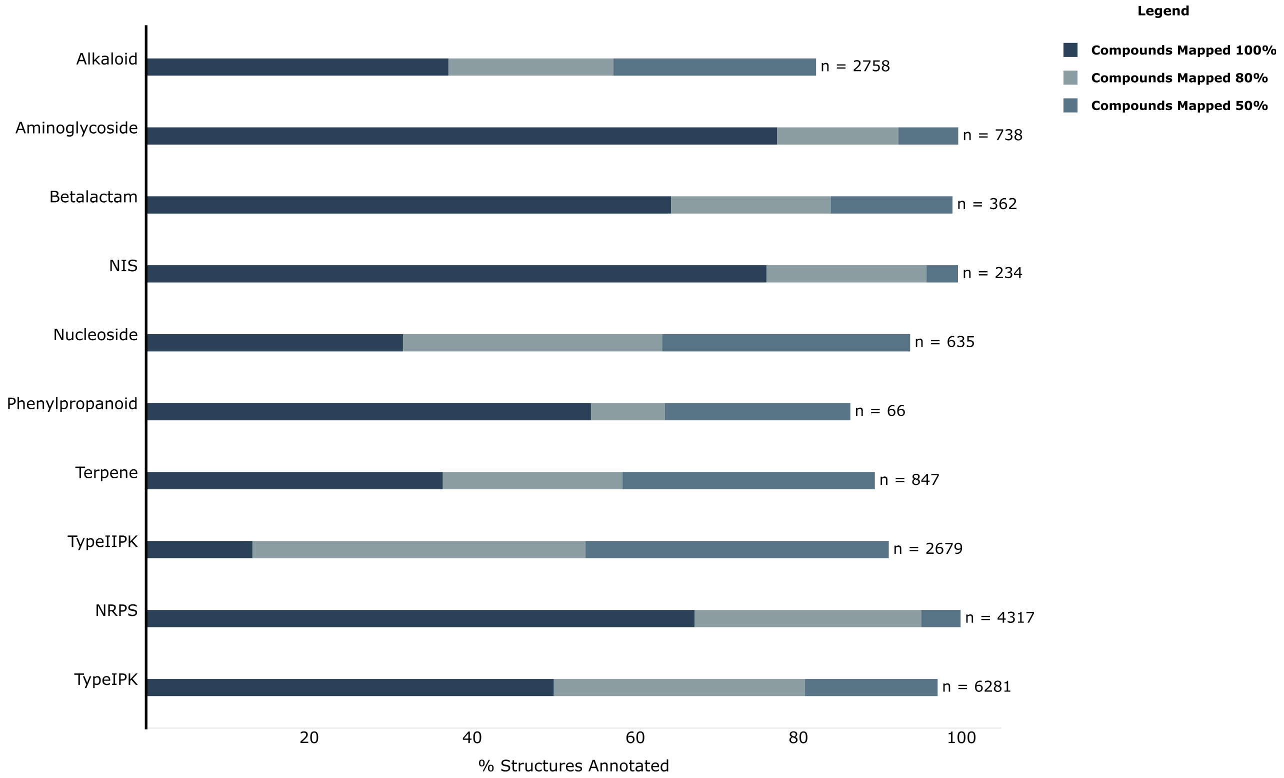


FIGURE 2.7 | **Results of full database atom coverage test.** The percentage of compounds mapped at 50%, 80% and 100% is shown in blue.

A comparison between Figure 2.5 and Figure 2.7 yields several observations. Firstly, the chemotypes which performed more poorly on the holdout test also performed more poorly on the full database test. Type2PKs, terpenes and nucleosides in particular performed worse than other chemotypes. This observation indicates that the biosynthetic pathways which the Bear unit library is based on may not sufficiently represent the totality of existing Type2PKs, terpenes and nucleosides. Conversely, aminoglycoside, NRPS-Independant Siderophores (NIS) and NRPS chemotypes demonstrated relatively strong performance on both tests. This is in accordance with the tendency for these chemotypes to be largely defined by repetitive chemical motifs. Interestingly, the betalactam chemotype demonstrated much better performance on the full database test, indicating that although it may not be resistant to unit library restrictions, the units contained within the full library are sufficient to map the majority of betalactams. Note, the phenylpropanoid chemotype was not included in the holdout test due to insufficient data, so its results on the full database have no comparator.

Validity of Unit Annotations

Full annotation of query structures is only truly purposeful if the units contained within the annotations are genetically valid. In other words, structures may be incorrectly annotated by units which would never come together within a BGC to form the respective structure. In order to test the validity of Bear's annotations, the set of genes corresponding to each Bear breakdown was compared to the expected genes from the BGC. If the units which make up a Bear breakdown of a known metabolite are fully representative of its biosynthesis, their genes should perfectly correspond with those in the cluster. Figure 2.8 illustrates the percentage of genes from a given BGC which are found within the unit-derived genes of the Bear breakdown, on a per-chemotype basis.

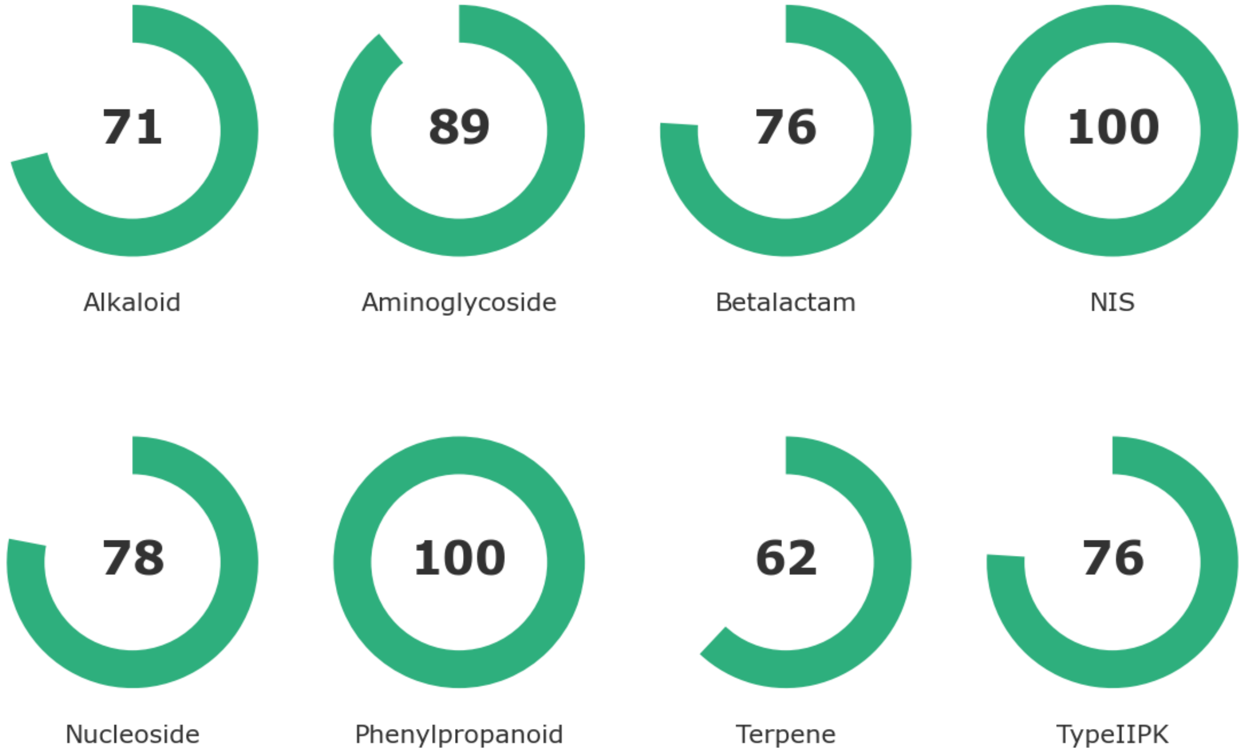


FIGURE 2.8 | **Percentage of genes called by Bear breakdowns.** Shown is the percentage of genes from BGCs which are found in the associated Bear breakdown.

The results in Figure 2.8 highlight several patterns regarding the unit annotations of Bear. Bear breakdowns within certain chemotypes, such as alkaloids and terpenes, are not associated with the totality of genes observed in the BGCs of these compounds. This is likely due to the fact that certain units are derived from intermediates of biosynthetic pathways where they are only connected to several genes from the BGCs. These units are still able to map the corresponding final metabolites perfectly, but the combination of units required to do so does not need to use all of the genes within the

BGC. The downstream effects of this situation can be better understood by predicting the encoded structures contained within BGCs, and searching for signs of decreased performance within these chemotypes.

All other chemotypes made use of significantly more genes from their respective BGCs, on average. Interestingly, the phenylpropanoid and NIS chemotypes used all genes from their BGCs. Results for modular chemotypes (NRPS and Type1PK) are not available, as their corresponding BGCs do not contain equivalent genes but rather specialized modular domains.

Although no direct comparator program exists, internal validation suggests that the Bear unit library is a promising initial work to examine most SMs in light of their biosynthetic composition. Further curation efforts are required to expand the unit library to cover specialized units within modular systems, as well as additional pathways in the chemotypes which are more difficult to fully annotate. Indeed, our group is currently undertaking such curation efforts in order to supplement the biosynthetic pathways leveraged by Bear with new data published in recent months.

2.6 Conclusion

The development of Bear presents an exciting opportunity to explore SMs by leveraging their novel encodings. Work remains to curate additional biosynthetic pathways and modular units, in order to increase Bear’s performance on the atom coverage test. Indeed, this curation effort is already underway in our lab, but remains a long-term, ongoing endeavour. In the meantime, the Bear tool can be extended to a number of potential downstream applications, including identifying new units from the totality of unknown units, connecting metabolites to BGCs, or annotating biosynthetic units with pharmacological activity data.

A use case which is of particular interest to our group is the reverse engineering of a Bear biosynthetic representation. At it stands, Bear is able to effectively map chemical data to genomic data. If this process were reversed, by way of mapping genomic data to chemical data, the chemical and genomic content of metabolites could be bridged into the same latent space. Theoretically, this should be an achievable goal given that the Bear library already contains gene hooks related to both chemical units and linker reactions. Therefore, BGCs could readily be screened by these gene hooks, and the detected units and reactions could be used to generate a predicted structure for the BGC. This predicted structure would be represented by biosynthetic units, thereby being readily comparable to other metabolites. The successful development of such a system would enable further downstream applications such as matching unknown BGCs to metabolites, or connecting genomic data directly to biological activity data.

Chapter 3

Predicting Encoded Chemistries Using *BearClaws*

3.1 Preface

The work presented in this chapter expands upon the collaborative Bear tool developed with the help of other lab members. Mathusan Gunabalasingam wrote most of the code for the BearClaws tool itself. I assisted with brainstorming and conceptual ideas related to the logic of the program. I generated pseudo clusters and ran all analyses described herein. Keshav Dial and Norman Spencer ran the PRISM program to generate its structural predictions. Dr. Nathan Magarvey provided oversight, mentorship and scientific expertise throughout this work.

