RANKING MICROBIAL METABOLOMIC AND GENOMIC LINKS USING COMPLEMENTARY SCORING FUNCTIONS

GRÍMUR HJÖRLEIFSSON ELDJÁRN

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

OCTOBER 20, 2021

Abstract

The rise of antimicrobial resistance has been identified as one of the major emerging healthcare problems of the 21st century. Microorganisms represent a great potential source of novel chemistry needed to combat the rise of this resistance. Identifying the correspondence between genomic regions and metabolites represents a vital step in the discovery and utilisation of such chemistry. One way in which this correspondence manifests is in the form of links between genomic regions and tandem mass spectra in metabolomic data.

This thesis presents an improvement on an established method of statistically correlating microbial genomic information with the results of metabolomic experiments, based on the correlation of strains in a population of microorganisms. We demonstrate how this method of correlation as currently used is adversely affected by the sizes of the objects under consideration and demonstrate how we can reduce the effect of object size by standardisation.

We also demonstrate how this approach can penalise genomic regions shared by horizontal gene transfer, giving preference to regions that are shared by inheritance. By taking into account the evolutionary relationship between the microorganisms, we present a way of reducing this bias.

Furthermore, we propose a new method of linking genomic and metabolomic data for individual microorganisms to metabolites using kernel functions to quantify their similarities. We demonstrate how this method can be applied to both genomic and metabolomic data individually to establish links to metabolites, and how two such models can be combined to establish links between genomic regions and mass spectra without the intermediary of a metabolite.

Finally, we demonstrate the complementarity of the two approaches. We also present a principled methods of combining the scores to make use of this complementarity to capture information only present in either one of the scores.

To evaluate the proposed models we compile two data sets. One is a database of paired spectra and genomic regions from established databases, combined using structural annotations from both databases. The other is a collection of microbial experiments, each with

genomic and metabolomic data for a collection of strains as well as a number of validated links between genomic regions and tandem mass spectra.

By improving existing methods, proposing new prediction models, and establishing ground truth data sets, the results presented in this thesis can help to further explore the metabolomic potential of microorganisms, both by facilitating the identification of links between genomic and metabolomic data for microorganisms and by providing the research community with much-needed benchmark data.

University of Glasgow

College of Science and Engineering

Statement of Originality to Accompany Thesis Submission

Name: Grímur Hjörleifsson Eldjárn
Registration Number:

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature:

Date: 22. 7. 2021

Author contributions

The principle behind the phylogenetic correction for the strain correlation scoring presented in Section 4.5 comes from Marnix Medema.

The phylogenetic tree for the *Salinispora* strains used to test the phylogenetic correction for the strain correlation score in Section 4.5 was constructed by Mohammad Alanjary.

The combined IOKR model described in section 5.12 is joint work with Juho Rousu. In particular the original derivation of Eq. 5.28 is Juho's work, although any typos in the derivation presented in this thesis are my own.

Code and data availability

The code for the experiments presented in this thesis is available at
`https://github.com/grimur/phd_code`

The MIBiG-GNPS data set described in Section 3.6 can be accessed at
`https://doi.org/10.1371/journal.pcbi.1008920.s006`

List of publications

The following publications serve as the basis for parts of the thesis.

Sections 3.6, 3.7, 4.3, 4.4, 5.10, 6.3, 6.4, 6.5.2:

> Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLOS Computational Biology. 202105;17(5):1-24.

Sections 2.3.1, 2.3.2, 4.1, 5.2:

> Soldatou S, Hjörleifsson Eldjárn G, Huerta-Uribe A, Rogers S, Duncan KR. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. FEMSMicrobiol Lett. 2019 Jul;366(13).

Sections 4.1 and 4.4:

> Soldatou S, Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Hughes AH, Rogers S, et al. Comparative Metabologenomics Analysis of Polar Actinomycetes. Marine Drugs.2021;19(2).

Acknowledgements

Thanks first and foremost to Simon Rogers for mentoring, cooperation, patience and hand-holding over the past four years.

Thanks also to my progress review committee, John Williamson and Ke Yuan, for useful critique. Particular thanks to Ke for stepping in and taking over formal administrative duties in time of need.

Sincere thanks to Katherine Duncan, Sylvia Soldatou and Justin van der Hooft for mentoring, support and cooperation, as well as to Joe Wandy, Rónán Daly and Andrew Ramsay for cooperation and support.

Thanks to Juho Rousu for cooperation on the IOKR scoring and discussions on kernel methods, and to Marnix Medema and Mohammad Alanjary for cooperation on the phylogenetic correction for the strain correlation scoring.

Along with Simon, Vinny Davies, Alexandra Pancheva, Fran Young, Katherine Duncan, Juho Rousu, Marnix Medema, Guðrún G Björnsdóttir and Kristján Eldjárn Hjörleifsson read drafts of various parts of the thesis in various stages of completion. I'm enormously grateful for their patience and help in clarifying my thinking and presentation, and for keeping my notation in check. Any remaining typos and abuses of notation are my sole responsibility.

I would also like to thank my fellow members of the IDI group as well as my office mates at various points: Francesco, Anders, Iulia, Natasha, Antoine, Adal, Josh, Val, Marco and Jasper.

I am grateful to the School of Computing Science for awarding me an Excellence in Computing Science bursary.

Last but not least I would like to sincerely thank my family, close and extended, and friends, for patience and encouragement over the last few years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

- AMR: antimicrobial resistance
- AUC: area under curve
- BGC: biosynthetic gene cluster
- GCF: gene cluster family
- HGT: horizontal gene transfer
- IOKR: input-output kernel regression
- KCB: known cluster BLAST
- MF: molecular family
- MKA: multiple kernel alignment
- MS: mass spectrometry
- MS2: tandem mass spectrometry
- Pfam: protein family
- PPK: product probability kernel

Chapter 1

Introduction

Arguably one of the most important events in the history of modern medicine is the discovery and adoption of antibiotics [1]. Many of the infectious diseases that had up to that point been the leading causes of morbidity and mortality were suddenly no longer as much of a threat, and the ability to manage bacterial infections paved the way for enormous advances in surgical medicine [2].

The dawn of the so-called Age of Antibiotics is typically considered to be Alexander Fleming's fortuitous discovery of penicillin in 1928, its mass-production starting in 1942, and the 1943 discovery of streptomycin [1]. However, the documented use of antibiotics stretches back further — arguably the first known methodical approach to drug discovery was Paul Ehrlich's 1904 large-scale systematic screening of compounds to discover a drug against syphilis [1], and the purposeful use of antibiotics in ancient Egypt, Greece, and China has been well documented [3] with evidence of the use of antimicrobials in Sudanese Nubia dating as far back as 350 CE [4, 5]. The decade between 1950 and 1960 has been termed the golden age of antibiotic discovery [6] and saw the discovery of over half of the antibiotics in common use today [6], but since then, the rate of discovery has slowed down significantly with most new antibiotics being modified versions of existing ones [1]. In fact, since the 1970s only three new classes of antibiotics have reached market [7]. A brief overview of the history of antibiotics can be found in [8].

While the theoretical potential for *antimicrobial resistance* (AMR) in microorganisms was acknowledged early on it was not considered a serious problem, as the actual development of such resistance was considered unlikely [6]. The intervening decades have shown conclusively that this optimism was sadly misplaced [9]. Already by the 1950s penicillin resistance was becoming a problem, and the emergence of methicillin-resistant *Staphylococcus aureus* (MRSA) in the mid-1960s and of vancomycin resistance in strains of coagulase-negative staphylococci in the early 1980s demonstrated that in general, microorganisms seemed to rapidly develop resistance to new antibiotics [9].

The ease with which microorganisms seem to develop resistance to antibiotics, coupled with the lack of new antibiotics, has lead the World Health Organisation (WHO) to label AMR as one of the top 10 global public health threats facing humanity [10, 11], with 10 million annual deaths projected by 2050 due to AMR [12]. The emergence of multi-drug or pan-drug resistant bacteria threatens to undermine many of the most significant medical advances of the last century, not only with regards to infectious diseases but also major surgeries, organ transplants and chemotherapy [13], as well as already putting a considerable economic strain on the healthcare systems of the world [9]. The need for new antimicrobial drugs is therefore great.

Microorganisms achieve resistance to antimicrobials in various ways: by limiting the uptake of the drug, by modifying or protecting the target, or by active efflux, i.e. actively pushing the molecules out of the organism [14, 15]. This resistance can be acquired in several ways: by overexpression or duplication of existing genes, by point mutations, insertions or deletions, or by acquiring genetic material from other microbes in the environment in a process known as *horizontal gene transfer* (HGT) [15].

Antimicrobial drugs can be broadly split into three categories: natural, semi-synthetic (based on natural molecules) and synthetic [16, 17]. Of these, the class of natural drugs is by far the largest, with 70% of commercially available antibiotics from microbial sources [7]. Although novel approaches are being developed for drug lead research [18], the so-called *microbial specialised metabolites* remains a rich source of antimicrobial drugs, both in pre-clinical and clinical stages [19, 20].

Microbial specialised metabolites (also known as *secondary metabolites*) are the compounds that a microorganism produces that are not strictly necessary for the survival of the organism but may confer an evolutionary advantage [21]. These include various molecules that the microorganism uses to affect its environment, including pigments, signalling molecules, and compounds with bactericidal activity [22, 23]. Specialised metabolites have for a long time been extremely important both for medical and industrial applications [24, 25]. For example, the majority of drugs are so-called small-molecule drugs (with a molecular weight less than 900 Dalton) and the majority of small-molecule drugs are either specialised metabolites (e.g. the anticancer drug paclitaxel), derived from specialised metabolites (e.g. the antibacterial drug tigecyclin), or inspired by specialised metabolites (e.g. the antiparasitic drug chloroquine) [26, 27]. A large number of these are of microbial origin, and in general, microbial specialised metabolism has been a great source of useful novel chemistry [25, 27].

In the last few years, research into microbial specialised metabolism has experienced a resurgence [23, 28]. Interest in the field declined sharply in the 90s, brought about by the difficulties of doing untargeted metabolomics, such as high rediscovery rates and low throughput of available methodologies [23]. Recently, however, advances in genomics, and in particular

advances in the prediction of metabolite-producing genomic regions, have brought about the realisation that researchers have still only scratched the surface of the metabolomic potential of microorganisms [23]. Recent research shows that it is common for the the microbial genome to contain a large number of metabolite-producing regions than are either not expressed at all, only expressed under specific conditions, or expressed at such a low level that when doing metabolomic analysis, the signal from the resulting metabolites is drowned out by metabolites that are produced in greater abundance. Such regions are called *cryptic* [29]. For instance, while many strains of *Streptomyces* contain between 25 and 50 potentially metabolite-producing regions, about 90% of these are cryptic under normal laboratory growth conditions [23]. A variety of methods exist to activate such cryptic regions, including heterologous expression, where the relevant region is inserted into a host microorganism, or promoter engineering [30, 31, 32].

The realisation that microorganisms have greater metabolomic potential than previously thought, along with the fact that most antibiotics in use today are of microbial origin in one way or another, has raised hopes of finding novel antibiotics among the metabolites that microorganisms either produce or *could* produce under the right conditions. However, mapping the correspondence between the genomic space, where the prediction of the producing region takes place, and the metabolomic space is still a laborious process. *A priori*, any one of the producing regions of the genome could be responsible for any one of the metabolites, giving a number of potential links that grows combinatorially with the number of genomic regions and the number of metabolites. For example, a microorganism that has 10 regions that might be producing secondary metabolites, and produces 10 distinct molecules, yields 100 potential links between a genomic region and a metabolite. Since many microorganisms are prolific producers of secondary metabolites, and have a potentially even higher number of genomic regions that are predicted to produce specialised metabolites, verifying each link by laboratory methods is obviously untenable. Instead, we would like to be able to prioritise the potential links for verification.

One way of approaching this problem is to develop methods to assign scores to these potential links and use the score to prioritise the links for further research in laboratories, for instance by heterologous expression or promoter engineering [30, 31, 32]. This prioritisation can take the form of either pairing regions producing novel metabolites with their corresponding data points in the metabolomics results, or identifying regions homologous (i.e. structurally similar and with a shared ancestry) to known regions and metabolites, thereby reducing the number of possibilities for the remaining regions and metabolites by a process of elimination.

1.1 Thesis statement

Current methods to rank potential links between genomic and metabolomic features in microbial data sets can be improved by the application of both statistical approaches and machine learning approaches. However, unified methodology and data sets are required to evaluate such methods and quantify improvements. Furthermore, different approaches to the problem are complementary, and combining them in a principled manner outperforms their separate application.

1.2 Thesis structure

This thesis consists of seven chapters. Following the introductory Chapter 1, Chapter 2 introduces necessary terminology and background, both in microbial genomics and metabolomics, with a particular focus on mass spectrometry. Chapter 3 describes how the results of subsequent chapters will be evaluated and the problems involved in obtaining ground truth data for such experiments. Furthermore, it describes the creation of the data sets used in the subsequent chapters and how they will be used to evaluate the models proposed.

Chapters 4 and 5 discuss each one of the two major approaches to the problem of linking metabolite-producing regions in microorganisms to their associated metabolites. Chapter 4 describes two improvements on a commonly-used statistical method while Chapter 5 introduces a novel method to link microbial metabolomic and genomic data using kernel methods.

Chapter 6 demonstrates how the two approaches described in the preceding chapters complement one another, and suggests how this complementarity can be used in a principled manner. The thesis concludes with Chapter 7, where the results presented in the preceding chapters are discussed in the context of further research, both ongoing and prospective.

Chapter 2

Background

The study of the molecules that are formed as a part of the metabolism of an organism is known as *metabolomics* [33]. As mentioned in the previous section, microorganisms have historically been a rich source of compounds with a variety of uses. Commonly known as *metabolites* or *natural products*, these molecules are usually classified into *primary metabolites*, which are the metabolites involved in normal growth, development, and reproduction of the organism, and *specialised metabolites*, also referred to as *secondary metabolites*, which are all other metabolites that the organism produces. As such, any compounds that the organism uses to change its environment, interact with other organisms or affect its appearance (e.g. via pigmentation) is a secondary metabolite.

In order to biosynthesize these microbial metabolites, knowing *how* the production takes place is of great value, since knowing this allows for potentially greater control of the synthesis. This control can be via altering expression levels, either upregulating the expression of the metabolite of interest or downregulating others, changing fundamentally the expression of the metabolite by e.g. inserting enhancer domains in the vicinity, or by transplanting the genomic region to other organisms where the expression can be controlled to a greater degree.

Figuring out which genomic region is responsible for the production of which metabolite can therefore play a key part in making use of the metabolite.

2.1 Metabolomics

2.1.1 Mass spectrometry

Given a chemical sample, for instance a sample of microorganisms that have been cultured in a plate and the metabolites extracted using an organic solvent, one of the primary ap-

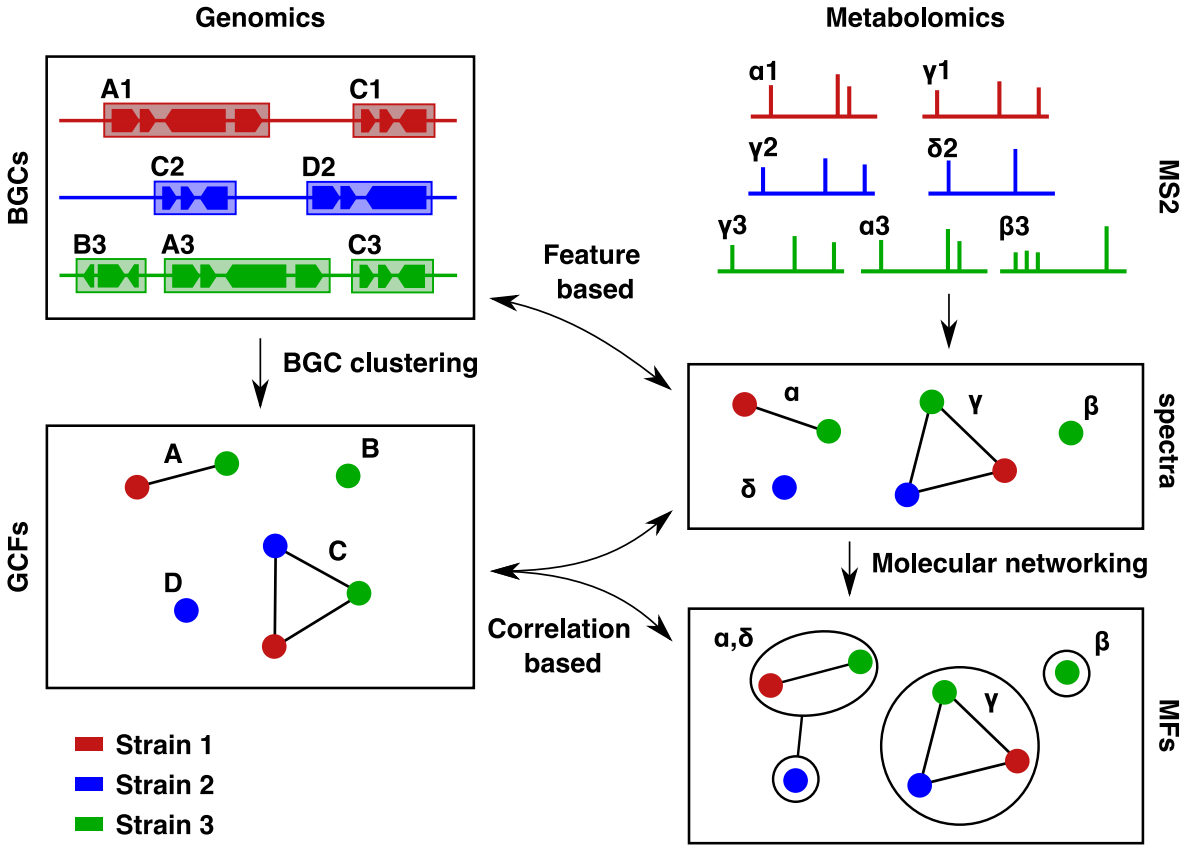


Figure 2.1: **The relationship between microbial genomics and metabolomics.** *Biosynthetic gene clusters* (BGCs) in microbial genomes are predicted, and clustered into *gene cluster families* (GCFs), containing BGCs from one or more strains. GCFs can also be considered as the set of strains contributing BGCs to the GCF. Identical MS2 spectra are identified across microbial samples and further clustered into *molecular families* (MFs). Spectra or MFs can also be considered sets of strains where the spectrum, or any spectrum in the MF, is observed in the sample for the strain. *Feature-based linking*, discussed in Chapter 5, can be used to link BGCs and spectra links, while *correlation-based linking*, discussed in Chapter 4, can be used to link GCFs to either MFs or spectra.

proaches to analysing the metabolite contents of the sample to figure out which molecules are present is *mass spectrometry* (MS) [34]. In principle, mass spectrometry involves imposing an electric charge on the molecules in the sample, which allows the researcher to measure the *mass-to-charge ratio* (*m/z*) of the molecule, i.e. the ratio between the atomic mass of the molecule and the electric charge that it carries [34].

The first instrument to make such measurements was built by J. J. Thomson in 1910 and soon proved a valuable tool in chemical analysis. To gain more structural information about the molecules, *tandem mass spectrometry* (MS2, MS/MS) was introduced in 1966. This involves two separate MS steps separated by a fragmentation step which breaks the molecules into component parts. The first step separates the molecules into classes by their *m/z*, after which

the molecules are fragmented and the fragments for each class sorted by their *m/z*. For a subset of the ionisable molecules in the original sample, this yields a fragmentation spectrum of *m/z* peaks which can serve to characterise the molecule. Various approaches exist to choose the *m/z* ranges of molecules that get fragmented, which are discussed in the next section. MS2 is currently one of the most common profiling tools in metabolomics [35]. An overview of the history of mass spectrometry can be found in [36].

2.1.2 LC-MS2

When multiple metabolites are ionised at the same time, the most abundant or the most easily ionisable metabolite can overwhelm the capacity of the ionisation source, meaning that less abundant or less easily ionisable metabolites do not get ionised and thus are not analysed by the mass spectrometer [37, 38]. To overcome this problem, known as *ion suppression*, a common first step in mass spectrometry-based metabolomics is separation of the metabolites in the input sample by the way of chromatography [35, 38, 37]. Usually, this is done by either *liquid chromatography* (LC) or *gas chromatography* (GC) although other methods exist [35]. Both LC and GC operate on the same principle: the sample is passed through a medium (*column*) which lets the molecules through at different speeds, so different metabolites come out of the column, or *elute*, at different times. The time at which a metabolite comes out of the column is known as *retention time*.

After passing through the column, the metabolites are ionised. When using LC, this is most commonly done using *electrospray ionisation* (ESI) although other methods are also possible [35, 38]. This usually involves adding a proton to or removing a proton from the molecule, although other charged particles can also be added or removed [39]. The molecule, after the change, is called an *adduct*. This addition affects the mass of the metabolite, so instead of measuring the *m/z* of the original metabolite the mass spectrometer measures the *m/z* of the adduct. This means that identical molecules can present as having more than one *m/z*, where each *m/z* corresponds to a different type of adduct, i.e. a different ion being added or removed [39]. Molecules containing different isotopes of atoms, i.e. atoms that have the same number of protons but a different number of neutrons, also present as different *m/z* peaks, where the difference in *m/z* corresponds to the difference in neutrons between the isotopes.

Metabolomics experiments are can be either *untargeted* or *targeted*. Targeted metabolomics aim to measure one or more known metabolites in a sample, and thus focus on particular, pre-calculated *m/z* values, while in untargeted metabolomics, all ions in a given *m/z* range are subject to analysis [40]. As targeted metabolomics focuses on known metabolites, the remainder of this section is concerned with untargeted metabolomics.

For untargeted metabolomics using tandem mass spectrometry, data are collected using either *data-dependent acquisition* (DDA) or *data-independent acquisition* (DIA) [41, 42]. In

LC-MS2 using DDA, an initial mass spectrometry step (MS1) is performed at intervals, analysing the material eluting from the LC column at the time of the scan [41]. The *m/z* of the metabolites is measured, giving an intensity signal along the *m/z* axis where the intensity corresponds to the the number of molecules at each *m/z* value. This signal is then processed into *peaks* corresponding to particular *m/z* values [43], i.e. a collection of pairs of *m/z*- and intensity values. The most relevant peaks — according to some heuristic, usually intensity [44, 45, 46] — are selected for fragmentation and subsequent mass spectrometry scan (MS2), which happens before the next MS1 scan of material currently eluting from the column [41]. The selected ions are called *precursor ions* or *parent ions* [38].

The most common algorithm for choosing which MS1 peaks to fragment involves fragmenting the n peaks of highest intensity in the last scan, where the number of peaks n, which is constant throughout the experiment, is set to ensure that the last MS2 step finishes before the next MS1 scan is scheduled to start [47]. This choice of peaks is normally also subject to some further criteria, intended to ensure for instance that the same metabolite is not immediately fragmented again [42].

The ions corresponding to each selected *m/z* value from the MS1 step are sent through a fragmentation step individually per selected *m/z* value. This fragmentation step normally takes the form of *collision-activated decomposition* (CAD, also known as *collision-induced dissociation* [CID]) [34]. This breaks chemical bonds in the metabolite and results in a so-called *product ion* and a *neutral fragment*, where the former retains the charge from the parent ion. The fragmentation can occur in various places on the metabolite, depending on the metabolite, and the product ions can fragment again into a new product ion and a neutral fragment. Only the product ions from this process are analysed in the subsequent MS2 step [34].

The product ions resulting from the fragmentation step are then each subject to another MS scan, which similarly to the first step sorts the ions from the fragmentation of each precursor ion by their *m/z* value, giving intensity values along the *m/z* axis. As is the case with the MS1 data, the intensity data along the *m/z* axis in the MS2 scan is converted into pairs of intensity and *m/z* values.

Once all the chosen ions have been sent through the MS2 step in this way, another MS1 scan is performed on the material currently eluting from the LC column, and the process repeated.

DIA is an alternative data acquisition mode to DDA. Unlike with DDA, instead of selecting specific narrow *m/z* ranges to fragment based on particular ions, either all ions in particular *m/z* ranges are fragmented or all ions at once (known as *all ions fragmentation* [AIF]). Therefore, DIA results in MS2 spectra that each may contain peaks from multiple ions in a particular *m/z* range. Further processing (*deconvolution*) is needed to extract the spectra for individual precursor ions from the combined spectra [41, 38].

There is a trade-off between the two data acquisition modes: while DIA captures all ions, within a pre-determined range, the calculations needed to extract the individual spectra are far from trivial and involve a considerable degree of uncertainty [41, 38]. Conversely, while DDA involves choosing which ions to fragment and generally does not fragment all the ions in the sample, the fragmentation spectra from DDA are generally from individual metabolites and can be used without further processing [41], although this is not always the case. However, the choice of ions to fragment in DDA is still an unsolved problem, and the choice of ions can be improved considerably [48].

The intensity signal along the retention time axis is influenced by the amount of material that elutes from the column and by how easily it ionises. [37, 34] Ideally, only one metabolite elutes from the column at a time, with the intensity signal along the retention time axis forming a *chromatographic peak* [37]. However, this is not always the case as several metabolites can elute from the column at the same time, even metabolites with similar molecular weight, which might then get fragmented together.

The final LC-MS2 data can be considered as consisting of MS1 data and MS2 data. The MS1 data takes the form of a collection of triads of retention time, *m/z*, and intensity of the ions in the sample (parent ions). The MS2 data, which exists for a subset of the triads in the MS1 data, consists of a collection of tuples of *m/z* and intensity values for the fragments of the corresponding data point in the MS1 signal [49, 50].

As retention time can vary greatly between experiments, based on conditions, standards and instruments, differences in retention time are difficult to interpret except in specific circumstances, such as when analysing multiple samples using the same instrumentation, or when analysing samples containing known metabolites, which can then be used to compensate for the difference in retention time [34, 51]. The data normally used for analysis of MS2 spectra are the triads composed of precursor *m/z*, fragment *m/z* and fragment intensity.

In general, there are a number of reasons why a metabolite can be present in a sample with no corresponding MS2 spectrum [42]. Firstly, the mass spectrometer only measures ionised metabolites, so if the ionisation step fails to ionise a particular metabolite, it does not get analysed. As different ionisation methods have their strengths and weaknesses, the particular ionisation method used can affect the output [34].

Secondly, using DDA only a subset of the peaks detected in the MS1 data get fragmented [44, 45, 46]. Although algorithms exist to try to select the most informative peaks to fragment, based on the peaks previously fragmented, this may lead to metabolites missing from the data. This can happen not only when different metabolites elute from the LC column at the same time, and therefore enter the mass spectrometer at the same time, but also if the mass spectrometer fails to identify MS1 peaks with similar *m/z* in consecutive MS1 scans as the same metabolite and fragments them again, at the cost of not fragmenting other precursor

ions.

Another reason for a metabolite not being analysed by the mass spectrometer is because the metabolite is overwhelmed by another metabolite that is eluting at a similar time. This can be either because it is expressed in a far greater abundance, or because it is more easily ionised. The result is the same in both cases: only, or overwhelmingly, the abundant or easily ionised metabolite gets ionised, and thus the rarer or less easily ionised metabolite does not get analysed by the mass spectrometer [52, 34].

Another problem is when two or more molecules are fragmented in a single MS2 scan, creating a *chimeric spectrum* [53]. This can happen when two molecules that elute from the LC column at the same time have a similar *m/z*, since although the MS1 peak that is chosen to fragment has a particular *m/z* value, in fact all ions in a particular *m/z* range around that value are fragmented [53].

The collision energy used to fragment the molecules can also impact the resulting spectra [54]. The same metabolite fragmented at different energy levels can yield very different spectra because of difference in how the metabolite fragments at different energy levels. Although the set of potential *m/z* values will generally be the same, representing all the possible ways in which the molecule can fragment, the peaks can have very different intensities depending on the collision energy, resulting in dissimilar spectra.

For a more thorough review of common problems in MS2 analysis of metabolites, the reader is referred to [55].

Ideally, this process would yield MS2 spectra for most or all of the metabolites in the sample, although in reality this is dependent on abundance and ionisation potential. Furthermore, as discussed earlier, the selection of ions to fragment is still not an adequately solved problem, and current methods have in fact been demonstrated to be sub-optimal [48]. To finally identify the metabolites in the sample, curated databases of MS2 spectra for known compounds [38], such as MassBank [56] or METLIN [57], are searched for spectra similar to the ones in the sample.

Similarity between MS2 spectra is usually calculated using *cosine similarity* [58]. To calculate the cosine similarity between two spectra, the spectra are first aligned, i.e. matching peaks between the two spectra are identified. If n peaks are matched between the two spectra, the intensities of those peaks form an n-dimensional vector for each spectrum, and the similarity between the two spectra is taken to be the cosine of the angle between the two vectors.

2.1.3 Molecular Networking

In natural product research, where many of the metabolites may be unknown, it can be useful to identify metabolites from different microbial strains that are identical or analogous, regardless of whether they are previously known. This can be done using *molecular networking* [51, 59, 60], which clusters the spectra across strains based on their similarity.

The key idea of molecular networking is to take the spectra for a given population of strains and cluster them such that analogous metabolites belong to the same cluster. The clusters resulting from the molecular networking are called *molecular families* (MF).

Various pipelines exist to perform molecular networking [51, 59, 60]. Two of the most popular are referred to as *feature-based molecular networking* (FBMN) [51] and *classical molecular networking* [59].

Both pipelines start by creating consensus spectra for precursor ions that are considered to correspond to the same metabolite. FBMN does this using MS1 data, by aligning the chromatographic peaks using exclusively MS1 data [51]. Retention time and MS1 ion *m/z* are used to cluster the MS1 ions into *peak sets*, where the peak sets represent the distinct ions in the experiment. For each peak set, a representative MS2 spectrum is chosen from among the available MS2 spectra for the peaks in the set, if any are available. This means that a metabolite in a particular sample can belong to a peak set, with a consensus spectrum, even if it was not fragmented in that sample, based exclusively on the MS1 data.

In contrast, classical molecular networking operates exclusively on MS2 data [59]. The pairwise cosine similarity between spectra is computed, and spectra that are very similar are deemed to be from the same metabolite. A consensus spectrum is then created by averaging the peak intensities over all of the spectra for the metabolite. This means that any metabolite that did not get fragmented in a particular sample is excluded from further analysis.

After establishing this set of distinct ions in the experiment for which MS2 is available, both pipelines proceed in the same fashion [51, 59]. The consensus spectra form a fully-connected graph, with edge weights set to a modified version of the cosine similarity between the nodes of the edge.

The goal of this *modified cosine similarity* [22] is to ensure that spectra for metabolites that differ by the addition or removal of a substructure are regarded as similar, given a simple model of the fragmentation of the metabolites, where the only effect of the fragmentation is to break the bonds in the molecule. This is done by allowing the peaks in either one of the spectra to shift by the difference in *m/z* between the precursor ions. After matching the peaks from both the un-shifted and shifted spectra, the final peak matches are chosen so that each peak is only matched once. The final result is that an MF can contain spectra with a range of parent ion *m/z*, potentially corresponding to different analogues of the metabolite.

The fully-connected graph is then pruned by removing edges with weights below a threshold value, and further by iteratively removing the lowest-scoring edges from the connected components of the graph until they fulfil a size criterion. The connected components of the graph are then taken to be the molecular families.

Molecular networking can be used to guide the identification of unidentified metabolites which belong to the same MF as known metabolites. This approach was used by Atencio and co-workers in [61] to identify two previously unknown molecules, dibromoalterochromides D/D', by their inclusion in an MF of bromoalterochromides. It has also been successfully used for dereplication [62, 63], i.e. the identification of known metabolites in a sample, by the inclusion of known metabolites in the molecular network. Finally, the composition of molecular families from related microbial strains can be used to statistically identify regions in their genome that may encode the production of the metabolite. This approach was used by Duncan and co-workers in [63] to identify the metabolite retimycin A and the BGC that encodes its production.

A discussion of the various uses of molecular networking can be found in [60].

FBMN has two main advantages over classical molecular networking. The first advantage is that it is less dependent on the data acquisition algorithm, i.e. the particular choice of ions to fragment at the MS1 stage, because it is enough for an ion to be fragmented in one of the samples in the experiment for a viable MF with a consensus spectrum to form for that ion. For instance, if a particular MS1 feature (i.e. a chromatographic peak and an *m/z* peak) appears in samples from three strains, but a corresponding MS2 feature appears in only two of them, meaning that the MS1 peak was only chosen for fragmentation in two of the samples, a consensus spectrum will be created using the two MS2 measurements, but all the MS1 ions belong to the peak set, so the consensus spectrum will be extended to all three strains. This means that the peak set — and by extension the MF — can contain ions that were detected in the MS1 scan but not fragmented. This makes the method less sensitive to the particular selection of ions from the MS1 scan for fragmenting.

Another advantage of FBMN is that using retention time, it is in some cases possible to distinguish between metabolites that have the same MS2 spectrum, e.g. stereoisomers [51]. If there is a difference in retention time between the isomers they will form two distinct consensus spectra, although as MS2 spectra do not distinguish between stereoisomers, the consensus spectra themselves will be very similar and thus end up in the same MF.

One drawback of FBMN is that as the clustering of the precursor ions is done using exclusively MS1 data, this will lead to the method missing identical metabolites with significant difference in the parent ion *m/z*, e.g. different adducts, and forming separate peak sets for the different types of adducts. These would, however, belong to the same MF.

Another drawback is that as the creation of consensus spectra is done using exclusively MS1

data, i.e. retention time and parent ion *m/z*, ions that happen to have the same retention time and parent ion *m/z* will be grouped together, regardless of how similar their MS2 spectra are. This can present a problem if two distinct metabolites have similar retention time and *m/z*, which is a possibility, particularly depending on the resolution of the MS1 scan [55].

A final drawback of FBMN, compared to classical molecular networking, is that the former is more likely to be affected by differences in experimental conditions or by batch effects and instrument settings, as these can influence MS1 measurements [51].

Both classical molecular networking and FBMN are currently integrated into the public *Global Natural Products Social* (GNPS) pipeline [64] which is further discussed in Section 3.6. The state of the art molecular networking pipeline is FBMN, but as it is quite a recent development, the microbial data sets considered in this paper have mostly been processed with classical molecular networking. However, both algorithms produce similar outputs: a collection of spectra, assumed to be from identical or closely related molecules, represented as tuples of *m/z* and intensity values, and the *m/z* of the precursor ion, along with a list indicating in which samples each spectrum was found.

2.2 Microbial genomics

As previously mentioned, since specialised metabolism is the part of the metabolism of an organism that is not strictly speaking necessary for survival, the specialised metabolites include various metabolites that the organism uses to affect its environment to confer an evolutionary advantage, including signalling molecules, pigments and molecules to kill other organisms.

While many species produce interesting and useful molecules as a part of their specialised metabolism, microbial specialised metabolites are of particular interest. This is mainly due to two things: firstly, microorganisms are extremely prolific producers of secondary metabolites, with many actinomycetes containing 20–50 regions thought to produce secondary metabolites [23]. Secondly, in microorganisms, all the genetic information for the production of a particular microbial secondary metabolite is usually in one place in the genome, making the relevant portions of the genome easier to detect and investigate [65].

Microbial secondary metabolism is in most cases controlled by *biosynthetic gene clusters* (BGCs), which are contiguous regions of adjacent genes which together encode for the production of a single molecule or a family of related molecules [66, 67]. This means that in most cases, everything that is needed for both the production and the control of expression of a molecule, i.e. the control of whether the molecule is produced at all or not, and in what quantity, are located in the same place in the genome [68]. This again makes microbial BGCs easier to analyse and affect than if they were spread over the genome, with remote parts of the genome interacting to affect the expression of the final product [65].