3.2 Abstract

Conventional methods for characterizing the chemical structures of bacterial metabolites from their biosynthetic gene clusters (BCGs) is resource-intensive and requires expert knowledge. Often, these pursuits may be undertaken only to conclude that a particular metabolite is not of interest. Computer tools have been released which can predict the structures of certain bacterial metabolites from their BGCs, but their robustness across all metabolite classes varies. There is a need for a fast

computational tool which can predict the encoded chemistries of *all* classes of bacterial metabolites. In this work, we present *BearClaws*, an extension to our previously-described tool, Bear, which leverages a library of *gene hooks* — combinations of genes which give rise to a chemical unit — in order to achieve this goal. Most importantly, BearClaws predicts chemical structures as series of biosynthetic units concurrent with those contained within Bear. This unified representation of metabolites readily allows for downstream analyses, which are discussed and suggested at the end of this work.

3.3 Introduction

Traditional efforts to identify novel specialized metabolites (SMs) and their chemical structures have been laborious and time consuming. Advancements in genomic sequencing technologies have enabled researchers to “mine” bacterial genomes for novel metabolite-producing biosynthetic gene clusters (BGCs)[25]. Even so, the subsequent production, isolation and characterization of a particular metabolite is resource intensive, and is only hampered by the increasing rate of rediscovery. In response, ambitious recent attempts have sought to leverage heuristics and computer models to increase the rate of novel SM discovery by predicting the encoded structures of metabolites directly from BGCs.

The *antibiotics & Secondary Metabolite Analysis Shell* (antiSMASH) and *PRediction Informations for Secondary Metabolomes (PRISM)* are two major software programs which annotate BGCs and predict the structure of encoded metabolites[23,26]. Both programs use hidden Markov models (HMMs) to identify genes contained within the BGCs. Using this gene annotation pipeline and a set of heuristics, both programs then predict the chemical structure of the metabolite synthesized by the BGC. While antiSMASH is able to predict the structure of modular SMs (such as peptide and polyketide-based metabolites synthesized from repeating chemical monomer units), PRISM takes this approach one step further by also predicting the structure of nonmodular SMs. Unlike their modular counterparts, these metabolites are synthesized through a series of reactions that build upon one another, thereby removing the modularity of units. Instead, PRISM contains a library of chemical reactions associated with genes, which it leverages in order to actively synthesize nonmodular components of a predicted structure.

Currently, PRISM is the most comprehensive program for predicting SM structures from BGCs. In the context of nonmodular SMs, however, PRISM's reaction-centred approach limits the program's generalizability and robustness. In particular, the program is susceptible to interruptions in its series of iterative reactions if the gene(s) required for one of the reactions is missing or not detected. In this case, the predicted structure would not contain any of the chemical changes associated with reactions after the missing reaction. Moreover, if incorrect reactions are applied, they may overwrite the correctly-predicted chemical structure up to that point. This is in contrast to a modular structure predicted from a series of units, where a portion of the units may be incorrect but the remainder are still accurate. Therefore, a unit-based approach to predicting the chemical structure of both modular and nonmodular SMs would not only bridge the two types of metabolites into the same biosynthetic space, but would also eliminate some of the shortcomings of current reaction-centred approaches.

Here, we present an extension to the previously-described Bear program, *BearClaws,* which leverages the genetically-traced biosynthetic units from the Bear library to predict encoded SM structures from BGCs in a unit-based manner. Each of these biosynthetic chemical units is connected to one or more so-called *gene hooks* — a combination of genes which is reported to give rise to the unit. Moreover, Bear also contains gene hooks for so-called *linker reactions* — chemical reactions which append two or more Bear chemical units together. Using these two sets of gene hooks, known BGCs belonging to both modular and nonmodular SMs were queried and their encoded structures predicted. The predicted structures were compared to both the true structure and those derived from PRISM. Finally, several examples were explored in order to identify strengths and weaknesses of BearClaws.

3.4 Methodology

BGC Annotations

The successful use of BearClaws relies on a BGC annotation pipeline which can identify biosynthetic genes from sequence data. Existing programs such as PRISM and antiSMASH do not annotate the entirety of genes contained within Bear’s 263 hand-curated biosynthetic pathways. AntiSMASH in particular focuses entirely on modular compounds. As a result, there are ongoing in-house efforts to create an updated, more comprehensive BGC annotation pipeline. In the absence of a finalized model at this time, pseudo BGC annotations were generated to test the BearClaws program.

Biosynthetic genes were sourced from the 263 specialized metabolic pathways contained in Bear. All of the genes from each pathway were pooled together and taken to be the respective BGC's set of genomic annotations, herein referred to as the *perfect pseudo clusters*.

In practice, BGC annotation programs are not perfect. Genomic annotations may be superlative, absent or mislabelled. As such, the perfect pseudo clusters may not be representative of the results that can be expected with real genomic annotations from sequence data. In order to account for this, *imperfect pseudo clusters* were generated by removing a random 10% of genes from each BGC. For BGCs where 10% represents less than one gene (those BGCs with <10 total genes), a single random gene was removed.

Unit and Reaction Calling

Gene Hooks

The gene hooks contained within the Bear library were used to call units and linker reactions from the perfect and imperfect pseudo clusters. Each unit gene hook was called if all of its required genes were present in the cluster. Each linker reaction gene hook was called if any one of its required genes were present in the cluster. This liberal approach to calling linker reactions was required given that not all linker reactions had a clearly unanimous pathway reaction source. We can afford a greater rate of false positive linker reaction calling if the units are called very strictly. This is because even if a false positive linker reaction is called, its effect will only be observed in the predicted product if all of its reactants are called as units. The entire set of called units and linker reactions for each perfect and imperfect cluster was then analyzed to synthesize a predicted structure.

Modular Domains

In contrast to the BGCs of nonmodular SMs, the order of detected genes in modular BGCs is very important and directly affects the structure of the encoded product. The open reading frames (ORFs) of modular BGCs are comprised of discrete modular genetic domains, each responsible for incorporating a distinct monomer unit into the growing backbone of the metabolite[39]. The order of these domains within an ORF, and the order of ORFs themselves, dictates the order in which the metabolite's modular backbone is synthesized. To account for this, BearClaws accepts annotations pertaining to these modular domains. These annotations are processed before the gene hook annotations, in order to account for the order of modular unit connectivity in the backbone.

Structure Synthesis

Upon successfully using gene hooks and modular domains to call units, linker reactions and modular backbones from a given cluster, BearClaws synthesizes the best predicted final structure. BearClaws will use all of the available units and reactions to synthesize every combination of predicted structures, such that the final set of units do not overlap in their requisite genes, and neither do the linker reactions. The genes of units are permitted to overlap with the genes of linker reactions, and vice versa. The predicted structures which makes use of the greatest number of genes are captured. In some cases, the predicted structure is composed of multiple fragments which failed to be connected together due to missing or unknown linker reactions. In these cases, the largest fragment is taken to be the predicted final structure used for subsequent analyses.

3.5 Results & Discussion

Accuracy of Nonmodular Predicted Metabolites

The principal metric of BearClaws' performance involves assessing the accuracy of its predicted structures. 95 BGCs belonging to known SMs which had been previously annotated by PRISM were annotated using BearClaws. Both perfect and imperfect pseudo clusters were used to predict structures using BearClaws. The structures predicted by BearClaws and PRISM were compared to the true structure of the SM using a Tanimoto similarity index[40]. The average Tanimoto similarity score based on BEARClaws with a perfect cluster, BearClaws with an imperfect cluster, and PRISM, were plotted per chemotype (Figure 3.1).

The results indicate that with respect to most nonmodular SMs, as represented by pseudo clusters, BearClaws' structural predictions share greater Tanimoto similarity to their true structures than those predicted by PRISM. Alkaloids and TypeIIPKs predicted by BearClaws demonstrated particularly higher scores than PRISM, regardless of whether the type of cluster used by BearClaws was perfect or imperfect. Results for Aminoglycosides favoured BearClaws regardless of cluster type, but to a lesser degree. The average Tanimoto similarity of Betalactams predicted by PRISM was sightly higher than those from BearClaws, although the PRISM-predicted structures demonstrated a significantly larger interquartile range (IQR). Results for Nucleosides were relatively similar regardless of the method used.

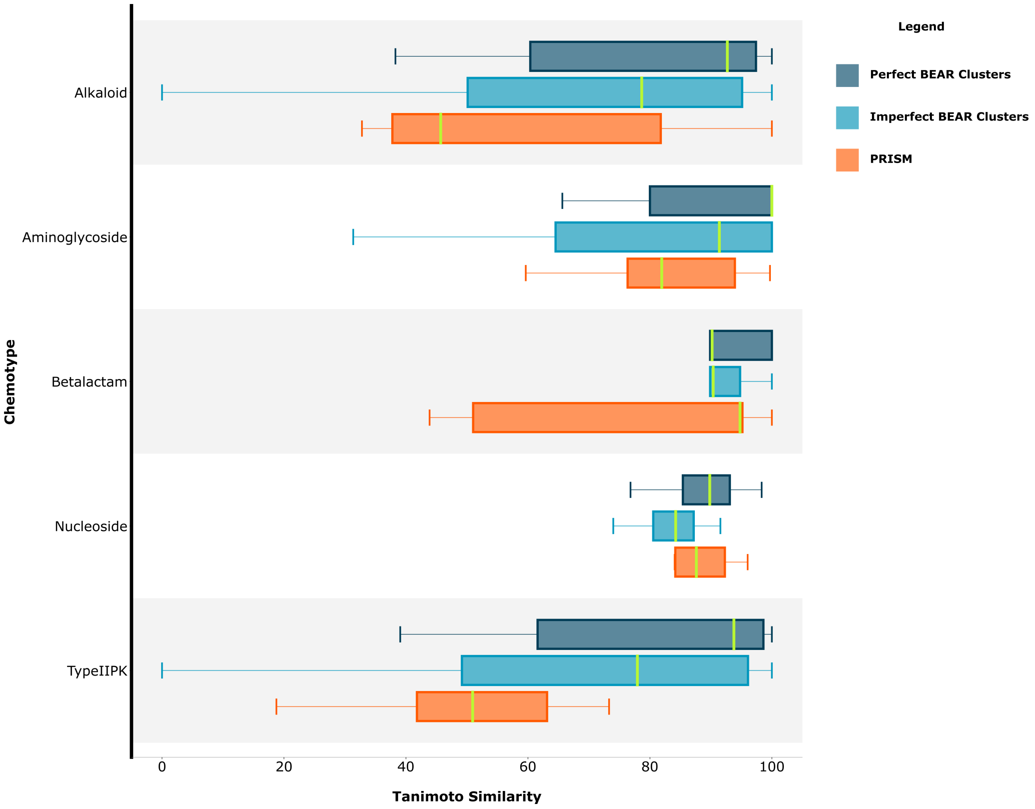


FIGURE 3.1 | **Average Tanimoto similarity of predicted structures compared to true structure.** Depicted are distributions of structures predicted from: BearClaws using perfect pseudo clusters (dark blue), BearClaws using imperfect pseudo clusters (light blue) and PRISM using genomic sequences (orange).

Comparison to PRISM

A more in-depth analysis of the predicted structures from both BearClaws and PRISM lends insight into the observed results. Figure 3.2 compares the predicted structures of a selected Alkaloid, Nucleoside and Type2PK, which correspond to *Welwitindolinone A Isonitrile*, *Blasticidin S* and *Lysolipin I*, respectively[41,42,43] (further examples of BearClaws-predicted structures are included in supplementary figure B1). The BearClaws-predicted structure of the Welwitindoinone alkaloid remained the same regardless of the cluster type used, and outperformed that from PRISM. All methods failed to incorporate a missing ketone group into the final structure. Conversely, PRISM was able to outperform BearClaws in the prediction of the Blasticidin nucleoside. Both BearClaws results missed a methylation, and the imperfect cluster interestingly incorporated an extra valine. With respect to the Lysolipin TypeIIPK, BearClaws was able to outperform PRISM. Although both BearClaws results missed several methylations, PRISM generated an entirely-incorrect backbone.