Metabolites, and by extension the BGCs that encode for their production, are sorted into *natural product classes* based on their properties and how they are synthesised. Three of these classes, nonribosomal peptides (NRPs), polyketides (PKs), and ribosomally synthetised and post-translationally modified peptides (RiPPs), have the distinction of lending themselves more readily than others, such as alkaloids an terpenes, to predicting the molecular structure of the resulting product. In the case of NRPs and PKs, this is because they are principally formed from preexisting modules by assembly line-like enzymes, while RiPPs are formed from so-called precursor peptides whose structure can in some cases be predicted from BGCs [69, 70, 71]. The availability of tools to predict metabolite structure for BGCs is therefore not the same for different natural product classes, with tools to predict metabolomic features for the so-called modular or multi-modular natural products most developed [71], i.e. for the metabolites where the assembly can to a large extent be seen as combining pre-defined parts with minimal post-translational alterations. These natural product classes include NRPs and PKs. Often, the differentiation of secondary metabolites by class is not clear-cut, and products can be classified as hybrids of more than one type, for instance NRP-PK hybrids.

2.2.1 BGC prediction

The resurgent interest in microbial specialised metabolites is to a large extent due to improvements in genomic sequencing, such as improved coverage and increased read length, along with advances in assembly, and subsequent identification and prediction of microbial BGCs. When the first complete genomic sequences of the genus *Streptomyces* became available, in 2002–2003, it became clear that the microorganisms had the potential to produce roughly 10 times the number of specialised metabolites that were known from fermentation studies, i.e. studies in which the bacterial colonies were grown and then subjected to chemical analysis. This large number of cryptic BGCs demonstrated that only a part of the metabolomic potential of each microorganism was already known, whether this is because the cryptic BGCs are not expressed at all, or only expressed at a low level, and thus drowned out by more abundant metabolites in the analysis [23, 72].

In the intervening years, the detection of BGCs has evolved into a fairly mature discipline. Tools such as antiSMASH [73], PRISM [74] and SMURF [75] have emerged as a standard for BGC detection in microorganisms and fungi. All of these tools are based on matching curated statistical models based on sequences of *protein family domains* (Pfam domains) to genomic sequences [76], and as they are explicitly modelled on known examples they tend to exhibit high specificity (low number of false positives) at the expense of relatively low sensitivity (high number of false negatives).

While BGC detection has traditionally been done using manually curated, rule-based logic to detect tell-tale composition of Pfam domains indicating the presence of a BGC, Hannigan

and co-workers recently introduced DeepBGC [77], which uses a text-mining-based recurrent neural network to detect BGCs from the Pfam data. However, all of the approaches, including DeepBGC, rely on detecting Pfam domains in genomic sequences by the alignment of *hidden Markov models* (HMMs).

The main advantage of DeepBGC over antiSMASH is the detection of novel BGCs, and BGCs less related to known ones. As such, DeepBGC is likely to have a higher rate of false positives than traditional, arguably more conservative, approaches, such as taken by antiSMASH.

Other more speculative algorithms exist to try to increase the sensitivity, such as ClusterFinder [78], MIDDAS-M [79] and MIPS-GC [80]. In recent versions, antiSMASH has incorporated ClusterFinder as an alternative prediction strategy.

For an in-depth review of BGC detection, the reader is referred to [81].

In our work with microbial data sets, we have restricted our research to BGCs detected by antiSMASH. There are three reasons for this. Firstly, the methods presented here are in principle independent of the tool used for BGC detection. Secondly, as parts of this work rely heavily on clustering BGCs into GCFs, and the only major tool to do this, BiG-SCAPE, does not accept PRISM or DeepBGC output at the moment, only a limited part of the results could be tested on other BGC detection approaches in any case. Lastly, as the MIBiG-GNPS data set (discussed in Section 3.6) is not dependent on antiSMASH-detected BGCs, evaluation on that data set ensures that we already have some method of evaluating our performance independent of antiSMASH.

2.2.2 BGC similarity evaluation and clustering

When working with multiple BGCs, often from multiple microbial strains, grouping them together into *gene cluster families* (GCFs) can be useful [82]. GCFs are defined as groups of BGCs such that that BGCs belonging to the same GCF produce the same or similar metabolites [82]. To be able to cluster BGCs, the space of BGCs must first be associated with some sort of similarity function, i.e. a real-valued function which accepts two BGCs as arguments and returns their similarity, the assumption being that similar BGCs encode similar metabolites.

Various similarity measures have been proposed for BGCs [83, 84, 82], but no clear consensus has been established yet. This is largely because measuring the performance of such functions is very difficult.

The clustering of BGCs builds upon early work by Lin and co-workers [83] on clustering homologous BGCs. They defined their distance function largely based on the decomposition of the BGC into Pfam domains, based on the Pfam database [85]. The entries in the Pfam

database are each a family of proteins, composed of functional regions called *domains*, along with multiple sequence alignments and HMMs, which are aligned to a sequence to evaluate its similarity to the Pfam domains.

The use of Pfam domains as building blocks for BGCs is a common thread in later work as well. Most similarity measures that have been proposed use the Pfam domain composition of the BGCs to a greater or lesser extent, although some also incorporate sequence similarity, either directly, as in [82], or by including the output scores of sequence alignment algorithms, as in [84]. Additionally, many similarity functions encode more complex features of the Pfam composition of the BGCs in various ways, e.g. accounting for duplications and ordering [82, 84, 83].

When defining similarity or distance functions in this way, it is not automatically clear if the resulting functions satisfy various conditions making them e.g. metrics or kernel functions. This applies in particular to the sequence similarity measures, both raw and alignment score based. This restricts the choice of machine learning approaches applied to the problem, unless extra steps are taken to make sure the similarity measures fulfill the necessary conditions for e.g. a kernel function. A notable exception from this is the recently-published pfam2vec algorithm [86], which projects BGCs into a shared latent space, where the appropriate choice of function on the latent space, combined with the function projecting the BGC into the latent space, can ensure that the combined function defined on the space of BGCs satisfies the desired conditions.

Once a similarity function has been defined, the BGCs can be clustered into GCFs. Various clustering algorithms have been used for this purpose: Lin and co-workers [83] used the distance-based neighbourhood-joining clustering method [87], Doroghazi and co-workers [84] used DBSCAN [88], while BiG-SCAPE [82] uses affinity propagation [89] for clustering. All of these approaches are to a greater or lesser extent influenced by one or more manually-tuned parameter affecting cluster size or number of clusters, which are in this context closely related, if not downright equivalent.

The state-of-the-art tool for BGC clustering is BiG-SCAPE [82]. Another recently-developed tool in the same vein is BiG-SLICE [90], which exhibits superior computational performance on large data sets but at the cost of decreased accuracy. In fact, the authors of BiG-SLICE recommend that BiG-SCAPE be used for more granular analysis of GCFs of interest, in addition to BiG-SLICE [90]. Since the data sets considered in this thesis are small enough that the running time of BiG-SCAPE does not pose a problem, we choose to do all GCF analysis using BiG-SCAPE.

When clustering BGCs, BiG-SCAPE restricts cluster members to a single natural product class, and in fact uses a different similarity measure depending on the class. The assignment of BGC to natural product class is based on the antiSMASH annotations for the BGC.

However, BiG-SCAPE does not specifically cover all natural product classes assigned by antiSMASH. Instead, it classifies BGCs into NRPS, type I PKS, other PKS, PKS-NRP hybrid, RiPP, Terpene, Saccharide, and Other. To accommodate hybrid natural product classes, BGCs that are assigned as hybrids get clustered along with each of the hybridised types, i.e. NRPS-PKS hybrids get clustered three times: once as NRPS, once as PKS, and once as NRPS-PKS hybrids.

One of the key features of BiG-SCAPE is their particular definition of a similarity measure on the space of BGCs. While many definitions of the similarity between BGCs exist, as described earlier, the BiG-SCAPE definition of similarity is unique in that the component weights vary by natural product class.

The BiG-SCAPE metric is composed of three components: Jaccard index (quantifying the overlap in Pfam domain composition), adjacency index (quantifying the difference in order in which pairs of Pfam domains occur) and direct sequence similarity (quantifying the sequence similarity between matched Pfam domains). The final similarity measure is a weighted combination of these terms, with different weights defined for each natural product class. For instance, similarity between RiPP BGCs is computed using all three components, while for NRPs, the BGC similarity is computed exclusively in terms of sequence similarity for the shared domains. This is because adenylation domain specificity, i.e. which amino acids are incorporated into the final product, which is a key characteristic of NRPs, is not captured by any of the other component metrics.

In addition, BiG-SCAPE can be made to cluster all BGCs together, as a special product class called ‘mix’. In this mode, the same coefficients are used for all comparisons, both within class and between classes.

Although only recently released, BiG-SCAPE has already been used for example by Männle and co-workers to assess the diversity of BGCs in *Nocardia*, and to identify the BGCs in *Nocardia* producing distinct structural types of nocobactin-like siderophores [91], as well as by Soldatou and co-workers to putatively identify the BGCs encoding for the production of ectoine and chloramphenicol in polar actinomycetes [92].

2.3 Linking genomic and metabolomic data

As previously stated, the combinatorial nature of the problem means that the number of potential links to be evaluated increases quadratically with the number of BGCs and MS2 spectra. Tools to aid in the prioritisation of the potential links are therefore indispensable to be able to focus verification efforts on the most likely links. Three recent reviews on linking genomic and metabolomic data are [93], [71], and [94].

Efforts to link genomic and metabolomic information for microorganisms can be broadly split into two approaches: *feature-based linking*, where predicted molecular features from the BGC are used to search the MS2 spectra for tell-tale signs of the predicted features, and *correlation-based linking*, where correspondences between BGC content and metabolomic output in multiple microbial strains are analysed to determine how likely they are to be connected.

Although the most concrete relationship is between a BGC and the spectrum of the metabolite encoded by that BGC in the MS2 results for that particular strain, the links thus defined generalise to GCFs and MFs as well. Since all BGCs in a GCF are assumed to encode for the production of the same or similar metabolites, and all spectra in an MF are assumed to arise from the same or similar metabolites, BGC-MS2 links generalise in a natural way to GCF-MF links, and vice versa. This also applies to links between GCFs and spectra, and BGCs and MFs.

2.3.1 Feature-based linking

At its root, feature based linking involves predicting some aspect of a metabolite from a BGC and then searching a set of MS2 spectra for signs corresponding to the molecular feature. A number of tools exists that can be used to this effect, although their integration into the metabolomics and genomics workflows is of a varying depth.

Some tools encompass the whole pipeline, taking genomic- and MS2 data as input and establishing links between BGCs in the genome and spectra in the MS2 data. Notably, MetaMiner [95] predicts the structure of RiPPs from genomic sequences and searches a set of spectra for corresponding features, while NRPquest [96] does the same for NRPs. Pep2path [97] predicts sequences of losses for peptidic natural products (i.e. NRPs and RiPPs) and looks for corresponding peaks and losses in MS2 spectra, while GNP [69] aims to link NRPs and PKs to LC-MS2 data.

Other tools are not as integrated, predicting various structural properties of the metabolites but leaving the interpretation of how these will appear in MS2 spectra to the researcher.

SANDPUMA ([98]) predicts substrate specificity for adenylation domains in NRPS BGCs, which makes it possible to looks for peaks or neutral losses indicative of the corresponding amino acids in the MS2 data. RODEO [99] predicts metabolomic features of RiPPs but is not yet integrated with MS2 data. Similar approaches, i.e. predicting features of a metabolite from a BGC and looking for corresponding features in metabolites are described by Panter, Krug and Müller [100] for PKs.

This process, of explicitly predicting features of peptidic metabolites (i.e. PKs, NRPs and RiPPs) from genomic data and looking for corresponding features in MS2 spectra, has been

termed *peptidogenomics* [101].

Some of these tools, such as GNP [69], do their own BGC detection, while others, such as MetaMiner [95] incorporate antiSMASH [73] for BGC detection.

The recently-published PRISM4 [102] incorporates structural predictions for a wide variety of natural product classes. This is an improvement from previous versions, since PRISM has historically mostly focused on NRPs and PKs. These structural predictions can be used to search the set of spectra from the data set for the best match, either by *in silico* MS2 fragmentation of the metabolite, or by using methods such as IOKR (see Section 5.5).

With the exception of PRISM4, all of the tools mentioned above are restricted to particular natural product class by design. While this is reasonable, as particularly PKs and NRPs are both diverse and historically important, matching spectra to BGCs one natural product class at a time can become problematic, if the same spectrum is assigned to different BGCs belonging to different natural product classes. In those cases, the strength of evidence for each match would have to be evaluated against the other matches. Such comparison is not straightforward.

Tools primarily developed for dereplication, i.e. using structural information, rather than spectral similarity, to identify various spectra as belonging a particular metabolite, can also be used to link BGCs and metabolites. Tools such as DEREPLICATOR [103] have been used to predict possible fragmentation patterns of a metabolite, which can then be matched to spectra in a sample. Based on similarity of a microbial BGC to a BGC with a known product, this approach has been used to tentatively link a novel BGC to its resulting spectrum, even for an unknown metabolite, based on the structural similarity of the metabolite to a known metabolite [104]. In this case the workflow involves using DEREPLICATOR to link a spectrum to a spectrum from a known metabolite, using MIBiG [66] or similar databases to find the BGC producing the known metabolite, and then searching the microbial BGCs for the BGCs most similar to the reference BGC.

2.3.2 Correlation-based linking

When working with samples from multiple strains it can be useful to consider each MS2 spectrum (or MF) as the set of strains where the spectrum (or any spectrum from the MF) occurs in the results for that strain [63, 84, 92]. This allows the comparison of strains based on metabolomic similarity, by looking at how they co-occur in these sets. In the same way, GCFs can be considered as the set of strains which have a BGC in the GCF.

Since microbial specialised metabolites are generally produced by BGCs, if a BGC is responsible for the production of a metabolite, then microorganisms that contain the BGC — or, equivalently, contain a BGC in the corresponding GCF — should be more likely to produce

the metabolite than microorganisms that do not contain the BGC. In other words, considering the set of strains from the GCF and the set of strains from the MF, the intersection of the two should be larger than expected if the two were not related.

This is the fundamental idea behind the correlation-based linking approach, where whole populations of microorganisms are analysed for similarities in the patterns of presence and absence of BGCs and spectra. Building upon the early work of Lin and co-workers [83] clustering homologous BGCs, Doroghazi and co-workers presented a scoring function for potential links based on the shared strain contents of GCFs and MFs, coining the term *metabologenomics* for the approach [84, 105]. Their scoring function is based on the intersection and differences of the sets of strains in the GCFs and the MF, rewarding correspondence between the two and penalising differences. Although in principle similar to *genome-wide association studies* (GWAS) [106, 107], this is done asymmetrically, to reflect the fact that a strain that contains a BGC does not necessarily produce the resulting metabolite in any measurable quantity, if the BGC is cryptic, while strains should only be able to produce the metabolites for which they have a corresponding BGC [84]. This scoring function is discussed in detail in Section 4.2.

2.3.3 Combined linking

Although two different approaches exist to the problem of linking BGCs to spectra, they are highly complementary and are indeed often used together [93]. Clustering BGCs can be used to infer previously established feature-based matches onto other members of the cluster, and conversely, shared predicted properties of BGCs can give indications of what features to search for in the metabolomic data.

While the two approaches are generally used together, no attempt has been made at systematically quantifying the extent to which they complement one another, nor has any principled investigation been done into how to best make use of their complementarity. This is discussed in detail in Chapter 6.

The relationship between BGCs, GCFs, MS2 spectra and MFs, and the different linking approaches, is shown in Fig. 2.1. For detailed reviews of linking microbial specialised metabolites to their producing BGCs, the reader is referred to [93] and [71].

Chapter 3

Evaluation of methods

3.1 Introduction

The following chapter is concerned with how the performance of methods to link BGCs and MS2 spectra are evaluated. Sections 3.2 and 3.3 discuss how such approaches are commonly evaluated, while Section 3.4 gives an overview of common metrics for information retrieval tasks and how they relate to the problem at hand. Section 3.4 also explains how performance evaluation will be carried out throughout the thesis.

To evaluate the models proposed in this thesis two data sets are compiled: Section 3.6 describes the construction of a database of BGC-spectrum links, while Section 3.7 describes the assembly of three microbial data sets with several validated GCF-MF links.

3.2 Establishing ground truth

A major problem in the development of methods to link BGCs to MS2 spectra is the lack of ground truth data [108, 38]. Although large databases exist of both MS2 spectra and microbial BGCs [109, 64, 67], databases of MS2 spectra with their associated BGCs or vice versa are small and hard to come by. Furthermore, although many validated BGC-MS2 links exist in the literature, resources to gather the information in a single place, along with their associated genomic and metabolomic data sets, have been scarce.

While the NPAtlas database [110] contains links to MIBiG and GNPS for a subset of entries, this is a considerably smaller subset than the one presented in Section 3.6, containing roughly 180 metabolite-spectrum pairs.

In any data set consisting of a population of strains, a number of BGCs per strain, and a number of MS2 spectra, there will as previously stated be a vast number of potential links

between a BGC and a spectrum. A very small number of the potential links will be *true*, meaning that the BGC encodes for the production of the metabolite to which the spectrum corresponds. Furthermore, in any given strain, not all BGCs present in the strain are necessarily expressed, particularly not in such quantity that the product appears in the MS2 analysis. This also depends on the growth conditions, as some BGCs are only expressed in certain situations, e.g. as a reaction to some sort of external stimulus. Finally, as discussed in section 2.1.2, the metabolite may be present in the sample but not analysed by the MS2, for instance if it does not ionise, if it comes through the LC column at the same time as a much higher-concentration or more easily ionisable metabolite, or if the data acquisition workflow does not choose that particular precursor ion to fragment.

From the above, it follows that if a strain has a BGC in a GCF that is in fact responsible for a particular metabolite, it is perfectly possible that the strain does not appear in the corresponding molecular family. Conversely, some of the spectra in a microbial sample may belong to either contaminants or primary metabolites, although depending on solvent choice, these will be much rarer than the spectra belonging to specialised metabolites [111, 112]. These spectra are not likely to be produced by any of the detected BGCs.

Even among the relatively few potential BGC-spectrum links that are in fact true, fewer still have actually been validated for any data set. For any of the data sets from microbial cultures, we therefore have a very small subset of true links within a very large subset of potential links. Of these true links, only an even smaller subset has been validated.

Furthermore, knowing when the whole set of true links has been discovered is almost impossible, unless either all BGCs or all spectra (or both) have been accounted for. This is a tall order, especially in microbes with a rich secondary metabolism: any cryptic BGCs would have to be identified to a high degree of certainty using genomic methods, and all spectra that do not have a validated link to a BGC would have to be matched to a library spectrum with very high similarity.

3.3 Current approaches to evaluation

As may be expected for an emerging field of study, no benchmark data sets for microbial BGCs and spectra have yet been compiled. By benchmark data sets we mean data sets with thoroughly validated links, ideally where all true links in the data set have been identified, and with as extensive annotations as possible. Once adopted by the research community, such data sets can be used to evaluate performance differences between different approaches.

This lack of benchmark data sets makes comparison of results from different research articles difficult. As the methods in many cases only have partial overlap, for instance because they are designed to work for different natural product classes, this is to some extent to be

expected. However, the sharing of experimental data has been lacklustre as well. The recent development of the *Paired omics Data Platform* [108], discussed further in Section 3.8, stands to remedy this latter point by making data sets easier to access, but the adoption of the platform is in the hands of the research community, and the problem remains of comparison between approaches with different but overlapping domains.

Traditionally, most articles presenting new ways of establishing links between BGCs and MS2 spectra demonstrate the utility of the proposed method by applying it to a microbial data set, finding a potential link, and validating the link. In this way, Kersten and co-workers introduced *natural product peptidogenomics* and used it to identify the BGC in *Streptomyces griseus* IFO 13350 encoding the known ribosomal peptide AmfS [101]. Panter, Krug and Muller introduced a novel statistical approach to identify the BGC in *Myxococcus fulvus* MCy9280 encoding the production of fulvuthiacene, which they verified by inactivating the BGC [100], and Johnston and co-workers introduced the *Genomes-to-Natural Products platform* (GNP), which they used to identify six genetically predicted metabolites in various microorganisms.

While this serves to demonstrate the validity of the methods, they are not generally compared to one another in terms of performance on a data set containing previously validated links.

Correlation-based approaches are not tied to particular natural product class and should therefore be somewhat easier to compare than feature-based approaches. To validate both their BGC clustering algorithm and their scoring scheme in [84], Doroghazi and co-workers assembled a microbial data set of MS2 spectra and predicted BGCs, for which they had several previously validated links. They then demonstrated that the validated links scored high within the distribution of all potential links in the data set.

To demonstrate the validity of their clustering approach in [82], Navarro-Muñoz and co-workers took a similar approach, by assembling a (different) microbial data set, with a set of 9 validated BGC-MS2 links, and demonstrating that a large proportion of these links appeared in the top end of the distribution of scores for all potential links. They also evaluated the clustering algorithm on a manually-curated set of clusters in a database of BGCs.

Easy access to such paired data sets of genomic and metabolomic information, along with a thoroughly annotated set of validated links, as provided by the Paired omics Data Platform [108], is crucial to be able to compare novel methods to existing ones.

3.4 Evaluating ranking results

Assume that we have an information retrieval model which returns from a set of documents a subset of results which are either relevant or not. We denote the number of true positives,

i.e. number of relevant results in the result set, as TP, and the number of false positives, i.e. the number of non-relevant results in the result set, as FP.

The problem of ranking BGC-MS2 links can be formulated as an information retrieval task in the following way. Given a BGC (query), we want to search a set of spectra (documents) for the spectra for the metabolite or metabolites encoded by the BGC (relevant documents).

Precision (P), recall (R) and F_β, which are all commonly used to evaluate information retrieval models, are defined as

$$P = \frac{TP}{TP + FP}, \tag{3.1}$$

$$R = \frac{TP}{TP + FN}, \text{ and} \tag{3.2}$$

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \tag{3.3}$$

where β^2 controls the trade-off between precision and recall in the evaluation of the model. Most commonly, the F_β measure is *balanced*, with $\beta^2 = 1$.

However, none of these scores are particularly well suited to the problem at hand. For most BGCs, there is a single correct spectrum, or at most very few, so unless the result set is very small, precision will always be fairly low. For example, if there is a single correct spectrum for a given BGC, a result set of size 2 would never give precision higher than 0.5 for that query. Recall would be less of a problem, as that does not penalise false positives in the result set, and for F_β, the parameter β could be chosen > 1, to skew the F_β score towards weighing recall higher than precision, but the question still remains what a good number of results would be, i.e. the appropriate size of the result set.

A common performance measure for ranking lists of results, which is *not* dependent on a pre-defined result set size, is *mean reciprocal rank* (MRR), which for a set of queries Q is defined as

$$\text{MRR}(Q) = \frac{1}{\#Q} \sum_{q=1}^{\#Q} \frac{1}{\text{rank}(q)} \tag{3.4}$$

where $\text{rank}(i)$ is the rank of the best-ranking true match for the i-th query, i.e. MRR is the mean of the inverse of the rank of the best-ranked true match.

Another common performance measure is *mean average precision* (MAP), defined as

$$\text{MAP}(Q) = \frac{\sum_{q=1}^{\#Q} \text{AveP}(q)}{\#Q} \tag{3.5}$$

where

$$\text{AveP}(q) = \frac{\sum_{k=1}^{\#K} P(k)\text{rel}(k)}{\#K_q} \tag{3.6}$$

is the average precision of the q-th query, with K as the set of all documents, $K_q \subseteq K$ the set of relevant documents for the q-th query, $\text{rel}(k)$ as an indicator function returning 1 if document k is relevant, and 0 otherwise, and $P(k)$ as the precision at index k, as defined in Eq. 3.1.

In cases where there is only ever one correct result, as is frequently the case in our situation, MAP simplifies to MRR: for each query q, the indicator function $\text{rel}(k)$ is 0 except exactly for the relevant document, so the product $P(k)\text{rel}(k)$ is non-zero only once, and the value of k for which $\text{rel}(k) = 1$ is the rank of the correct match.

Following [113], for evaluating IOKR performance, we generally use top-n performance. For the problem of ranking a list with a single correct match, this measures for what proportion of cases the correct match is among the top n entries on the list, i.e.

$$\text{top-}n(Q) = \frac{\#\{q \in Q \mid \text{rank}(q) \leq n\}}{\#Q}. \tag{3.7}$$

As a more general measure, we also compute the *area under curve* (AUC) for increasing n,

$$\text{AUC} = \sum_{n=1}^{\#N} \frac{\text{top-}n(Q)}{\#N} \tag{3.8}$$

$$= \frac{1}{\#N} \int_0^{\#N} \text{top-}n(Q)dn. \tag{3.9}$$

A perfect ranking, which always ranks the correct match highest, would have an AUC of 1, while a random ranking would have an AUC of 0.5. Put differently, the AUC is the area under the curve formed by plotting n (x-axis) against the top-n performance (y-axis). In this way, a top-1 performance of 100% would give an AUC of 1.0.

Assuming one-to-one correspondence between the objects and the candidate set, i.e. each query object has exactly one match in the candidate set, an AUC of 0.5 corresponds to random assignment. Assume that the ranking is randomly ordered, and the number of candidates is N. For $k < N$, the proportion of query objects that have the correct match among the k highest-ranking assignments is therefore $\frac{k}{N}$. The AUC is then

$$\frac{1}{N}\sum_{k=1}^{N}\frac{k}{N} = \frac{1}{N^2}\sum_{k=1}^{N}k \tag{3.10}$$

$$= \frac{1}{N^2}\frac{N(N+1)}{2} \tag{3.11}$$

$$= \frac{N+1}{2N} \tag{3.12}$$

$$= \frac{1}{2} + \frac{1}{2N}. \tag{3.13}$$

The last term tends to 0 as $N \to \infty$, so the AUC tends to $\frac{1}{2}$ with increasing N.

The above can also be demonstrated in terms of expected values. Taking N to be the number of candidate objects, and M to be the number of query objects, the expected number of query objects x with the correct match ranked at i (denoted here as rank$(x) = i$) is

$$E[\#\{x \mid \text{rank}(x) = i\}] = E\left[\sum_{i}^{M} Y_j\right] \tag{3.14}$$

where Y_j are i.i.d. binomial variables, indicating that the j-th example has rank i, with $E[Y_j] = \frac{1}{N}$. We have therefore

$$E\left[\#\{x \mid \text{rank}(x) = i\}\right] = E\left[\sum_{i=1}^{M} Y_j\right] \tag{3.15}$$

$$= \sum_{i=1}^{M} E[Y_i] \tag{3.16}$$

$$= \sum_{i=1}^{M} \frac{1}{N} \tag{3.17}$$

$$= \frac{M}{N} \tag{3.18}$$

giving

$$AUC = \sum_{i=1}^{N} \frac{E[\#\{x \mid \text{rank}(x) \leq i\}]}{MN} \tag{3.19}$$

$$= \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{i} E[\#\{x \mid \text{rank}(x) = i\}] \tag{3.20}$$

$$= \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{i} \frac{M}{N} \tag{3.21}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{i} 1 \tag{3.22}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} i \tag{3.23}$$

$$= \frac{1}{N^2} \frac{1}{2} N^2 \tag{3.24}$$

$$= \frac{1}{2}. \tag{3.25}$$

When calculating the AUC, the summation is done over the number of candidates in the candidate set. Of course, the top-n value is not informative for values of n larger than the size of the candidate set, so in order for the AUC calculations to be valid, they must all be done in relation to the same candidate set. In order to calculate the top-n AUC for different candidate sets N_q for each query $q \in Q$, we can use the relative rank instead,

$$\text{rrank}(q) = \frac{\text{rank}(q)}{\#N_q}. \tag{3.26}$$

In this formulation, the n in top-n is replaced by r, with $r \in [0, 1]$, giving

$$\text{top-}r(Q) = \frac{\#\{q \in Q \mid \text{rrank}(q) \leq r\}}{\#Q} \tag{3.27}$$

and we can define rAUC similarly to AUC, with the summation as before, but with the caveat that the intervals along the x-axis will no longer have equal length. Equivalently, we can consider rAUC as an integral:

$$\text{rAUC} = \int_0^1 \text{top-}r(Q)dr. \tag{3.28}$$

For $N_q = N$, where all the candidate sets are identical, this simplifies to AUC as defined in Eqs. 3.8 and 3.9.

For a candidate set N and BGC set Q, we can express the AUC differently, by doing the

summation along the x-axis instead of the y-axis, i.e. summing over the size of the candidate set first and then the proportion of the queries, instead of summing first over the proportion of the queries and then the size of the candidate set, as in Eq. 3.8.

Taking $\delta_{\text{rank}(b),i}$ to be the Kronecker delta function returning one when $\text{rank}(b) = i$ and zero otherwise, we note first that

$$\sum_{n=1}^{\#N}\sum_{i=1}^{n}\delta_{\text{rank}(b),i} = \sum_{n=1}^{\text{rank}(b)-1}\sum_{i=1}^{n}\delta_{\text{rank}(b),i} + \sum_{n=\text{rank}(b)}^{\#N}\sum_{i=1}^{n}\delta_{\text{rank}(b),i} \tag{3.29}$$

$$= \sum_{n=\text{rank}(b)}^{\#N} 1 \tag{3.30}$$

$$= \#M - \text{rank}(b) \tag{3.31}$$

where the first term on the right hand side of Eq. 3.29 sums to zero because $i < \text{rank}(b)$, and in the second term, $i = \text{rank}(b)$ is true exactly once for every value of n.

The AUC can then be expressed as

$$\text{AUC} = \frac{1}{\#N}\sum_{n=1}^{\#N}\frac{\#\{b \in Q \mid \text{rank}(b) < n\}}{\#Q} \tag{3.32}$$

$$= \frac{1}{\#Q\#N}\sum_{n=1}^{\#N}\sum_{i=1}^{n}\#\{b \in Q \mid \text{rank}(b) = i\} \tag{3.33}$$

$$= \frac{1}{\#Q\#N}\sum_{n=1}^{\#N}\sum_{i=1}^{n}\sum_{b\in Q}\delta_{\text{rank}(b),i} \tag{3.34}$$

$$= \frac{1}{\#Q\#N}\sum_{b\in Q}\sum_{n=1}^{\#N}\sum_{i=1}^{n}\delta_{\text{rank}(b),i} \tag{3.35}$$

$$= \frac{1}{\#Q\#N}\sum_{b\in Q}(\#N - \text{rank}(b)) \tag{3.36}$$

$$= \frac{1}{\#Q\#N}\left(\#N\#Q - \sum_{b\in Q}\text{rank}(b)\right) \tag{3.37}$$

$$= 1 - \frac{1}{\#Q}\sum_{b\in Q}\frac{\text{rank}(b)}{\#N}. \tag{3.38}$$

Considered thus, the AUC ends up being the complement of the average relative rank taken over all BGCs, i.e. one minus the average relative rank.

While a BGC will occasionally have a number of associated metabolites, we choose to look at the rank of the top associated metabolite for our top-n calculations. In that way, the top-n AUC is similar to the MRR, which only considers the top-ranked match for each query.

As demonstrated above, AUC is the complement of the average relative rank, while MRR is the average of the reciprocal of the rank. While the two are closely related, they are not equivalent, not even when considered as orderings, as demonstrated by the following counterexample.

Recall that

$$\text{MRR} = \frac{1}{\#\text{BGC}} \sum_{b \in \text{BGC}} \frac{1}{\text{rank}(b)} \tag{3.39}$$

and

$$\text{AUC} = 1 - \frac{1}{\#M\#\text{BGC}} \sum_{b \in \text{BGC}} \text{rank}(b), \tag{3.40}$$

so for the two scores to be in agreement for any two models, we would need $a + b < c + d$ if and only if $\frac{1}{a} + \frac{1}{b} > \frac{1}{c} + \frac{1}{d}$ to hold. As a counterexample, set $a = 100$, $b = 1$, $c = 50$ and $d = 50$. $\frac{101}{100} > \frac{2}{50}$, but $101 > 100$.

In particular AUC is sensitive to changes in performance only relative to the magnitude of the change, while MRR is also sensitive to the ranking itself, with the changes in good rankings having a higher effect than changes in bad rankings, i.e. if a BGC changes rank from one to two, it affects the MRR much more than a change from 101 to 102, while both affect the AUC in the same way.

For an overview of performance evaluation for information retrieval tasks, the reader is referred to [114].

3.5 Creating data sets for evaluation

The remainder of this Section is dedicated to compiling the data sets needed to evaluate the approaches proposed in the following chapters, as well as to explain how this performance evaluation will be carried out.

The data used to validate the approaches proposed in this thesis take two forms. The first is a library of BGC-spectrum pairs, compiled from databases of BGCs and MS2 spectra without regards to source strain, by using molecular structures. This was created by merging two databases, one of BGCs and the other of MS2 spectra, using the molecular structure annotations from both to establish the links.

The other is a collection of microbial data sets, each with genomic information, MS2 spectra, and a small number of experimentally validated links. While numerous such data sets exist, it is only with the recent development of the Paired omics Data Platform [108] that these data sets, and in particular the validated links, have become easily accessible. The metabolomic and genomic data sets from the platform, along with the validated links, can be used individually or in combinations to investigate the ratio of validated links to the total number of links in various subsets of the potential links.

3.6 The MIBiG-GNPS data set

One way to evaluate the performance of link scoring functions is using a set of validated BGC-spectrum pairs, irrespective of the rest of the BGCs in the genome and the spectra from the experiment. While this is only useful for feature-based scoring functions, this provides us with a much larger set of validated links than would be available for a data set of microbial strains.

Several databases of microbial BGCs exist. In particular, three databases are currently actively maintained and growing: MIBiG, IMG-ABC and antiSMASH DB [115]. MIBiG [66, 67] contains exclusively known, experimentally validated BGCs with known metabolites and detailed, high-quality annotations. In contrast, both IMG-ABC and antiSMASH DB contain a large number of computationally predicted BGC, as well as validated ones. Of the three, MIBiG is therefore best suited to serve as a source of ground truth. Since its release in 2015, it has become the primary repository of characterised microbial BGCs and currently contains roughly 2000 community-curated BGCs, with extensive metadata.

The MIBiG database is integrated into many BGC analysis tools. Notably, both antiSMASH and BiG-SCAPE use MIBiG to characterise the similarity of detected BGCs and GCFs, respectively, to BGCs with known metabolites.

As a part of the analysis of each detected BGC, antiSMASH computes the similarity of the detected BGC to all entries in MIBiG, in terms of the BLAST [116] score of the component Pfam domains, and returns for each BGC a (possibly empty) list of MIBiG BGCs showing considerable similarity. Assuming that similar BGCs produce similar metabolites, this list can be used to assign zero or more metabolites to the detected BGC, depending on how many MIBiG-matches are reported.

To assign MIBiG matches to GCFs, BiG-SCAPE can include the entries from the MIBiG database in its input of BGCs to cluster, thus assigning the associated metabolites for the MIBiG BGC to the GCF in which they are placed.

Although MIBiG contains structural annotations for a large majority of its entries, it does

not contain MS2 spectra for the entries. While tools exist to predict MS2 spectra *in silico* [117], such tools are generally intended to find the correct metabolite for a given spectrum from a set of candidate metabolites, by simulating the fragmentation of the metabolite and matching the predicted spectra to the input spectrum. This means that given a metabolite, the spectrum that these tools predict cannot necessarily be taken as an indicator of what a spectrum for the metabolite would look like in reality, but only that enough of the features of the spectrum are similar to the actual spectrum such that the distance between the predicted and actual spectrum is the least distance among the candidates.

While these *in silico* fragmentation tools could nevertheless be used to find putative matches for the predicted BGC structures, by taking an entire database of MS2 spectra as a candidate set, we prefer the higher accuracy afforded by the fact that databases of MS2 spectra also have structural annotations. These annotations can be matched to the structural annotations in MIBiG, and used to link the MIBiG BGCs to the manually curated MS2 spectra.

Two databases in particular exist for MS2 data, MetaboLights and the *Global Natural Products Social* (GNPS) [115]. Both serve as repositories for MS2 results from microbial data sets, where researchers can upload their data. In addition, GNPS has a curated subset of spectra with known structural annotations, and is the largest freely available such library. The GNPS database was introduced in 2016 [64] and is focused on the community-sharing of MS2 data. It contains 147647 spectra (as of March 2020), 77102 of which have structural annotations.

The structural annotations of the entries from the MIBiG and GNPS databases were used to link the two. MIBiG contains structural annotations as SMILES strings [118], and GNPS variously as SMILES or InChI [119]. This presents two problems. Since SMILES strings are not uniquely determined (i.e. the same molecule can be represented by multiple valid SMILES strings) they cannot be used for comparison without further processing. Secondly, both SMILES and InChI contain information about molecular properties such as chirality, which would not be visible in MS2 analysis, and should therefore be ignored in the matching.

InChIKey [119] is a fingerprint representation of a molecule, in the form of a set of hashes based on the molecular structure. It is split up into two blocks, where the first block encodes information about the molecular skeleton of the metabolite, while the second block encodes various types of isomerism [120]. Therefore, metabolites with identical first block, and differing second block, would be expected to have identical or very similar spectra, while metabolites that have different first block of the InChIKey would be unlikely to have identical spectra. For an investigation into the collision resistance of InChIKey, please see [120].

Since InChI and SMILES are really two ways of encoding the same information, the InChIKey of a metabolite can be computed from either SMILES or InChI representation. Only the first block of the InChIKey of each entry was used to link the MIBiG and GNPS databases, since

molecules sharing the first block are not expected to be distinguishable from one another using MS2 spectra. For each entry of either database, the InChIKey was computed, and each entry from the MIBiG database paired with all entries from the GNPS database that had a matching first block of the InChIKey.

This resulted in 2966 BGC-spectrum pairs, where each pair has two associated structures, one from MIBiG and the other from GNPS. In 1355 cases these structures matched completely, i.e. the entire InChIKey matched. These pairs consist of 2069 unique spectra and 242 unique BGCs, the natural product classes of which can be seen in Table 3.1. While the NPAtlas database [110] contains links to MIBiG and GNPS for a subset of entries, this is a considerably smaller subset than the one presented here, containing roughly 180 metabolites.

As the same BGC can produce multiple structurally related metabolites, and each metabolite can have multiple spectra in GNPS, each MIBiG BGC can be linked to many spectra. Conversely, the same metabolite can be produced by multiple MIBiG BGCs, e.g. when the same metabolite is produced by more than one strain. The relationship is therefore potentially many-to-many.

By taking for instance a BGC, considering all potential links between the BGC and any of the spectra, and ranking the list of those links, or vice versa, this list of validated BGC-spectrum pairs can be used to evaluate feature-based link-scoring algorithms, by observing the position of the validated links wihtin the collection of all potential links.

Number	Natural product class
46	NRP
30	NRP, Polyketide
1	NRP, Other
60	Polyketide
5	Polyketide, Saccharide
1	Polyketide, Terpene
1	Other, Polyketide, NRP
19	Saccharide
2	Saccharide, Other
16	Terpene
2	Terpene, Alkaloid
6	Alkaloid
44	Other
16	Unknown

Table 3.1: Natural product classes of BGCs in the MIBiG-GNPS data set

3.7 Microbial data sets

Microbial samples generally contain many metabolites, products of both primary and specialised metabolism. In our analysis, we are mostly interested in the spectra for secondary metabolites. This is both because the space of secondary metabolites contains a greater number of useful metabolites [25], and because the production of primary metabolites is generally not encoded by BGCs [66]. However, the processing of the sample greatly affects the results of the metabolomics analysis. In particular, the choice of solvent dictates the type of molecules that will be analysed, since no single solvent can extract all types of metabolites from the sample [111, 112]. Therefore, by making a judicious choice of solvent, it can be assumed that the majority of the spectra in the MS2 data belongs to secondary metabolites, and should therefore have a corresponding BGC.

For a population of strains with a given set of BGCs or GCFs and spectra or MFs the number of potential links is vast. If the strains are active producers of secondary metabolites, as many microbes are, a number of these potential links are true, meaning that the BGCs in the GCF are responsible for the production of a metabolite in the MF.

When scoring the potential links, the ideal scoring function would rank all true links higher than any of the other potential links. Therefore, the proportion of true links to all links in the top quantiles of the distributions can be used to evaluate two scoring functions against one another. Taking the number of links that have a score above a given percentile for each scoring function, and looking at the ratio of true links to all links, should give a higher ratio for the better-performing scoring function.

However, for any given data set, only a small subset of the true links are likely to have been validated. Complicating matters further, the number of true links in the data set is unknown, although a rough upper bound of the number of true links can be estimated from the number of GCFs, since most BGCs are only responsible for the production of a single metabolite, and the number of MFs.

Even so, looking at the proportion of *validated* links in the top quantiles of the distributions of scores can be used as a measure of performance, since by increasing the proportion of *true* links in the top quantiles of the distribution, we would be increasing the proportion of *validated* links. In this way, we can use the small subset of validated links to evaluate the performance of scoring functions even without knowing the set of true links.

Considering only a subset of the true links may create the illusion of a performance difference between two ideal scoring functions, by ranking the subset of validated links relatively higher or lower within the subset of true links. This can be avoided by considering a large enough set of potential links: if the set of high-scoring potential links used for the performance evaluation is larger than the set of true links, then two ideal scoring functions would

demonstrate the same performance in this metric.

3.8 Paired omics Data Platform

One of the major obstacles to the development of new computational methods to link BGCs to metabolites is the difficulty of accessing ground truth data for performance evaluation. Such ground truth data would ideally be in the form of paired MS2- and genomic data for a collection of strains, along with a set of validated links between the two. Although plenty of such paired data experiments exist [108], the data has been scattered in the various genomic and metabolomic databases with no clear way of systematically identifying corresponding entries. Furthermore, as no standardised way has existed to specify the correspondence between BGCs and spectra, re-establishing the link within the data set has required considerable effort.

To improve the situation, Schorn and co-workers recently developed the *Paired omics Data Platform* (PoDP) [108]. The purpose of the platform is to document the location of microbial genomic and metabolomic data sets, in particular data sets with information about BGCs and MS2 spectra, which entries in the genomic databases correspond to which entries in the metabolomic databases, along with information on which exact BGCs are linked to which exact spectra or MFs. This gives ground truth data which can be used to compare both feature-based and correlation-based scoring functions as described above.

From the PoDP, three data sets in particular were selected for evaluation: MSV000078836 [121], MSV000085018 [122] and MSV000085038 [123], hereafter referred to as the Crüsemann, Gross and Leão data sets, respectively. As the platform has only recently been developed, the choice of data sets was in part dictated by the presence of data on the platform, and in part by how easily the reported BGC-spectrum links could be resolved to the genomic and metabolomic data, i.e. if the validated links had already been entered into the platform by the submitter.

Table 3.2 shows the sizes of the data sets obtained from the PoDP, along with the types of links included, while Table 3.3 shows the natural product class of the validated links in each data set, along with the BGC sizes.

3.8.1 Mapping validated links to detected BGCs

The links from the PoDP are characterised by MIBiG IDs on the genomic side, and IDs of either individual spectra or MFs on the metabolomic side. As long as the molecular networking data set used for the analysis is provided, resolving the metabolomic IDs to the underlying data is a relatively simple matter.

On the genomic side, the links are defined by BGCs in specific strains, characterised by MIBiG IDs. After accessing the sequence for the strains, and running antiSMASH to detect the BGCs, the problem remains to identify which of the BGCs detected by antiSMASH is the annotated BGC from the link. This can be done in either of two ways (see Fig. 3.1):

- *Using BiG-SCAPE.* When BiG-SCAPE clusters the BGCs into GCFs, it can include the MIBiG reference BGCs in the clustering. These get assigned to GCFs, and can then be used to tag the corresponding BGCs. If a BGC from the strain is in a GCF that includes the MIBiG entry involved in the link, then that BGC can be assumed to correspond to the BGC associated with the link.

- *Using antiSMASH known cluster BLAST.* When running antiSMASH to detect BGCs, the similarity of the detected clusters to MIBiG entries can be evaluated, by using the so-called *known cluster BLAST* mode (KCB). For each MIBiG entry, this aligns the sequence of each Pfam domain in the entry to the detected BGC using BLAST [116], quantifying how similar the detected BGC is to the MIBiG BGC. The outcome can be evaluated as either the cumulative BLAST score, or the average similarity for domains in the MIBiG BGC where there is significant similarity. This can be used to assign each detected BGC zero or more MIBiG matches, possibly with a threshold on the score to reduce noise, and the strain involved in the link can then be searched for BGCs that show high similarity to the MIBiG ID implicated in the link.

Using BiG-SCAPE to assign links from the PoDP to GCFs is not without risks. As BiG-SCAPE assigns MIBiG labels by embedding the MIBiG database in the data set being processed, and clustering the database entries along with other BGCs, each MIBiG entry belongs to exactly one (possibly singleton) GCF. If BiG-SCAPE happens to split a GCF in two (for instance because of a bad choice of parameters for the clustering algorithm) only one of the GCFs gets tagged by the MIBiG entry. This can be seen in action for instance in the Crüsemann data set, where antiSMASH identifies far more BGCs as having significant homology to MIBiG ID BGC0000827 (staurosporine) than BiG-SCAPE places in the corresponding GCF. This holds true for all cutoff values tested in BiG-SCAPE, indicating a potential discrepancy that merits further investigation.

For GCFs associated with MIBiG IDs, antiSMASH can in fact be used as an alternative strategy to define GCFs, by taking all BGCs that have a significant KCB hit for the MIBiG ID to belong to the GCF. In theory, this should be equivalent to using BiG-SCAPE to cluster the BGCs, including MIBiG BGCs, and taking the relevant GCF to be the one including the particular MIBiG BGC.

Given a MIBiG BGC, the corresponding GCF as defined by BiG-SCAPE should contain BGCs that are similar to each other, where similarity is evaluated using the similarity func-

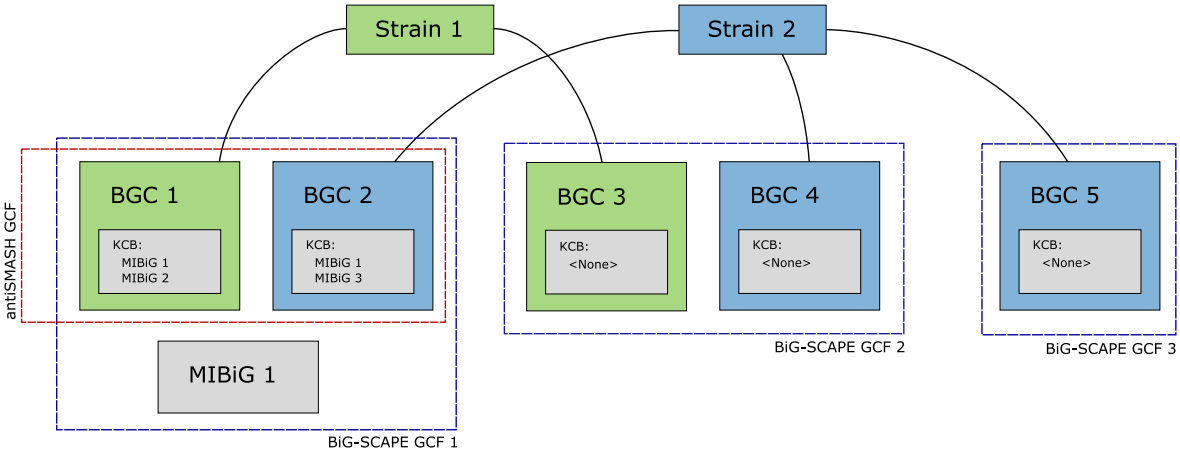


Figure 3.1: **Using antiSMASH and BiG-SCAPE to tag GCFs and BGCs.** Strain 1 contains BGCs 1 and 3, while Strain 2 contains BGCs 2, 4 and 5. BiG-SCAPE clusters BGCs 1 and 2 into a GCF, 3 and 4 into another GCF, and BGC 5 into a GCF of its own. antiSMASH tags BGC 1 as similar to MIBiG 1 and 2, and BGC 2 as similar to MIBiG 1 and 3. BGCs 1 and 2 can be assembled into a GCF corresponding to MIBiG 1 by either considering the inclusion of MIBiG 1 in the GCF by BiG-SCAPE, or the fact that antiSMASH annotates both as similar to MIBiG 1.

tion defined by BiG-SCAPE. Let $\sim$ denote similarity relation between BGCs, i.e. we say that a and b are similar if $a \sim b$. If we assume that the similarity relation is transitive, i.e. $a \sim b$ and $b \sim c$ implies $a \sim c$ for BGCs a, b, and c, then all of the BGCs in the GCF are similar to the MIBiG BGC.[1] This is exactly the condition that would be used to define the corresponding GCF using antiSMASH, albeit using a different similarity function, so assuming the similarity functions used by BiG-SCAPE and antiSMASH show similar behaviour, the resulting GCFs should generally be identical.

However, as described previously, this is not always the case.

When BiG-SCAPE clusters BGCs, it does so separately for each natural product class, as described in Section 2.2.2. When dealing with BGCs that do not have an unambiguously assigned natural product class, such as hybrid PK-NRP BGCs, they are classified three times: once as PK, once as NRP and once as hybrid. After clustering, identical GCFs resulting from different product classes are removed, so that a given BGC of the "PKS-NRP Hybrid" product class can belong variously to one, two or three GCFs.

For instance, the MIBiG ID BGC0000137 has a validated link against two MFs in the Crüsemann data set. Two BGCs from the relevant strain show considerable homology to the

[1] Obviously this is not generally true if we consider for instance similarities as inverse distances in $\mathbb{R}^n$, with $a \sim b$ if the similarity exceeds a given threshold. By considering a, b and c as points on a line in $\mathbb{R}^n$, it is easy to envision a case where $a \sim b$, $b \sim c$ but $a \not\sim c$. However, in our case we can justify this assumption by observing that the space of metabolites is discrete, and we are trying to cluster things that are relatively similar, i.e. in general we expect the intra-cluster similarity to be much higher than the inter-cluster similarity. We can therefore define the relation $\sim$ in such a way that the transitivity holds.

	Crüsemann	Leão	Gross
ID	MSV000078836	MSV000085038	MSV000085018
Strains	118	4	7
Links	8	5	9
Type of links	BGC-MF	BGC-spectrum	BGC-MF
BGCs	3316	147	131
BGCs w. structure	2242	57	83
GCFs	454	110	120
GCFs w. structure	323	54	77
Spectra	6246	173	9593
MFs	3094	90	6518

Table 3.2: Size and properties of the microbial data sets. A subset of BGCs are assigned molecular structures using antiSMASH, and correspondingly a subset of GCFs are are assigned molecular structures from their BGCs.

MIBiG BGC, according to antiSMASH. As BGC0000137 belongs PKS-NRP Hybrid natural product class, each of these gets clustered in three different natural product classes. After merging identical GCFs, one of the BGCs ends up in three GCFs, while the other one ends up in two. The result is five distinct GCFs being linked to two distinct MFs, for a total of 10 validated links.

3.8.2 Extending validated links with similarity matching

For many applications, especially for prioritising potential links for verification, similarity searches have been used: a spectrum that closely resembles that of a known metabolite in a database, and a BGC that has a high similarity to a BGC that is known to produce the metabolite from the database, are assumed to be linked [124, 125, 61]. This would present a way of expanding the set of validated links in the microbial data sets.

However, since there is no way of knowing how exhaustive the coverage of the set of validated links is, and in particular no way of knowing if all true links have been discovered, adding the similarity-based links to the set of validated links would not change the fact that the only method of comparison between two link ranking approaches would be the relative enrichment of validated samples in the top quantiles. Expanding the set of validated links in this way and would therefore not be of great value, while increasing the danger of incorrect links being classified as validated, due to possible false positives in the similarity search. We therefore choose to only use experimentally validated links in the subsequent work.

Crüsemann	Size (nt)	Type	Genes
BGC0000827	24300	Alkaloid	17
BGC0000333	47477	NRP	23
BGC0000940	7328	Other	6
BGC0000241	62231	Polyketide (Other), Saccharide (hybrid/tailoring)	58
BGC0001228	37781	NRP (Cyclic depsipeptide)	23
BGC0000137	91573	Polyketide	39
BGC0001830	64171	Polyketide	23
Leão	**Size (nt)**	**Type**	**Genes**
BGC0001165	87427	NRP, Polyketide (other)	14
BGC0000962	40156	NRP, Polyketide (other)	12
BGC0001000	41964	NRP (Other lipopeptide), Polyketide (other)	8
BGC0001001	69900	NRP, Polyketide (other)	26
BGC0001560	28792	NRP, Polyketide	12
Gross	**Size (nt)**	**Type**	**Genes**
BGC0000632	16122	Terpene, Saccharide	13
BGC0001381	210303	Polyketide	102
BGC0001842	42814	NRP (other lipopeptide)	7
BGC0000463	43207	NRP (other lipopeptide)	4
BGC0001116	57999	NRP, Polyketide	18
BGC0000399	47020	NRP (cyclic depsipeptide)	17
BGC0000296	66086	NRP	28
BGC0001298	41001	Polyketide	9

Table 3.3: Natural product classes of the validated links in the microbial data sets.

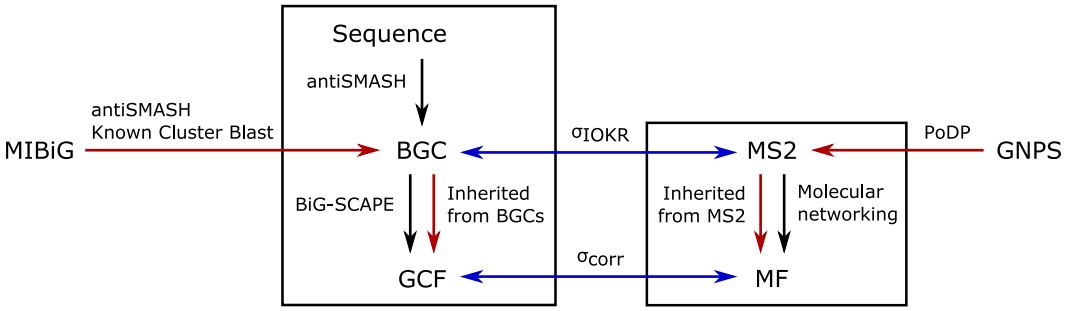


Figure 3.2: **Assigning structures to BGCs and spectra.** As well as describing the annotation of the data set discussed in Section 3.7, this diagram shows the processing of the data discussed in subsequent chapters. Black arrows indicate processing of data, red arrows indicate assignment of molecular structures to objects, and blue arrows indicate scoring functions. Structures are assigned to BGCs from MIBiG entries using antiSMASH, and to MS2 spectra using PoDP, while GCFs inherit their structures from their component BGCs and MFs from their component MS2 spectra. The scoring functions σ_{corr} and σ_{IOKR} are discussed further in Chapters 4 and 5.

3.9 Summary

In this section, we have defined the two types of data sets that will be used for the verification of the methods presented in this thesis, and discussed how these data sets will be used to evaluate the performance of the proposed methods.

We have created a new data set of BGC-spectrum links by joining the MIBiG and GNPS databases, using annotations of molecular structure present in both databases. This gives us a set of thoroughly validated links, albeit without the context of strain information.

Finally, using the PoDP, we have collected a set of three microbial experiments, each containing genomic and metabolomic results for a different population of strains, along with several validated links for each experiment.

Chapter 4

Strain correlation scoring

4.1 Introduction

If an organism produces a particular specialised metabolite, the production of the metabolite must be somehow encoded within its genome. In particular, since microbial secondary metabolism is most commonly controlled by BGCs, any microbe that produces a given specialised metabolite can be expected to contain a specific BGC encoding the production of that metabolite.

This chapter is concerned with the use of patterns of metabolite production and patterns of genomic similarities across microbial strains to establish correspondence between BGCs and spectra. Section 4.2 introduces a common method for such correlation and discusses its benefits and shortcomings. Two shortcomings in particular are discussed: the difficulty in comparing correlation of different objects, and the tendency of the method to prefer BGCs shared between strains because of common ancestry.

Section 4.3 derives the formula for calculating the statistical significance of such approaches.

The last two sections of the chapter are concerned with correcting the two problems pointed out in Section 4.2. Section 4.4 demonstrates how standardising the strain correlation score fixes the problem of comparing objects of different sizes, while Section 4.5 investigates the effect of shared ancestry on the score and suggests a way to mitigate this effect.

4.2 Metabologenomics

When working with a population of (closely related) microbes, this offers a way of ranking prospective BGC-spectrum links according to the patterns of BGC similarities on one hand, and metabolite expression on the other: after clustering the BGCs from the strains in the

population into GCFs, and possibly the spectra into MFs, each one can be considered as a set of the strains making a contribution to the GCF and MF. For a given GCF-MF pair, considerable overlap in the strain sets would constitute evidence that the GCF is responsible for the production of the metabolite.

This approach, termed *metabologenomics*, has been used to good effect, for instance in [105] to identify the metabolite tambromycin, in [63] to identify the metabolite retimycin A, and in [126] to identify the tyrobetaine class of metabolites, and to link each of them to their producing BGC.

As mentioned in Section 2.3.2, metabologenomics can be considered a close relative of GWAS, which have been a mainstay of human genomics for a long time [127], and have seen increased use in microbial research in recent years [107]. Broadly speaking, GWAS aims to correlate patterns of genotypes to patterns of phenotoypes by presence or absence [106]. In recent years, microbial GWAS (mGWAS) approaches have been developed to manage problems particular to correlating phenotypes and genotypes for microbial data sets, such as the impact of the evolutionary relationship between the strains [107]. While mGWAS is closely related to the problem of correlating BGCs and MS2 spectra, the asymmetry introduced by cryptic BGCs compounds the problem. mGWAS approaches are generally symmetrical, treating the absence of a phenotype in the presence of a genotype equivalently to the presence of a phenotype in the absence of a genotype. This is in contrast to the desired behaviour of a scoring function for BGC-MS2 links, where the absence of a spectrum (phenotype) in the presence of a BGC (genotype) can merely imply that the BGC is cryptic, and should therefore be far less harshly penalised than the presence of the spectrum in the absence of the BGC. Since a large proportion of microbial BGCs are in fact cryptic this severely limits the applicability of mGWAS tools to the problem of linking BGCs and MS2 spectra [23, 29, 72].

In mathematical notation, given sets of strains N, GCFs $\mathcal{G}$ and MFs $\mathcal{M}$, each GCF and MF can be considered as the set of strains that have a BGC in the GCF, or a spectrum in the MF, i.e. for all GCFs $G \in \mathcal{G}$ and MFs $M \in \mathcal{M}$, we can consider them as subsets of N, $G \subseteq N$ and $M \subseteq N$. We can then use the overlap $M \cap G$ and the set differences $G \setminus M$ and $M \setminus G$ to evaluate how likely it is that the GCF G is responsible for the production of the metabolite corresponding to M. We denote the *size*, or *cardinality*, of a set M by $\#M$.

The most common scoring function in metabologenomics is defined in [84]. For a given GCF-MF pair, the score is calculated by starting from zero and considering each strain in turn:

- add 10 to the score if the strain is in both the GCF and the MF,
- add 1 to the score if the strain is neither in the GCF nor the MF,

- subtract 10 from the score if the strain is in the MF but not the GCF,
- leave the score unchanged if the strain is in the GCF but not the MF.

In subsequent discussion, we refer to this scoring function as *strain correlation score* or σ_{corr}, writing $\sigma_{\text{corr}} : \mathcal{M} \times \mathcal{G} \to \mathbb{R}$, where $\mathcal{G}$ is the set of all GCFs and $\mathcal{M}$ the set of all MFs. Considering the GCF as the set G of strains that contribute a BGC to the GCF, and the MF as the set M of strains in which a spectrum in the MF is observed, σ_{corr} depends on $\#G$, $\#M$, the size $\#(M \cap G)$ of their overlap, and the total size of the population, $\#N$, as well as a set of weights α, β, γ and δ, here set to 10, -10, 1 and 0. In mathematical notation, it can be written as

$$\sigma_{\text{corr}} = \alpha\#(M \cap G) + \beta\#(M \setminus G) + \gamma\#(N \setminus (G \cup M)) + \delta\#(G \setminus M). \tag{4.1}$$

Leaving aside the exact choice of weights, this scoring scheme makes intuitive sense. It highly rewards overlap in the strain set, severely penalises strains that produce the metabolite but do not contribute to the GCF, slightly rewards shared absences and slightly penalises (relatively) non-expressed BGCs.

The first two choices are fairly obvious, but other two are also sensible. In particular, shared absence is mildly indicative that the prospective link is true, while a particular strain contributing to the GCF but not the MF is not strong evidence against the link, since the BGC might be cryptic or not strongly expressed enough to be detected in the MS2 data.

Given a population N, a GCF G and MF M, and assuming that G and M are independent, the size of the overlap $\#(M \cap G)$ follows a hypergeometric distribution with the total number of strains $\#N$ as the population size, the size of the molecular family $\#M$ as the number of positive cases in the population, the size of the GCF $\#G$ as the number of draws from the population, and the size of the overlap $\#(M \cap G)$ as the number of positive cases in the draw.

One way in which this independence assumption can be violated is if the GCF G is actually encoding for the production of the metabolite M, in which case the overlap between the two would be higher than expected. This is what the strain correlation score is intended to detect. However, the strain correlation score is highly dependent upon GCF and MF sizes, as further discussed in Section 4.4.

This independence assumption is also affected by *strain phylogeny*, i.e. how closely related the strains are in evolutionary terms. If the strains are closely related, they would be expected to have similar metabolism, and therefore show similar patterns of presence or absence in GCFs and MFs, that is, if a strain is present or absent in a particular GCF or MF, a closely related strain is more likely to show the same pattern of presence or absence than a distantly

related one. The effect of strain phylogeny on the correlation score is discussed in Section 4.5.

To fix the former, we propose standardising the score to reduce the impact of the sizes of the elements under consideration, and to fix the latter we propose taking strain phylogeny into account when calculating the score.

The next section is dedicated to calculating the statistical significance of the strain correlation score, while the following sections explain in greater detail our proposed solutions to each of the problems mentioned above.

4.3 Significance tests for strain correlation scores

When working with the strain correlation scoring function for a given GCF-MF pair, a reasonable question to ask is what the p-value of a given score is, i.e. what is the probability of the pair getting assigned this score or a better one, assuming that the GCF and MF are independent. Recall that for a population N of strains, a GCF G and an MF M with $\#N = n$, $\#M = m$, $\#G = g$ and $\#(M \cap G) = o$, the strain correlation function σ_{corr} is defined in Eq. 4.1 as

$$\sigma_{\text{corr}}(M, G) = \alpha\#(M \cap G) + \beta\#(M \setminus G) + \gamma\#(N \setminus (M \cup G)) + \delta\#(G \setminus M) \tag{4.2}$$

$$= \alpha o + \beta(m - o) + \gamma(n - (m + g - o)) + \delta(g - o) \tag{4.3}$$

where the weights α, β, γ and δ are set at $\alpha = 10$, corresponding to the strain being in both the GCF and the MF, $\beta = -10$, corresponding to the strain being in the MF but not the GCF, $\gamma = 1$, corresponding to the strain being in neither the GCF nor the MF, and $\delta = 0$, corresponding to the strain being in the GCF but not the MF.

For the strain correlation score σ_{corr} we can define a corresponding function $\sigma'_{\text{corr}} : \mathbb{N}^4 \to \mathbb{R}$, which depends on the cardinality of the sets. For a MF $M \in \mathcal{M}$ and GCF $G \in \mathcal{G}$, and setting $m = \#M$, $g = \#G$, $o = \#(M \cap G)$ and $n = \#N$, we define σ'_{corr} as

$$\sigma'_{\text{corr}}(m, g, n, o) = \sigma_{\text{corr}}(M, G). \tag{4.4}$$

By making use of the fact that given a GCF G and an MF M, the cardinalities m and g are constant, the strain correlation score $\sigma_{\text{corr}}(M, G) = \sigma'_{\text{corr}}(m, g, o, n)$ is determined by the cardinality of the overlap o, which follows the hypergeometric distribution as previously stated. Given an MF M and a GCF G, we can therefore define a second scoring function

$\hat{\sigma}_{\text{corr}}(o)$ which depends only on the overlap o. This can be used to calculate the p-value of a given score.

Given a GCF G and an MF M, and given a value s, we want to calculate $p(\sigma_{\text{corr}}(M,G) > s)$, assuming that M and G are independent. Since $\#M$ and $\#G$ are given, we have

$$p(\sigma_{\text{corr}}(M,G) = s) = \sum_{o \in \mathbb{N}} p(\#(M \cap G) = o) p(\hat{\sigma}_{\text{corr}}(o) = s) \tag{4.5}$$

We can then calculate the probability that the score is higher than s, i.e. the p-value, by calculating

$$\begin{aligned} p(\sigma_{\text{corr}}(M,G) > s) &= \sum_{s' > s} p(\sigma_{\text{corr}}(M,G) = s') && (4.6)\\ &= \sum_{s' > s} p(\sigma_{\text{corr}}(M,G) = s' \mid \#M, \#G, \#N) && (4.7)\\ &= \sum_{s' > s} \sum_{o \in \mathbb{N}} p(\#(M \cap G) = o) p(\hat{\sigma}_{\text{corr}}(o) = s). && (4.8) \end{aligned}$$

Since $\hat{\sigma}_{\text{corr}}$ is completely determined by o, the expression $p(\hat{\sigma}_{\text{corr}}(o) = s)$ is always either 1 or 0, and equals one exactly for those values of o where $\hat{\sigma}_{\text{corr}}(o) = s'$.

Since we assume M and G to be independent, the first factor of Eq. 4.8,

$$p(\#(M \cap G) = o) = p(o \mid n, m, g), \tag{4.9}$$

follows hypergeometric distribution, with the population size n, the size of the MF, m, as the number of 'successes' in the population, the size of the GCF, g, as the number of draws, and the overlap, o, as the number of 'successes' in the draw.

The two sums of Eq. 4.8 can therefore be taken together as

$$p(\sigma_{\text{corr}}(M,G) > s) = \sum_{o \mid \hat{\sigma}(o) > s} p(o \mid m, g, n) \tag{4.10}$$

where $p(o \mid m, g, n)$ follows hypergeometric distribution.

To calculate the sum in Eq. 4.10 requires calculating the scores for a range of possible overlaps, in order to determine which of them give a score higher than s. This is made manageable by the fact that the range of values that o can take is limited by m and g, the sizes of the MF and GCF, as well as n, the total population size. In fact, the lower bound of o is $o_{\text{min}} = \max(0, (m+g) - n)$ and the upper bound of o is $o_{\text{max}} = \min(m, g)$.

While calculations carried out in this Section are done using σ_{corr} as defined in Eq. 4.1, they generalise to any scoring function that can be expressed as a function of $\#M$, $\#G$, $\#(M \cap G)$ and $\#N$. In particular, the same equations can be used to calculate the significance level for the standardised strain correlation function defined in Section 4.4, substituting $\bar{\sigma}_{\text{corr}}$ for σ_{corr}.

4.4 Standardising the strain correlation score

The strain correlation score is highly dependent not only on population size, but also the size of the GCF and MF being considered at each time. This makes not only comparison between links from different data sets difficult, but even within the same data set, comparing two links containing GCFs and MFs of different sizes does not always yield the expected results.

Figure 4.1 (A) shows two GCF-MF pairs, where each box represents a strain in the population, and shaded boxes indicate that the strain has a BGC in the GCF or a spectrum in the MF. Since the pattern of strain membership for the GCF and the MF in the lower example is exactly the same, the evidence for that link is arguably stronger than for the top link. However, the strain correlation score assigns the top link a score of 30, and the bottom link 26, ranking the top link higher than the bottom link. In general, a link containing a large GCF and a moderately-sized MF can easily outscore a link between a smaller GCF and MF, even if the pattern of strain membership for the smaller GCF and MF is much more similar, and the evidence for the link therefore much stronger.

By making use of the fact that given a GCF G and an MF M, the cardinalities $m = \#M$ and $g = \#G$ are constant, the strain correlation score $\sigma_{\text{corr}}(M, G) = \sigma'_{\text{corr}}(m, g, o, n)$ is determined by the size of the overlap $o = \#(M \cap G)$. Assuming that M and G are independent, o follows the hypergeometric distribution as previously stated.

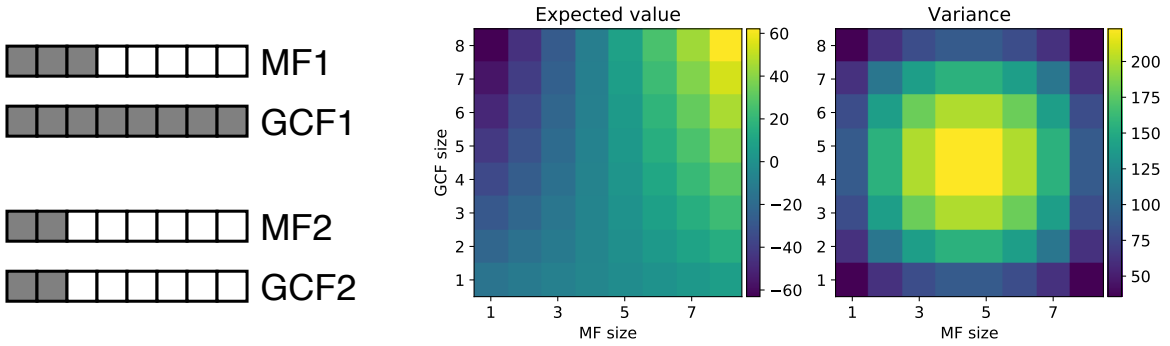
For given sizes of G and M, and assuming their independence, we can therefore compute the expected value $E[\sigma_{\text{corr}}(M, G)]$ of the strain correlation score as

$$E[\sigma_{\text{corr}}(M, G)] = \sum_k p(o = k)\sigma'_{\text{corr}}(m, g, k, n) \tag{4.11}$$

with k running over all possible values of the overlap o and where $p(o = k)$ is the probability of the overlap o being of size k, which follows the hypergeometric distribution as previously stated.

The variance $Var[\sigma_{\text{corr}}(M, G)]$ can then be computed as

$$Var[\sigma_{\text{corr}}(M, G)] = E[\sigma_{\text{corr}}(M, G)^2] - E[\sigma_{\text{corr}}(M, G)]^2. \tag{4.12}$$



(A) *(B)*

Figure 4.1: **The effect of GCF and MF size on the strain correlation score.** *(A)* Example of the impact of GCF and MF sizes on link score. Each box represents a strain, with the filled boxes indicating the presence of the strain in the GCF or MF. Using the unmodified version of the strain correlation score, the top pair is assigned a score of 30, while the bottom pair is assigned a score of 26. This is in contradiction to the obviously closer correspondence between the GCF and MF in the bottom pair. *(B)* The expected value and variance of the strain correlation score for varying sizes of GCF and MF, in a population of size 8, as in the example in *(A)*. Comparison between links involving GCFs and MFs of different sizes are virtually meaningless, because both expected value and variance have a considerable range. For instance, a GCF and MF of size 6 could easily get a score of 40 or higher by chance, while for a GCF and MF of size 2, a score of 40 would be highly significant.

The expected value and variance of σ_{corr} as a function of m and g can be seen in Figure 4.1 (B) for a population of size $n = 8$, as in the example in Figure 4.1 (A). Both show considerable variation across the range, with the expected value being particularly variable for larger GCF sizes.

To reduce the size effect on the strain correlation score, we propose a *standardised strain correlation score*, $\bar{\sigma}_{\text{corr}}(M, F)$, defined as

$$\bar{\sigma}_{\text{corr}}(M, F) = \frac{\sigma(M, G) - E[\sigma(M, G)]}{Var[\sigma(M, G)]}. \tag{4.13}$$

This has the effect of setting the expected value to 0 and the variance to 1 across the whole domain of the function, negating the size effect of the MF and the GCF.