The comparison of BearClaws and PRISM predicted structures sheds light on the strengths and weaknesses of each approach. More specifically, PRISM is less resistant to missing or incorrect annotations (Welwitindlinone, Lysolipin), but yields quite accurate results when it works (Blasticidin). This may be largely in part due to PRISM's iterative reaction-based synthesis of predicted molecules. In this approach, missing or incorrect reaction affects the product of all downstream reactions, and is directly observed as an inaccuracy in the final molecule.

BEARClaws, on the other hand, calls units first and only uses reactions to append units which have been called. Misannotations more often manifest in discrete units or regions of the final molecule (i.e. the additional valine predicted in Blasticidin). Although these regions may not be correct, the remaining backbone structure often is. PRISM's iterative reaction approach, on the other hand, may manifest more often in misannotated backbone structures (i.e. Welwitindlinone, Lysolipin). Conversely, when annotations are accurate, PRISM results appear quite accurate. The predicted structure of Blasticidin S, for example, is exactly the same as the expected structure. Compared to PRISM, BearClaws sometimes misses annotations even when a perfect gene cluster is used (i.e. missing methylations in Blasticidin and Lysolipin). Overall, both programs have their individual strengths and weaknesses, with BearClaws perhaps being poised to more consistently predict some aspects of the correct structure of nonmodular SMs.

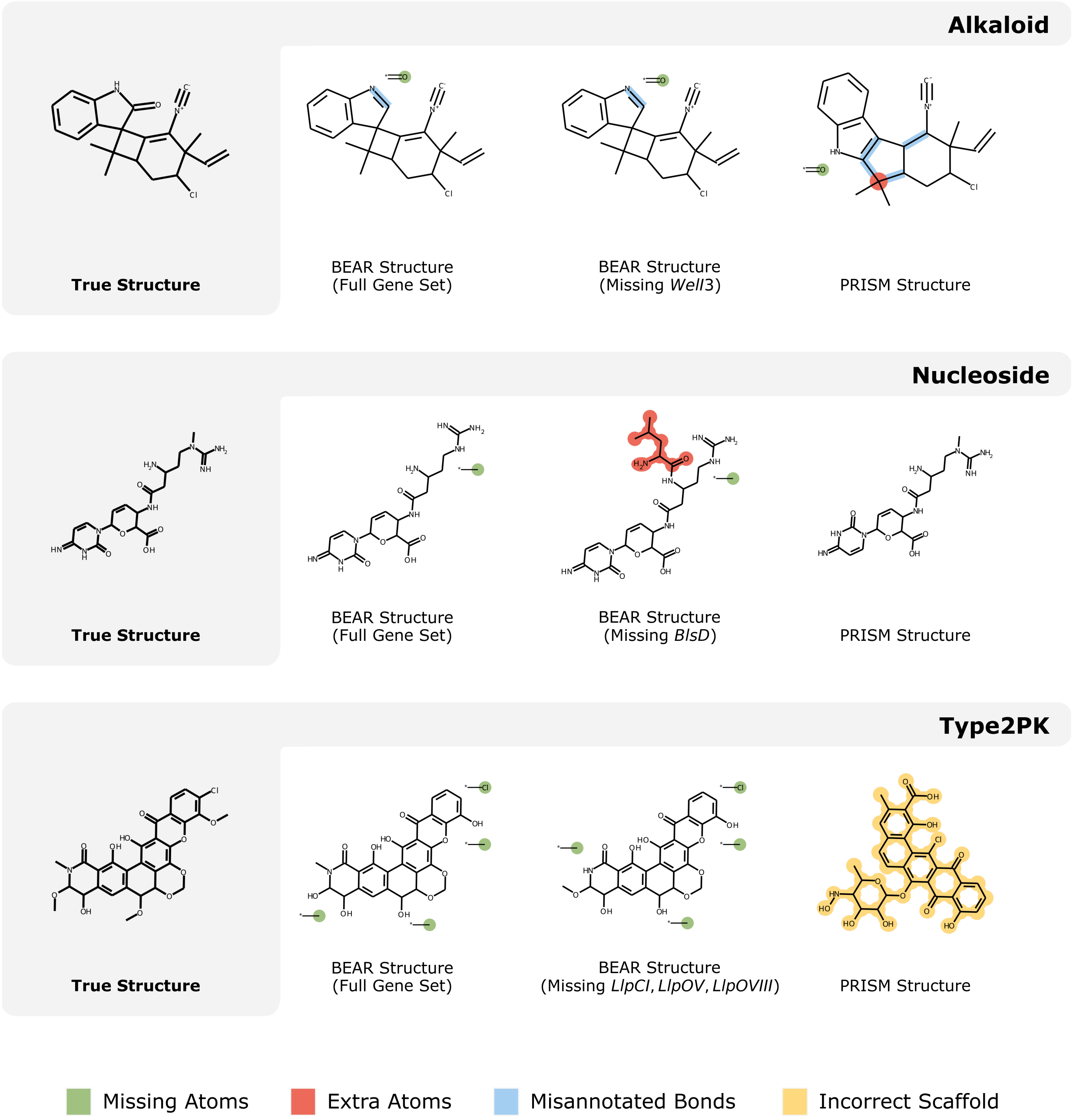


FIGURE 3.2 | **Selected examples of predicted structures from BearClaws and PRISM.** Illustrated are the predicted structures of *Welwitindolinone A Isonitrile* (Alkaloid), *Blasticidin S* (Nucleoside) and *Lysolipin I* (Type2PK).

Perfect vs Imperfect Cluster Results

The effect of imperfect clusters on BearClaws results varied widely. In the case of Welwitindolinone, there was no effect. The use of an imperfect Blasticidin cluster, however, manifested in the superfluous attachment of a valine unit. Interestingly, the use on an imperfect Lysolipin cluster caused a correct methylation to be swapped for another correct methylation that was missing in at another locus.

These effects may be the result of missing or incorrect units and reactions stemming from incomplete gene annotations in the imperfect clusters. As an explorative next step, the units and reactions called from perfect and imperfect clusters were compared (Table 3.1). The results indicate that imperfect clusters were missing a maximum of 1 unit as compared to perfect clusters. However, for most chemotypes this average was closer to 0.5, indicating there were some imperfect clusters which did not miss any units at all. More interestingly, some imperfect clusters led to the addition of extra units as compared to perfect clusters. This explains situations such as the Blasticidin nucleoside described in Figure 3.2, where an extra valine was found in the imperfect cluster case. These additional units may be called due to the configuration of available genes. In other words, the missing gene(s) in the imperfect cluster yield a missing unit, and the gene(s) which would have gone towards calling that unit are used to call the superfluous unit instead. This situation was unique to calling units, however, and did not occur with reactions. Imperfect clusters did miss on average 1-2 reactions as compared to perfect clusters. This may have led to units being called which were not added to the backbone of the predicted structure. This was not, however, responsible for the lack of certain units such as the missing methyls in Lysolipin (Figure 3.2), as these were due to the units themselves not being called. Overall, a comparison of the units and reactions called using perfect and imperfect clusters demonstrates that those from imperfect clusters fall short of those from perfect clusters. However, the difference is not enough to explain some misannotations in BearClaws, namely the additional units added using imperfect clusters and the missing units observed even with perfect clusters. Evidently, further refinement to the BearClaws algorithm is needed to improve results in these specific cases.

	Units			**Reactions**		
Chemotype	*Missing*	*Present*	*Extra*	*Missing*	*Present*	*Extra*
Alkaloid	**1**	**3.1**	**0.6**	**1.3**	**7.2**	**0**
Aminoglycoside	**0.5**	**1.9**	**0.5**	**2.1**	**12.9**	**0**
Betalactam	**0.7**	**4**	**0.7**	**1.5**	**12.3**	**0**
Nucleoside	**0.9**	**4.8**	**0.8**	**1.6**	**10.1**	**0**
Type2PK	**0.6**	**3.9**	**0.5**	**1.4**	**10.2**	**0**

Table 3.1 | **Unit and reaction calling rates between perfect and imperfect clusters.** The average number of missing, present and extra units and reactions called per imperfect cluster compared to perfect cluster, per chemotype, are shown. *Missing* units and reactions refer to those called by BearClaws when using a perfect cluster, but not when using the imperfect counterpart. *Present* units and reactions are called by both types of clusters. *Extra* units and reactions are called by the imperfect cluster, but not the perfect cluster counterpart. Values indicate the average number per cluster.

Accuracy of Modular Metabolite Prediction

Unlike the BGCs of nonmodular specialized metabolites, those of modular SMs are highly dependant on the order of their constituent genes. Given that most modular SMs are cyclic or branched, it is nearly impossible to create pseudo clusters for modular SMs with the correct organization of their genes and corresponding ORFs. In the absence of pseudo clusters for modular SMs, a test case was manually created using the hybrid NRPS-Type1PK SM, *griseoviridin*[44]. This metabolite was chosen as it would allow testing of both NRPS and TypeIPK biosynthesis, as well as that of a hybrid structure where NRPS and TypeIPK genes are interspersed between each other. The published BGC of griseoviridin was manually reviewed and converted to the input required by BearClaws. Figure 3.3a illustrates the structure predicted by BearClaws, as compared to the true structure and that predicted by PRISM.