Taking the synthetic example in Figure 4.1 (A), we can calculate σ_{corr} and $\bar{\sigma}_{\text{corr}}$ for both potential links. Let m_1, m_2, g_1 and g_2 be MF1, MF2, GCF1 and GCF2, respectively. For both links, we have $n = 8$, and for the link between MF1 and GCF1 we have $m = 3$, $g = 8$ and $o = 3$, while for the link between MF2 and GCF2, we have $m = 2$, $g = 2$ and $o = 2$. For the strain correlation score σ_{corr}, we therefore have, by Eq. 4.3,

$$\sigma_{\text{corr}}(m_1, g_1) = 10 \times 3 - 10 \times 0 + 1 \times 0 = 30, \tag{4.14}$$

$$\sigma_{\text{corr}}(m_2, g_2) = 10 \times 2 - 10 \times 0 + 1 \times 6 = 26. \tag{4.15}$$

To calculate $\bar{\sigma}_{\text{corr}}(m_1, g_1)$ and $\bar{\sigma}_{\text{corr}}(m_2, g_2)$, we calculate $E[\bar{\sigma}_{\text{corr}}(m_1, g_1)]$, $E[\bar{\sigma}_{\text{corr}}(m_2, g_2)]$, $Var[\bar{\sigma}_{\text{corr}}(m_1, g_1)]$ and $Var[\bar{\sigma}_{\text{corr}}(m_2, g_2)]$ using Eqns. 4.11 and 4.12, giving

$$E[\bar{\sigma}_{\text{corr}}(m_1, g_1)] = 30, \tag{4.16}$$

$$Var[\bar{\sigma}_{\text{corr}}(m_1, g_1)] = 1, \tag{4.17}$$

$$E[\bar{\sigma}_{\text{corr}}(m_2, g_2)] = -5.5, \tag{4.18}$$

$$Var[\bar{\sigma}_{\text{corr}}(m_2, g_2)] = 141.75. \tag{4.19}$$

Using the calculated expected value and variance, by Eq. 4.13, we get

$$\bar{\sigma}_{\text{corr}}(m_1, g_1) = 0, \tag{4.20}$$

$$\bar{\sigma}_{\text{corr}}(m_2, g_2) = 2.645, \tag{4.21}$$

giving a higher standardised strain correlation score, $\bar{\sigma}_{\text{corr}}$, for the lower example in Figure 4.1 *(A)*, as intended.

Results on a microbial data set To observe the effect of standardisation more generally, we can consider the relative abundance of verified links among all potential links for the three data sets introduced in Section 3.7. In particular, we can pay attention to the Crüsemann data set, as that has the highest number of strains, which means that out of the three, the strain correlation score gives the richest information for that data set.

An ideal scoring function would assign all true links a score higher than all incorrect links. Failing that, in general, true links should score higher than incorrect ones. Even for an imperfect scoring function, since the vast majority of the potential links are incorrect, the mean score of true links should be higher than the mean score of all links. The same should be true for the subset of validated links. Tables 4.1 and 4.2 show the mean score of the validated links compared to the mean score of all links within the data set for the three microbial data sets, along with the p-value for t-test using the null hypothesis that the distributions of all potential links, on the one hand, and validated links, on the other hand, have identical means.

	Mean σ_{corr} all	Mean σ_{corr} valid	p-value
Crüsemann	83.5144	14.6667	0.0001
Leão	-1.9843	12.625	0.0001
Gross	-0.7386	0.6	0.7929

Table 4.1: Mean raw correlation score for links in the microbial data sets

	Mean $\bar{\sigma}_{\text{corr}}$ all	Mean $\bar{\sigma}_{\text{corr}}$ valid	p-value
Crüsemann	-0.0060	3.6717	6.8302e-64
Leão	-0.0218	1.4962	1.1887e-05
Gross	0.0092	1.6149	6.0056e-06

Table 4.2: Mean standardised correlation score for links in the microbial data sets

Considering first the raw strain correlation score σ_{corr} displayed in Table 4.1, the mean score for validated links in the Crüsemann data set is actually lower than the mean score for all links, underscoring the impact of the size of the GCFs and MFs on the scores. While the mean score for the validated links is higher than the mean score for all links in the other two data sets, the difference is only statistically significant for one of the data sets.

Turning to the standardised strain correlation score $\bar{\sigma}_{\text{corr}}$ displayed in Table 4.2, the mean score for the validated links is significantly higher than the mean score for all links in all of the data sets.

Another way of considering the relative performance of link scoring function is to observe the relative enrichment of validated links in the upper percentiles of the scoring functions, as described in Section 3.7. Table 4.3 shows the proportion of verified links compared to all links, first among all potential links, and then filtered to links scoring above the 90th percentile for the raw and standardised correlation scores. Figure 4.2 shows the distribution of scores for the potential links for the three data sets, with the verified links indicated in black below the histograms.

As can be seen in both the table and the figure, the verified links are more concentrated towards the higher end of the distribution using the standardised strain correlation score than

	Crüsemann	Leão	Gross
Total	1.50×10^{-5}	0.00086	9.96×10^{-6}
$> 90\%$ σ_{corr}	6.22×10^{-5}	0.00266	3.48×10^{-5}
$> 90\%$ $\bar{\sigma}_{\text{corr}}$	0.00013	0.00385	7.69×10^{-5}

Table 4.3: Proportion of links that are validated in the whole data set, and scoring above the 90th percentile for the raw and standardised correlation scores.

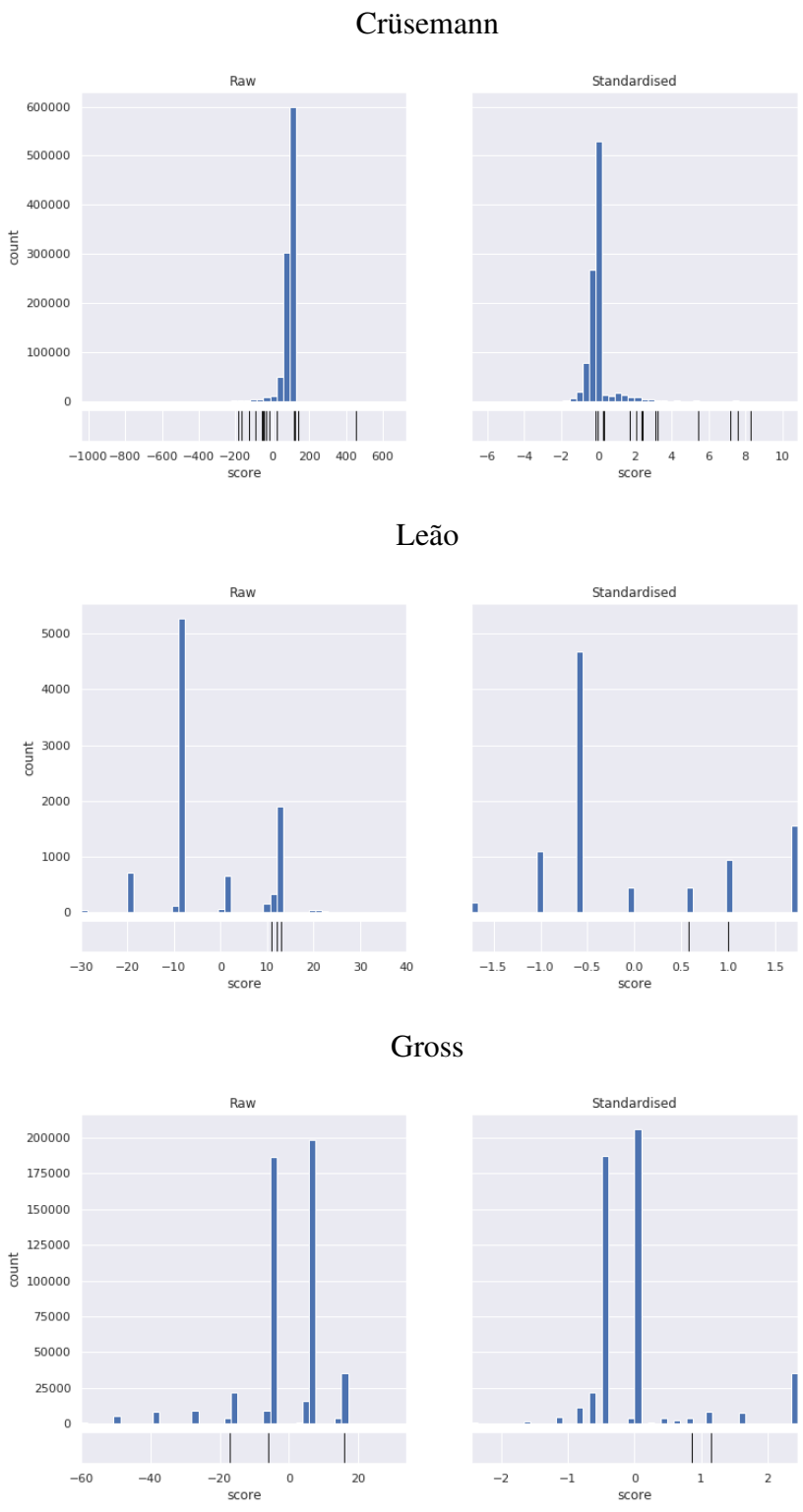


Figure 4.2: **Standardised vs. raw strain correlation score.** Distribution of standardised and raw strain correlation scores for the three microbial data sets. Verified links are indicated in black below the histogram.

the raw strain correlation score. This demonstrates how standardising the strain correlation score makes it more useful to rank potential GCF-MF links for further verification.

4.5 Phylogenetic correction for the strain correlation score

We now turn our attention to a more subtle property of the strain correlation score. While the previous sections were concerned with correcting a fundamental flaw in the strain correlation score, the impact of the evolutionary context between strains, or *strain phylogeny*, on the score is not a flaw in the score itself, but rather a question of trying to emphasise links of particular interest, that might otherwise be drowned out.

The evolutionary relationship of organisms is quantified using *phylogenetics*, where observed traits of the organisms, such as genomic sequence, are used to infer the evolutionary relation between the organisms [128, 129]. The degree of similarity between organisms is often referred to as *phylogenetic distance* [129].

The two major factors involved in shared genetic traits between microbes are common ancestry, where two microbes share an ancestor, and *horizontal* (or *lateral*) *gene transfer* (HGT), where a microbe acquires a BGC from another microbe in the environment [128]. The acquisition of a BGC by HGT can indicate that there is evolutionary pressure to acquire the trait encoded by the BGC [130]. As this has been observed for BGCs responsible for e.g. bactericidal activity or antimicrobial resistance, BGCs acquired by HGT can in some cases be of particular interest [131, 132].

If two strains are closely related in evolutionary terms, their genomes are generally more similar than those of more distantly related strains. This means they are more likely to share BGCs, and therefore end up in the same GCFs. This can lead the strain correlation score to reward GCF-spectrum links where closely related strains co-occur in the GCF.

Consider a population consisting of three strains A, B and C, where two of the strains, A and B, are closely related, and the third, C, has a larger phylogenetic distance from the other two. Because of their closely-shared ancestry, A and B have in common a BGC which encodes a particular metabolite, while C has an unique BGC which encodes a different metabolite. For the links between the GCFs and the corresponding metabolites, the former link would be assigned a higher strain correlation score than the latter, based mostly on the fact that two of the strains are phylogenetically more similar than the others.

In some cases, this bias in the strain correlation score caused by phylogenetic distance could cause the signal for BGCs shared because of common ancestry to drown out the signal for

BGCs shared because of HGT. Since BGCs acquired by HGT may be of particular interest, as alluded to earlier, this can potentially be a problem.

While mGWAS approaches, such as those discussed in Section 4.2 have been developed precisely to take into account factors such as phylogeny when doing association studies of this sort, these have limited applicability to the problem at hand in their current form because of the asymmetry introduced by cryptic BGCs, as discussed in Sections 2.2.1 and 4.2.

By incorporating the phylogenetic distance between strains into the strain correlation score, we hope to decrease this phylogenetic bias towards GCFs and MFs containing closely-related strains. We propose to do this by considering the cases where two phylogenetically related strains show the same pattern of membership with respect to a particular GCF and MF as no more informative than if we had a single strain showing the same pattern of membership. In doing so we construct a new, synthetic collection of strains by starting with the actual collection of strains and, taking into account the phylogenetic relationship between them, assigning the new, synthetic strains GCF and MF memberships according to the original strains.

We now describe the method to generate this new synthetic data set.

In keeping with standard graph theory terminology, we consider a *graph* to be a tuple (V, E) of sets, with V a set of *nodes* (or *vertices*) and E of *edges*, and with each edge in E connecting two nodes in V.[1] The *degree* of a node is the number of edges containing that node. A *path* between two nodes a and b is an ordered sequence of edges $\{e_i\}_0^n$, where e_i is the edge between nodes $v_{i,0}$ and $v_{i,1}$, where $v_{0,0} = a$, $v_{n,1} = b$, and $v_{i,1} = v_{i+1,0}$ for $i \in \{0, 1, \ldots, n-1\}$, and no edges are repeated. The *distance* between two nodes is the shortest path between them.

A *tree* is a graph such that between any two nodes there exists exactly one path. A node in a tree is called a *leaf node* if it has degree one. Two leaf nodes in a tree are *siblings* if they each have an edge connecting them to the same *parent node*. This is equivalent to the distance between the nodes being equal to two.

The phylogenetic relationship between multiple organisms can be expressed as a tree, where the leaf nodes of the tree represent the organisms being analysed, and the branches of the tree represent evolutionary divergence between them [128, 129]. In this setting, the distance between two leaf nodes in the tree can be though of as a marker of the phylogenetic distance between the corresponding organisms, i.e. how evolutionarily related they are relative to the other organisms in the tree.

For an overview of phylogenetic trees in the context of natural products, the reader is referred to [133], while a more detailed explanation of the general principles of phylogenetic tree

[1] The graphs considered here are *undirected*, i.e. the order of the tuples of nodes that make up an edge does not matter, and we take them to be presented in the order most convenient at any given time.

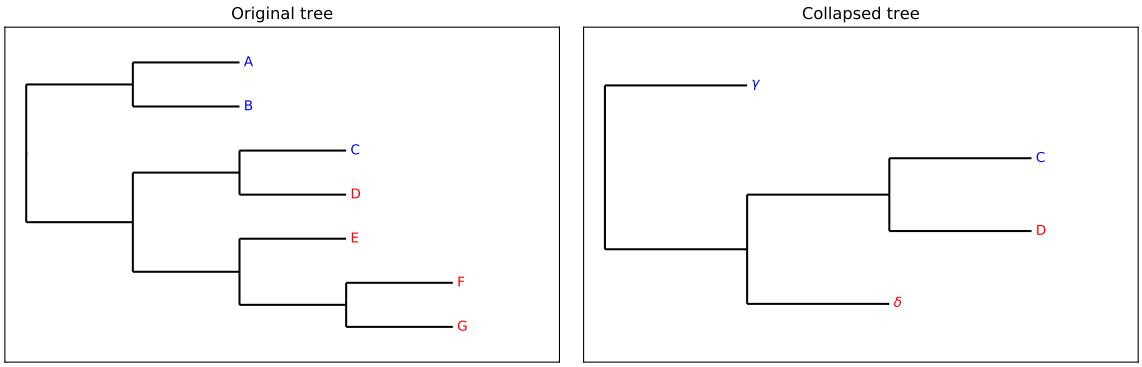


Figure 4.3: **Collapsing the phylogenetic tree.** The strains labelled A, B and C on the initial tree (left) have the same phenotype with regards to membership patterns in a particular GCF and MF, as do the strains labelled D, E, F and G. On the collapsed tree (right), strains A and B have been collapsed into the synthetic strain γ, with the same phenotype as A and B, and strains E, F and G have been collapsed into the synthetic strain δ by first collapsing strains F and G, and then collapsing the resulting synthetic strain with strain E. The synthetic strain δ has the same phenotype as strains E, F and G. Strains C and D are not collapsed, since they do not have identical phenotype, and neither γ nor δ are further collapsed since neither has a sibling with the same phenotype.

building can be found in [128].

To mitigate the effect of strain phylogeny, we can use the phylogenetic tree for the strains to reduce the number of closely-related nodes with identical properties. Given a GCF G and a MF M, a strain in the population can be said to have a certain *membership pattern* with regards to its membership in G and M. For each strain, four patterns are possible: the strain being in both G and M, neither of them, or only in one of the two.

Recall that the leaf nodes of the tree are precisely the strains in the population. To perform the phylogenetic correction, two leaf nodes in the tree that have the same membership pattern are collapsed into a single artificial strain having the same membership pattern as the original strains. That strain replaces the shared parent node of the original strains, which are removed, leaving the artificial strain as a new leaf node in the tree. This process, which we refer to as *phylogenetic collapsing*, is repeated until no sibling leaf nodes in the tree have the same membership pattern, at which point we consider the tree *fully collapsed*. Figure 4.3 shows an example of collapsing the phylogenetic tree according to membership patterns.

To calculate the phylogenetically corrected strain correlation score for the (M, G) link, we take as an updated population set N' the set of leaf nodes from the fully collapsed tree, and define the updated MF M' and GCF G' as the members of N' with the appropriate memberships. We then proceed to calculate the standardised strain correlation score $\bar{\sigma}_{\text{corr}}$ in the usual fashion, but using N', M' and G' in place of N, M and G.

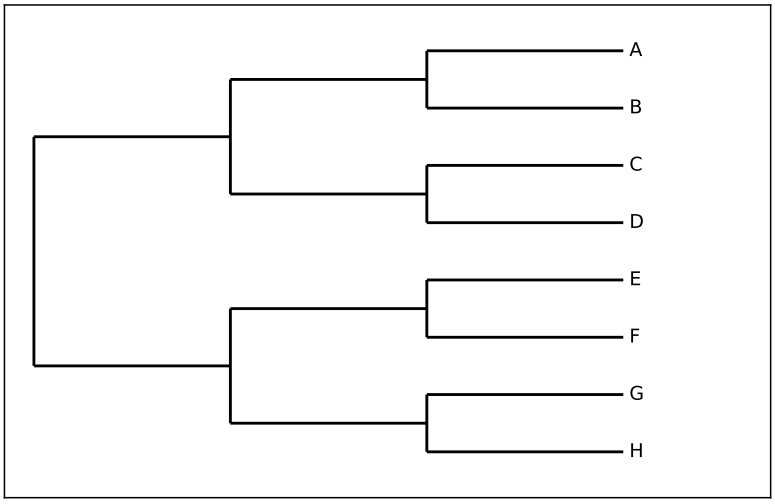


Figure 4.4: **Phylogenic tree of the strains in the synthetic example.** The tree represents the phylogenetic relationship between the strains with the GCF and MF memberships shown in Table 4.4.

It should be pointed out that, as demonstrated in Section 4, the strain correlation score is greatly affected by the population size, and the sizes of the MF and GCF being scored. As the proposed phylogenetic correction works by altering the strain count of the population, and therefore also the strain count of the MF and GCF, the scores before and after the correction would not be comparable using the raw strain correlation score. Therefore, standardising the strain correlation score, as discussed in section 4.4, is necessary to make the scores comparable.

Synthetic example To observe the reordering of links afforded by phylogenetic correction, and the prioritising of BGCs acquired by HGT, we can take a look at a synthetic example.

Assume eight strains, labelled A through H, with the phylogenetic tree shown in Figure 4.4. Assume also three MFs α, β and γ, with corresponding GCFs G_α, G_β and G_γ. The strain membership for the MFs and GCFs is shown in Table 4.4, with G_β being cryptic in strain G and G_γ cryptic in strain H, i.e. the strain in the GCF but not in the MF, but the BGCs being otherwise expressed.

Note that if the overlap between the GCF and the MF is complete, the standardised strain correlation score is constant given the size of the population, and the value varies with population size due to the statistical power of the hypergeometric test. The variability in the phylogenetically corrected score is therefore due to the different population sizes after collapsing the phylogenetic tree.

We take the membership of GCF α to be purely due to homology, i.e. the shared ancestry

	A	B	C	D	E	F	G	H
α	x	x	x	x				
β			x		x	x	o	x
γ		x				x		o

Table 4.4: Strain membership of the GCFs G_α, G_β and G_γ, and the corresponding MFs. 'x' denotes that the strain is a member of the GCF and the MF, while 'o' denotes that the strain is only a member of the GCF, i.e. the BGC is cryptic.

Raw		Standardised		Phyl.corr.	
α	44	α	2.6458	γ	1.7889
β	43	β	2.0494	β	1.5811
γ	25	γ	1.9720	α	1.0000

Table 4.5: Raw correlation score, standardised correlation score and phylogenetically corrected correlation score for the links. The phylogenetic correction penalises the link α, as it only occurs in phylogenetically related strains. This penalty changes the ranking of the links, so that γ replaces α as the highest ranking link.

of strains A through D. For GCF β, the membership of strains E through H is also due to shared ancestry, with a corresponding BGC in C due to HGT. Finally, for GCF γ, we assume that B, F and H have all acquired the BGC through HGT from an unknown source.

While it could be argued that the plausibility of γ being in H because of HGT is perhaps at odds with it being cryptic, the example is supposed to represent the outcome of a culture, and the BGC may just not have been expressed under the growth conditions.

Table 4.5 shows the score of the true links in the data set (the other potential links are ranked below the true links using all of the scoring methods). The phylogenetic correction penalises the α link relatively, owing to the fact that it's only present in phylogenetically related strains. Conversely, since the γ link is not present in any of the strains most related to B, F and I, it scores relatively higher than β, even if β is on the whole present in more strains. The ranking of the links is therefore changed, so that γ replaces α as the highest-scoring link.

Continuing with the synthetic example, we can analyse the correlation between phylogenetic distance and standardised strain correlation score $\bar{\sigma}_{\text{corr}}$ using pairs of strains selected from the population defined above. Let A and B be microbial strains, G_A be the set of GCFs involving A, and M_B be the set of MFs involving B. Let $\bar{\sigma}_{\text{corr}}(A, B)$ be the highest score of any GCF-MF link involving a GCF from G_A and a MF from M_B, and $d(A, B)$ be the phylogenetic distance between them, calculated as path length in the phylogenetic tree. In this case, the phylogenetic bias should manifest in a negative correlation between $\bar{\sigma}_{\text{corr}}(A, B)$ and $d(A, B)$, i.e. a higher phylogenetic distance should result in a lower $\bar{\sigma}_{\text{corr}}$.

Table 4.6 shows the correlation coefficients for $\bar{\sigma}_{\text{corr}}(A, B)$ and $d(A, B)$ for all pairs of strains

	Corr.coeff.	p-value
Uncorrected	-0.7592	2.5771×10^{-10}
Corrected	-0.3435	0.0157

Table 4.6: Spearman's R for the correlation of distance between pairs of strains in the phylogenetic tree, and the highest scoring GCF-MF link, where one strain is a member of the GCF and the other a member of the MF

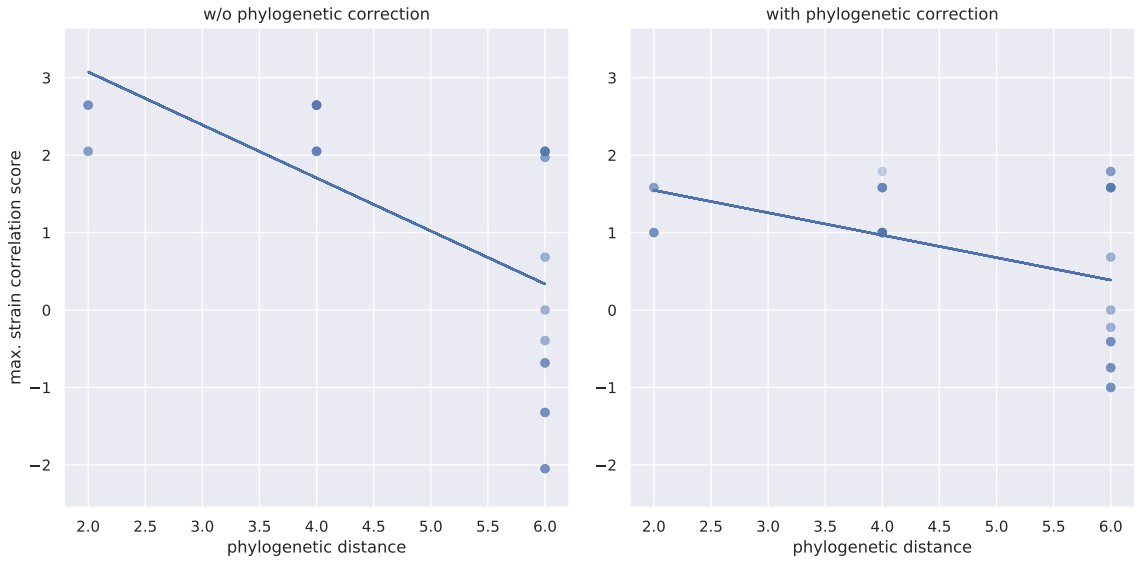


Figure 4.5: **Correlation between phylogenetic distance and standardised strain correlation score.** The relationship between the phylogenetic distance and the standardised strain correlation score without (left) and with (right) phylogenetic correction. While not removing it completely, correcting for phylogeny in the way proposed here reduces the correlation between the two.

from the population defined above, calculating the correlation between the distance in the phylogenetic tree and the highest score for any GCF-MF combination where one strain is a member of the GCF and the other a member of the MF. Since the number of strains, and number of GCFs, is very small, the distribution of the scores is far removed from the normal distribution, so Pearson's R is not appropriate, and calculations are done using Spearman's R. Figure 4.5 shows the relationship between the standardised strain correlation score without (left) and with (right) the phylogenetic correction and the phylogenetic distance.

The phylogenetic correction greatly reduces the correlation between the two, however, it does not remove it completely. This may be due to the uneven distribution of GCFs among strains, the small size of the data set, or both.

Example on microbial data set To demonstrate the impact of phylogeny on link score in real-world data, we can consider the Crüsemann data set. The data set contains several dif-

ferent species, with a number of strains each. If the score were not affected by phylogenetic distance, choosing a GCF and an MF at random, associated with the same species, should yield the same distribution of scores as if the GCF and MF were associated with different species.

Figure 4.6 shows the phylogenetic tree for the *Salinispora arenicola* and *Salinispora pacifica* from the Crüsemann data set, generated using autoMLST [134]. The parameters used to generate the tree can be found in Appendix A. As expected, the two exist on completely different branches of the phylogenetic tree, so the separation of strains into *S. arenicola* and *S.pacifica* can be used as a rough proxy for phylogenetic distance.

Using *S. arenicola* and *S. pacifica* from the Crüsemann data set as test cases, we restrict the base sets $\mathcal{G}$ of GCFs and $\mathcal{M}$ of MFs to the sets $\mathcal{G}_{\text{are}}$, where the GCF contains at least one *S. arenicola* strain, $\mathcal{G}_{\text{pac}}$, where the GCF contains at least one *S. pacifica* strain, and $\mathcal{M}_{\text{are}}$ and $\mathcal{M}_{\text{pac}}$ where the MFs are similarly filtered. From each combination of base sets, we choose 10000 random pairs of GCF G and MF M and calculate the strain correlation score. Since choosing the pair from "matching" base sets, i.e. both associated with *S. arenicola* or both with *S. pacifica*, would give a higher chance of strain overlap than choosing from "mismatched" base sets, we impose the condition that the intersection between the MF and the GCF is non-empty, to avoid exaggerating the bias by virtue of a non-empty intersection between strains being more likely if both strains are selected from the same base set.

As strains belonging to the same species are more related than strains belonging to different species, i.e. the phylogenetic distance between them is smaller, the phylogenetic bias should result in a higher number of high-scoring links between an MF and a GCF from strains belonging to the same species than an MF and a GCF from strains belonging to different species.

	Without correction	With correction
$M \in \mathcal{M}_{\text{pac}}, G \in \mathcal{G}_{\text{are}}$	0.0351	0.0459
$M \in \mathcal{M}_{\text{are}}, G \in \mathcal{G}_{\text{pac}}$	0.0253	0.0408
$M \in \mathcal{M}_{\text{pac}}, G \in \mathcal{G}_{\text{pac}}$	0.0803	0.0654
$M \in \mathcal{M}_{\text{are}}, G \in \mathcal{G}_{\text{are}}$	0.0547	0.0479

Table 4.7: Percentage of high-scoring links associated with strains from different combinations of base sets, as a proportion of all links.

To demonstrate the phylogenetic bias, we can observe Table 4.7, which shows the percentage of links scoring above the 95th percentile of all links, broken down by the combination of base sets. Selecting both GCF and MF from base sets associated with the same species (rows 3 and 4) results in a considerably higher proportion of the links scoring above the 95th

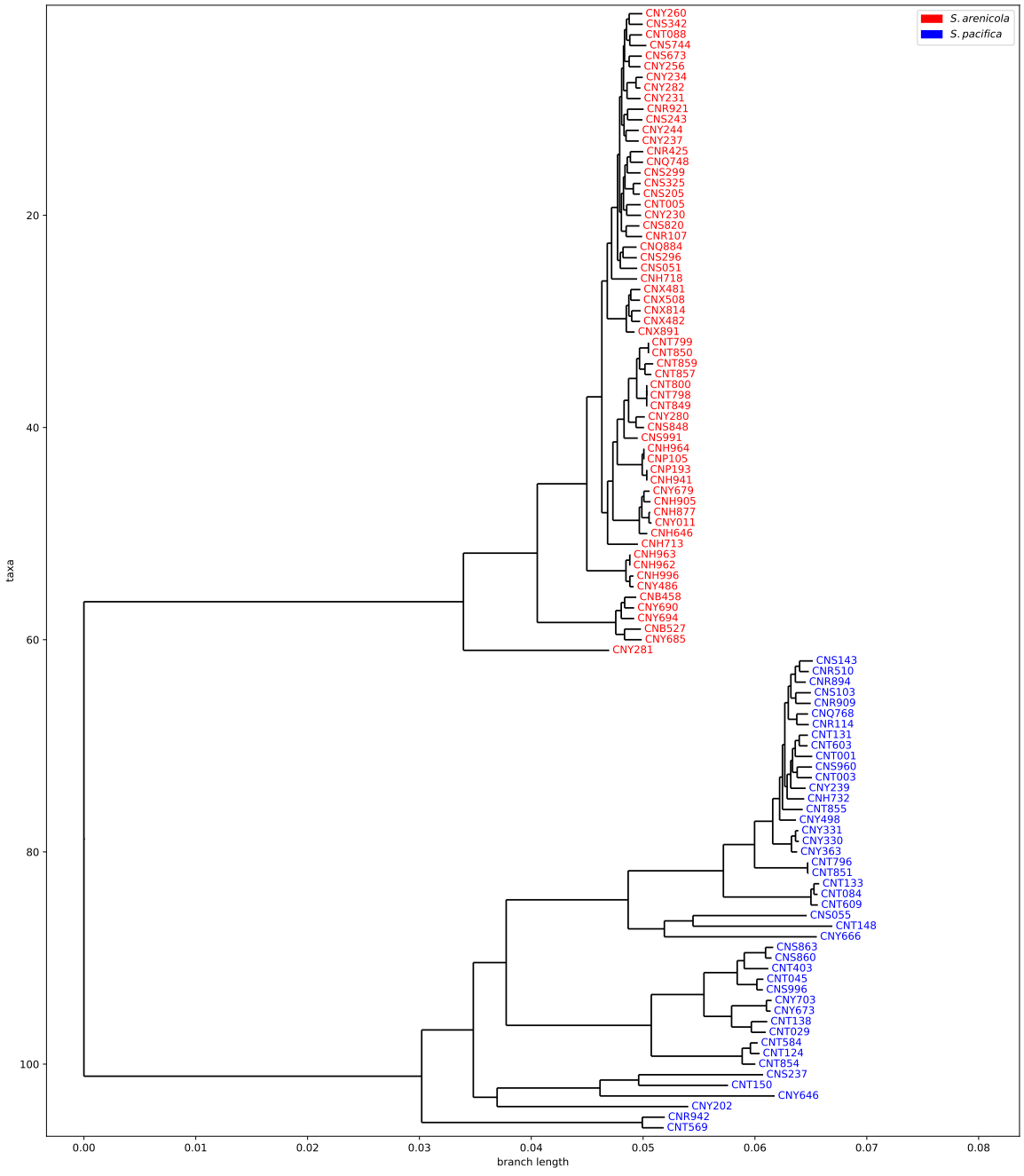


Figure 4.6: **Phylogenetic tree of the *S. arenicola* and *S. pacifica* strains from the Crüsemann data set.** Strains belonging to *S. arenicola* are coloured red, while strains belonging to *S. pacifica* are coloured blue.

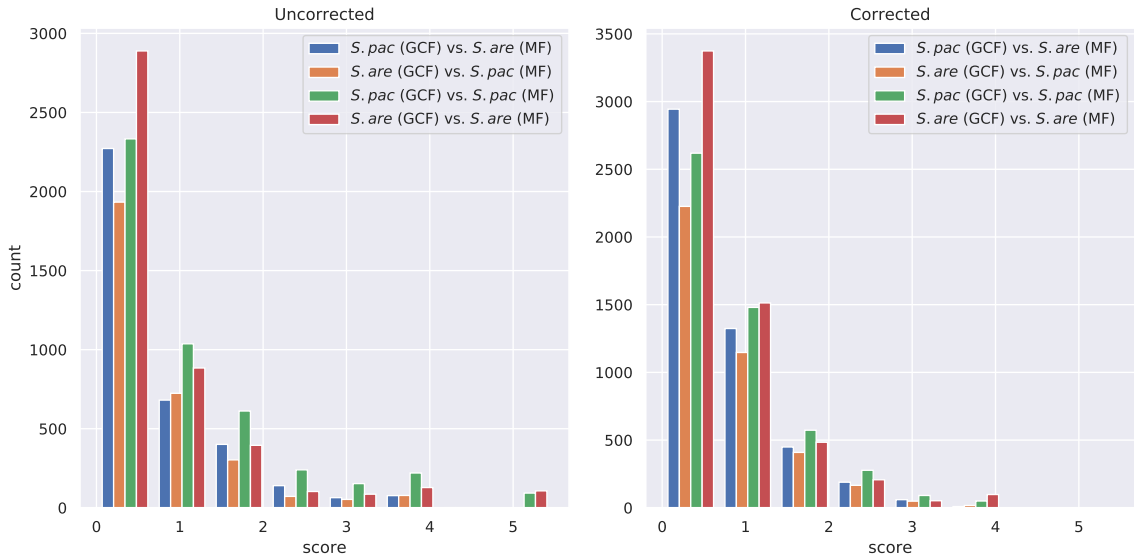


Figure 4.7: **Distribution of scores for potential links broken down by base set.** For comparison purposes, both scores are standardised to have mean zero and standard deviation one, and only links that score above zero are considered. Without correction for phylogeny (left), GCFs and MFs containing strains from the same base set (red and green bars) yield more high-scoring links than from different base sets (blue and yellow bars). This difference is reduced when correcting for strain phylogeny (right).

percentile on the raw strain correlation score (column 1) than selecting the GCF and the MF from different base sets (rows 1 and 2).

The second column of table 4.7 shows the same statistics for the strain correlation score after correcting for strain pylogeny. As can be seen, the bias is much less pronounced after applying the proprosed correction.

The distribution of the scores for the sampled links, broken down by base set combinations, is shown in Figure 4.7. Using the strain correlation score without phylogenetic correction, choosing both strains from the same base set yields a relatively higher number of high-scoring links compared to choosing the strains from different base sets. This difference is reduced by applying the proposed phylogenetic correction to the score. In other words, a high-scoring link in the uncorrected score is relatively more likely to be from matching base sets than a high-scoring link in the corrected score.