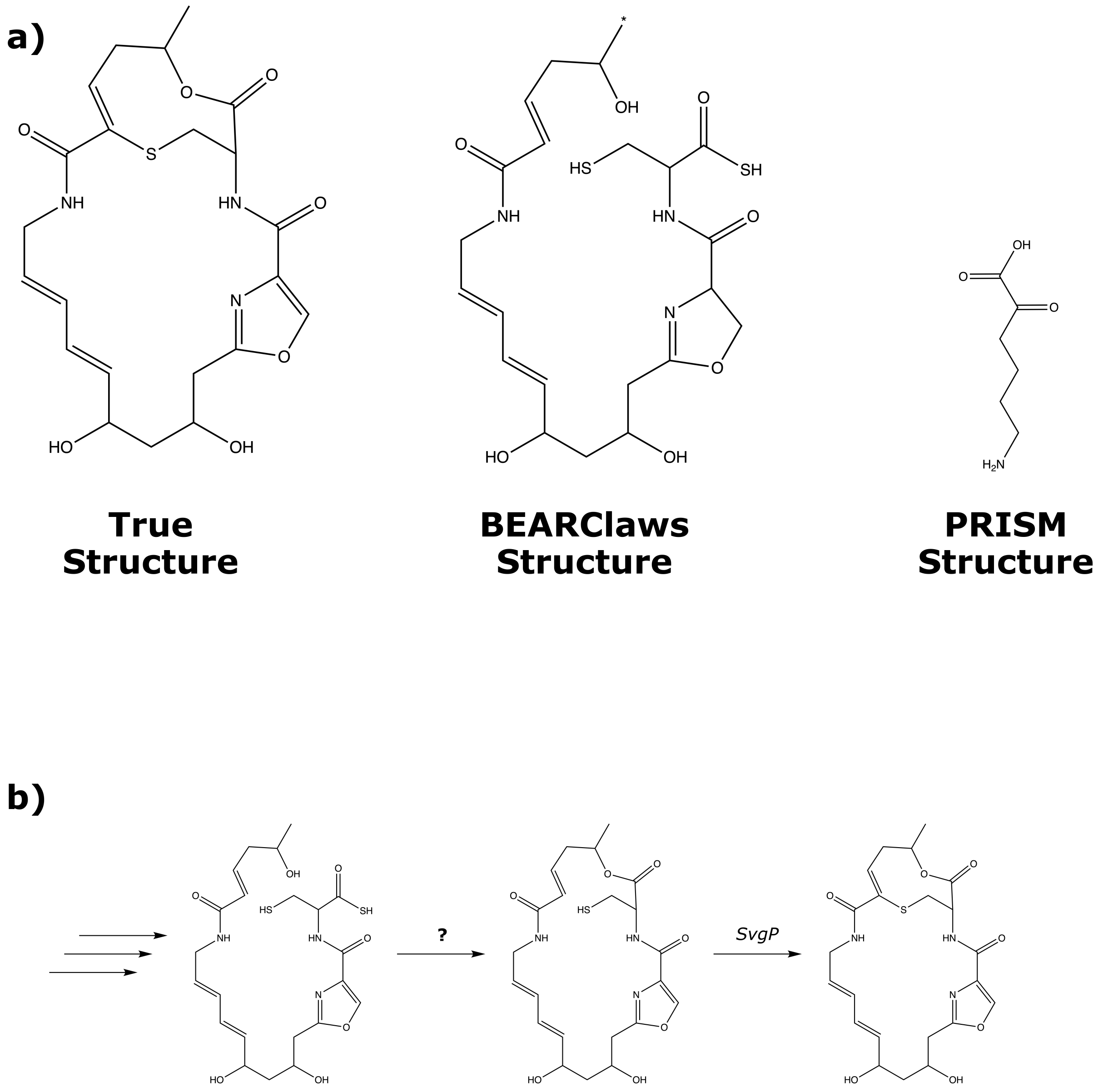


FIGURE 3.3 | **Predicted structure of g*riseoviridin* by BearClaws and PRISM compared to the true structure.** (a) The griseoviridin structures predicted by BearClaws and PRISM are illustrated adjacent to its true structure. (b) The final two reactions in the biosynthetic pathway of griseoviridin are depicted. The gene(s) responsible for the second last reaction is not yet known, therefore its reaction cannot be detected by BearClaws. The subsequent reaction is indeed detected by BearClaws, but is never actually applied to the structure as the preceding intermediate is never formed.

Compared to the true structure of griseoviridin, the one predicted by BearClaws is quite similar overall but differs are three loci. Firstly, the azole ring formation in griseoviridin contains two dehydrated bonds, whereas the predicted structure only contains one. This is due to the fact that azole ring formation in modular SMs does not always follow a clear genetic pattern. In some cases, azoles are formed in their fully hydrated form and a dehydratase enzyme catalyzes the dehydration of a bond[45]. In other cases, such as that of griseoviridin, the dehydrated azole is formed from the start and no dehydratase domains are detected nearby[44]. Although the Bear library comprehensively covers all dehydration states of azoles, their differential detection within BGCs remains a challenge.

The second discrepancy noted in the predicted structure of BearClaws is the lack of macrocyclization. This is due to a lack of reported genes corresponding to this reaction in the literature (Figure 3.3b)[44]. As a result, the reaction itself is never be called from the BGC, and the product never cyclizes. This limitation highlights the dependancy of BearClaws on the body of published literature.

The final thioether formation in griseoviridin (Figure 3.3b) is also absent from the structure predicted by BearClaws. This reaction, however, is linked to a gene (SgvP) and is detected by our program. However, the SMARTS pattern for the reaction necessitates the intermediate from the previous reaction as input. Due to the fact that the previous reaction fails to be called, this intermediate is never formed and the thioether reaction does not occur. This demonstrates a PRISM-like limitation of sequential reaction calling within BearClaws. However, these cases are limited to specialized modular post-modification situations[46]. In other words, the core scaffold of a given modular unit can still be predicted by BearClaws and it is just tailoring modifications at the end of synthesis that may be absent in the program. This limitation is not observed with nonmodular-type SMs in BearClaws, however it is inherent in PRISM.

Although the structure of griseoviridin predicted by PRISM differs significantly from the true structure and that predicted by BearClaws, it is important to note PRISM's use of real genomic data. In the absence of a unified genomic annotation pipeline, BearClaws' literature-derived griseoviridin pseudo cluster offers the program an unfair advantage. When annotating genomic data directly, misannotations are a regular occurrence, particularly with regards to mislabelled modular amino acids. PRISM's structure prediction engine could not be tested separately, as there was no way to pass perfect genomic annotations to just the structural prediction component of the software. Nonetheless, it is reasonable to expect that PRISM would perform better if presented with a perfect pseudo cluster, and BearClaws would perform worse if presented with real genomic sequence annotations. In the future, it will be important to test the extent to which this is true, once our in-house genomic annotation pipeline is completed.

3.6 Conclusion

Here, we presented one of the many potential extensions to the Bear program and library, BearClaws. This program leveraged gene hooks within the Bear library to successfully predict the structures for 95 SMs with considerable accuracy and outperform the state-of-the-art structure prediction program, PRISM. However, the absence of an appropriate genomic annotation pipeline to detect these gene hooks rendered BearClaws dependant on pseudo clusters not fully representative of the expected set of genomic annotations. Consequently, the results of the comparison between BearClaws and PRISM should only be interpreted as support for the proof-of-concept of BearClaws, and not a true competitive analysis against PRISM. This experiment will be completed once more, upon our finalization of a genomic annotation pipeline, in order to establish BearClaws' robustness when working directly from genomic sequence data.

The ability to successfully bridge the chemical and genomic encodings of metabolites into the same space presents boundless opportunities for downstream analyses. For example, the predicted biosynthetic representations of encoded metabolites can be compared to the Bear representations of query metabolites in order to match unknown BGCs to metabolites. This is of particular interest to our group, given the potential to expand the library of known BGC-metabolite connections. These new connections would serve as additional training data for analyses such as the exploration of unknown Bear units, and the expansion of the Bear library. These exercises could serve as a means to improve the Bear suite of software tools, which would in turn yield improved results.

Chapter 4

Activity-Guided Gene Hooks

4.1 Preface

The work presented within this chapter described my own attempt to connect Bear’s biosynthetic units to siderophore activity. Of course, the Bear and BearClaws programs were collaborative tools developed with the help of other lab members, as described in the preface of Chapters 2 and 3, respectively. Within this work, I formatted all data, trained all models and validated the results. Denesh Kumar helped curate the additional siderophore dataset. Mathusan Gunabalasingam kindly offered assistance and mentorship with many of the tools and methodologies used here. Dr. Nathan Magarvey provided oversight, mentorship and scientific expertise throughout this work.

4.2 Abstract

Despite historical success repurposing specialized metabolites (SMs) for pharmaceutical applications, the isolation of clinically-relevant novel SMs remains infrequent and difficult. Newly-identified SMs require extensive activity testing, which even if successful offers no guarantee that the respective metabolite can be isolated in sufficient quantity. There is a need for a more-targeted tool which can suggest potential active metabolites directly from the biosynthetic gene clusters (BGCs) which produce them. In this work, I expand upon our previously-described tools, *Bear* and *BearClaws*, to introduce *activity-guided gene hooks (AGGHs)* — combinations of genes observable within BGCs which may be associated with a particular biological activity. Deep learning models were trained to classify broad activities of SMs, and augment known active metabolites with additional examples predicted by the model. These structures were then subjected to analysis by Bear, and the most frequently observed biosynthetic units for a particular activity were pooled. This protocol was more deeply explored and validated with siderophore activity as a specific test case. Downstream applications and future validation of this work are suggested.

4.3 Introduction

Bacterial specialized metabolites (SMs) have long been sought for their pharmacological and biological activities. In fact, nearly 50% of FDA-approved drugs in recent years trace their roots to SMs[47]. However, despite new SMs being routinely isolated, their clinic use is rarely observed. The lengthy drug approval process in most jurisdictions certainly plays a role in this situation, but there are other contributors[48]. Isolating a novel SM brings with it no guarantee of biological activity. Even in cases where a particular SM does display the desired activity, its large scale use is dependant on sufficient production. With the exception of rare and expensive synthetic methods, SM production remains contingent on inducing an organism's biosynthetic gene cluster (BGC) to produce the corresponding metabolite[49]. It logically follows, then, that an efficient methodology for identifying novel active metabolites would involve inferring activity directly from genomic data.

In previous work, we presented a model, *Bear*, which was able to infer genomic data from chemical structures of metabolites. Using a library of biosynthetic units, Bear mapped metabolites as series of chemical units connected to their genetic origins. We reversed the capabilities of Bear with *BearClaws*, which was able to infer chemical structures directly from genomic data. Using a library of *gene hooks* — combinations of genes which give rise to Bear's biosynthetic units — we were able to predict the chemical structures of 95 known SMs directly from their BGCs, with considerable accuracy. In this work, I build upon both Bear and BearClaws, in order to expand the methodology into the activity space.

I trained a deep learning classifier for each of 4 major biological activities, in order to predict if a metabolite is active directly from its chemical structure. I chose siderophore activity as a specific test case in order to explore the model's robustness and predictive validity. From there, I pooled all manually-labelled and computationally-predicted siderophores and ran Bear on each of their structures to generate their biosynthetic unit representation. The most frequently observed units were deemed *siderophore-related units*, and their associated gene hooks were pooled to generate *activity-guided gene hooks (AGGHs)*— combinations of genes associated with particular activity, in this case siderophore activity. This library of AGGHs is presented, and future steps for validation and exploration of this methodology are suggested.

4.4 Methodology

4.4.1 — Activity Prediction Models

Deep learning models were trained for each of the following activities: *antibacterial, anti fungal, antiviral, siderophore*. Each model was built upon the open-access *Chemberta* model[50]. Models were engineered to take as input the SMILES string corresponding to a chemical structure, and provide a binary output identifying whether the structure is predicted to be active or not.

Data

Training and testing data was sourced from our in-house database of SMs. For each activity, metabolites labelled with the particular activity were grouped into the active class. The inactive class was composed of all other SMs in the database with other activity labels, as well as those without any activity labels which were sufficiently different from all metabolites in the active class (< 0.7 Tanimoto similarity). Using this protocol, SMs with no labelled activity but which shared significant structural similarity to any other active SM were excluded. This was due to the fact that it was not possible to discern whether these SMs were truly inactive, or simply had not been tested yet. Therefore, it was assumed that SMs sharing significant structural similarity to active counterparts were more likely untested rather than truly inactive. Table 4.1 contains the size of active and inactive classes for each activity.

Activity	Size of Active Class	Size of Inactive Class
Antibacterial	**18,670**	**14,999**
Antifungal	**12,952**	**20,734**
Antiviral	**2,727**	**31,653**
Siderophore	**379**	**46,127**

TABLE 4.1 | **Sizes of active and inactive classes for each activity model.** These value for the active siderophore class includes the entries from the augmented dataset.

Using the active and inactive class of each activity label, a train/test/validation dataset split was performed. This split corresponded to 70% of data points being used to train the model, 20% of data points being used to test the model's accuracy during training, and 10% of data points being reserved for validation of the model's performance after training. A stratified data split was performed in order to preserve the relative ratio of active to inactive SMs within each dataset.