To better quantify the impact of the phylogenetic bias, we can estimate the correlation between the strain correlation score $\bar{\sigma}_{\text{corr}}$ and the phylogenetic distance, in a similar way as with the synthetic example above. Letting A and B be strains, we take G_A to be the set of GCFs involving strain A, and M_B to be the set of MFs involving strain B. Let $\bar{\sigma}_{\text{corr}}(A, B)$ be the highest score of any GCF-MF link involving a GCF from G_A and a MF from M_B, and

	Corr.coeff.	p-value
Spearman without correction	-0.5122	5.1240×10^{-8}
Spearman with correction	-0.3147	0.0014
Pearson without correction	-0.5053	8×10^{-8}
Pearson with correction	-0.3102	0.0017

Table 4.8: Correlation coefficients between the corrected and uncorrected form of the strain correlation score and the phylogenetic distance. The phylogenetic correction reduces both Spearmann and Pearson correlation coefficients.

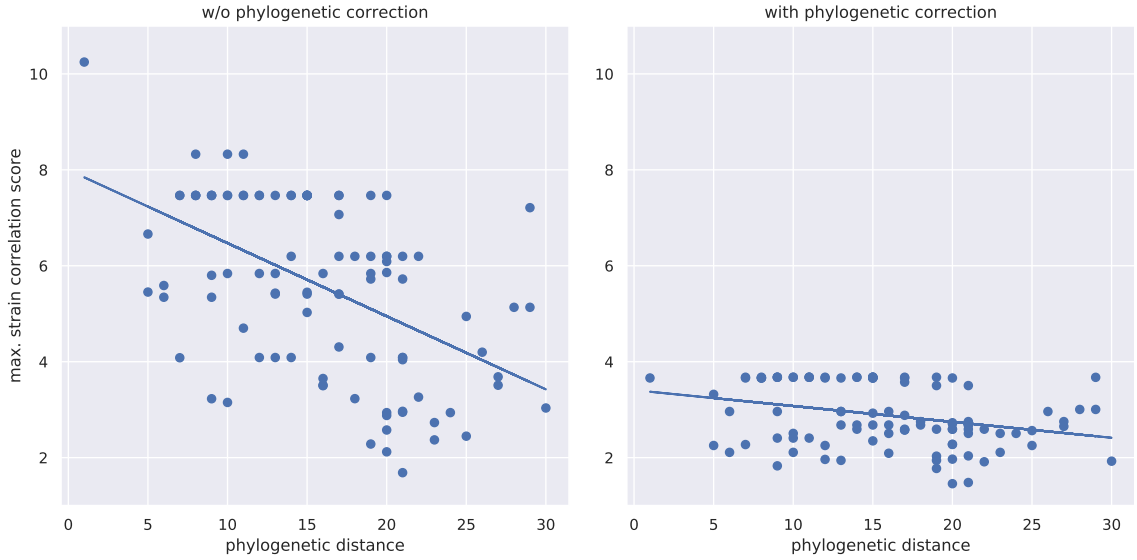


Figure 4.8: **Correlation between phylogenetic distance and strain correlation score.** Highest strain correlation score between a GCF and a MF associated with a pair of strains, without phylogenetic correction (left) and with phylogenetic correction (right), and phylogenetic distance between the strains, for 100 randomly chosen strain pairs from the Crüsemann data set. The right plot shows less correlation between the two axes than the right plot, as demonstrated by the correlation coefficients in Table 4.8.

$d(A, B)$ the phylogenetic distance between A and B.

As before, the phylogenetic bias should manifest in negative correlation between $\bar{\sigma}_{\text{corr}}(A, B)$ and $d(A, B)$.

Figure 4.8 shows $\bar{\sigma}_{\text{corr}}(A, B)$ with and without phylogenetic correction, plotted against $d(A, B)$ for 100 randomly sampled pairs of A and B. The correlation coefficients for the for $\bar{\sigma}_{\text{corr}}(A, B)$ and $d(A, B)$ can be seen in Table 4.8, with the non-phylogenetically-corrected form of the strain correlation score in rows 1 and 3, and the phylogenetically corrected strain correlation score in rows 2 and 4. While still significant, the correlation is much weaker for the

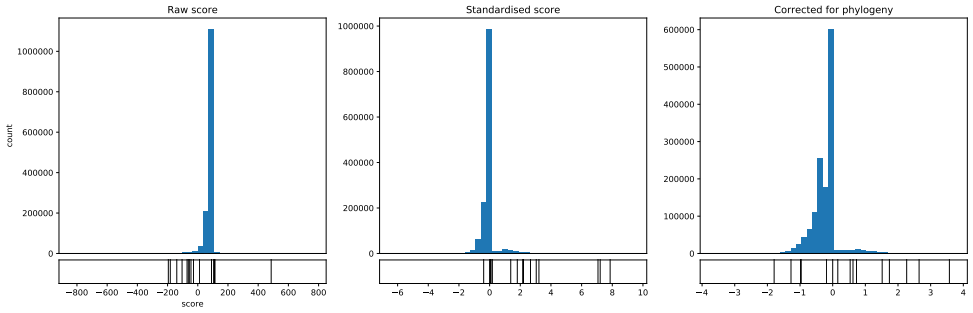


Figure 4.9: **Distribution of validated links in the distribution of scores for all potential links.** Validated links are indicated in black beneath the score histogram.

phylogenetically corrected score than for the raw score.

This demonstrates that adjusting the strain correlation score to correct for phylogenetic distance by collapsing nodes showing identical patterns of strain membership can dramatically reduce the correlation between the phylogenetic distance and the strain correlation score. This can serve to emphasise BGCs shared by virtue of HGT rather than inheritance, which might otherwise in some cases be drowned out, despite their potential interest.

Distribution of validated links Observing the effect of the phylogeny correction on the distribution of the validated links in the Crüsemann dataset, some reordering of the top links takes place, and a small decrease in the number of validated links scoring above the 95th percentile.

Figure 4.9 shows the distribution of scores for the verified links within the distribution of scores for all hypothetical links, while Table 4.9 shows the proportion of verified links scoring above the 95th percentile for the different scoring methods. Although the phylogenetically corrected score puts a smaller proportion of the verified links above the 95th percentile, this may be compensated for by the ordering of the links, which is shown in Table 4.11.

Note that a drawback of the strain correlation score, particularly as it relates to BGC clustering, is that it is very sensitive to the clustering algorithm splitting clusters up erroneously. For instance, in the case of BGC0000137, merging all GCFs that get tagged with that MIBiG entry by antiSMASH known cluster blast might change the position of BGC0000137 within the ranking, particularly as strains missing from a GCF but present in a MF incur a large penalty.

To verify that the behaviour of the phylogenetic correction is as expected, we can consider how the relative ranking of the validated links changes as we apply the phylogenetic correction to the strain correlation score, and consider if the sharing of the BGC amongst the strains is likely to be due to inheritance or HGT.

Scoring function	Ratio
Raw strain correlation score	8.4899×10^{-5}
Standardised strain correlation score	0.0001423
Phylogenetically corrected strain correlation score	0.0001139

Table 4.9: Ratio of validated links to all links scoring above the 95th percentile

Metabolite name	MIBiG ID	Std. rank	Phyl.corr. rank	Rel. change
cyclomarin D	BGC0000333	3216	2793	0.1315
retimycin A	BGC0001228	275167	263920	0.0409
staurosporine	BGC0000827	3789	885	0.7664
rosamicin	BGC0001830	1543	5056	-2.2767
lomaiviticin	BGC0000241	14676	66419	-3.5257
rifamycin	BGC0000137	108063	1354052	-11.5302
rifamycin	BGC0000137	105046	1357570	-11.9236
rifamycin	BGC0000137	21594	88610	-3.1035
rifamycin	BGC0000137	1240045	1402964	-0.1314
rifamycin	BGC0000137	28765	14965	0.4797
rifamycin	BGC0000137	15879	10275	0.3529
rifamycin	BGC0000137	52362	698671	-12.3431
rifamycin	BGC0000137	107105	1390611	-11.9836
rifamycin	BGC0000137	39428	53325	-0.3525
rifamycin	BGC0000137	28971	59940	-1.0690

Table 4.10: Change in ranking of potential links between standardised strain correlation score and phylogenetically corrected strain correlation score. Change in rank is estimated relative to the rank using the standardised strain correlation score.

Raw			Standardised			corrected		
Name	ID	Score	Name	ID	Score	Name	ID	Score
staurosporine	BGC0000827	486	rosamicin	BGC0001830	7.8598	staurosporine	BGC0000827	3.5736
cyclomarin D	BGC0000333	114	cyclomarin D	BGC0000333	7.2111	cyclomarin D	BGC0000333	2.6458
rosamicin	BGC0001830	111	staurosporine	BGC0000827	7.0682	rosamicin	BGC0001830	2.2657
retimycin A	BGC0001228	105	lomaiviticin	BGC0000241	3.2167	rifamycin	BGC0000137	1.7363
lomaiviticin	BGC0000241	91	rifamycin	BGC0000137	3.0407	rifamycin	BGC0000137	1.5152
rifamycin	BGC0000137	12	rifamycin	BGC0000137	2.6653	rifamycin	BGC0000137	0.7280
rifamycin	BGC0000137	-28	rifamycin	BGC0000137	2.2103	rifamycin	BGC0000137	0.6217
rifamycin	BGC0000137	-45	rifamycin	BGC0000137	2.1581	lomaiviticin	BGC0000241	0.5345
rifamycin	BGC0000137	-56	rifamycin	BGC0000137	1.8053	rifamycin	BGC0000137	0.1564
rifamycin	BGC0000137	-63	rifamycin	BGC0000137	1.3776	retimycin A	BGC0001228	0.0000
rifamycin	BGC0000137	-71	rifamycin	BGC0000137	0.1790	rifamycin	BGC0000137	-0.1877
rifamycin	BGC0000137	-104	rifamycin	BGC0000137	0.0930	rifamycin	BGC0000137	-0.9641
rifamycin	BGC0000137	-139	rifamycin	BGC0000137	0.0549	rifamycin	BGC0000137	-0.9877
rifamycin	BGC0000137	-182	retimycin A	BGC0001228	0.0000	rifamycin	BGC0000137	-1.2770
rifamycin	BGC0000137	-195	rifamycin	BGC0000137	-0.3868	rifamycin	BGC0000137	-1.7936

Table 4.11: Order of verified links for metabolites by strain correlation score

While concrete methods exist to determine this [135, 136], their application to this problem remains an avenue for future work. In the interim, we can try to determine, by looking at the phylogenetic tree, if the change in relative ranking of validated links is consistent with the phylogenetic tree.

Such comparison is by its nature rather vague, unless we can construct a way to quantify how phylogenetically diverse a GCF is, a notion wich could be called the "phylogenetic area" of the GCF. In contrast to e.g. the diameter of a set, such a measure would need to capture more of the topology of the set, i.e. how many of the strains which are closely related to members of the GCF are not themselves in the GCF. This will not be attempted here.

In general, we would expect some of the verified links to score relatively higher (those shared partly because of HGT) and some relatively lower (those shared mostly because of shared ancestry). Table 4.10 shows how the rank of the verified links is affected by the application of the phylogenetic score. For verification purposes, we can look at BGC0000333 and BGC0000827 as examples of links that are positively affected, and BGC0001830 and BGC0000241 as examples of links that are negatively affected. Our hope would be that the two former show greater phylogenetic diversity, while the latter show more phylogenetic homogeneity, in the sense that whole branches of the tree have the same phenotype with respect to GCF and MF membership.

Figure 4.10 shows the phylogenetic tree for the GCF-MF link associated with MIBiG ID BGC0000333. The two strains that are in the GCF are not collapsed at all, while most of the rest of the tree gets collapsed.

Figure 4.11 shows the phylogenetic tree for the GCF-MF link associated with MIBiG ID BGC0000827. Although the GCF seems to be faulty, in that many of the strains are in the MF but not the GCF, the phylogenetic tree is diverse enough not to be collapsed to a great extent, and while the link does get penalised where the strains in the MF are not appearing in the GCF, it is still fairly phylogenetically diverse.

Figures 4.12 and 4.13 show the phylogenetic tree for the GCF-MF links associated with MIBiG IDs BGC0000241 and BGC0001830, respectively. Each phenotype occurs in a very well-defined branch of the tree, indicating that the BGC is shared largely because of shared ancestry, which causes the ranking of the links to be lower.

Discussion While further investigation is needed into the changes in ranking of the individual links, the reduced correlation between the strain correlation score and phylogenetic distance, combined with the limited impact on the enrichment of validated links in the top percentile of the scores, nevertheless indicates that this approach to compensating for the phylogenetic bias in the strain correlation score holds considerable promise.

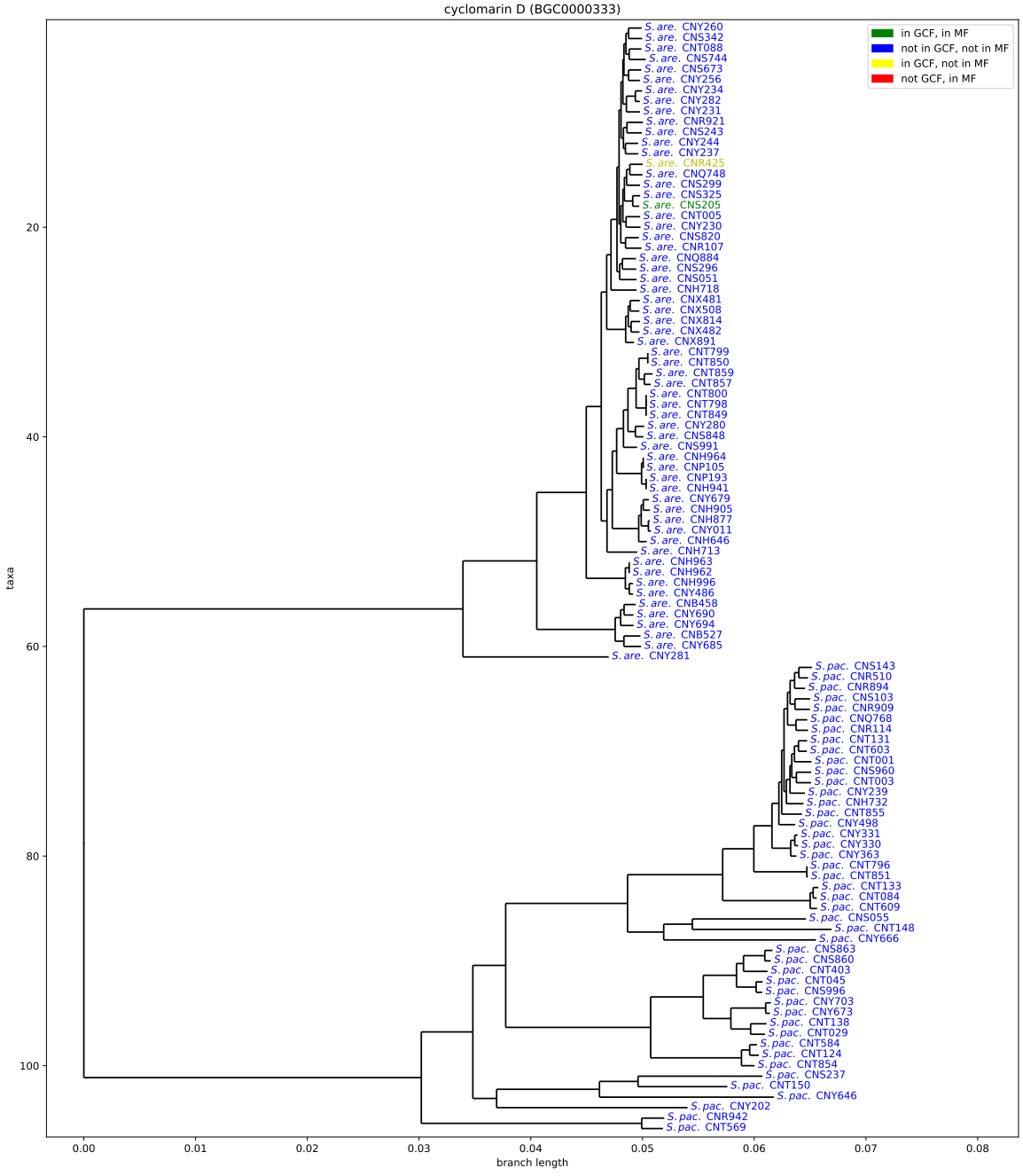


Figure 4.10: **Phylogenetic tree for the upgraded GCF-MF link BGC0000333, cyclomarin D**

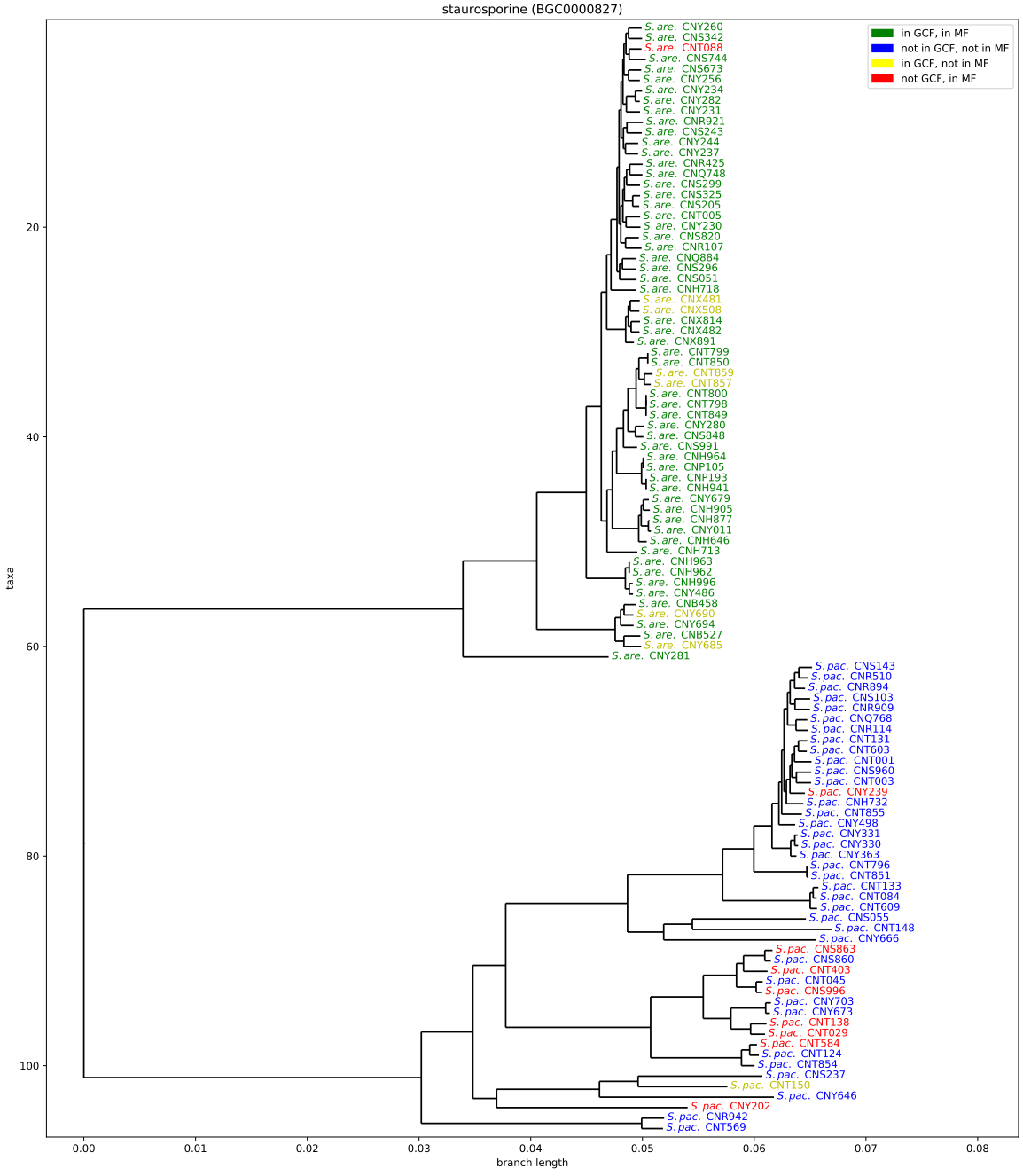


Figure 4.11: **Phylogenetic tree for the upgraded GCF-MF link BGC0000827, staurosporine**

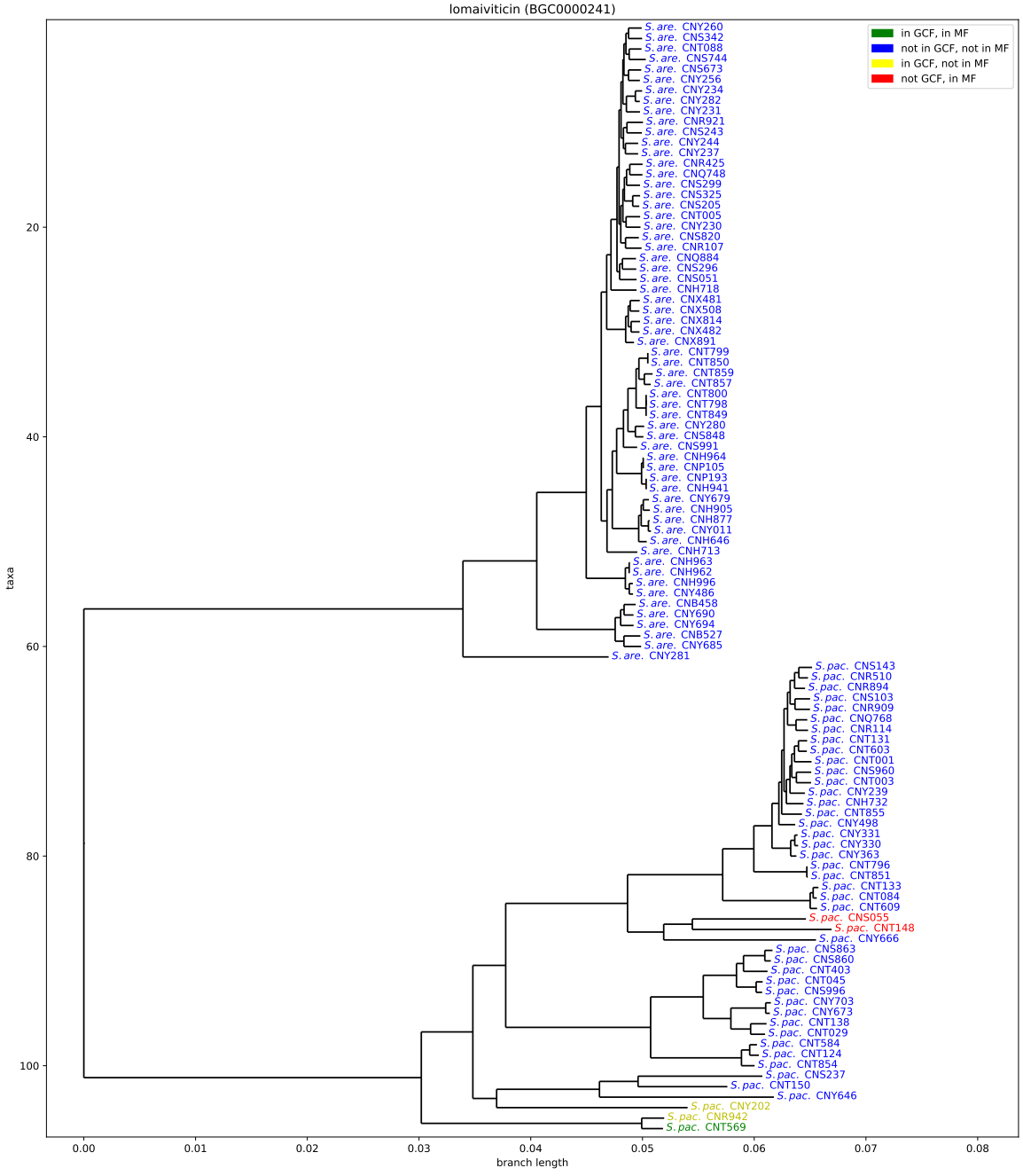


Figure 4.12: **Phylogenetic tree for the downgraded GCF-MF link BGC0000241, lomaiviticin**

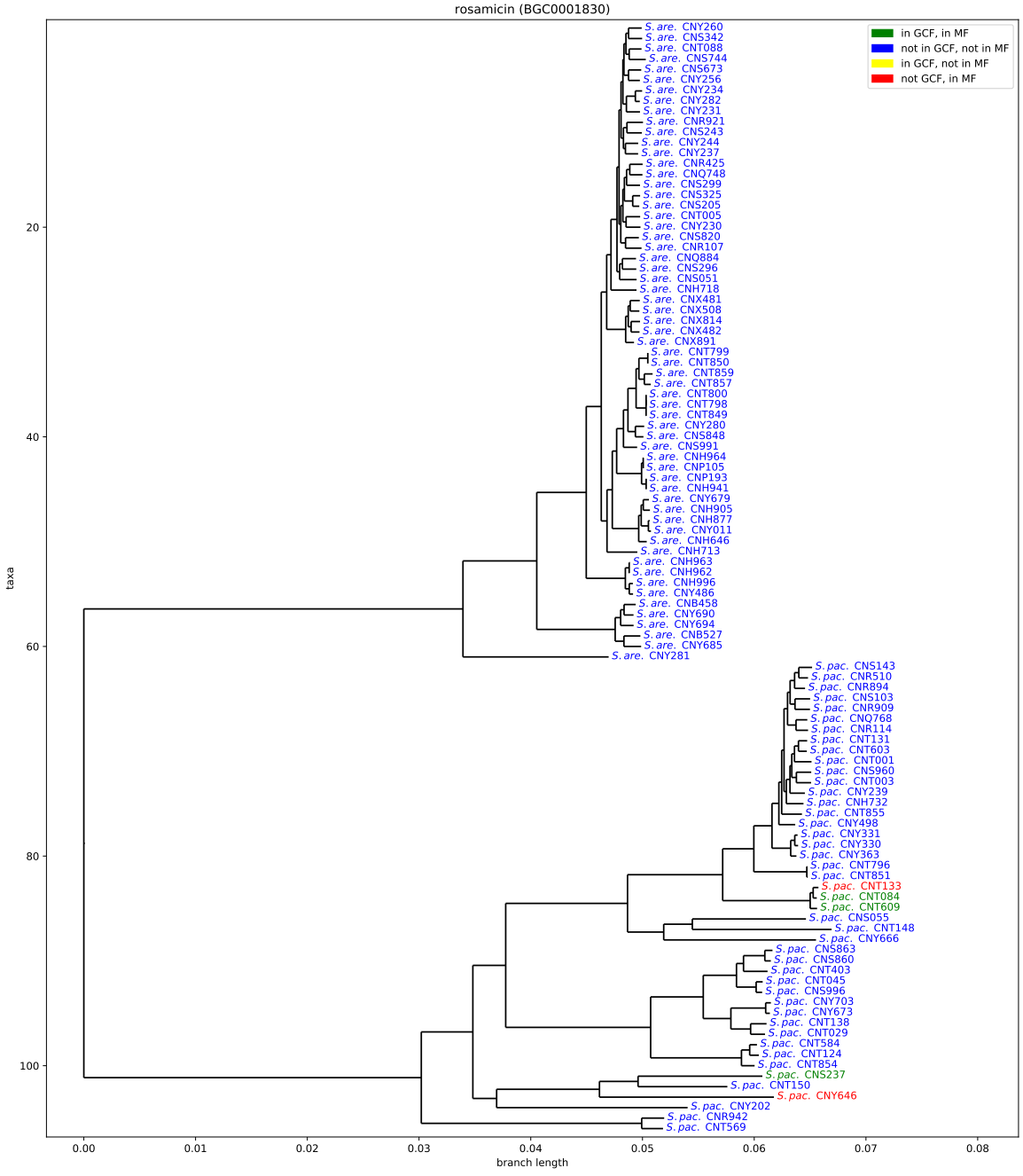


Figure 4.13: **Phylogenetic tree for the downgraded GCF-MF link BGC0001830, rosamicin**

4.6 Summary

While the strain correlation score defined in [65] has been successfully used in its raw form, we have demonstrated that it has serious drawbacks when used to compare potential links, even within a data set. We have proposed a way of fixing these by standardising the score, and demonstrated how standardising the score improves its performance in prioritising links for further investigation.

Furthermore, we have made the case that BGCs shared between strains by HGT can be of particular interest, and demonstrated how the strain correlation score favours BGCs and MFs which contain evolutionarily related strains at the expense of BGCs shared by virtue of HGT. We have proposed a way to remedy this and demonstrated its efficacy in broad terms, although further validation is needed.

Chapter 5

Feature-based scoring

5.1 Introduction

A common approach to link microbial genomic and metabolomic information is to predict genomic features from metabolomic data, or to predict metabolomic features from genomic data.

Section 5.2 gives an overview of feature-based approaches to linking genomic and metabolomic data. Section 5.3 discusses this approach in the context of untargeted metabolomics, and highlights a kernel-based algorithm used to match MS2 spectra to molecular structures.

Sections 5.4, 5.5 and 5.6 contain further discussion of the algorithm, while Sections 5.7, 5.8 and 5.9 present how this approach can be applied to genomic and metabolomic data.

The final three sections of the chapter describe three variants of this model. Section 5.10 presents a model to link metabolomic information to molecular structures, including an investigation into how the algorithm makes its predictions. Section 5.11 presents a similar model to link genomic information to molecular structures, and finally Section 5.12 joins two such models together to present a model to link genomic and metabolomic information.

5.2 Feature-based scoring

One of the central problems of genomics is identifying the role of particular parts of the genome. This is known as *functional genomics* [137]. For microbial BGCs, which are, to reiterate, the regions in the microbial genome predicted to produce specialised metabolites, this question leads naturally to the problem of predicting molecular properties from genomic data.

Many approaches exist to do this, with various domains. As different natural product classes are synthesised in different ways, most existing tools for structural prediction focus on a particular class. So far, NRPs and PKs have lent themselves particularly to structural prediction, as these are modular in structure, i.e. central to the synthesis of these molecules is assembling previously-existing components (such as amino acids) to form the final product [71].

However, even for NRPs and PKs, complete structural prediction for a novel BGC is difficult, as the metabolites can undergo post-translational modifications in the form of e.g. tailoring enzymes and cyclization [71]. Predicting the existence and outcome of such modification for a given metabolite is difficult for BGCs not closely related to known BGCs [71].

Instead of trying to predict the entire structure of the metabolite, it is in some cases possible to make predictions about parts of the moleclue, or *substructures*. For NRPs, and PKs, the fact that they are modular to begin with makes identification of substructures easier, provided that they are not affected by potential post-translational modifications to the extent that they become unrecognisable. For most natural product classes other than NRPs and PKs, predicting the molecular structure, or even the core scaffold — i.e. the general large-scale structure of the molecule — is non-trivial [71]. However, common sub-clusters within the BGCs can sometimes be identified, some of which are responsible for the production of specific molecular sub-structures [71]. This can allow prediction of particular chemical features in isolation from the rest of the molecular structure.

As tandem mass spectrometry involves fragmenting the molecules and measuring the *m/z* of the resulting fragments, some properties — like amino acid or peptide composition — are likely to be observable in the resulting spectrum [97, 95, 101] and some — such as stereochemistry — are not [71]. In this context, the modular specialised metabolites are particularly interesting, as the component modules tend to show up as characteristic peaks (corresponding to the *m/z* of the module) or differences between peaks (losses, corresponding to the loss of the module, i.e. one peak including the molecule and the other not). Identifying e.g. the amino acid composition of a molecule produced by a BGC, as Pep2Path does for NRPs and RiPPs [97], allows searching of spectra for tell-tale peaks or losses, or even sequences of peaks or losses, characteristic of the metabolite.

Many of the tools used to detect BGCs also incorporate some predictions of the properties of the product. antiSMASH incorporates both various predictors of amino acid composition of the resulting molecules, such as SANDPUMA [138] for predicting NRP amino acid content, but also its own predictive logic for predicting some structural properties of NRPs and PKs [139]. RODEO [99] aims specifically to detect RiPP BGCs, particularly lassopeptides, but only makes structural predictions as far as predicting the precursor peptides for the eventual product.

The most recent versions of both antiSMASH [73] and PRISM [102] both claim improvements in structural predictions for their predicted BGCs. However, attempts at using structures predicted by antiSMASH to match BGCs to spectra via the intermediary of molecular fingerprints have not fared well. To some extent, this is because many of the structural predictions are only partial, i.e. only specific parts of the metabolite are predicted. This does not translate well to molecular fingerprints, as the absence of a molecular feature in a partial prediction cannot be used to rule out the feature in the rest of the metabolite. It is therefore unclear how partial structural predictions can be used to predict molecular fingerprints except to a limited degree.

Some tools predict specific features that are directly observable in spectra. PRISM [74, 102] predicts the sequence of monomers that are incorporated into the molecule for NRPs and type I and II PKs, although the predictions only extend to how the features would be expressed in LC-MS data (but not LC-MS2). NRPquest [96] detects NRPS BGCs and makes predictions about the fragmentation spectra for their associated products, while MetaMiner [95] does the same for RiPPs, and GNP [69] links NRPS and PKS BGCs to LC-MS2 spectra. For peptidic natural products (NRPs and RiPPs), Pep2Path [97] predicts amino acid sequences from hypothetical series of losses in the spectra, and matches those to previously detected BGCs.

5.3 Identifying metabolites from spectra

A central problem in untargeted metabolomics is finding the molecule associated with a given spectrum from a candidate set of molecules [140]. This is generally done by using the spectrum to predict properties of the molecule, and then searching the candidate set for the molecule most similar to the predicted molecule.

A common approach to quantifying the similarity of molecules is the use of *molecular fingerprints* [141, 142, 143, 144]. Molecular fingerprints are binary vectors that encode the structure of a molecule. As an infinite variety of possible dimensions exist to model the fingerprint on, and as there is no single correct way of defining similarity between molecules, various similarity measures have proven useful in different contexts, and many different molecular fingerprints have been defined. In many common approaches, each dimension denotes the presence or absence of certain molecular substructures, such as in the Klekota-Roth fingerprint [142] commonly used to characterise the biological activity of molecules, but other molecular fingerprints, with less obvious interpretations, are common as well, such as *Extended Connectivity Fingerprints* (ECFP) [143, 144].

The state of the art in such systems to map spectra to molecules is CSI:FingerID [145], which computes a fragmentation tree for a given spectrum, uses the fragmentation tree to predict

the molecular fingerprint for the spectrum using a separate binary classifier for each bit of the fingerprint, and finally searches a database of candidate molecules for the closest matching fingerprint.

While prediction times for CSI:FingerID are low, it suffers from long training time, as central to the method is training a SVM to predict each bit of the molecular fingerprint from the spectrum, and depending on the fingerprint, these can number in the thousands.

Brouard and co-workers have suggested *input-output kernel regression* (IOKR) [113, 146] as an alternative. In this approach, starting with a spectrum and a candidate set of molecules, the objects are mapped into a shared *latent space*, an abstract space which encodes a representation of the relevant features of the original objects, where the distances between the spectrum and the candidate molecules are computed. This distance can be used to rank the candidate set so that the molecules with the shortest distance to the spectrum in the latent space having the best rank. In this case, the shared latent space is the space of molecular fingerprints, with a pair consisting of a spectrum and a molecule being assigned a high score if the predicted fingerprint for the spectrum is similar to the fingerprint for the molecule. Using this model, Brouard and co-workers demonstrate a slight improvement in performance over CSI:FingerID, with greatly improved training- and prediction times.

Matching MS2 spectra to metabolites is, generally speaking, a very hard problem. Even for state of the art approaches, such as CSI:FingerID and IOKR, top-1 accuracy above 20% is highly non-trivial, unless the candidate set is severely restricted, which to a degree defeats the purpose of the tool.

The use of kernel functions is central to IOKR. We now turn to a brief overview of kernels, after which we discuss the IOKR model in detail.

5.4 Kernel functions

Let X be an arbitrary set, and Y be a vector space with an inner product $\langle \cdot, \cdot \rangle$. A *kernel* on the set X is a function $k : X \times X \to \mathbb{R}$ that satisfies $k(x, z) = \langle \phi(x), \phi(z) \rangle$, where $\phi : X \to Y$, i.e. the computing the kernel product k of two elements x and y is equivalent to projecting the elements into a given vector space using a particular function ϕ and taking the inner product of the resulting vectors $\phi(x)$ and $\phi(y)$.