Canonical SMILES strings were used for each metabolite entry. For metabolites where isomeric SMILES strings were available, these were added to the respective train/test/validation dataset after the split, in order to avoid cross contamination of a given metabolite structure by way of having its canonical SMILES in one dataset and its isomeric SMILES in another.

Additional Augmentation of Siderophore Data

In order to ensure the most accurate siderophore model possible, our in-house siderophore dataset was augmented with external resources. A catalogue of over 300 siderophore molecules was analyzed[51]. The name of each siderophore was cross-referenced to our database, and any matching entry in our database was annotated with a siderophore activity label if not already present. Any siderophores not present in our database were added to the total dataset. The values in Table 4.1 include these augmented siderophores.

Training Protocol

Activity models were built upon the pre-trained Chemberta model available through HuggingFace[52]. For each activity label, the original Chemberta model was fine-tuned using the dataset specific to that activity. Each model was trained for 20 epochs using a batch size of 64 on two graphics-processing units (GPUs).

Validating Activity Models

The performance of each model was validated using the holdout validation dataset. The overall accuracy of each model was computed and can be found in *4.5 Results & Discussion.*

In an effort to investigate the robustness of our siderophore model, we investigated its performance on a series of situations. The siderophore f*errocin A* was used as a test case[53]. Using its Bear representation, we substituted biosynthetic units from its structure with glycine and retested its siderophore activity. Unit substitution was chosen in favour of unit subtraction so that the model's performance would not be influenced by large fluctuations in the size of the structure. Glycine was chosen as the substitute unit due to its neutral structure and lack of significant functional groups. Based on these results, a subset of units whose absence yielded a loss of siderophore activity in ferrocin A were considered to be the model's interpretation of the molecular pharmacophore — the chemical region responsible for the molecule's activity. This model-inferred pharmacophore was compared to literature-derived chemical moieties frequently associated with siderophore activity.

4.4.2 — Generating Activity-Guided Gene Hooks

Upon exploration and validation of the model's ability to understanding chemical structures associated with siderophore activity, AGGHs were generated. Firstly, all SMs with labelled or model-predicted siderophore activity were pooled together. Next, these SMs were subjected to analysis by Bear. The units composing their biosynthetic representation were pooled, and the most frequently observed units were proposed to be siderophore-related units. Units were only deemed siderophore-related units if they appeared in at least 5 siderophores in our dataset. The gene hooks corresponding to these units were labelled as AGGHs specific to siderophore activity.

4.5 Results & Discussion

Validation of Activity Models

The performance of each deep learning model on classifying the respective validation dataset varied based on the activity label (Table 4.2). Performance of the antibacterial and antifungal models suffered compared to that of the antiviral and siderophore models. This is likely due to the fact antibacterial and antifungal activity can be attributed to a wide variety of chemical backbones and pharmacophores. In turn, the model may experience difficulty understanding and quantizing these different chemical patterns. Conversely, the smaller training dataset size of the antiviral model may indicate that fewer distinct pharmacophores are contained within this dataset. As a result, the model is better poised to comprehensively understand the nature of all of the corresponding chemistries. Siderophores are a particular class of molecules whose activity has been directly attributed to certain chemical moieties, such as that of *N5-acetyl-N5-hydroxyornithine*[54]. The diversity of siderophore structures is consequently less than that of other classes, so the siderophore model is undoubtedly faced with a less challenging task than the other models. Nonetheless, is provides a simple proof-of-concept of this methodology and is more easily explored in greater depth due to the clearly defined pharmacophore regions of siderophores. It is for this reason in particular that the siderophore model was chosen for a more in-depth analysis of its predictive patterns.

Activity	Model Accuracy	Model F1 Score
Antibacterial	**0.59**	**0.59**
Antifungal	**0.64**	**0.64**
Antiviral	**0.79**	**0.82**
Siderophore	**0.99**	**0.99**

TABLE 4.2 | **Accuracy and F1 score of activity models**. Values were calculated using the *sklearn* package in Python.

Ferrocin A was selected as a test case siderophore. Its predicted activity upon the substitution of various Bear units in its structure is illustrated in Figure 4.1. Remarkably, the substitution of an acetyl, hydroxyl and an ornithine (corresponding to N5-acetyl-N5-

hydroxyornithine) at 3 different loci all yielded a loss in activity. This observation is in direct accordance with literature reporting N5-acetyl-N5-hydroxyornithine as a key constituent in many different siderophores[54]. It could have been the case, however, that any modifications to the structure of ferrocin A would have caused the model to register a loss in activity. To test whether this was the case, a serine and a valine were each substituted at separate occasions. Both substitutions retained the molecule's activity, suggesting that the model understands at least some facet regarding the key chemical motifs of siderophores. Interestingly, the substitution of a fatty acid moiety with 3 glycines totalling a similar number of atoms, eliminated the structure's predicted activity. No published work has directly investigated the role of this fatty acid region in the activity of ferrocin A. However, fatty acid side chains have been reported to exhibit significant positive effects in synthetic siderophore mimics comprised of N5-acetyl-N5-hydroxyornithine moieties[55].

Taken together, these observations indicate that our siderophore model indeed understands not only the chemical language of SMILES, but more remarkably the chemical nature of siderophores themselves. It is important to note, however, that these conclusions are limited to the siderophore model specifically, and offer no indications regarding other models. Our analysis of the siderophore model may owe its success to the fact that siderophores themselves are not as chemically diverse as other active metabolites, such as broadly-labelled antibacterials. Nonetheless, this successful proof of concept may inspire similar analyses of other activity models. In order for this endeavour to be successfully undertaken, however, more training data will need to be curated for other activities so that model validation accuracy and F1 score increase.

It is equally important to note that the protocol presented here is not without its biases, which future work should aim to address. Specifically, the use of glycine as a substitute unit is merely one potential method amongst thousands of others. A more comprehensive analysis leveraging many biosynthetic units as substitutes may offer more detailed insight into the model's prediction logic.

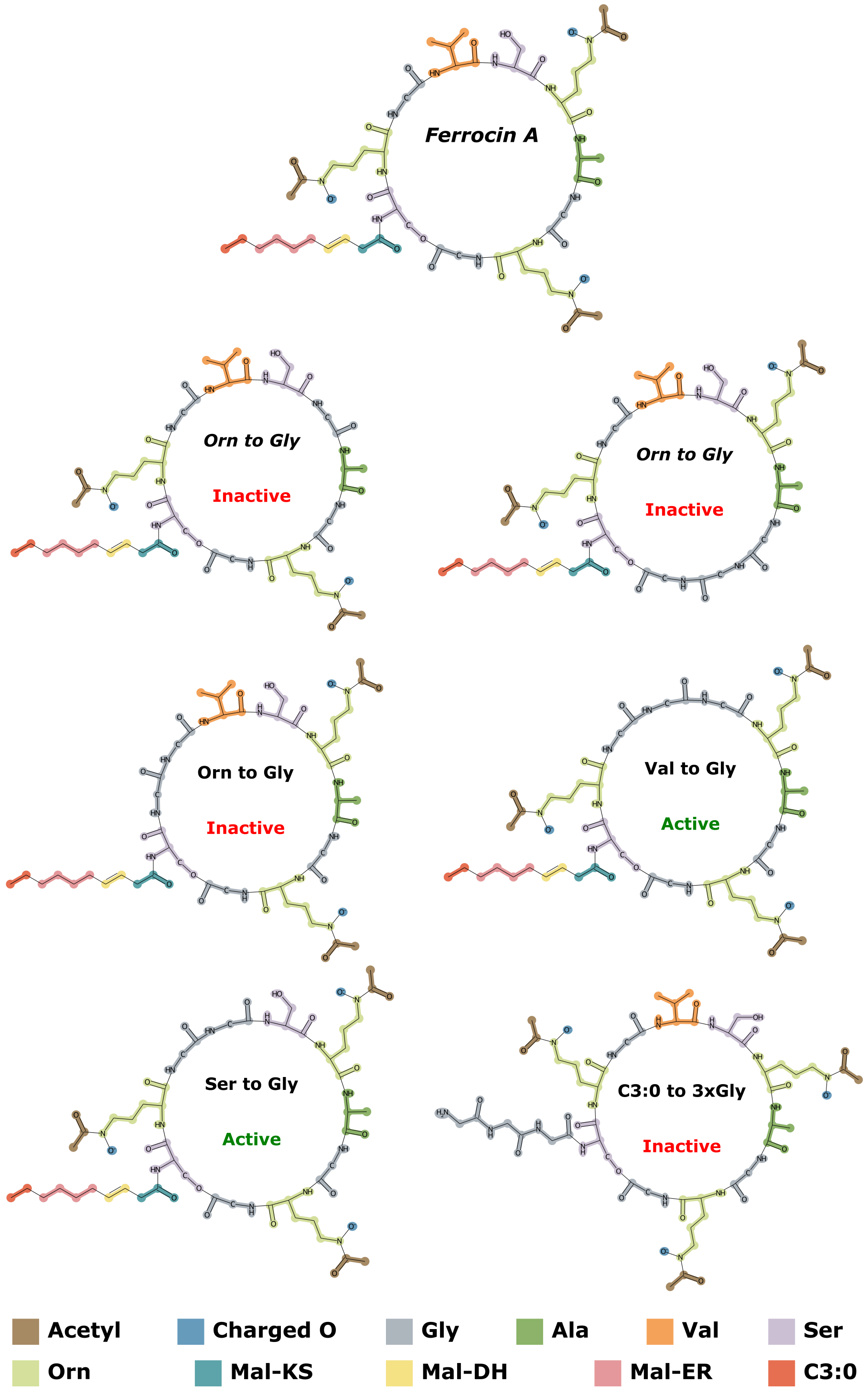


FIGURE 4.1 | **Effect of biosynthetic unit substitution on predicted siderophore activity in ferrocin A**. Note: ***Orn*** refers to N5-acetyl-N5-hydroxyornithine comprised of the *Acetyl, Charged O* and *Orn* biosynthetic units, while ***FA*** refers to the fatty acid moiety comprised of the *Mal-KS, Mal-DH, Mal-ER* and *C3:0* biosynthetic units.

Activity-Guided Gene Hook Library

Bear Biosynthetic Unit	Percentage of Siderophores Containing Unit
Serine	**32%**
Hydroxyl	**24%**
Acetyl	**23%**
OH-Ornithine	**19%**
2,3-Dihydroxybenzoic Acid	**18%**
Threonine	**17%**
Lysine	**15%**
OH-Aspartate	**13%**
Glycine	**12%**
Ornithine	**11%**

TABLE 4.3 | **Top 10 most common Bear units observed in siderophores.**

Bear biosynthetic units used to represent each siderophore were analyzed and the top 10 most commonly observed units within siderophores are presented in Table 4.3. Supplementary Table C1 contains selected AGGHs for these units. Amongst the most frequent units were *Acetyl*, *Hydroxyl* and *Ornithine/OH-Ornithine*, corresponding to the known siderophore-related moiety N5-acetyl-N5-hydroxylornithine. Similarly, the fifth most common unit, *2,3-Dihydroxybenzoic acid*, is also a known siderophore-related moiety[56]. Interestingly, many of the other units contain a hydrogen-donating hydroxyl group (*Serine*, *Threonine*, *OH-Aspartate*). These units may mimic similar effects to N5-acetyl-N5-hydroxyornithine and 2,3-Dihydroxybenzoic acid, or may work cooperatively with these units.

Beyond the readily-observed patterns described above, however, it is difficult to manually identify other, more complex patterns. For example, it may be the case that certain *combinations* of units, rather can single instances themselves, may be better associated with siderophore activity. Likewise, it may be possible that certain combinations of genes not perfectly corresponding to any reported gene hooks may give rise to the necessary chemistries of certain siderophores. As such, future work would benefit from an informatics strategy which can infer its own patterns from the data generated within this work.

4.6 Conclusion

Here, we were able to introduce the concept of AGGHs by leveraging the Bear library of biosynthetic units from our previous work. Further inquiry into siderophore activity specifically indicated that our model was able to capture and understand some of the chemical nuances inherent in these molecules. Literature-base validation of the generated siderophore-related Bear units suggests promising potential for the use of their associated activity-related gene hooks.

Future work, however, should focus on validating these AGGHs in the lab. Experiments may involve using gene hooks to identify new candidate siderophores, and subjecting them to activity assay testing. These results could then be used to validate the AGGHs. Moreover, the BGCs of candidate siderophores could be analyzed by BearClaws to predict their structure. The corresponding SMILES could be passed to the activity model, and its prediction cross-referenced with assay results.

In vitro validation of our AGGHs may inspire subsequent work exploring this methodology across all biological activities. In order for this to be feasible, the performance of the non-siderophore activity models presented here would need to be increased through additional training using an expanded dataset. Further, this methodology could be repeated with more specific activity labels corresponding to specific organism or molecular targets. Indeed, ongoing efforts in our lab aim to curate additional activity-related data in order to develop more targeted activity prediction models.

Chapter 5

Significance & Future Perspective

Bacterial specialized metabolites (SMs) are chemically-unique molecules which have long drawn the interest of researchers and clinicians. Owing to their diverse chemistries, SMs have been pursued since the early-mid 1900s for their potential as potent therapeutics. Early on, the so-called golden age in SM discovery saw hundreds of novel therapeutics be uncovered at will, and in a short period of time. Since then, the rate of re-discovery of these previously-identified metabolites has increased due to a lack of sufficiently targeted discovery approaches. Advancements in next-generation sequencing and liquid chromatograph-mass spectrometry technologies have partially sustained the momentum of SM discovery, but impending antibiotic resistance and emerging diseases dictate the need for something greater. Today, the outdated method of searching for metabolites in an untargeted manner still prevails, despite failing to leverage the abundant data available from previous decades of research.

The work described within this thesis is positioned to be a foundation for future inquiries upon which to pursue more targeted inquiry into specialized metabolites. For my first aim, I developed a universal encoding method which could represent all metabolites in a biosynthetically-informed manner. This novel method, presented within the *Bear* program, leverages a library of chemical units derived from the curation of all known bacterial metabolic pathways. Moreover, these chemical units were connected to their genomic requisites. No other methodology has analyzed the totality of bacterial metabolism to generate a library of biosynthetic units.

This library allowed me to pursue my second aim: the prediction of SM structures directly from the biosynthetic gene clusters (BGCs) which encode them. Towards this aim, a second program, *BearClaws*, was developed. Owing to the novel methodology using biosynthetic units, BearClaws was able to demonstrate comparable results to a state-of-the-art competitor. The two complimentary programs of Bear and BearClaws present the first opportunity to seamlessly integrate chemical and genomic metabolite data into the same latent space.

Finally, this chemical and genomic data was supplemented with biological activity information in my third aim. Using a curated dataset of siderophores, I trained a deep learning model to predict siderophore activity from SMILES. This model was used to generate a finalized set of known and predicted siderophores whose Bear biosynthetic representations were analyzed. Commonly observed units and their associated genes were used to introduce *activity-guided gene hooks* — combinations of genes associated with a particular activity.

Future work should be directed towards further validating the methodologies presented here and building upon these initial proof-of-concepts. Firstly, a long-term strategy should be established for the continual curation of biosynthetic pathway data. This would ensure the most comprehensive Bear library possible and improve the accuracy of biosynthetic representations. Indeed, efforts are already underway within our lab to establish such a protocol. Secondly, the BearClaws tool should be re-tested with genomic annotations from real biological sequence data. It is highly likely that these results will differ from those using pseudo clusters, and the extent to which this is true must be established. To this end, colleagues of mine are currently finalizing a comprehensive genomic annotation pipeline which will readily integrate with BearClaws. Finally, the activity-guided gene hooks introduced in this work should be expanded upon to include both comprehensive broad activities (such as antibacterial, antiviral and anti fungal), as well as more specific molecular and organism target activities. To achieve this, additional data curation and cleaning will be required. Efforts are currently underway in our lab to generate a large-scale activity dataset which can be leveraged by the activity-guided gene hook methodology.

Taken together, the data generated using the tools presented here can easily accumulate to levels of significant complexity. The intimate coupling of activity data to genomic and chemical data from the Bear suite of programs adds additional layers of information and complexity to the situation. With this in mind, it may no longer be feasible to analyze these information streams manually or using simple statistical metrics. Instead, deep learning models may be leveraged in the future to identify patterns within this data not previously-observed. For example, new connections between metabolites and BGCs may be identified, or more nuanced combinations of genes may be attributed to certain activities. Graph network models seem particularly well-suited for this task, given their propensity to integrate many layers of data and identify only key patterns within them. The work presented in this thesis provides a foundation upon which these subsequent developments may lie, and offers hope that modern developments may continue to uproot SM discovery in this unprecedented landscape.

List of References

1. Rook G, Backhed F, Levin BR, McFall-Nagi MJ, McLean AR. Evolution, human-microbe interactions, and life history plasticity. *Lancet* 2017;**390**:521-530. doi:10.1016/S0140-6736(17)30566-4
2. Gary MW. Mitochondrial Evolution. *Cold Spring Harb Prospect Biol* 2012;**4**. doi:10.1101/cshperspect.a011403
3. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* 2016;**14**:8. doi:10.1371/journal.pbio.1002533
4. Nelson ML, Dinardo A, Hochberg J, Armelagos GJ. Brief communication: Mass spectroscopic characterization of tetracycline in the skeletal remains of an ancient population from Sudanese Nubia 350-550 CE. *Am J Phys Anthropol* 2010;**143**:151-54. doi:10.1002/ajpa.21340
5. Nelson ML & Levy SB. The history of the tetracyclines. *Ann N Y Acad Sci* 2011;**1241**:17-32. doi:10.1111/j.1749-6632.2011.06354.x
6. Fisberg M & Machado R. History of yogurt and current patterns of consumption. *Nutrition Reviews* 2015;**73**:4-7. doi:10.1093/nutrit/nuv020
7. Katz L & Baltz RH. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology and Biotechnology* 2016;**43**:155-176. doi:10.1007/s10295-015-1723-5
8. Demain AL. Microbial production of primary metabolites. *Naturwissenschaften* 1980;**67**:582-587. doi:10.1007/BF00396537
9. Hellinga HW, Evans PR. Nucleotide sequence and high-level expression of the major Escherichia coli phosphofructokinase. *Eur J Biochem* 1982;**149**:363-373. doi:10.1111/j.1432-1033.1985.tb08934.x
10. Borodina I, Siebring J, Zhang J, Smith CP, van Keulen G, Dijkuizen L, et al. Antibiotic overproduction in Streptomyces coelicolor A3 2 mediated by phosphofructokinase deletion. *J Biol Chem* 2008;**283**:25186-25199. doi: 10.1074/jbc.M803105200
11. Altschul AF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;**215**:403-410. doi:10.1016/S0022-2836(05)80360-2
12. Chang A, Jeske L, Ulbrich S, Hoffman J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 2021;**49**:D498-D508. doi:10.1093/nar/gkaa1025
13. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;**49**:D545-D551. doi:10.1093/nar/gkaa970
14. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection

of Pathway/Genome Databases. *Nucleic Acids Res* 2014;**42**:D459-D471. doi:10.1093/nar/gkt1103

15. Starke C, Wegner A. MetAMDB: Metabolic Atom Mapping Database. *Metabolites* 2022;**12**:122. doi:10.3390/metabo12020122
16. Davies J. Specialized microbial metabolites: functions and origins. *The Journal of Antibiotics* 2013;**66**:361-364. doi:10.1038/ja.2013.61
17. Atanasov AG, Zotchev SB, Dirsch VM, the International Natural Product Sciences Taskforce, Supuran CT. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* 2021;**20**:200-216. doi:10.1038/s41573-020-00114-z
18. Baltz RH. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J Ind Microbiol Biotechnol* 2019;**46**:281-299. doi:10.1007/s10295-018-2115-4
19. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS. Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol* 2008;**11**:198-204. doi:10.1016/j.mib.2008.05.006
20. Jensen PR. Natural Products and the Gene Cluster Revolution. *Trends Microbiol* 2016;**24**:968-977. doi:10.1016/j.tim.2016.07.006
21. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ., et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* 2020;**48**:D454-D458. doi:10.1093/nar/gkz882
22. Hyatt D, Chen G, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119. doi:10.1186/1471-2105-11-119
23. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Weezel VP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 2021;**49**:W29-W35. doi:10.1093/nar/gkab335
24. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *PNAS* 2014;**111**:9259-9264. doi:10.1073/pnas.1401734111
25. Johnston CW, Skinnider MA, Wyatt MA, Li X, Ranieri MRM, Yang L. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nature Communications* 2015;**6**:8421. doi:10.1038/ncomms9421
26. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature Communications* 2020;**11**:6058. doi:10.1038/s41467-020-19986-1
27. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019:**47**:e110. doi:10.1093/nar/gkz654

28. Sepulveda M, Pirozzolo I, Alegre M. Impact of the Microbiota on Solid Organ Transplant Rejection. *Curr Opin Organ Transplant* 2019;24:679-686. doi:10.1097/MOT.0000000000000702

29. Bartman C, Chong AS, Alegre M. The influence of the microbiota on the immune response to transplantation. *Curr Opin Organ Transplan*t 2015;**20**:1-7. doi:10.1097/MOT.0000000000000150

30. Quiros M, Grazulis S, Girdzijauskaitė S, Merkys A, Vaitkus A. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J Cheminform* 2018;**10**:23. doi:10.1186/s13321-018-0279-6

31. Martin E, Cao E. Euclidean chemical spaces from molecular fingerprints: Hamming distance and Hempel's ravens. *J Comput Aided Mol Des* 2015;**29**:387-395. doi:10.1007/s10822-014-9819-y

32. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res* 2008;**36**:D326-D331. doi:10.1093/nar/gkm792

33. Rahman SA, Torrance G, Baldacci L, Cuesta SM, Fenninger F, Gopal N, *et al.* Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* 2016;**32**:2065-2066. doi:10.1093/bioinformatics/btw096

34. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference* 2008:11-15.