By this definition, and the properties of the inner product, it follows that k is symmetric, i.e. $k(x, y) = k(y, x)$.

If $S = \{x_1, \ldots, x_n\}$ is a finite subset of X, the matrix G where $G_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ is called the *Gram matrix*, or the *kernel matrix*, of the kernel (on the training set S).

A symmetric $n \times n$ matrix $M \in \mathbb{R}^{n \times x}$ is said to be *positive semi-definite* if $\mathbf{x}^T M \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. A matrix is positive semi-definite if and only if all its eigenvalues are non-negative.

Let X be a set. A function $k : X \times X \to \mathbb{R}$ is a kernel if and only if the kernel matrix K with $K_{ij} = k(x_i, x_j)$ is positive semi-definite (and therefore symmetric) for all finite subsets $S = \{x_1, \ldots, x_n\} \subseteq X$. In particular, if X is finite, this applies to the set $S = X$.

For a detailed background on kernel methods in machine learning, the reader is referred to Shawe-Taylor and Cristianini [147].

Various approaches exist to turn arbitrary functions, such as functions describing similarity, into kernel functions. When constructing kernels from similarity functions, which may not fulfill the properties of kernel functions, care must be taken when extending the derived kernel to new samples, so as not to break the kernel properties.

The use of general similarity functions as kernels is discussed further in Section 5.11.1, and in more detail in [148].

5.5 Input-output kernel regression (IOKR)

Let $\mathcal{X}$ be the space of MS2 spectra, $\mathcal{Y}$ be the space of metabolites, and $f : \mathcal{X} \to \mathcal{Y}$ be the function mapping a spectrum to its corresponding metabolite. Let $\mathcal{F}$ be the shared latent space of molecular fingerprints with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Let also $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$ be kernel functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Furthermore, let $\phi : \mathcal{Y} \to \mathcal{F}$ be the mapping $\mathcal{K}(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{F}}$ expressing the kernel function $\mathcal{K}_{\mathcal{Y}}$ as an inner product in $\mathcal{F}$, and $h : \mathcal{X} \to \mathcal{F}$ be the function mapping elements $x \in \mathcal{X}$ to the image $\phi(y)$ of the element $y \in \mathcal{Y}$ which corresponds to the spectrum x. An arrow diagram of the IOKR model is shown in Figure 5.1.

In the training phase, IOKR uses subsets $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ of paired examples (x, y) with $x \in X$ and $y \in Y$, where each $x \in X$ has a corresponding $y \in Y$ and vice versa, to learn an approximation $\hat{h}$ of the function h. This approximation is given as

$$\hat{h} : \mathcal{X} \to \mathcal{F}, x \mapsto \sum_{x_j \in X} \mathcal{K}_{\mathcal{X}}(x, x_j) \mathbf{c}_j \tag{5.1}$$

where $\mathbf{c}_j \in \mathcal{F}$ [113]. If $\mathcal{F}$ is finite-dimensional, the solution for $\mathbf{c}_j$ using the regularised least-squares loss function with regularisation parameter λ can be written in closed form as

$$\mathbf{c}_i = (\lambda I + K_X)^{-1} \phi(y_i) \tag{5.2}$$

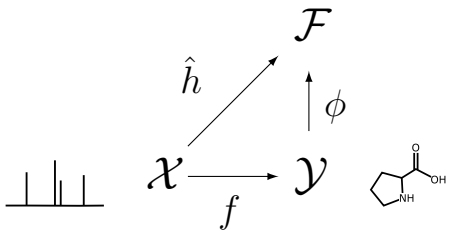


Figure 5.1: **Arrow diagram of the input-output kernel regression (IOKR) framework.** $\mathcal{X}$ is the space of MS2 spectra, $\mathcal{Y}$ is the space of molecules, and $\mathcal{F}$ is the space of molecular fingerprints. ϕ is the mapping that sends a molecule $y \in \mathcal{Y}$ to its molecular fingerprint $\phi(y) \in \mathcal{F}$, the shared latent space. $\hat{h}$ is the learned approximation to the function h which takes a spectrum $x \in \mathcal{X}$ to the molecular fingerprint $\phi(y')$ of the molecule y' that gave rise to the spectrum. Given an element $x \in \mathcal{X}$ a candidate set $Y \subseteq \mathcal{Y}$ is ranked for match to x using the similarity between the images of the elements in $\mathcal{F}$, i.e. the similarity between $\hat{h}(x)$ and $\phi(y)$. The dimensionality of $\mathcal{F}$ is controlled by the choice of molecular fingerprint.

where K_X is the Gram matrix of the kernel $\mathcal{K}_\mathcal{X}$ on the training set [113, 149, 150]. This task is called *output kernel regression*.

To rank a set $Y^* \subset \mathcal{Y}$ of candidate structures for a spectrum $x \in \mathcal{X}$, the set of candidates is ordered by decreasing value of the expression

$$\sigma_{\text{IOKR}}(x, y) = \langle \hat{h}(x), \phi(y) \rangle_\mathcal{F} \tag{5.3}$$

where $y \in Y^*$, so that the element getting assigned the highest value is taken to be the most likely candidate.

Taking K_X and K_Y to be the Gram matrices of the kernels $\mathcal{K}_\mathcal{X}$ and $\mathcal{K}_\mathcal{Y}$ on the training set, and $\mathbf{k}_X^x$ and $\mathbf{k}_Y^y$ to be the column vectors composed of $\mathcal{K}_\mathcal{X}(x_i, x)$ and $\mathcal{K}_\mathcal{Y}(y_i, y)$ for x_i and y_i in the training set, and substituting $\hat{h}(x)$ with Eqs. 5.1 and 5.2, Eq. 5.3 can be rewritten as

$$\sigma_{\text{IOKR}}(x, y) = (\mathbf{k}_Y^y)^T (\lambda I + K_X)^{-1} \mathbf{k}_X^x \tag{5.4}$$

where λ is a regularisation parameter and I is an appropriately sized identity matrix.

5.6 Computational complexity

In the formulation presented here, the computational complexity of IOKR is dependent on two factors: the size of the training set and the computational complexity of the kernel functions.

In the training phase, the calculation of the kernel matrix in Eqn. 5.4 requires evaluating the kernel functions for all pairs of elements from the training set, along with the inversion of the kernel matrix.

In the prediction phase, the kernel function on the input space is evaluated against all elements of the training set, while in the output space, the kernel function is evaluated for each of the elements in the candidate set against all of the training set elements.

While the kernel matrix grows quadratically with the size of the training set, with corresponding impact on the matrix inversion, the prediction phase scales linearly with the size of the training set. Both aspects depend heavily upon the computational complexity of the kernel functions being used, in particular the training phase, which requires pairwise computation of kernel values between elements of the training set.

However, the current size of the training data sets makes these computations very manageable. Furthermore, as the current training data sets include metabolites from a wide variety of sources, optimisation of the training data set to focus on specific parts of the metabolite space, e.g. on metabolites of microbial origin, may allow significant reduction of the training set size without loss of precision for the metabolites of interest.

Given reasonable optimisations on the calculation of the kernel functions, the scalability of the algorithms presented here should therefore not present a problem for the foreseeable future.

5.7 Kernel functions for spectra and metabolites

As with all kernel methods, the choice of kernel function $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$ is a very important consideration for the performance of IOKR. For spectra, Brouard and co-workers use a combination of kernels, the single best-performing of which is the *recalibrated product probability kernel* (PPKr) [113, 151], with a top-1 performance of roughly 22%, compared with roughly 30% for the best-performing combination.

For the molecules, Brouard and co-workers used the Gaussian kernel

$$k(y, y') = exp(-\gamma||c(y) - c(y')||^2), \tag{5.5}$$

where y and y' are molecules, and c is the function mapping a molecule to its fingerprint vector.

In a follow-up article, however, Brouard and co-workers went on to show that substituting the norm in the Gaussian kernel with the Tanimoto similarity for the molecular fingerprint

vectors yields a slight improvement in performance [152]. For binary vectors X and Y, the *Gaussian-Tanimoto kernel* is given by

$$d(X,Y) = \exp\left(\frac{-T(X,Y)}{2\sigma^2}\right) \tag{5.6}$$

where

$$T(X,Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)} \tag{5.7}$$

is the Tanimoto similarity between X and Y.

In our context, as we take the Gaussian of the Tanimoto similarity as the inner product on $\mathcal{F}$, ϕ is simply the mapping of $y \in \mathcal{Y}$ to its molecular fingerprint, and h is the mapping of a spectrum to its molecular fingerprint, i.e. the molecular fingerprint of the metabolite that gives rise to the spectrum.

The structure of the IOKR model is described in detail by Brouard and co-workers in [146], and its application to metabolomics is further described in [113], where the relevant equations are derived.

5.8 Using IOKR to score links

As described previously, the idea behind IOKR is to project the various elements into a shared latent space and rank the matches based on similarity in the latent space. A link between a spectrum x and a metabolite y will be assigned a high score if the fingerprints predicted for the spectrum and calculated for the metabolite are similar. Although [113] presents IOKR as a method to rank candidate molecules for spectra, it can also be used to rank candidate molecules for BGCs, as long as we have an appropriate kernel function on the space of BGCs. In this way, we can break the problem of mapping BGCs to spectra into two components: mapping BGCs to metabolites, and mapping spectra to metabolites.

Each of these models can be trained individually and evaluated in isolation, and then coupled together to link BGCs and spectra, bypassing the need for a candidate molecule set.

The following sections are dedicated to these problems in turn. Section 5.10 covers the IOKR model linking MS2 spectra to metabolites, Section 5.11 covers the IOKR model linking BGCs to metabolites, and Section 5.12 covers the combined model linking BGCs to MS2 spectra.

Besides adhering more closely to the structure of the model presented in [113], breaking the problem in two in this way, and training each part of the model to use the space of molecular

fingerprints as an intermediary, has the added benefit of a greater amount of available training data. Whereas a model predicting the relations between BGCs and spectra directly would be trained on the combined MIBiG-GNPS data set, which only has 242 unique BGCs and 2069 unique spectra in 2966 pairs, we can use the GNPS data set from [113] to train the model linking spectra to metabolites, and the whole MIBiG database to train the model linking BGCs to metabolites. For the GNPS data set, this means 4138 spectrum-metabolite pairs with 4138 unique spectra and 2919 unique molecules. The MIBiG data set contains 2966 pairs consisting of 249 unique BGCs and 2069 unique spectra. By using the same molecular fingerprints to train the two models, we ensure that they are compatible and can be coupled to rank BGC-spectrum links.

5.9 Using IOKR to rank GCF-MF links

As IOKR ranks links between spectra and BGCs, and the strain correlation score ranks links between GCFs and MFs (or spectra), the two are not directly comparable. To make them comparable, therefore, we must either generalise the strain correlation score to individual BGCs, or the IOKR score to GCF-MF links. The former can be achieved by assigning the score for the GCF-MF link to each individual BGC-spectrum pair where the BGC is in the GCF and the spectrum is in the MF. On the other hand, to generalise the IOKR score to GCF-MF pairs, we take the IOKR score for a GCF-MF link to be the highest IOKR score for a BGC-spectrum pair where the BGC is in the GCF and the spectrum is in the MF, i.e. for a GCF G and a MF M, and a scoring function σ evaluating BGC-spectrum links, we define the generalised scoring function $\hat{\sigma}$ as

$$\hat{\sigma} = max_{m \in M, g \in G} \sigma(m, g). \tag{5.8}$$

In a slight abuse of notation, we will henceforth omit the $\hat{}$ and simply write σ for the generalised scoring function as well, with the understanding that the domain of the function is clear from the context.

5.10 MS2-MIBiG IOKR

To apply the IOKR model as defined in [113] to match BGCs to MS2 spectra, we would like to take the set of BGCs as a candidate set for each spectrum, and rank the links according to the scores obtained by evaluating expression 5.4. However, to rank the candidate set, we need molecular predictions for the BGCs. As mentioned before, for novel BGCs, these are not easily obtained. We must therefore restrict ourselves to BGCs with known structure, or

BGCs that have significant similarities to such BGCs. In practice, this means limiting our candidate set to BGCs that are in MIBiG, or are similar to BGCs therein.

Although the application of IOKR to the BGC-MS2 linking problem is limited to BGCs that are in MIBiG, it can still be considered useful in two ways: Firstly, searching for BGCs with products that are similar to known metabolites is a common starting point in the search for novel metabolites. This approach was for instance used by Duncan and co-workers in [63] to identify the BGC encoding the production of retimycin A. Secondly, assigning BGCs with known products to their spectra can aid in dereplication, i.e. identifying spectra belonging to these known products, and thus help focusing the search for novel metabolites elsewhere by a process of elimination. This is the idea behind DEREPLICATOR, developed by Mohimani and co-workers [103].

To assign metabolite structures to BGCs, we use the antiSMASH KCB results, as described in section 3.8.1. To estimate the similarity of MIBiG entries to a detected BGC, the Pfam domains making up the MIBiG entry are aligned against the sequence of the detected BGC using BLAST [73]. If the sum of scores of the BLAST hits for the individual Pfam domains exceeds 10000 for a given MIBiG entry, we assign any molecular structures associated with the MIBiG entry to the BGC. That means that a single BGC can have zero or more KCB matches, any of which can have one or more associated structures. This results in each BGC in the data set being assigned zero or more molecular structures. The cutoff is needed in order to reduce the noise from low-quality KCB matches, but the data sets do seem stable with regards to the exact value of the cutoff.

The set of all molecular structures that are assigned to a BGC in the data set becomes our candidate set Y^* for IOKR, and we can define a scoring function σ_{IOKR} between a BGC g with an associated set of molecular structures $G' \subset \mathcal{Y}$, and a spectrum $m \in \mathcal{X}$, as

$$\sigma_{\text{IOKR}}(m, g) = max_{g' \in G'} \langle \hat{h}(m), \phi(g') \rangle_{\mathcal{F}} \tag{5.9}$$

5.10.1 Training data

The IOKR model is trained on the same data as in [113], namely, spectra and structural annotations from GNPS [64]. In total, the data set consist of 4138 spectrum-metabolite pairs, with the molecular fingerprints being computed from the InChI or SMILES annotations using the Python bindings for Chemistry Development Kit (CDK) [153]. The fingerprint vector used is a concatenation of three fingerprint vectors: *CDK substructure*, *PubChem Substructure* and *Klekota-Roth*. Taken together, these cover a large majority of the features used in [113], and result in similar performance on the test set used in [113]. The metabolites in the training set come from various sources, and include microbial, plant and human metabolites.

While this data set only represents a small subset of GNPS spectra with structural annotations, this set contains high-quality MS2 data, generated with similar device configurations and high-quality structural annotations. Furthermore, as the same data set has been used in numerous previous experiments (such as [145] and [146]), using this subset as our training data makes comparison with existing work easier.

5.10.2 Kernel choices

In order to be able to evaluate our model on data not included in the training set, we need to be able to evaluate the kernel functions between the new input data and the training data. In [113], Brouard *et al.* use the PPKr kernel on the MS2 spectra, while the kernel function on the metabolites in [152] is the Gaussian-Tanimoto kernel in the molecular fingerprint space, and we follow their example to a large extent.

A common preprocessing step in *in-silico* metabolite identification is to apply some kind of filtering to the input spectra. The PPKr as used by Brouard and co-workers in [113] involves computing a fragmentation tree for each spectrum, using the computed tree to filter the spectrum to only those masses appearing in the tree, and then calculating the PPK as defined in [154] on the filtered spectra. To avoid computing the fragmentation tree for each input spectrum, which is computationally expensive, we instead assemble a collection of all fragment *m/z* occurring in the training set, for which we have precomputed fragmentation trees, and filter the input spectra to only include peaks that match those *m/z* values.

While it may be argued that this biases the model towards peaks seen in the training data, peaks that are not present in the training data can by definition not be used by the predictive model, and cannot therefore provide any information. In addition, the training set appears to be large enough to contain enough ions to build a robust model, and the effects of the bias are small, as confirmed by comparing our implementation of PPKr to the fragmentation-tree-based filtering approach used by PPKr in [113] on the data set used in that paper.

After filtering, the PPK of the resulting spectra is computed as follows.

Given two spectra $\mathbf{x}$ and $\mathbf{x}'$, the PPK kernel [151, 154] on the spectra is defined by modelling each of the spectra as a mixture of Gaussians $p(\mathbf{x})$ and $p'(\mathbf{x}')$ [154]. The peaks $\mathbf{x}_i$ of each spectrum are modelled as $(\text{mass}, \text{intensity})$ tuples, with each $\mathbf{x}_i$ corresponding to a non-isotropic two-variate normal distribution $p_i(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i, \Sigma)$ over the mass and intensity of peak i. Assuming the mass and intensity to be independent, we set $\Sigma' = \Sigma = [\sigma_{\text{mass}}, 0; 0, \sigma_{\text{int}}]$ for both p and p'. Following Brouard and co-workers, we set $\sigma_{\text{int}} = 100000$ and $\sigma_{\text{mass}} = 0.00001$ [113].

Letting k be the kernel function on the distributions, by [154], the kernel function k' on the spectra is computed as

$$k(x, x') = \int_{\mathbb{R}^2} p(\mathbf{x}) p'(\mathbf{x}') d\mathbf{x} \tag{5.10}$$

$$= \frac{1}{k}\frac{1}{k'} \sum_{i,j}^{k,k'} \frac{1}{2\pi |\Sigma|^{\frac{1}{2}} |\Sigma'|^{\frac{1}{2}} |\Sigma_*|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} \left(\mathbf{x}_i^T \Sigma^{-1} \mathbf{x}_i + \mathbf{x}_j'^T \Sigma'^{-1} \mathbf{x}'_j + \mathbf{x}_{*ij}^T \Sigma_*^{-1} \mathbf{x}_{*ij} \right) \right) \tag{5.11}$$

where

$$\Sigma_* = \Sigma^{-1} + \Sigma'^{-1}, \tag{5.12}$$

$$\mathbf{x}_{*ij} = \Sigma^{-1} \mathbf{x}_i + \Sigma'^{-1} \mathbf{x}'_j \tag{5.13}$$

and $\mathbf{x}_i$ and $\mathbf{x}'_j$ are the features from x and x' respectively.

When working with MS2 spectra, the PPK kernel is usually computed for both the spectra and the neutral losses. To compute the kernel for the neutral losses, for each pair of peaks from the original spectrum, the mass is set to the *m/z* difference between the peaks, and the intensity to the average of the two intensities. The values of the kernel function k for the peaks and the losses are computed separately and added together for the final PPK value for the two MS2 spectra [154].

On the space of metabolites, the kernel used is the Gaussian-Tanimoto kernel on the space of molecular fingerprints, i.e. the Gaussian kernel with the l_2-norm replaced by the Tanimoto coefficient, as described in Section 5.7.

5.10.3 Results

Evaluation on MIBiG data As previously mentioned, the data set used to train the IOKR model in [113] contains metabolites from various sources, and performance evaluation is not broken down by metabolite source. To evaluate the performance of the model on microbial secondary metabolites specifically, we can use the paired MIBiG-GNPS data set introduced in Chapter 3.6.

Taking as candidate set the set of all molecular structures associated with a BGC in the data set, IOKR assigns a score to each, which we use to rank the metabolites in the candidate set. To translate this into the rank of a particular BGC, the rank of the BGC is taken to be the number of distinct BGCs associated with any molecule that scores higher than the highest-scoring molecule for the BGC, i.e. if $\sigma_{\text{IOKR}}(m, s)$ is the IOKR score between a metabolite m

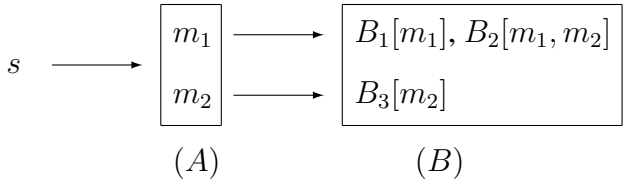


Figure 5.2: **Ranking BGCs based on ranking of metabolites.** (A) The metabolite m_1 has better rank against the spectrum s than the metabolite m_2. (B) Based on the ranking of their associated metabolites, the BGCs B_1 and B_2 are tied for best match with the spectrum s, while B_3 ranks lower than both, despite sharing the associated metabolite m_2 with B_2.

	Top-1	Top-5	Top-10	Top-20	Top-200	AUC
Data	0.1208	0.1708	0.1870	0.2121	0.2946	0.6534
Random	0.0	0.0014	0.0044	0.0103	0.1486	0.5209

Table 5.1: Top-n accuracy and AUC of IOKR on the MIBiG-GNPS data set.

and a spectrum s, and $\mathcal{B}$ is the set of BGCs, where each $B \in \mathcal{B}$ is a set of the metabolites m associated with it, then the rank of the link between a BGC $B \in \mathcal{B}$ and a spectrum s is given by the expression

$$\#\left\{B' \in \mathcal{B} \,|\, max_{m' \in B'} \sigma_{\text{IOKR}}(m', s) > max_{m \in B} \sigma_{\text{IOKR}}(m, s)\right\}. \tag{5.14}$$

As an example, considering Figure 5.2, assume that s is a spectrum, m_1 and m_2 are metabolites which σ_{IOKR} ranks in this order. Assume also that B_1, B_2 and B_3 are BGCs, with m_1 associated with B_1, m_1 and m_2 associated with B_2 and m_2 associated with B_3. By the definition above, B_1 and B_2 are tied at rank 0, while B_3 has rank 2, as both B_1 and B_2 are ranked higher, because m_1 has better rank than m_2, and is used to place both b_1 and b_2 within the ranking.

To estimate a baseline performance for comparison, the rank of the metabolites was randomised, and the same logic used to rank the BGCs.

Table 5.1 shows the top-n performance of IOKR on the paired MIBiG-GNPS data, i.e. how often the correct BGC for a spectrum is among the top-n BGC matches for the spectrum, for a selection of n, as ranked by IOKR.

The last column of Table 5.1 is the AUC for the proportion of spectra correctly identified as a function of increasing n. Recall that AUC of 1.0 would in this context mean that the top-1 accuracy was 100%, i.e. for all spectra, a molecule associated with the correct BGC would be the top ranked molecule.

The proportion of MS2 spectra with a correct match in the top-n ranked BGC candidates can be seen as a function of n in Figure 5.3. The AUC for IOKR is 0.6534, compared with

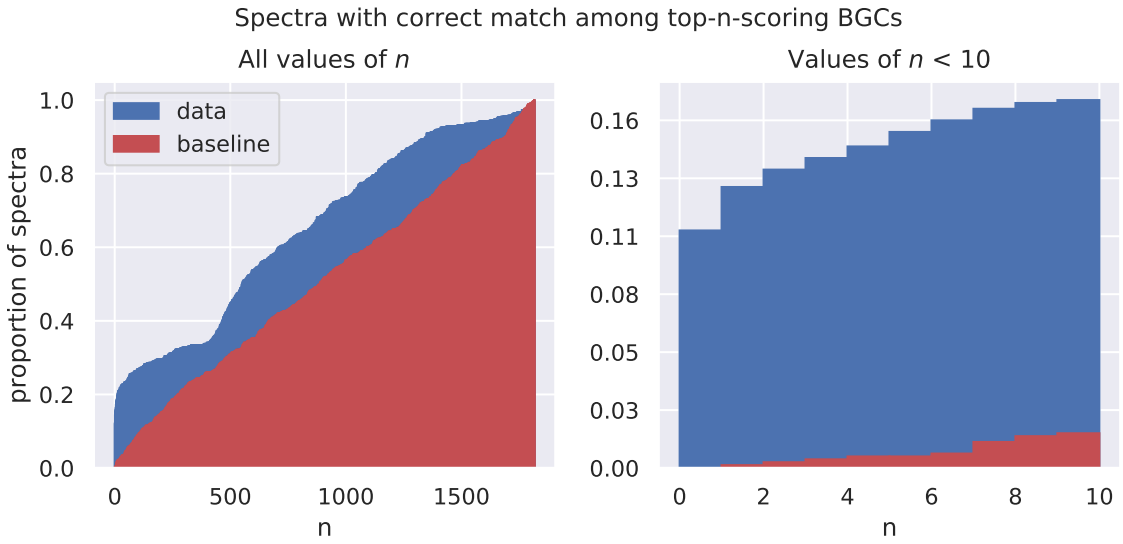


Figure 5.3: **IOKR performance on MIBiG data** Top-n accuracy of IOKR on MIBiG data, i.e. how often the correct BGC for a given spectrum is among the n top ranked BGCs, compared to a baseline acquired by permuting the IOKR scores before ranking. IOKR outperforms random ordering by a rather considerable margin, relatively, especially for low values of n, i.e. when only the top few candidates are considered.

0.5209 for the baseline, and the difference between IOKR and baseline for a selection of n is shown in Table 5.1. IOKR outperforms the baseline by a considerable margin, particularly for low values of n, where the performance of IOKR is relatively much higher than of the randomised baseline.

Evaluation on microbial data sets To evaluate the performance on microbial data sets, we turn to the data sets introduced in Chapter 3.7. The distribution of IOKR scores for all potential links in the data sets can be seen in Figure 5.4, with validated links indicated in black below the histograms.

Looking at the mean IOKR scores for the validated links, and the mean scores for all of the links, an effective scoring function should have a higher mean score for the validated links than for all links. The mean IOKR score for the three microbial data sets is shown in Table 5.2. For all of the data sets, the mean for the validated links is higher than the mean for all links, although the difference is only statistically significant for the Crüsemann data set.

Looking at the relative enrichment of the top quantiles for validated links — i.e. if the proportion of verified links scoring above a certain percentile is higher than would be expected if the scores were random — both the sets of links scoring above 90th and 95th percentiles of all data sets are enriched for validated links relative to the whole set, but this difference is only statistically significant for the Crüsemann data set.

Data set	Mean score all	Mean score valid	p-value
Crüsemann	0.0105	0.0364	1.7968e-9
Leão	0.0014	0.0038	0.3922
Gross	0.02721	0.037020	0.5155

Table 5.2: Mean IOKR score by microbial data set

		Crüsemann	Leão	Gross	total
All	Verified	15	8	5	28
	Total	999362	9342	501886	1510590
	Ratio	1.5010×10^{-5}	0.0009	9.9624×10^{-6}	1.8536×10^{-5}
$> 95\%$ σ_{IOKR}	Verified	6	1	1	8
	Total	49970	437	25095	75502
	Ratio	0.0001	0.0023	3.9849×10^{-5}	0.0001
	p-value	0.0003	0.3378	0.2538	0.0002
$> 90\%$ σ_{IOKR}	Verified	6	1	1	8
	Total	99937	935	50189	151061
	Ratio	6.0038×10^{-5}	0.0011	1.9925×10^{-5}	5.2959×10^{-5}
	p-value	0.009	0.5764	0.4355	0.0139

Table 5.3: Number of validated links with IOKR score above 95th and 90th percentile, and significance according to Fisher's Exact Test, compared to all links.

To increase the statistical power of the tests, we can combine the data sets. As we cannot rely on the distribution of the scores to be identical across the data sets, due to for instance different instrumentation in the MS2 analysis [155], simply combining the scores is not feasible. However, by looking at the number of links that score above a given threshold in each data set, we can pool the counts of validated and potential links, and consider the relative overrepresentation of validated links in the top percentiles of the distribution of the scores both for individual data sets and for the combined link count.

The number of links scoring above the 95th and 90th percentiles for all data sets can be found in Table 5.3. Adding up the data sets, i.e. the number of validated links in the top percentiles and the total number of links across all data sets, validated links are significantly overrepresented above both percentiles.

5.10.4 Investigating the function of IOKR

As with many machine learning approaches, the function of IOKR is somewhat of a black box. In order to further validate IOKR, we therefore tried, for a few BGC-spectrum pairs, to identify specific peaks in the MS2 spectrum as belonging to specific substructures of the

metabolites. We then removed these peaks from the spectra, in order to see if we could identify substructures (represented by peaks) that had a particular influence on the IOKR score of the match relative to other matches, where removing the peak lead to a change in the ranking of the BGC-spectrum pair, and furthermore if that particular substructure could be considered characteristic of the metabolite in some way.

The starting point for the verification was a BGC-spectrum pair, where the BGC has an associated metabolite. To identify the peaks in the MS2 spectrum that corresponded to a particular substructure, the metabolite was fragmented using MetFrag [156]. Using the accurate mass reported by name-based search for the metabolite in NPAtlas [157], a candidate set for MetFrag was generated by searching the `NPAtlas_Aug2019` database in MetFrag for molecules with the same neutral mass. Ideally, this should only return one candidate, but in cases where multiple candidates were returned, only candidates where the name of the candidate matched the name of the metabolite were analysed further. Finally, the candidate molecules were fragmented using the MetFrag *in-silico* fragmentation algorithm with default settings.

The relevant spectra from the MF were extracted using the Metabolomics Spectrum Resolver [158] and their peaks matched to the hypothetical fragmentation spectra for the candidate molecules. Peaks that matched were then filtered one by one from the spectra, and the change in IOKR score against the metabolite, relative to the change in IOKR score against other metabolites, observed. A decrease in the score relative to other potential links manifests as a worse ranking for the BGC-spectrum pair using the filtered spectrum relative to the unfiltered one.

Results The links subjected to this analysis were all the validated links in the Crüsemann data set, along with two unverified links with high σ_{IOKR} and $\bar{\sigma}_{\text{corr}}$ scores. For three examples, a single peak identified as corresponding to a particular substructure had a large impact on the ranking of the spectrum against the BGC, among the candidate set of other spectra.

The metabolite rosamicin, which as a product of the MIBiG BGC BGC0001830 has a validated link with spectrum 93193, has an accurate mass of 581.3564, as reported by NPAtlas. This yielded an unique match for neutral mass in the `NPAtlas_Aug2019` database in MetFrag. Comparing the predicted spectrum for that match to spectrum 93193 from the Metabolomics Spectrum Resolver, MetFrag matched 12 out of 31 peaks with a raw score of 67.1248. Out of the matched peaks, removing the peak at *m/z* 158.117004 (which matched a peak at 158.11762 Da in the predicted spectrum) changed the (0-based) rank of the metabolite against the spectrum from 16 to 163. The predicted peak represents a biological subunit, aminosugar, of the metabolite. The spectrum from the Metabolomics Spectrum Resolver, and the metabolite from MetFrag, can be seen in Figure 5.5, with the identified subunit of

the molecule highlighted in green.

The metabolite grisochelin, which the NPAtlas database reports as having accurate mass of 568.4339, is hypothetically linked to spectrum 51165. Searching the `NPAtlas_Aug2019` database in MetFrag for molecules with neutral mass equal to the reported accurate mass yielded two candidates. Curiously, only one result line showed up in the interface. This metabolite was annotated as zincophorine, which is another name for grisochelin [159], and so was used for the analysis. Comparing the predicted spectrum for the metabolite to the spectrum from the Metabolomics Spectrum Resolver yielded two matched peaks out of 18, with a raw score of 2.9841. Removing one of those peaks, the peak at *m/z* 95.086995 (which matched a peak at 95.08558 Da in the predicted spectrum) changed the (0-based) rank of the metabolite against the spectrum from 0 to 3. The spectrum and the metabolite from MetFrag can be seen in Figure 5.6, with the identified subunit of the molecule highlighted in green.

The metabolite staurosporine, which as a product of the MIBiG BGC BGC0000827 has a validated link with spectrum 44982, has an accurate mass of 466.2005, as reported by NPAtlas. Using this as a selection criteria for neutral mass in the `NPAtlas_Aug2019` database in MetFrag yielded seven candidates, only one of which was annotated as staurosporine, and which was used for subsequent analysis. Comparing the predicted spectrum to spectrum 44982 from the Metabolomics Spectrum Resolver, MetFrag matched seven out of 22 peaks with a raw score of 159.7769. Out of the matched peaks, removing the peak at *m/z* 56.049999 (which matched a peak at 56.0495 Da in the predicted spectrum) changed the (0-based) rank of the metabolite against the spectrum from 1 to 67. The spectrum and the metabolite from MetFrag can be seen in Figure 5.7, with the identified subunit of the molecule highlighted in green.

Discussion For the three examples discussed, we have identified an ion corresponding to a particular substructure which IOKR considers important in matching the spectrum to the metabolite. Being able to identify the substructure as characteristic of the metabolite further validates the method, as this reinforces the validity of the link, as well as validating the emphasis IOKR puts on that particular feature.

5.11 BGC-MIBiG IOKR

Although presented as a way to map spectra to molecules in [113], IOKR can just as easily be trained to map *BGCs* to molecules, given a suitable kernel function on the BGCs. Doing so would give the second piece needed to map MS2 spectra to BGCs, in addition to the MS2-MIBiG IOKR discussed in the previous section.

In this formulation, instead of being the space of MS2 spectra, $\mathcal{X}$ is the space of BGCs. $\mathcal{Y}$ remains the space of metabolites, and $\mathcal{F}$ the feature space of molecular fingerprints.

The calculations follow the exact same blueprint: in the first phase (training phase), we approximate a function $h : \mathcal{X} \to \mathcal{F}$ mapping the input BGCs to their corresponding fingerprint vectors with a function

$$\hat{h} : \mathcal{X} \to \mathcal{F}, x \mapsto \sum_{x_j \in X} \mathcal{K}_{\mathcal{X}}(x, x_j)\mathbf{c}_j \tag{5.15}$$

while in the second phase (prediction phase) we search a candidate set $Y^* \subset \mathcal{Y}$ for the molecule $y \in Y^*$ that maximises the expression

$$\langle \hat{h}(x), \phi(y) \rangle_{\mathcal{F}} \tag{5.16}$$

for a given x.

5.11.1 Kernel choices

As discussed in Section 2.2.2, defining a similarity measure on the space of BGCs is still an open question. The most advanced such method, used by BiG-SCAPE [82], is based on a great deal of expert knowledge, and incorporates many of the components of the distance measures defined previously, such as Pfam domain composition (used in e.g. [83]) and sequence similarity (used in e.g. [84], although they use a different way of evaluating the similarity).

However, IOKR requires the similarity function on the input space to be a kernel function, and there is no reason to believe that the similarity function used by BiG-SCAPE satisfies the conditions. In fact, the Gram-matrix for the Crüsemann data set has non-negative eigenvalues, demonstrating that the BiG-SCAPE similarity function is not in fact a kernel function. While the use of such *indefinite* (or *non-positive*) kernels in optimisation problems has been studied to some extent, the results have been mixed, and indicated that in many cases, the global optimality of the solution cannot be guaranteed [148, 160, 161, 162].

This leaves two options: either use a different similarity function on the BGCs, that is guaranteed to be a kernel, or convert the BiG-SCAPE similarity function into a kernel function.

A straightforward way of defining a similarity function on the BGCs is to consider each BGC as a *word vector*, i.e. establish a "dictionary" of all the possible Pfam domains and regard each BGC as a vector in the space where each Pfam domain has a corresponding axis, and the magnitude of the vector along any given axis is equal to the number of times that the Pfam domain corresponding to the axis appears in the BGC. Taking this a step further, we can

apply *tf-idf correction* to the word vector, which can in many cases improve the performance of machine learning tasks [163, 164].

Both the word count vector and the tf-idf vector are vectors in $\mathbb{R}^n$, where n is the size of the dictionary, i.e. the number of distinct Pfam domains in the set of BGCs. A wide variety of kernels exist defined on $\mathbb{R}^n$. The most common one is the Gaussian kernel in the l_2-norm, $k(v, v') = exp(-\gamma||v - v'||^2)$, which we use for both the word count and tf-idf vectors going forward [147].

Another option to define a kernel on the space of BGCs is to convert the BiG-SCAPE similarity function into a kernel function. This can be done in several ways. An important consideration for the purpose of using this kernelisation in our context is out-of-sample extension, i.e. given that the kernel function is defined in terms of the training set, we require that it be extensible to new BGCs while still keeping the kernel property.