35. Landrum G. RDKit: Open-source cheminformatics. 2022. doi:10.5281/zenodo.591637

36. Lee AC. SMARTS Approach to Chemical Data Mining and Physicochemical Property Prediction. PhD [dissertation]. Michigan: University of Michigan; 2009. Available from: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/64627/adamclee_1.pdf?sequence=1

37. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J Anal Test* 2018;**2**:249-262. doi:10.1007/s41664-018-0068-2

38. Zhang Z, Pan H, Tang G. New insights into bacterial type II polyketide biosynthesis. *F100Res* 2017;**6**:172. doi:10.12688/f1000research.10466.1

39. Behsaz B, Bode E, Gurevich A, Shi Y, Grundmann F, Acharya D, *et al.* Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. *Nature Communications* 2021;**12**:3225. doi:10.1038/s41467-021-23502-4

40. Racz A, Bajusz D, Heberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J Cheminform* 2018;**10**:48. doi:10.1186/s13321-018-0302-y

41. Hillwig ML, Fuhrman HA, Ittiamornkul K, Sevco TJ, Kwak DH, Liu X. Identification and Characterization of A Welwitindolinone Alkaloid Biosynthetic Gene Cluster in Stigonematalean Cyanobacterium Hapalosiphon welwitschii. *Chembiochem* 2015;**15**:665-669. doi:10.1002/cbic.201300794
42. Cone MC, Yin X, Grochowski LL, Parker MR, Zabriskie TM. The blasticidin S biosynthesis gene cluster from Streptomyces griseochromogenes: sequence analysis, organization, and initial characterization. *Chembiochem* 2003;4:821-828. doi:10.1002/cbic.200300583
43. Lopez P, Hornung A, Welzel K, Unsin C, Wohlleben W, Weber T, *et al.* Isolation of the lysolipin gene cluster of Streptomyces tendae Tü 4042. *Gene* 2010;**461**:5-14. doi:10.1016/j.gene.2010.03.016
44. Xie Y, Wang B, Liu J, Zhou J, Ma J, Huang H, *et al.* Identification of the Biosynthetic Gene Cluster and Regulatory Cascade for the Synergistic Antibacterial Antibiotics Griseoviridin and Viridogrisein in Streptomyces griseoviridis. *Chembiochem* 2012;**13**:2745-2757. doi:10.1002/cbic.201200584
45. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, *et al.* Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. *ACS Cent Sci* 2016;**2**:99-108. doi:10.1021/acscentsci.5b00331
46. Walsh CT, Chen H, Keating TA, Hubbard BK, Losey HC, Luo L, *et al.* Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Curr Opin Chem Biol* 2001;**5**:525-534. doi:10.1016/s1367-5931(00)00235-0
47. Newman DJ & Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod* 2020;**83:**770-803. doi:10.1021/acs.jnatprod.9b01285
48. Patridge E, Gareiss P, Kinch MS, Hoyer D. An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discov Today* 2016;**21**:204-207. doi: 10.1016/j.drudis.2015.01.009
49. Park D, Swayambhu G, Lyga T, Pfeifer BA. Complex natural product production methods and options. *Synthetic and Systems Biotechnology* 2021;6:1-11. doi:10.1016/j.synbio.2020.12.001
50. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* [cs.LG] 2020. doi:arXiv:2010.09885
51. Hider RC, Kong X. Chemistry and biology of siderophores. *Nat Prod Rep* 2010;**27**:637-657. doi:10.1039/b906679a
52. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C., Moi A., *et al*. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* [cs.CL] 2020. doi:10.48550/arXiv.1910.03771

53. Katayama N, Nozaki Y, Okonogi K, Harada S, Ono H. Ferrocins, new iron-containing peptide antibiotics produced by bacteria. Taxonomy, fermentation and biological activity. *J Antibiot* 1993;**46**:65-70. doi:10.7164/antibiotics.46.65

54. Dolence EK, Lin CE, Miller MJ, Payne SM. Synthesis and siderophore activity of albomycin-like peptides derived from N5-acetyl-N5-hydroxy-L-ornithine. *J Med Chem* 1991;**34**:956-968. doi:10.1021/jm00107a013

55. Lin YM, Miller MJ, Mollmann U. The remarkable hydrophobic effect of a fatty acid side chain on the microbial growth promoting activity of a synthetic siderophore. *Biometals* 2001;**14**:153-157. doi:10.1023/a:1016666414848

56. Lopez-Goni I, Moriyon I, Neilands JB. Identification of 2,3-dihydroxybenzoic acid as a Brucella abortus siderophore. *Infection and Immunity* 1992;**60**:11. doi:10.1128/iai.60.11.4496-4503.1992

Appendix A

Chapter 2 Supplementary Figures

Substrate	SMILES
Ethylmalonyl-CoA	SC(C(C(O)=O)CC)=O
R2-Ethylmalonyl-CoA	SC(C(C(O)=O)CC[*])=O
Hydroxymalonyl-CoA	SC(C(C(O)=O)O)=O
Isobuteryl-CoA	SC(C(C(O)=O)(C)C)=O
Ketomalonic-CoA	SC(C(C(O)=O)=O)=O
Isobutylmalonic-CoA	CC(C)CC(C(S)=O)C(O)=O
Malonyl-CoA	SC(CC(O)=O)=O
Methoxymalonyl-CoA	SC(C(OC)C(O)=O)=O
R-Methylmalonyl-CoA	SC(C(C(O)=O)([*])C)=O
Methylmalonyl-CoA	SC(C(C)C(O)=O)=O
R2-Methylmalonyl-CoA	SC(C(C[*])C(O)=O)=O
Epoxmalonyl-CoA	SC(C1(C(O)=O)CO1)=O
R-Malonyl-CoA	SC(C(C(O)=O)[*])=O
2R-Malanyl-CoA	SC(C([*])([*])C(O)=O)=O
Hydroxy-Methylmalonyl-CoA	SC(C(C(O)=O)CO)=O
Hydroxy-Methoxymalonyl-CoA	SC(C(CO)(O)C(O)=O)=O

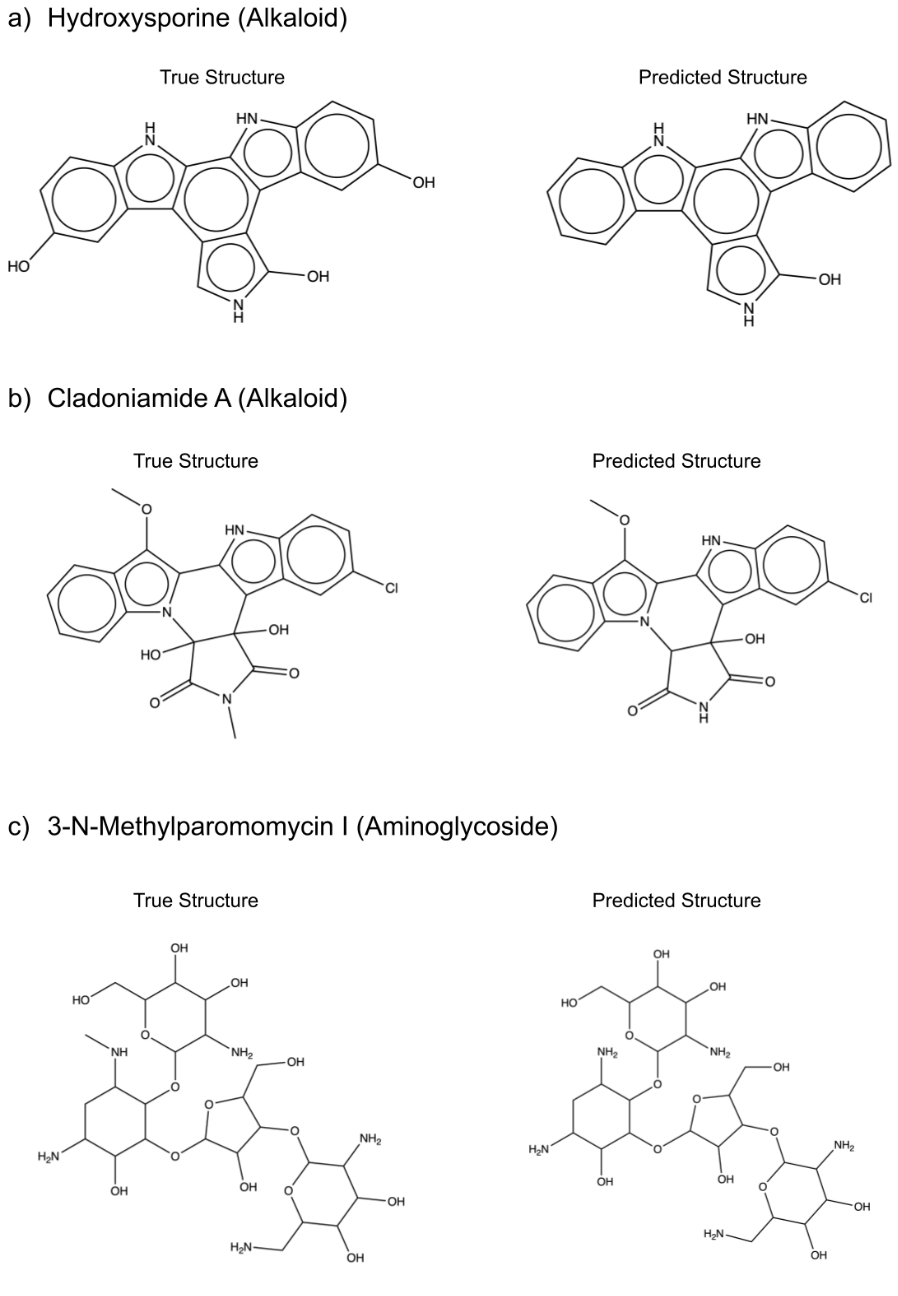
FIGURE A1 | **Polyketide substrates for Bear unit library.** Depicted are SMILES of the original substrate in the form it exists in before incorporation in a natural product.

Rule	Order of Priority
Greatest number of atoms annotated by units	**1**
Fewest number of flexible units used in annotation	**2**
Greatest number of sugars used in annotation	**3**
Fewest number of polyketide unit R-groups	**4**
Fewest number of specialized polyketide units	**5**
Prioritize if any polyketide starter units are used	**6**
Prioritize if any polyketide terminal units are used	**7**
Greatest number of amino acids used in annotation	**8**
Fewest number of units used in annotation	**9**

FIGURE A2 | **Prioritization rules for finding the optimal solution in Bear.** Lower numbers indicate a more significant rule that is prioritized first.

Appendix B

Chapter 3 Supplementary Figures

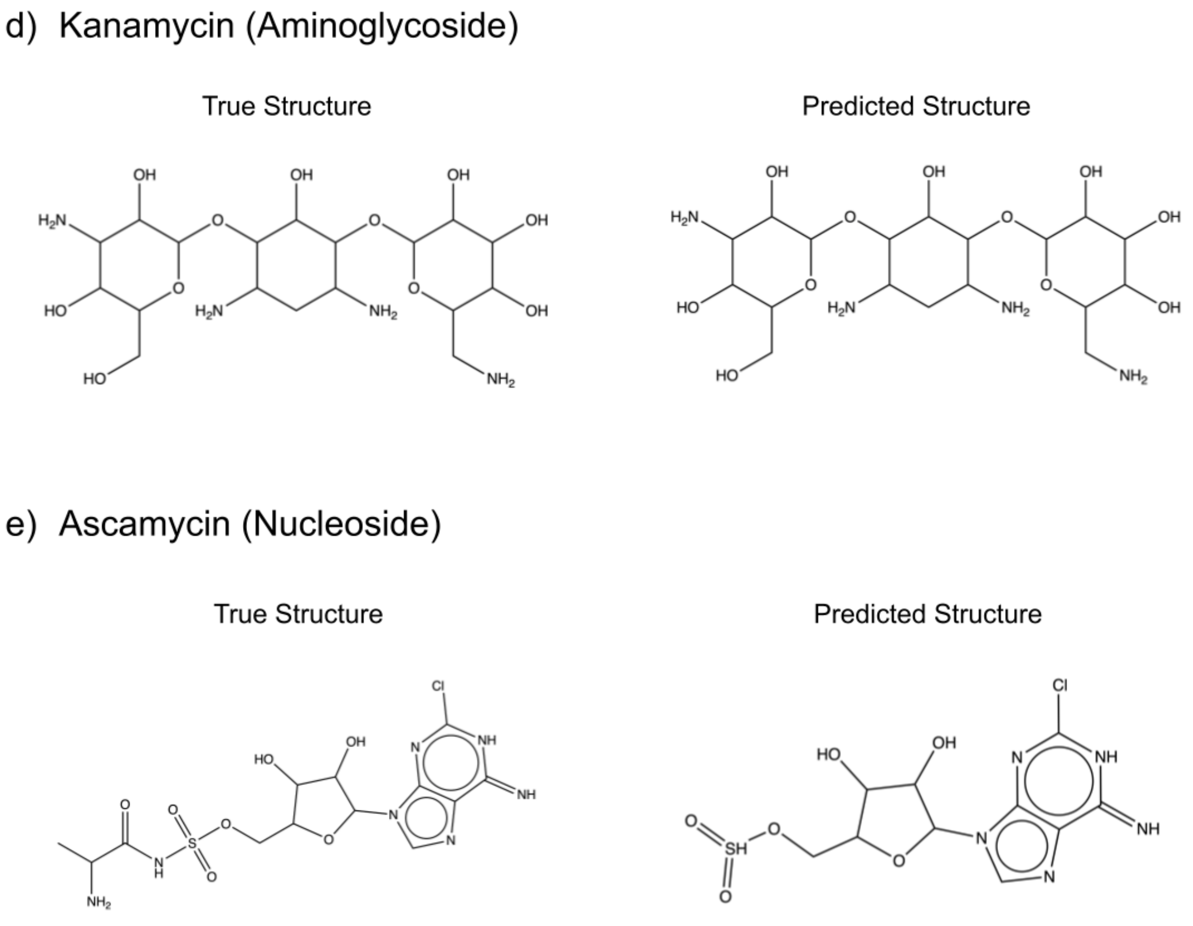


a) Hydroxysporine (Alkaloid)

b) Cladoniamide A (Alkaloid)

c) 3-N-Methylparomomycin I (Aminoglycoside)

FIGURE B1 | **Additional examples of BearClaws predicted structures.**

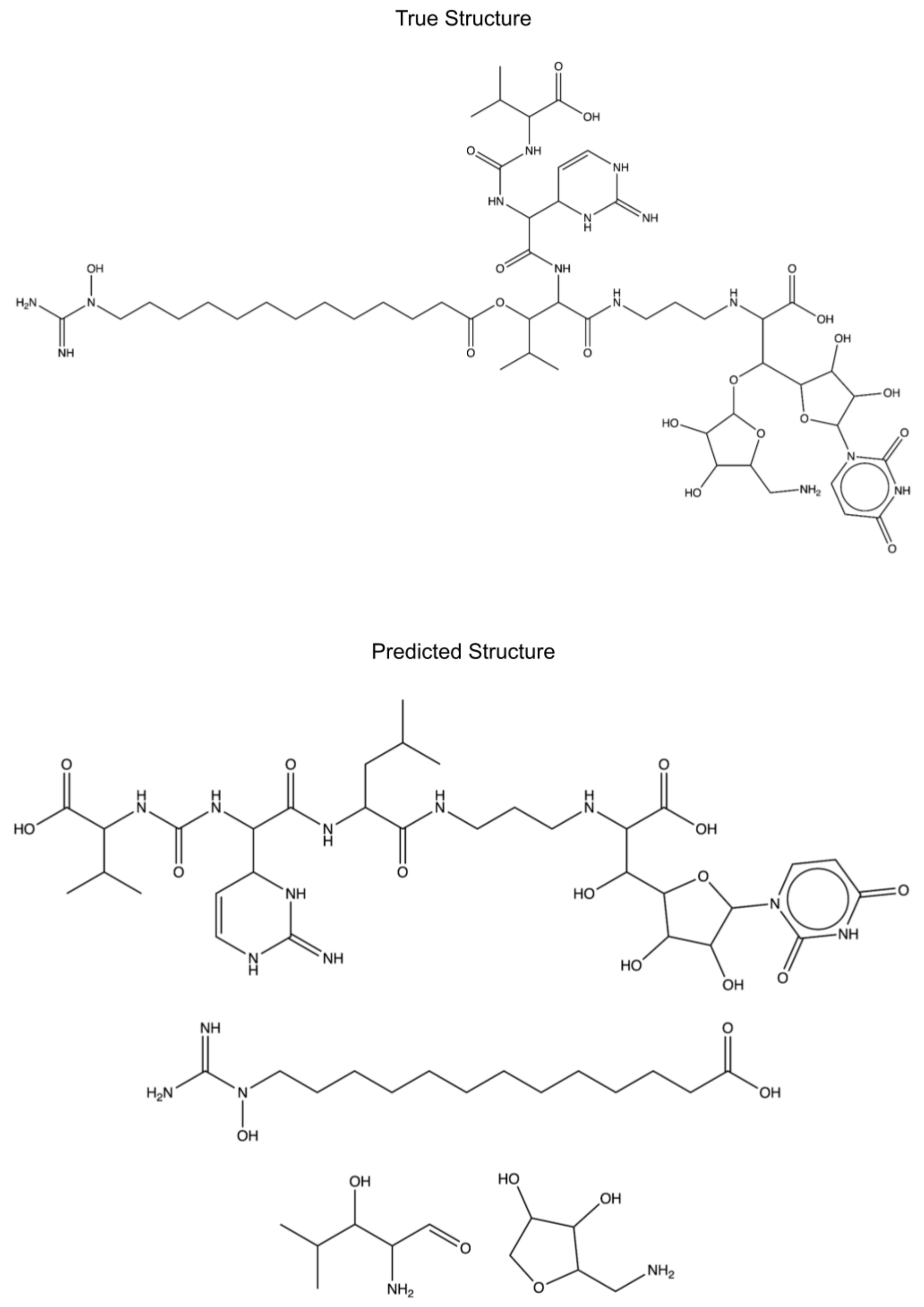


d) Kanamycin (Aminoglycoside)

e) Ascamycin (Nucleoside)

FIGURE B1 (continued) | **Additional examples of BearClaws predicted structures.**

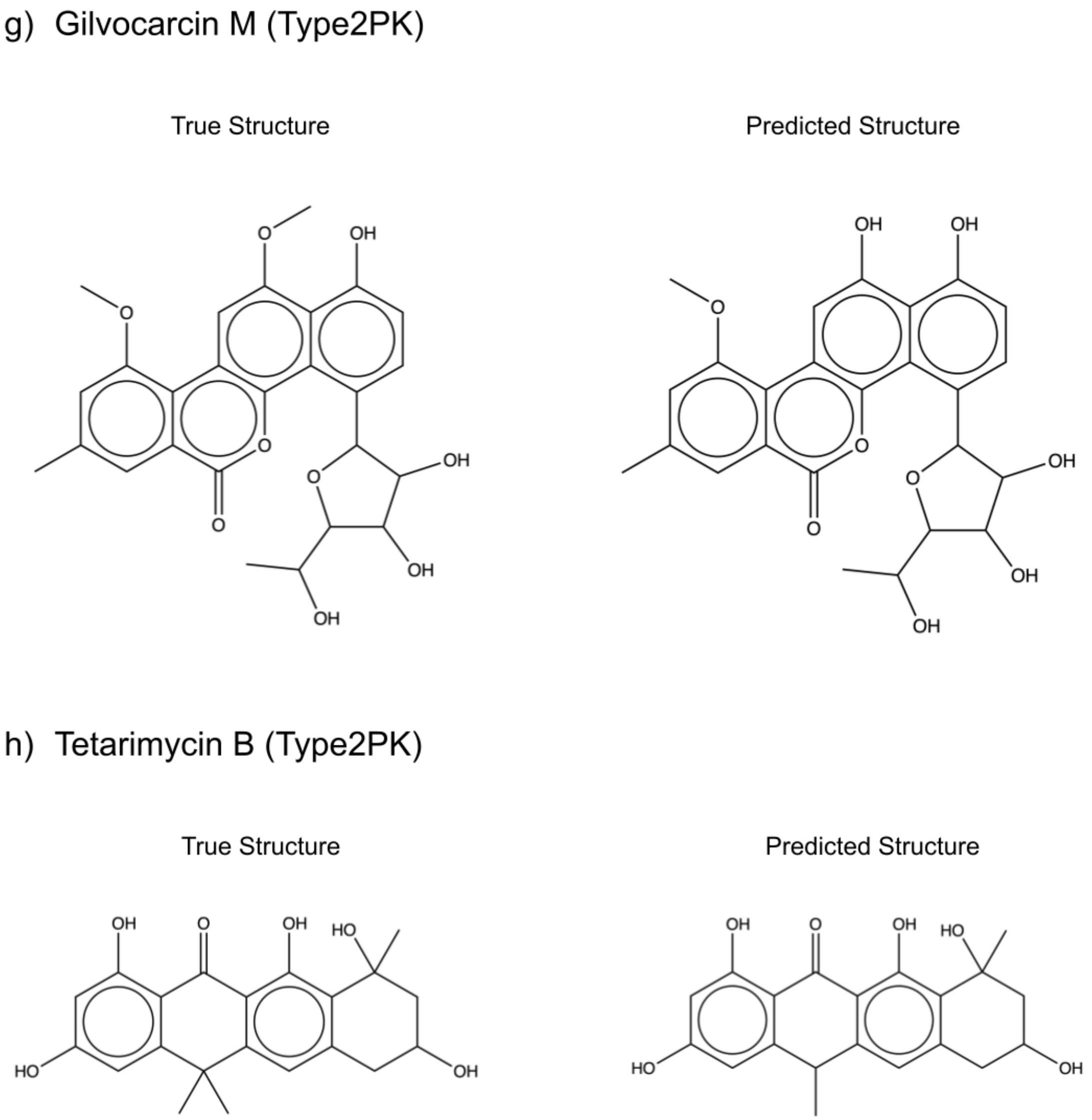
f) Muraymycin (Nucleoside)



True Structure

Predicted Structure

FIGURE B1 (continued) | **Additional examples of BearClaws predicted structures.**



g) Gilvocarcin M (Type2PK)

h) Tetarimycin B (Type2PK)

FIGURE B1 (continued) | **Additional examples of BearClaws predicted structures.**

Appendix C

Chapter 4 Supplementary Figures

Bear Unit	***Activity-Guided Gene Hook***
Serine	*NocA, NocB*
Hydroxyl	*PenG*
Acetyl	*Pur6*
OH-Ornithine	*PenG, A(Orn), T, C*
2,3-Dihydroxybenzoic Acid	*ObaD, ObaE*
Threonine	*SimA4*
Lysine	*A(Lys), T, C*
OH-Aspartate	*PenG, A(Asp), T, C*
Glycine	*TcpD, TcpP*
Ornithine	*A(Orn), T, C*

FIGURE C1 | **Selected activity-guided gene hooks for siderophore activity.** Due to the number of AGGHs generated, not all could be included. One example is illustrated for each of the top ten Bear units most frequently associated wth siderophore activity.