Chen and co-workers suggest four approaches to turning a similarity function into a kernel function, given the Gram matrix of the training data set [148]. Two of these can be represented as linear transformations defined in terms of the training set, and are shown to extend to new data points in a consistent fashion, meaning that the transformation can be computed once, at training time, and new data points be transformed accordingly without need to recompute the eigendecomposition of the Gram matrix to incorporate the new points.

Both of these involve directly manipulating the eigenvalues of the matrix. The first approach (*clip*) sets all negative eigenvalues to zero, while the second (*flip*) replaces negative eigenvalues with their magnitude, i.e. if λ is a negative eigenvalue, it is replaced with $-\lambda$. Furthermore, as opposed to the other approaches considered in [148] and [165], Chen and co-workers show that the two methods translate easily to new samples via linear transformations [148], and both of them have some theoretical justification – in particular, it can be shown that for a given matrix X, that is not positive semidefinite, the matrix X_{clip} where all the negative eigenvalues have been set to 0, is actually the closest positive semidefinite matrix to X in the Frobenius norm

$$||Y||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|y_{ij}|^2} \tag{5.17}$$

where Y is a $m \times n$ matrix [148].

Recall from section 2.2.2 that the BiG-SCAPE similarity function is defined for BGCs b_1 and b_2 as

$$\sigma_{\text{BS}}(b_1, b_2) = \alpha JI(b_1, b_2) + \beta AI(b_1, b_2) + \gamma DSS(b_1, b_2) \tag{5.18}$$

Shared BGC	No shared BGC	p-value
0.9731	0.7451	8.9567×10^{-65}

Table 5.4: The mean similarity for the molecular fingerprints for pairs of SMILES depending on if they are annotated as products of the same BGC. The p-value is for the t-test for the mean of the two distributions being the same.

where JI is the Jaccard index, i.e. the percentage overlap in Pfam domains, DSS is the direct sequence similarity for matched Pfam domains, and AI is the similarity of pairs of domains, and α, β and γ are coefficients, potentially with different values according to product class.

As the linear combination of kernel functions is a kernel function, and the BiG-SCAPE similarity function is a linear combination of functions, the kernelisation can be applied in two places: Individually to each component function, or as a last step, to the final similarity function. Furthermore, if the component functions are kernelised individually, the weights may no longer be optimal. We therefore turn to *multiple kernel alignment* (MKA) [166] to recompute the weights of the component functions and compare the performance of these recalibrated weights with the original BiG-SCAPE weights.

In short, given a matrix Y and a set of matrices X_i, MKA solves a quadratic programming problem to determine the vector $\boldsymbol{\alpha} = \{\alpha_i\}$ such that the linear combination of matrices $\sum \alpha_i X_i$ is as close to Y as possible in the Frobenius norm.

5.11.2 Training set

As MIBiG contains structural annotations for the majority of the entries, the BGC-MIBiG IOKR model can be trained on a large part of the MIBiG database. Recall from section 5.5 that the training data for the IOKR model consists of paired data points (x, y) where $x \in X \subset \mathcal{X}$ and $y \in Y \subset \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the sets of BGCs and metabolites respectively, and the training phase approximates a function $h : \mathcal{X} \to \mathcal{F}$ into the set $\mathcal{F}$ of molecular fingerprints.

This model is not completely accurate, as the same BGC can have more than one associated metabolite. Such instances can be found in the MIBiG training data, where for example multiple structural annotations exist for BGC0000241, which codes for four different variants of lomaiviticin. However, as can be seen in Table 5.4, different metabolites produced by a single BGC generally have very similar fingerprints. We therefore choose to ignore this discrepancy and treat each BGC as producing a single molecule in the IOKR model.

The instances of multiple metabolites corresponding to a single BGC in the training data are handled by treating each pair of BGC and metabolite as a different data point in the training

set, i.e. the BGC x producing the metabolites y_1 and y_2 results in two points (x, y_1) and (x, y_2) in the training data set.

As BiG-SCAPE requires a specially annotated version of MIBiG to correctly sort the entries into natural product classes, and these annotations are not at the time of writing available for MiBiG version 2.0, the analysis in this section is done using MIBiG version 1.4.

For version 1.4, the MIBiG training data set includes 1205 BGCs producing 1452 distinct metabolites. The relationship between BGC and metabolite can be many-to-many, i.e. multiple BGCs can produce the same metabolite, and each BGC can produce multiple metabolites. The final data set, after expanding all BGCs into a separate data point for each of their metabolites, consists of 1692 BGC-metabolite pairs.

Performance is estimated using 10-fold cross validation on the set of BGCs, where the randomisation is done in such a way that all BGCs that share a SMILES string belong to the same fold.

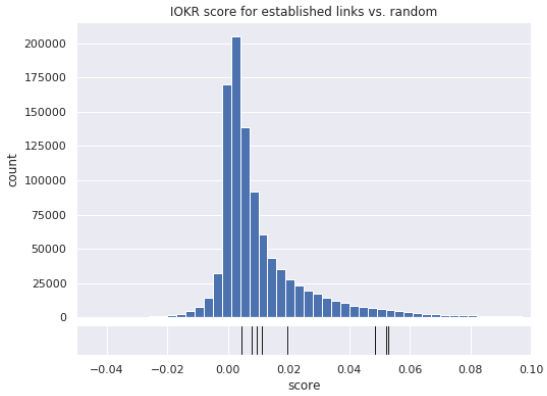
5.11.3 Results

As a simple measure of the total accuracy of the model, we can look at the AUC measurements of the top-n accuracy with varying n, as discussed in Section 5.10.3. Recall that AUC of 0.5 corresponds to random, and AUC of 1.0 corresponds to the highest-ranking result being correct in all cases.

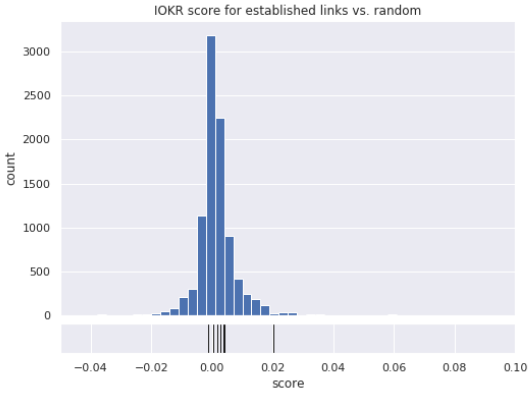
Table 5.5 shows the AUC for the different kernels, broken down by natural product class, with the 'mix' class including all of the product classes. Similar data for the top-n performance for $n = 1$, $n = 5$ and $n = 20$ can be found in Table 5.7. Looking at the AUC results, the best-performing kernel on the BGC space is the Gaussian kernel on the Pfam word vectors, which outperforms the other kernels, including the tf-idf adjusted word vectors, for six out of eight product classes, as well as on the 'mix' class. Furthermore, when kernelising the component pars of the BiG-SCAPE similarity function, recalibrating the weights for the individual metrics does not seem to significantly impact the performance (see the "recal" columns in Tables 5.5 and 5.7). Since the weights resulting from the MKA optimisation are not particularly close to the weights BiG-SCAPE itself uses, the similarity function appears to be relatively robust to changes in coefficients, at least for this use.

Considering the kernelisation of the composite BiG-SCAPE similarity function, *clip* outperforms *flip* for all product categories, while both approaches outperform the model trained on the raw BiG-SCAPE similarity values (which the model accepts, even though they do not fulfill the criteria for being kernel functions). Kernelising the component functions individually performs slightly better than kernelising the combined function for six out of eight natu-

Crüsemann



Leão



Gross

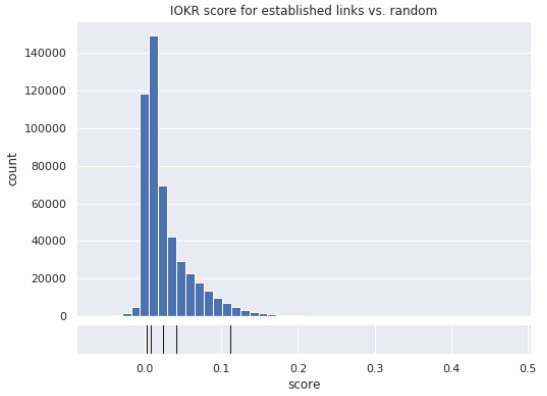


Figure 5.4: **Distribution of validated links in IOKR scores of all potential links** Crüsemann, Leão and Gross data sets.

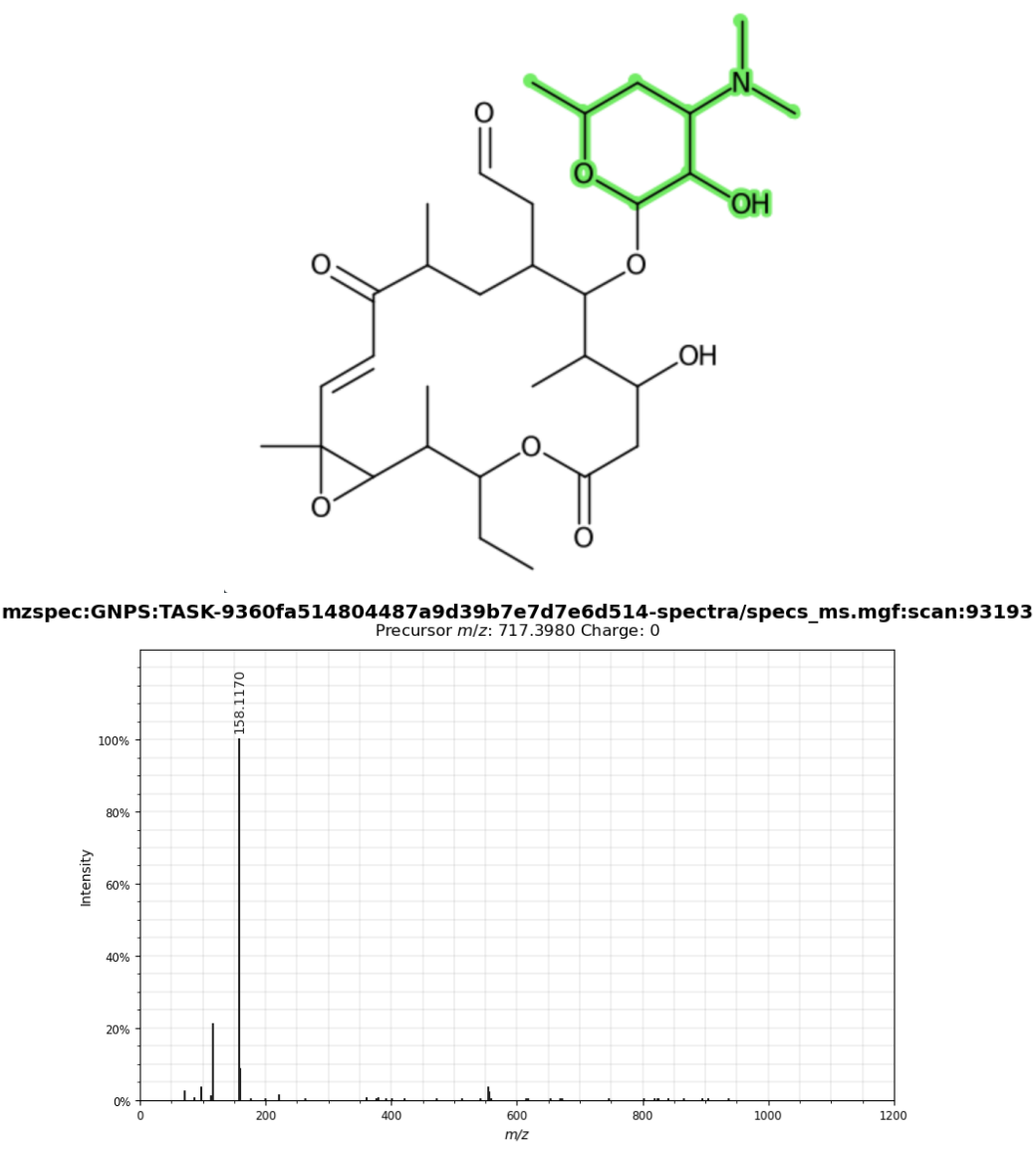


Figure 5.5: **Substructure of rosamicin identified as influencing the IOKR score.** The metabolite rosamicin, and the verified spectrum linked with the metabolite. The substructure identified as particularly influencing the ranking is indicated in green and can be identified as an aminosugar. The difference in reported precursor *m/z* in the image and the mass discussed in the main text is due to each MF potentially containing several different precursor ion *m/z*, and σ_{IOKR} being calculated for a MF using the highest-scoring spectrum from the MF, regardless of the precursor *m/z*.

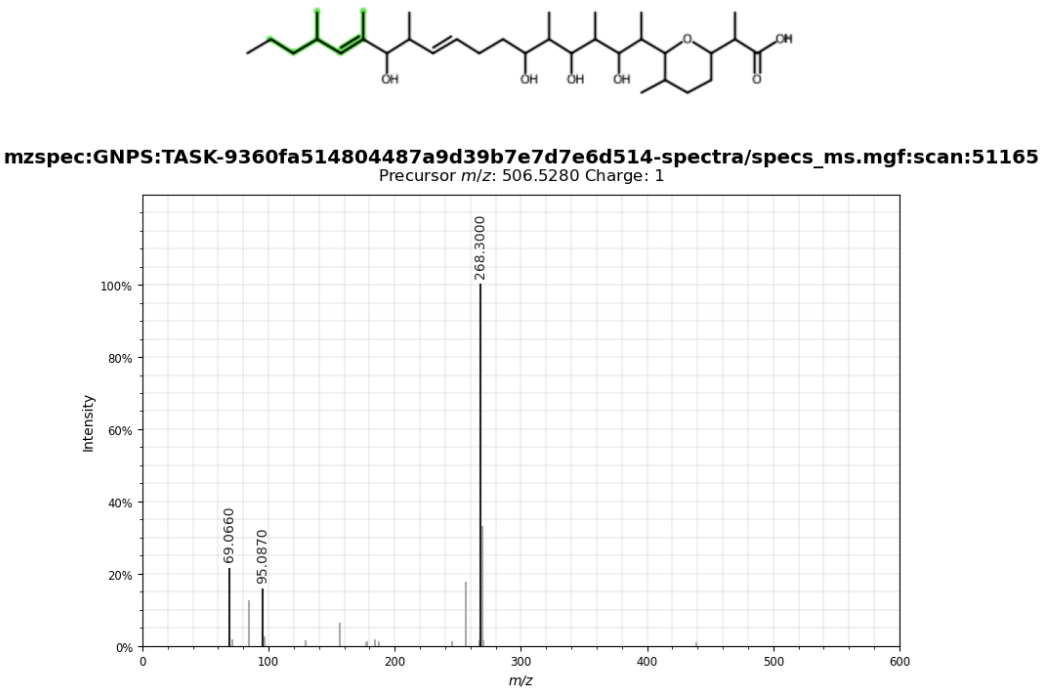


Figure 5.6: **Substructure of grisochelin identified as influencing the IOKR score.** The metabolite grisochelin, and the spectrum potentially linked with the metabolite. The substructure of the metabolite identified as particularly influencing the ranking is indicated in green.

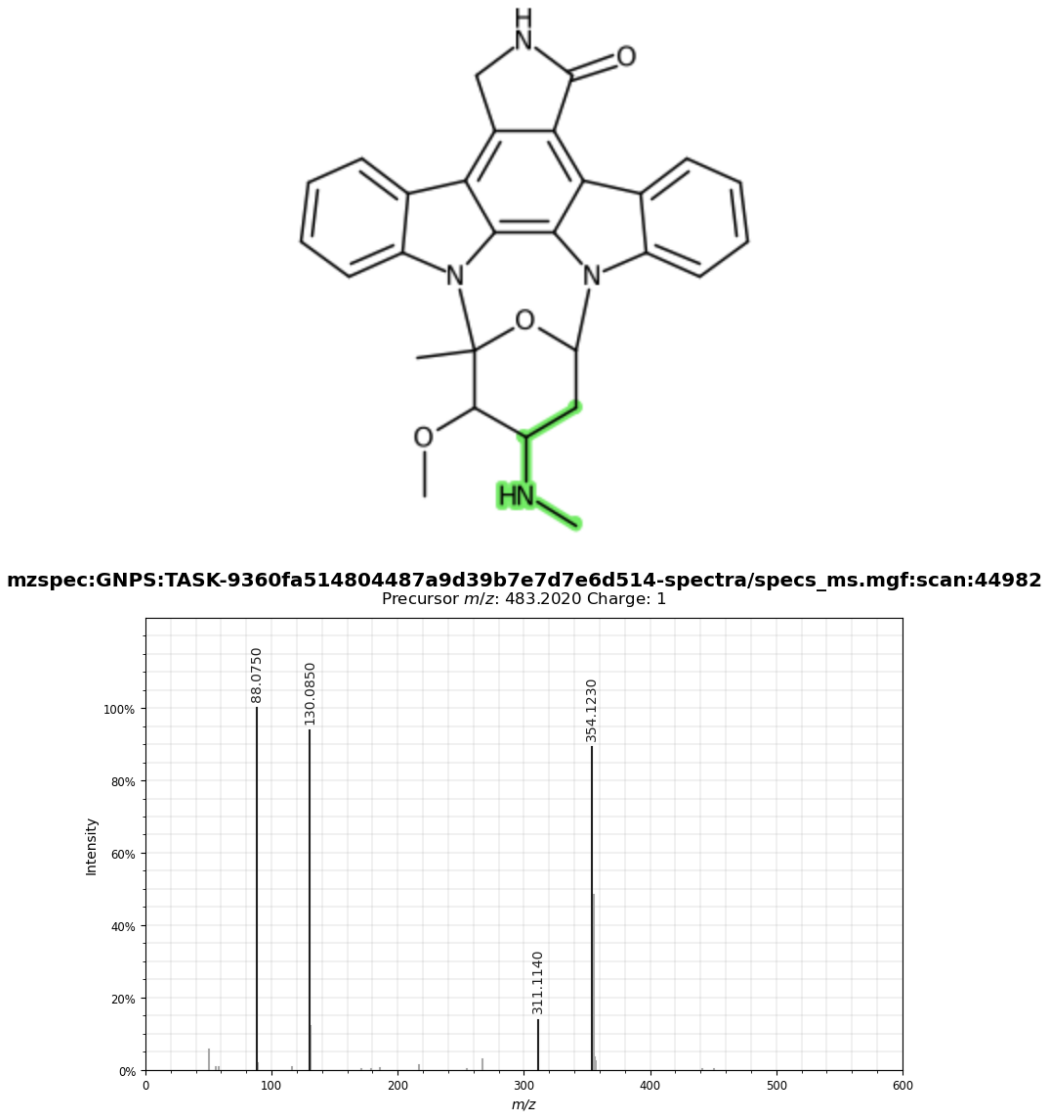


Figure 5.7: **Substructure of staurosporine identified as influencing the IOKR score.** The metabolite staurosporine, and the verified spectrum linked with the metabolite. The substructure identified as particularly influencing the ranking is indicated in green.

	bs	bs clip	bs flip	wv	tfidf	bs raw clip	bs recal clip	bs raw flip	bs recal flip
NRPS	0.568	0.735	0.646	**0.819**	0.780	0.754	0.757	0.754	0.628
Others	0.606	0.688	0.650	**0.778**	0.739	0.638	0.644	0.638	0.573
PKSI	0.547	0.741	0.646	**0.794**	0.758	0.773	0.768	0.726	0.604
PKS-NRP Hybrids	0.596	0.725	0.618	**0.798**	0.770	0.772	0.772	0.772	0.691
PKSother	0.575	0.693	0.627	**0.841**	0.817	0.708	0.713	0.708	0.638
RiPPs	0.666	0.762	0.718	0.757	0.764	0.857	0.850	**0.860**	0.768
Saccharides	0.697	0.720	0.704	0.768	0.772	0.766	**0.810**	0.766	0.673
Terpene	0.538	0.627	0.595	**0.697**	0.678	0.564	0.558	0.531	0.487
mix	0.544	0.752	0.691	**0.855**	0.821	0.733	0.732	0.698	0.613

Table 5.5: AUC values for BGC-MIBiG IOKR for different kernel functions, broken down by natural product class. Highest AUC per class is in bold. Using a Gaussian kernel on a simple Pfam word vector yields the highest AUC in six out of eight natural product classes, as well as for the combined set of all BGCs (‘mix’).

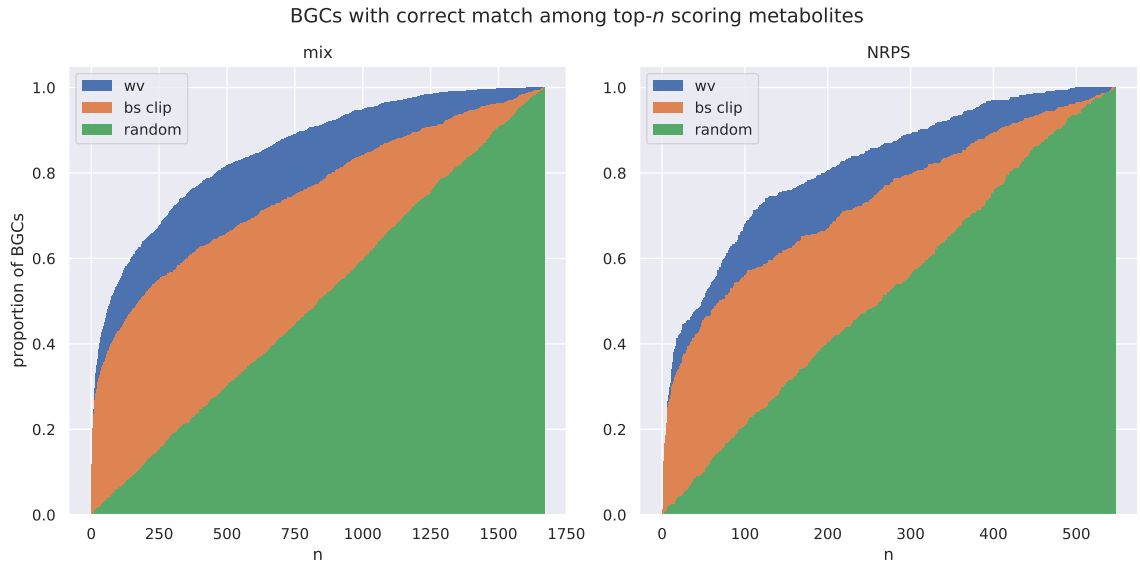


Figure 5.8: Top-n performance for varying n for the word vector kernel, and the kernel derived from BiG-SCAPE using the *clip* method, compared to a randomised baseline. While both kernels perform better than the baseline, the word vector kernel outperforms the BiG-SCAPE kernelised similarities.

ral product classes, with negligible difference in performance between the two kernelisation methods applied to the component functions individually.

Considering the Top-n results presented in Table 5.7, the best-scoring method is less clear. The Pfam word vector and the tf-idf corrected vector show similar performance for the values of n considered, with the kernelised BiG-SCAPE metric showing comparable performance using the *clip* method at $n = 1$.

A graph of the Top-n accuracy for varying n can be seen in Figure 5.8, for the word vector kernel and BiG-SCAPE similarity, kernelised using the *clip* method, on the 'mix' and 'NRPS' natural product classes, along with randomised baseline. While both kernels outperform the randomised baseline, the word vector kernel outperforms the BiG-SCAPE-derived kernel by a considerable margin, as demostrated in Tables 5.7 and 5.5.

Another way of extending the BiG-SCAPE metric to the 'mix' class is to use the 'mix' similarity measurement only for intra-class similarities, i.e. where the elements being compared belong to different classes, and use the class-specific similarity measurements for within-class comparisons. Based on the performance of the BiG-SCAPE kernelised versions of the individual classes, we use the 'clip' method to kernelise the similarity matrix, and refer to the measurement as BiG-SCAPE composite.

Table 5.6 shows the top-1, top-5 and top-20 accuracies for word vector, tf-idf and BiG-SCAPE composite kernels, as well as the AUC for the kernels. Similarly as for the case of individual classes, the BiG-SCAPE-derived kernel seems to shows similar performance to

	Top-1	Top-5	Top-20	AUC
bs composite	0.0054	0.1580	0.2902	0.7461
tf-idf	0.0054	0.1502	0.3303	0.8189
wv	0.0072	0.1544	0.3339	0.8544
random	0.0006	0.0014	0.0116	0.5024

Table 5.6: Top-n performance, and AUC, of three of the best-performing kernels The BiG-SCAPE composite kernel consists of all BGCs, where inter-class distances are computed using the 'mix' distance metric, while intra-class distances are computed using the class-appropriate metric. The simple word vector metric achieves the best performance for $n = 1$ and $n = 20$, as well as the highest AUC.

	bs	bs clip	bs flip	wv	tfidf	bs raw clip	bs recal clip	bs raw flip	bs recal flip
NRPS	0.002	**0.013**	0.004	0.005	0.004	0.009	0.002	0.009	0.002
Others	0.002	0.004	0.008	0.004	0.006	**0.012**	0.004	0.008	0.004
PKSI	0.000	0.011	0.004	**0.015**	0.013	0.011	0.013	0.009	0.000
PKS-NRP Hybrids	0.000	**0.004**	**0.004**	**0.004**	0.000	**0.004**	**0.004**	**0.004**	**0.004**
PKSother	0.000	**0.003**	0.000	**0.003**	**0.003**	0.000	**0.003**	0.000	0.000
RiPPs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Saccharides	0.000	0.000	**0.015**	**0.015**	0.000	0.000	0.000	0.000	0.000
Terpene	0.000	**0.013**	0.000	**0.013**	0.000	0.006	0.006	0.006	**0.013**
mix	0.002	**0.013**	0.004	0.008	0.005	0.009	0.010	0.006	0.001

	bs	bs clip	bs flip	wv	tfidf	bs raw clip	bs recal clip	bs raw flip	bs recal flip
NRPS	0.031	**0.186**	0.084	0.184	0.173	0.157	0.159	0.157	0.033
Others	0.064	**0.153**	0.085	0.137	0.147	0.145	0.145	0.145	0.048
PKSI	0.002	0.176	0.067	0.191	**0.194**	0.165	0.169	0.135	0.022
PKS-NRP Hybrids	0.046	0.212	0.108	0.229	**0.233**	0.192	0.204	0.192	0.058
PKSother	0.048	0.168	0.082	**0.175**	0.148	0.134	0.148	0.134	0.045
RiPPs	0.268	0.408	0.268	0.352	0.423	**0.437**	0.380	0.394	0.155
Saccharides	0.388	0.403	0.358	0.388	0.358	0.358	**0.418**	0.358	0.164
Terpene	0.051	0.083	0.038	0.083	**0.109**	0.071	0.077	0.058	0.019
mix	0.016	**0.169**	0.077	0.160	0.156	0.155	0.151	0.124	0.039

	bs	bs clip	bs flip	wv	tfidf	bs raw clip	bs recal clip	bs raw flip	bs recal flip
NRPS	0.088	0.338	0.234	**0.414**	0.365	0.308	0.308	0.308	0.115
Others	0.128	0.263	0.209	0.333	**0.337**	0.242	0.255	0.251	0.147
PKSI	0.041	0.296	0.169	0.357	**0.365**	0.302	0.306	0.252	0.085
PKS-NRP Hybrids	0.146	0.400	0.296	**0.479**	0.446	0.367	0.362	0.367	0.204
PKSother	0.127	0.326	0.210	0.447	**0.467**	0.278	0.278	0.278	0.220
RiPPs	0.549	0.676	0.648	0.634	0.690	0.761	0.789	**0.817**	0.690
Saccharides	0.597	0.582	0.567	0.597	0.612	0.657	**0.731**	0.657	0.478
Terpene	0.154	0.224	0.167	0.263	**0.340**	0.231	0.224	0.173	0.128
mix	0.045	0.291	0.184	**0.344**	0.336	0.275	0.272	0.246	0.121

Table 5.7: Top-n performance for the BGC kernel by natural product class, for $n = 1$, $n = 5$ and $n = 20$, with the best performance for each class in bold.

the word vector kernels for low values of n, but with a significantly lower AUC. Which one is preferable depends on how far down the list of links we are prepared to search for correct links.

Results on microbial data sets In the microbial data sets defined in Section 3.7, the links are defined using MIBiG IDs. A reasonable choice of candidate set of metabolites would be all metabolites linked to MIBiG entries, or all metabolites linked to a MIBiG entry with significant homology to a BGC in the data set.

Either way, as the data set would include no metabolites not included in the preceding analysis on MIBiG data, we consider analysis on the microbial data sets to be largely redundant. However, implicit verification on microbial data sets will be carried out in Section 5.12, as part of the combined IOKR model.

5.11.4 Discussion

The results presented in the preceding section show the feasibility of training an IOKR model to map BGCs to metabolites, i.e. to rank a candidate set of metabolites for a given BGC. As previously mentioned, the performance of IOKR is greatly influenced by the choice of kernels, both on the space of BGCs and metabolites.

Although the method is evaluated here both on the individual natural product classes, and the collected set of all BGCs (the 'mix' row in Table 5.5) this is to accommodate the BiG-SCAPE similarity function, which has a different weight vector depending on the natural product class. Interestingly, not only does the word vector kernel on the combined set of all BGCs outperform the other kernels in seven out of nine instances, including on the set of all BGCs, the word vector kernel on the set of all BGCs outperforms the majority of the other kernels, including the product-type-specific BiG-SCAPE similarities. While the BiG-SCAPE similarities may be adversely affected by the kernelisation, it is also interesting to note that for this method, the individual metrics do not seem to provide much additional information over the metric for the combined classes, as evident by the very similar performance of the composite BiG-SCAPE metric and the unaltered BiG-SCAPE metric on the 'mix' class.

Different kernels than ones considered here have shown considerable promise in other applications, for instance the pfam2vec embedding defined in [86], which has been successfully used for BGC detection. While this method is not directly applicable to sequences of Pfam domains, the word2vec model on which it is based [167] has been extended from words to sentences (e.g. [168, 169]), and similar approaches might be applicable to pfam2vec. However, as the present work is mostly intended as a proof of concept, we leave the exact choice

of kernels and the optimisation of parameters for future work.

5.12 MS2-BGC IOKR

The two previous sections introduced IOKR as a method to rank a candidate set of metabolites against an MS2 spectrum (Section 5.10) and a BGC (Section 5.11). While these approaches can be useful in their own right [92, 170] the requirement of a candidate set of metabolites limits their utility. When matching MS2 spectra to a candidate set of BGCs, this restricts the use to BGCs with considerable homology to known BGCs, and for matching BGCs to a candidate set of MS2 spectra, this restricts the use to MS2 spectra which have been identified as belonging to a particular metabolite.

Ideally, we would like to rank a set of BGCs against a set of spectra (or vice versa) without the intermediary of a candidate set of metabolites. To do this, we propose to join the individual models developed in the two previous sections. We refer to this as a *combined IOKR* model.

5.12.1 Combining IOKR models

Recall that in Section 5.5, the IOKR model was defined in terms of an input space $\mathcal{X}$, metabolite space $\mathcal{Y}$ and feature space $\mathcal{F}$ of molecular fingerprints with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. The function $\phi : \mathcal{Y} \to \mathcal{F}$ was the function sending metabolites to their molecular fingerprints, and $h : \mathcal{X} \to \mathcal{F}$ the function sending input objects to their molecular fingerprints. Based on training data, we learned an approximation $\hat{h} : \mathcal{X} \to \mathcal{F}$ of h, which was then used to rank links by computing $\sigma_{\text{IOKR}}(x, y) = \langle \hat{h}(x), \phi(y) \rangle_{\mathcal{F}}$. Section 5.10 described this model where $\mathcal{X}$ was the space of MS2 spectra, while Section 5.11 did the same where $\mathcal{X}$ was the space of BGCs.

We consider the combined model to consist of two sub-models, a MS2-metabolite model, and a BGC-metabolite model. For the combined model, let $\mathcal{Y}$ be the space of metabolites, as before, and $\mathcal{F}$ the space of molecular fingerprints. We take $\mathcal{X}$ to be the space of MS2 spectra, and $\mathcal{Z}$ to be the space of BGCs. We denote by $h_{\mathcal{X}} : \mathcal{X} \to \mathcal{Y}$ the function mapping MS2 spectra to their corresponding molecular fingerprints, and by $h_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Y}$ the function mapping BGCs to their corresponding molecular fingerprint. A diagram of the model can be seen in Figure 5.9.

The combined IOKR model calculates the scores for the BGC-MS2 pairs in the same way as for the individual models, by computing the similarity between the predicted molecular fingerprints, but instead of computing the similarity between a single predicted molecular fingerprint and a candidate set of fingerprints that can be exactly calculated from known

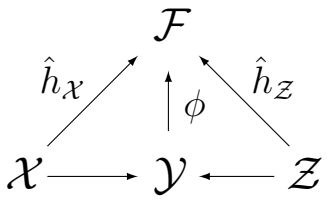


Figure 5.9: **Arrow diagram of the combined IOKR framework.** ϕ is the mapping that sends a molecule y in the space $\mathcal{Y}$ of all molecules to its molecular fingerprint $\phi(y) \in \mathcal{F}$, the shared latent space. $\hat{h}_{\mathcal{X}}$ and $\hat{h}_{\mathcal{Z}}$ are the learned approximation to the functions $h_{\mathcal{X}}$ and $h_{\mathcal{Y}}$ which take a spectrum x and a BGC z to the molecular fingerprint of the corresponding metabolite. Given a spectrum $x \in \mathcal{X}$ and a BGC $z \in \mathcal{Z}$, the potential link between them is assigned a score by computing the similarity between $\hat{h}_{\mathcal{X}}(x)$ and $\hat{h}_{\mathcal{Z}}(z)$, the images of x and y in $\mathcal{F}$ under the learned functions.

metabolites, the scores in the combined model are similarities between two predicted molecular fingerprints, one for the MS2 spectrum and one for the BGC, with pairs that have similar predicted fingerprints assigned a higher score than pairs that have dissimilar predicted fingerprints.

We define the training data sets for the two models, MS2-metabolite and BGC-metabolite, separately. The training data for the MS2-metabolite half of the model consists of a set S_X of tuples (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Similarly, the training data for the BGC-metabolite half of the model consists of a set S_Z of tuples (z, y) where $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$. Importantly, the training data metabolites do not need to be the same for the S_X and S_Y sets. This matters a great deal because otherwise the training set would in effect be restricted to the combined MIBiG-GNPS data set described in Section 3.6. Being able to train the model on separate data sets affords us a much bigger training data set.

Let $\mathcal{K}_{\mathcal{X}}$ be the kernel on the space $\mathcal{X}$, and $\mathcal{K}_{\mathcal{Z}}$ be the kernel on the space $\mathcal{Z}$, with $\psi_{\mathcal{X}}$ as the function such that $\mathcal{K}_{\mathcal{X}}(x_i, x_j) = \langle \psi_{\mathcal{X}}(x_i), \psi_{\mathcal{X}}(x_j) \rangle_{\mathcal{F}}$ for all $x_i, x_j \in \mathcal{X}$, and $\psi_{\mathcal{Z}}$ the function such that $\mathcal{K}_{\mathcal{Z}}(z_i, z_j) = \langle \psi_{\mathcal{Z}}(z_i), \psi_{\mathcal{Z}}(z_j) \rangle_{\mathcal{F}}$ for all $z_i, z_j \in \mathcal{Z}$. Let furthermore M_X and $M_{Y_{\mathcal{X}}}$ be the matrices of $[\psi_{\mathcal{X}}(x_1), \ldots, \psi_{\mathcal{X}}(x_n)]$ and $[\phi(y_1), \ldots, \phi(y_n)]$ respectively, where $(x_i, y_i) \in S_X$, i.e. the column vectors of the matrices are the images in the molecular fingerprint space of the training set points for the MS2 half of the model. Let M_Z and $M_{Y_{\mathcal{Z}}}$ be the corresponding matrices for the BGC half. Finally, let K_X and K_Z be the Gram matrices on the training sets for the spectrum part and the BGC part, respectively.

Similarly to Section 5.5, we aim to learn an approximation to the functions $h_{\mathcal{X}}$ and $h_{\mathcal{Z}}$ by minimising the regularised least-squares loss function with regularisation parameters $\lambda_{\mathcal{X}}$ and $\lambda_{\mathcal{Z}}$, respectively. As demonstrated in [150], the approximation to $h_{\mathcal{X}}$ has the form $\hat{h}_{\mathcal{X}}(x) = W_X \psi_{\mathcal{X}}(x)$, which has the closed-form solution

$$W_X = M_Y(K_X + \lambda_{\mathcal{X}} I)^{-1} M_X^T \tag{5.19}$$

so

$$\hat{h}_{\mathcal{X}}(x) = M_{Y_{\mathcal{X}}}(K_X + \lambda_{\mathcal{X}} I)^{-1} M_X^T \psi(x) \tag{5.20}$$
$$= M_{Y_{\mathcal{X}}}(K_X + \lambda_{\mathcal{X}} I)^{-1} \mathbf{k}_X(x) \tag{5.21}$$

with $\mathbf{k}_X(x)$ as the vector with elements $\mathcal{K}_{\mathcal{X}}(x_i, x) = \langle \psi(x_i), \psi(x) \rangle_{\mathcal{F}}$ for x_i from the training set S_X. Correspondingly, the approximation to $h_{\mathcal{Z}}$ becomes

$$\hat{h}_{\mathcal{Z}}(z) = M_{Y_{\mathcal{Z}}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} M_Z^T \psi(z) \tag{5.22}$$
$$= M_{Y_{\mathcal{Z}}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} \mathbf{k}_Z(z). \tag{5.23}$$

We can then compute the score between a spectrum x and a BGC z as

$$\sigma_{\text{IOKR}}(x, y) = \langle \hat{h}_{\mathcal{X}}(x), \hat{h}_{\mathcal{Z}}(z) \rangle_{\mathcal{F}} \tag{5.24}$$
$$= \langle M_{Y_{\mathcal{X}}}(K_X + \lambda_{\mathcal{X}} I)^{-1} \mathbf{k}_X(x), M_{Y_{\mathcal{Z}}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} \mathbf{k}_Z(z) \rangle_{\mathcal{F}} \tag{5.25}$$
$$= M_{Y_{\mathcal{X}}}(K_X + \lambda_{\mathcal{X}} I)^{-1} \mathbf{k}_X(x) M_{Y_{\mathcal{Z}}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} \mathbf{k}_Z(z) \tag{5.26}$$
$$= (K_X + \lambda_{\mathcal{X}} I)^{-1} \mathbf{k}_X(x) M_{Y_{\mathcal{X}}}^T M_{Y_{\mathcal{Z}}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} \mathbf{k}_Z(z) \tag{5.27}$$
$$= (K_X + \lambda_{\mathcal{X}} I)^{-1} \mathbf{k}_X(x) K_{\mathcal{Y}_{X,Z}}(K_Z + \lambda_{\mathcal{Z}} I)^{-1} \mathbf{k}_Z(z) \tag{5.28}$$

where $K_{\mathcal{Y}_{X,Z}}$ is the matrix of kernel values $k_{xy} = \langle \phi(y_x), \phi(y_z) \rangle$ where y_x belongs to a tuple in the training set S_X for MS2 spectra, and y_z belongs to a tuple in the training set S_Z for BGCs.

5.12.2 Results

Matching an MS2 spectrum to a candidate set of BGCs Using the MIBiG-GNPS data set of matched MS2 spectra and BGCs introduced in Section 3.6, we can match each individual MS2 spectrum to the candidate set of unique BGCs, ranking the BGCs by score.

When establishing the rank of the correct links in the MIBiG-GNPS data set, we again encounter the problem of the true relationship between objects being many-to-many, i.e. each BGC can have multiple associated spectra, and each spectrum can have several associated

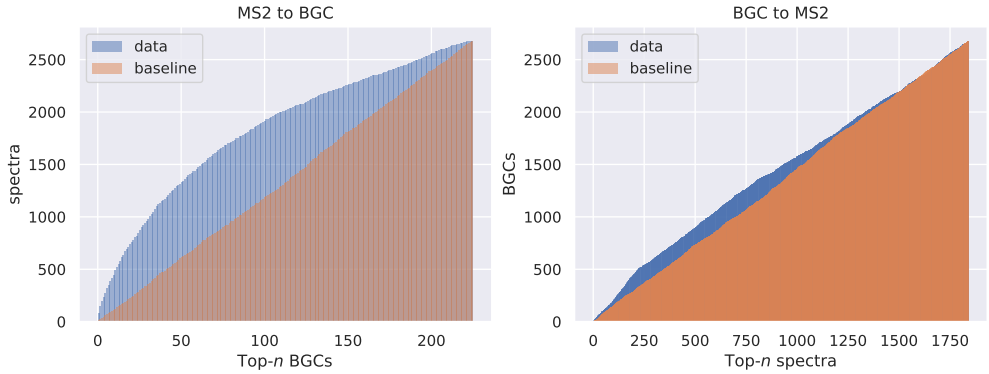


Figure 5.10: Top-n accuracy of IOKR using individual links, matching BGCs to candidate sets of MS2 spectra, and matching MS2 spectra to candidate sets of BGCs.

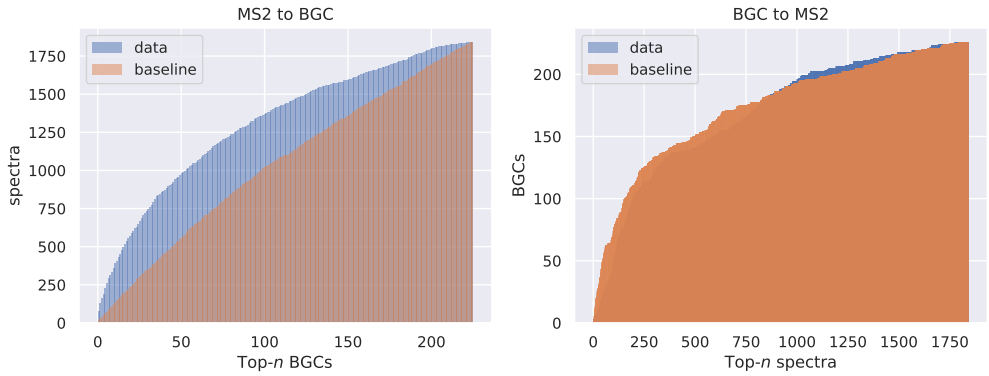


Figure 5.11: Top-n accuracy of IOKR using equivalence classes, matching BGCs to candidate sets of MS2 spectra, and matching MS2 spectra to candidate sets of BGCs.

BGCs. In this case, given an MS2 spectrum, we can have multiple BGCs linked to the spectrum in the ground truth data set. In such cases, we can either take the ranks for the individual links, counting each BGC that is linked to the spectrum as one link, or treat all the BGCs that are matched to the spectrum as an equivalence class, and taking the highest rank of a BGC belonging to the equivalence class corresponding to the spectrum, yielding a single correct link for the spectrum. To distinguish between the two, we refer to these approaches as using *individual links* or using *equivalence classes*.

The latter approach is better indicative of how far the researcher would need to go down the ranked list of links before encountering a correct link for the query object. However, when the subset of correct matches in the candidate set is relatively large, as is the case for matching BGCs to a candidate set of MS2 spectra, this can easily drown out weak signal when randomising the scores.

Table 5.8 shows the Top-n performance and AUC for the combined IOKR model matching MS2 spectra to a candidate set of BGCs individually, i.e. the former of the two matching approaches, while Table 5.9 shows the same data using only the top-ranking BGC for each

	Top-1	Top-5	Top-10	Top-20	Top-200	AUC
Data	0.0299	0.1030	0.1657	0.2746	0.9485	0.6818
Random	0.0052	0.0220	0.0504	0.0922	0.8795	0.4934

Table 5.8: Top-n performance and AUC matching MS2 spectra to a candidate set of unique BGCs, ranking individual links.

	Top-1	Top-5	Top-10	Top-20	Top-200	AUC
Data	0.0407	0.1243	0.1933	0.3094	0.9712	0.7081
Random	0.0109	0.0385	0.0657	0.1200	0.9142	0.5592

Table 5.9: Top-n performance and AUC matching MS2 spectra to a candidate set of unique BGCs, using the highest scoring member of each BGC equivalence class.

spectrum. The randomised data set is created by shuffling the vector of scores for each spectrum before extracting the scores for the correct matches.

Both approaches outperform the randomised baseline. The left-side graphs in Figures 5.10 and 5.11 show the comparison between the performance of the combined IOKR model on the actual and randomised data for individual and equivalence class links, respectively.

Matching a BGC to a candidate set of MS2 spectra Using the MIBiG-GNPS data set to match BGCs to a candidate set of MS2 spectra is complicated by the fact that there are an order of magnitude more MS2 spectra than BGCs. Many of the BGCs have a large number of associated spectra, and therefore treating the spectra as equivalence classes and only choosing the best-ranked spectrum for each BGC risks drowning the signal.

Tables 5.10 and 5.11 show the top-n performance and AUC of the combined IOKR model matcing BGCs to a candidate set of MS2 spectra. Ranking each BGC-spectrum link individually, the model shows a modest performance improvement over the randomised baseline, which is created by randomising the score vector for each BGC before extracting the scores for the correct spectra. However, this performance improvement disappears when treating the spectra as equivalence classes, possibly because of the reasons discussed above.

The performance of the IOKR model on the real data compared with the randomised baseline can be seen on the right-hand side of Figures 5.10 and 5.11 for the individual and equivalence class links, respectively. Note that the unexpectedly high AUC value for the randomised data in Table 5.11 is likely due to the large size of the equivalence classes for the different BGCs.

Microbial datasets To evaluate the performance of the combined IOKR model on the microbial data sets, we again have to choose at which level we want to evaluate the links. Since the validated links are variously at the BGC-MF or BGC-MS2 level, we choose to

	Top-1	Top-5	Top-10	Top-20	Top-200	AUC
Data	0.0011	0.0060	0.0097	0.0194	0.1683	0.5366
Random	0.0004	0.0041	0.0052	0.0112	0.1168	0.4973

Table 5.10: Top-n performance and AUC matching BGCs to a candidate set of unique MS2 spectra, ranking individual links.

	Top-1	Top-5	Top-10	Top-20	Top-200	AUC
Data	0.0133	0.0177	0.0265	0.0398	0.4469	0.7485
Random	0.0044	0.0265	0.0575	0.0973	0.4823	0.7513

Table 5.11: Top-n performance and AUC matching BGCs to a candidate set of unique MS2 spectra, using the highest scoring member of each MS2 equivalence class.

do the matching using MFs rather than individual spectra, for consistency between data sets. This makes it possible for us to pool the results to enhance the statistical power of the analysis.

Furthermore, we choose to evaluate GCF-MF links, rather than BGC-MF links. This is both to stay consistent with the analysis in previous sections, but also to reduce the number of true-but-unverified links in the data set. Because many of the microbes in the data sets are closely related, they are likely to harbour similar BGCs with similar products, yielding potentially high-scoring, true but unverified links that might obscure the signal.

To propagate the BGC-spectrum scores to GCFs and MFs, we take the maximum score over all pairs of BGCs and spectra in the GCF and MF respectively, as in Section 5.9.

Table 5.12 shows the proportion of validated links scoring above the 95th and 90th percentile for each data set, compared to the proportion of validated links in the whole data set. Using the 95th percentile as cutoff, two out of three data sets are significantly enriched for validated links towards the top end of the distribution, while the enrichment is only significant for one out of the three using the 90th percentile. The enrichment is significant for both cutoffs, however, when pooling the results.

Notably, comparing Tables 5.12 and 5.3, in the Leão data set, both the 95th and 90th percentiles are relatively more enriched for validated links than in the MS2-MIBiG IOKR model. This may be because analysis for the MS2-MIBiG model was restricted to BGCs that show considerable homology to MIBiG BGCs. This means that the data set under consideration here might contain more novel predictions, which at the same time might be of lower quality. This could potentially result in a higher number of low-scoring links for the BGCs that are included in the current analysis, but not in the analysis for the MS2-MIBiG model, while all GCFs involved in the set of validated links have considerable homology to MIBiG BGCs by their nature.

		Crüsemann	Leão	Gross	total
All	Verified	15	8	5	28
	Total	1404676	9900	782160	2196736
	Ratio	1.0678×10^{-5}	0.0008	6.3926×10^{-6}	1.2746×10^{-5}
$> 95\%$ σ_{IOKR}	Verified	4	3	1	8
	Total	70234	495	39108	109837
	Ratio	5.6952×10^{-5}	0.0061	2.5570×10^{-5}	7.2835×10^{-5}
	p-value	0.0112	0.0135	0.2538	0.0002
$> 90\%$ σ_{IOKR}	Verified	7	3	1	11
	Total	140468	990	78126	219584
	Ratio	4.9833×10^{-5}	0.0030	1.27998×10^{-5}	5.0095×10^{-5}
	p-value	0.0026	0.0714	0.4355	0.0005

Table 5.12: Number of validated links with IOKR score above 95th and 90th percentile, and significance according to Fisher's Exact Test, compared to all links.

One way to verify this would be to compare the distributions of scores for all links involving GCFs with homology to MIBiG BGCs and for all links involving GCFs that do not have considerable homology to MIBiG GCFs. A lower average score for the latter would support this hypothesis.

Figure 5.12 shows the distribution of validated links in the histogram of scores for all potential links for the three data sets, demonstrating the concentration of the validated links in the upper part of the distribution.

5.13 Summary

In this section, we have proposed two IOKR models, one linking MS2 spectra to known BGCs and the other linking BGCs to metabolites. We then described a way to combine the models to extend them to novel BGCs and spectra, with minimal loss of performance compared to the individual models. Results were demonstrated both on the combined MIBiG-GNPS data set, and on microbial data sets, as applicable. Used as an end-to-end model, the combined IOKR model has the benefit over other models of not being natural product class specific.

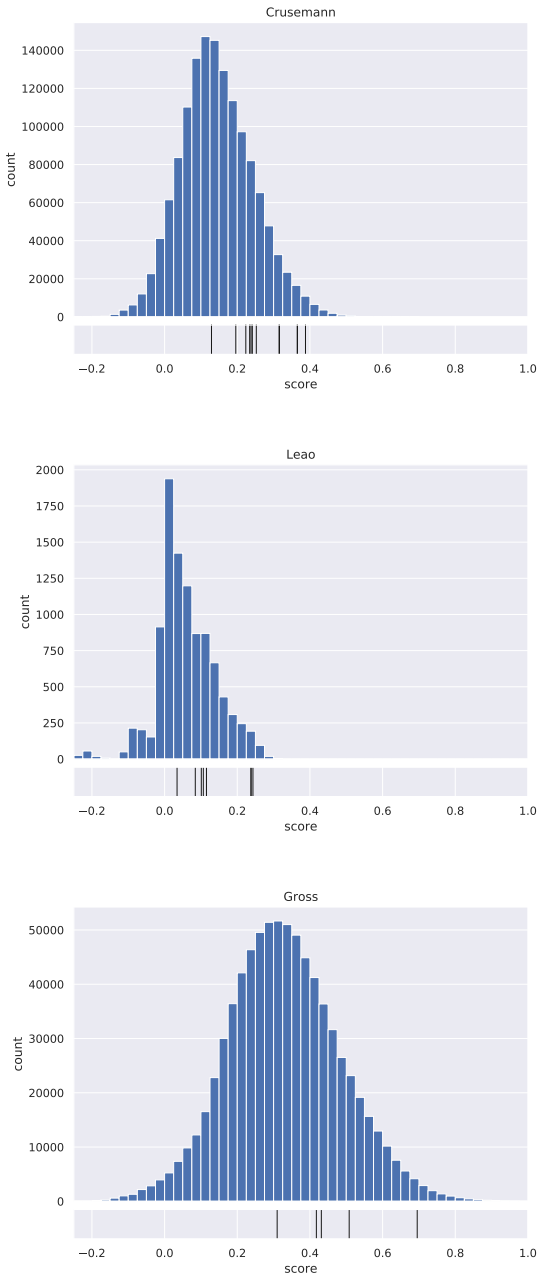


Figure 5.12: Histograms of the distribution of scores for all potential links, and validated links using the combined IOKR model.

Chapter 6

Using multiple scoring functions

6.1 Introduction

The scoring functions described in Chapters 4 and 5 can be considered as representatives of two fundamentally different approaches to the problem of linking genomic and metabolomic information. While the two approaches have been used in tandem, the relation between them has not been previously investigated.

Section 6.2 discusses the strengths and weaknesses of the two approaches to linking, while Sections 6.3 and 6.4 demonstrate their how they each capture different aspects of the problem and are therefore complementary. Finally, Sections 6.5 and 6.6 present a way of combining the scoring functions into a single function which outperforms each of the individual scoring functions on the data sets under consideration.

6.2 Comparing correlation-based and feature-based linking

While the goal of both strain correlation-based and feature-based scoring methods is to rank links between BGCs and MS2 spectra, the approaches discussed in the two preceding chapters obviously rely on a very different set of assumptions, and are based on very different data.

As we have already discussed, both approaches have been used to good effect, and each has their own advantages and disadvantages.

The metabologenomics approach does not rely upon any predictions about the structure of the BGC, and can therefore be said to be more straightforward than the the feature-based approaches. The underlying idea is also simple to understand. However, it relies on having

a considerable number of strains in the population, with larger population size translating into greater statistical power. Furthermore, it relies on the BGC not being cryptic in too many cases. Therefore, it is very sensitive to culture conditions and is not very well suited to conclusively identify rare BGCs or BGCs that are only rarely expressed. Furthermore, the granularity of the metabologenomics approach is by its nature limited, in that it ca not be used to prioritise potential links that show the same correlation pattern. For instance, it offers no way of prioritising potential links involving collections of singleton GCFs and MFs, i.e. GCFs and MFs which only contain BGCs and spectra from one strain.

On the other hand, feature-based approaches are usually limited to a single natural product class. The prediction of adenylation specificity used in Pep2Path [97], for instance, is only applicable to NRPs and RiPPs, and GNP [69] focuses only on NRPs and PKs. In addition, while feature-based methods ideally would be able to predict unique, characteristic features for each BGC, the same features are often predicted for multiple BGCs, leading to identical scores for more than one BGC-spectrum link. Furthermore, not all molecular properties can be detected in MS2 at all, notably stereochemistry.

As the two approaches rely on different underlying data to draw their conclusions, it stands to reason that they may very well be complementary. While this complementarity has in fact been widely implicitly assumed, for instance when using strain correlation analysis as a filtering tool for further analysis, as in [63] and [92], it has neither been quantified nor used in a systematic manner to compensate for potential shortcomings in the different scoring functions.

In the following sections we demonstrate how the standardised strain correlation score and IOKR scores complement one another, and discuss how they can be used in conjunction to better prioritise potential links.

Because the IOKR model presented in the previous Section was initially developed for MS2-MIBiG analysis, as that application was closest to the application presented in [113], the analysis in this chapter is mostly done using the MS2-MIBiG IOKR model. However, Section 6.6 briefly demonstrates similar results for the combined IOKR model introduced in Section 5.12, indicating that the results presented here for MS2-MIBiG IOKR model apply at least partially to the combined IOKR model.

6.3 Complimentarity of correlation- and feature based scores

To demonstrate the complementarity of the standardised strain correlation- and IOKR scores, we can take an approach similar to the one used to demonstrate the efficacy of the individual

		Gross	Leão	Crüsemann	all
all	verified	5	8	15	28
	total	501886	9342	999362	1510590
	ratio	9.96×10^{-6}	0.00086	1.50×10^{-5}	1.85×10^{-5}
$> 95\%$ σ_{IOKR}	verified	1	1	6	8
	total	25095	437	49970	75502
	ratio	3.98×10^{-5}	0.00211	0.00012	0.00011
$> 95\%$ $\bar{\sigma}_{\text{corr}}$	verified	2	6	10	18
	total	35333	1560	50224	87117
	ratio	5.66×10^{-5}	0.00385	0.00020	0.00021
$> 95\%$ both	verified	0	1	4	5
	total	1537	77	2517	4131
	ratio	0	0.01299	0.00159	0.00121

Table 6.1: The number of links (potential and verified) scoring above the 95th percentile for the microbial data sets.

scores. We consider a microbial data set and look at the relative enrichment of validated links in the top quantiles of the individual scores, and of the joint top quantiles of the scores, i.e. links that simultaneously score higher than a given quantile on both scores.

Tables 6.1 and 6.2 show the number of links from the microbial data sets described in Section 3.7 scoring above the 95th and 90th percentiles for IOKR, standardised strain correlation, and both scores, while Figures 6.1-6.3 show the distribution of IOKR- and standardised strain correlation scores for all potential links in the three microbial data sets. The histograms for the individual scoring functions are along the axes. Validated links are indicated with red dots on the plots, and black lines on the histograms.

While the number of strains in the data sets means that the strain correlation score is only particularly informative for the Crüsemann data set, as borne out by Tables 6.1 and 6.2, the upper quantiles of both scores are significantly enriched for validated links, i.e. a higher proportion of the links with the highest scores are validated links than the proportion of validated links in the set of all potential links. Figures 6.1–6.3 demonstrate how the validated links for all data sets are concentrated in quadrant I of the coordinate system, i.e. the quadrant where both x- and y-axis are positive. Furthermore, some of the validated links score relatively higher on the IOKR score than on the standardised correlation score, while the opposite is true for others, again demonstrating the complementarity of the two scoring functions.

As demonstrated by Tables 6.1 and 6.2, the top percentiles for both of the scoring methods contain a relatively higher ratio of validated links than the whole set. In addition, this ratio is higher still when considering the links that score above the top percentiles on both functions.

		Gross	Leão	Crüsemann	all
all	verified	5	8	15	28
	total	501886	9342	999362	1510590
	ratio	9.96×10^{-6}	0.00086	1.50×10^{-5}	1.85×10^{-5}
$> 90\%$ σ_{IOKR}	verified	1	1	6	8
	total	50189	935	99937	151061
	ratio	1.99×10^{-5}	0.00107	6.00×10^{-5}	5.30×10^{-5}
$> 90\%$ $\bar{\sigma}_{\text{corr}}$	verified	4	6	13	23
	total	52014	1560	100494	154068
	ratio	7.69×10^{-5}	0.00385	0.00013	0.00015
$> 90\%$ both	verified	1	1	5	7
	total	5313	147	10836	16296
	ratio	0.00019	0.00680	0.00046	0.00043

Table 6.2: The number of links (potential and verified) scoring above the 90th percentile for the microbial data sets.

While the proportions of validated links in the top scoring links, even for both scoring functions, is small, this is due to the low number of validated links in the data sets. The number of *true* links in the data set — and in particular in the top percentiles — is likely to be much higher.

To overcome the small number of validated links, and to get a clearer sense of the statistical significance of the overrepresentation of links in the upper percentiles of the scoring functions, we can pool the microbial data sets by ranking the potential links in each data set individually and adding up the numbers of total links and validated links in the top percentiles, as discussed in Section 5.10.3. This is done in the last row of Tables 6.1 and 6.2. Considering links scoring above the 90th percentile, the enrichment is significant for both IOKR and the standardised strain correlation scores (with p-values of 0.0139 and 2.483×10^{-11}, respectively), and the enrichment in the combined top percentile for both scores is significantly higher again than for either score alone (with p-value of 2.633×10^{-4} compared with IOKR alone, and 0.0208 compared with the standardised strain correlation score alone).

These results demonstrate the complementarity of the IOKR- and standardised strain correlation scores, at least in the data sets investigated here.

6.4 Links starting from individual BGCs

A common starting point for establishing correspondence between GCFs and metabolites in microbial data sets is to start with a GCF which has significant homology to a known

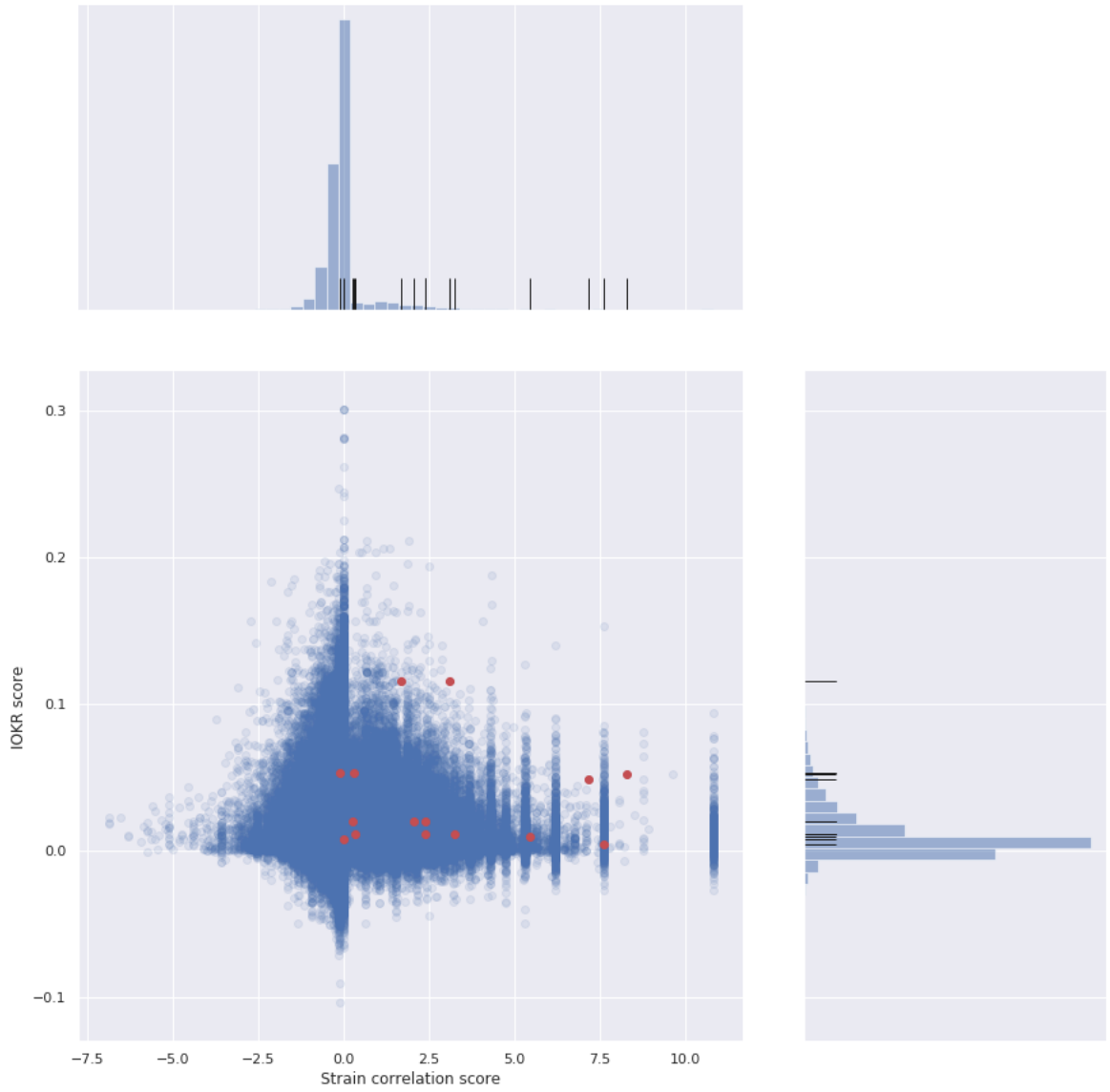


Figure 6.1: **Distribution of IOKR- and standardised strain correlation scores for the Crüsemann data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

BGC, and instead of considering the whole set of potential GCF-MF links, only ranking links involving that particular GCF. As mentioned before, this is the approach Duncan and co-workers used in [63] to identify retimycin A. We now turn our attention to how the scoring functions perform in this setting, both individually and combined.

Table 6.3 shows the rankings of the validated links in the Crüsemann data set, using the relevant GCFs as a starting point. Starting with a particular GCF, a validated link involving that GCF, and the set of potential links that consist of that GCF and each MF in the data set, the table shows the ranking of the validated link within that set.

MIBiG ID	BGC	GCF	MF	Rank of validated link					
				σ_{IOKR}	$\bar{\sigma}_{\text{corr}}$	$\ell_{\frac{1}{2}}$	ℓ_1	ℓ_2	σ_{IOKR} and $\bar{\sigma}_{\text{corr}}$
BGC0001228	990	161	492	384	1152	**253**	271	317	141
BGC0000241	2688	377	381	119	16	**5**	12	14	1
BGC0000333	739	132	489	1810	25	32	**12**	**12**	12
BGC0000827	1642	232	206	6	**1**	2	**1**	**1**	1
BGC0001830	2104	295	309	673	**7**	**7**	9	24	4
BGC0000137	394	71	353	80	16	**2**	14	49	1
BGC0000137	394	333	353	78	189	**12**	29	66	3
BGC0000137	394	71	358	**798**	1968	1232	997	920	534
BGC0000137	394	333	358	636	612	**409**	642	735	120
BGC0000137	711	48	353	**6**	287	36	214	261	1
BGC0000137	711	123	353	**6**	126	14	105	122	1
BGC0000137	711	367	353	6	216	**4**	85	203	1
BGC0000137	711	48	358	184	282	**146**	271	285	17
BGC0000137	711	123	358	184	101	**28**	92	99	7
BGC0000137	711	367	358	184	90	**19**	89	88	7
Best ranks				3	2	**10**	2	2	

Table 6.3: **Scoring function performance.** Columns one through four show MIBiG ID and BGC, GCF and MF IDs of the validated links. Columns five and six show the rank of the validated link using σ_{IOKR} and $\bar{\sigma}_{\text{corr}}$. Columns seven through nine show the rank of the validated link using the ℓ_p scoring functions, and column eight using the product order. Lower ranking is better, and the best rank for each link (excluding the product order) is indicated in bold.

For the Crüsemann data set, this translates into 3094 potential links for each GCF, with one potential link for each MF in the data set. For each validated link, Table 6.3 shows the IDs of the GCF and MF in the validated link, as well as the ranking of that particular link using the various scoring methods. In particular, columns five and six show the ranking according to the IOKR- and standardised strain correlation scores, while the last column shows the rank of the validated link using both scores, which gives an indication of how many links score higher on both scores than the particular link.

In particular, while only one of the validated links is ranked highest for the relevant GCF using a single scoring function (BGC0000827 using $\bar{\sigma}_{\text{corr}}$), for six out of 15 links, no link scores higher than the validated link on both scores.

Columns seven to nine give the ranking of the validated link using a combination of the scoring functions. The results in columns seven to ten are discussed in detail in Section 6.5.

Figures 6.4 – 6.7 show the score of the validated links, indicated with red, within the dis-

tribution of scores for all links involving the relevant GCF for the microbial data sets. The last column of images shows the placement of the validated link in the joint distribution of the scores. As for the scores for all links in the data sets, shown in Figures 6.1 – 6.3, the validated links fall in quadrant I of the coordinate system, but in some instances, validated links with relatively high IOKR score have a relatively low strain correlation score, and vice versa.

6.5 Combining scores

6.5.1 Ranking in multiple dimensions

Assigning a score to a set of items can be considered as positioning the items on the line of real numbers, $\mathbb{R}$. When considering more than one scoring function, the score of an item can be considered as a point in $\mathbb{R}^n$, where n is the number of scoring functions. As we are interested in using multiple scoring functions to rank objects, we need to consider ordered sets in general.

A *totally ordered* set A is a set such that for all elements $x, y \in A$, where $x \neq y$, either $x < y$ or $y < x$, while for a *partially ordered* set there may exist elements $x, y \in A$ such that $x \neq y$ but neither $x < y$ nor $y < x$. An example of a totally ordered set is the set of real numbers $\mathbb{R}$, while ordering the set of points in $\mathbb{R}$ by their distance from 0 would not make for a total order, since more than one point on the line can have the same distance from 0.

The Cartesian product of two sets A and B is the set $A \times B$ with elements (a, b), where $a \in A$ and $b \in B$. The set of real numbers in two dimensions $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the Cartesian product of two instances of $\mathbb{R}$.

In the preceding sections, scoring functions have been used implicitly to define an order on a set. Given a set A and a scoring function $\sigma : A \to \mathbb{R}$, we have used this function to define an order on the set A in an intuitive way by setting $x < y$ if and only if $\sigma(x) < \sigma(y)$ for $x, y \in A$. Note, however, that since σ is not injective (not one-to-one, i.e. there may be elements $a, b \in A$, with $a \neq b$, such that $\sigma(a) = \sigma(b)$), the order thus defined is a partial order on the set, i.e. there may exist elements x and y in A such that neither $x < y$ nor $y < x$, but $x \neq y$. These are exactly the elements x and y in A where $\sigma(x) = \sigma(y)$.

While using a scoring function to rank objects in this way is straightforward, there is no canonical way to generalise such rankings to take more scoring functions into account. For two scoring functions, such generalisation is equivalent to defining an order on $\mathbb{R}^2$, i.e. an order on real numbers in two dimensions, or more generally, for n scoring functions, defining an order on $\mathbb{R}^n$.

Given two ordered sets A and B, various orders can be defined on the Cartesian product $A \times B$. Two common ways of extending the orders on A and B to $A \times B$ are the *lexicographical order*, defined by

$$(a_1, b_1) < (a_2, b_2) \iff \begin{cases} a_1 < a_2 \text{ or} \\ a_1 = a_2 \text{ and } b_1 < b_2, \end{cases} \tag{6.1}$$

and the *product order* defined by

$$(a_1, b_1) < (a_2, b_2) \iff a_1 < a_2 \text{ and } b_1 < b_2. \tag{6.2}$$

If the orders on A and B are total orders, the lexicographical order on $A \times B$ is a total order as well, while this is not the case for the product order, as (a_1, b_1) and (a_2, b_2) cannot be compared if $a_1 < a_2$ but $b_2 < b_1$, or vice versa.

In the preceding section, the complementarity of $\bar{\sigma}_{\text{corr}}$ and σ_{IOKR} is demonstrated using the product order. However, although useful in demonstrating the complementarity of the scores, the product order is not particularly useful in prioritising novel links for further investigation. This is because as a partial order, it leaves many elements incomparable with one another, the product order even more so than many other partial orders. Given two elements $(a_1, b_1), (a_2, b_2) \in A \times B$, they can only be compared in the product order if both $a_1 < a_2$ and $b_1 < b_2$, or $a_2 < a_1$ and $b_2 < b_1$. As we hope to use the complementarity of the two scoring functions to compensate for the cases where one or the other fails, we want to minimise the number of incomparable elements.

One potential way to prioritise links using the product order would be to consider links whose internal order cannot be determined in the product order as equivalence classes, rank the equivalence classes, and then try to find a different way to internally rank each equivalence class.

The lexicographical order does not suffer from this problem to the same degree. In fact, given two totally ordered sets A and B, the lexicographical order inherited by the Cartesian product $A \times B$ is a total order as well, so if the sets are not totally ordered, the only ties inherited by that order are elements that are tied in the individual orders on the sets A and B. However, informally speaking, the lexicographical order involves giving precedence to one of the scoring functions and only using the other scoring function in case of ties in the former. This means that, for instance, a high IOKR score *only* has meaning within the context of two links having the same strain correlation score. Since we hope to use the two scoring functions to complement one another, looking at Figures 6.1 – 6.3, this is likely not optimal behaviour, and we hope to be able to do better.

6.5.2 Combined scoring functions

As stated in the previous section, for a set X, any function $f : X \to \mathbb{R}$ can be used to define a (possbily partial) order on the set. In particular, if X and Y are sets with scoring functions σ_x and σ_y, this applies to the Cartesian product $X \times Y$, so we can define a new real-valued function, with domain $X \times Y$, by combining the individual scoring functions as we see fit, and use that function to define an order on the set.

If $X = Y$, we can use this to define a new order on the set X, based on the two functions: if $\sigma_x : X \to \mathbb{R}$ and $\sigma_y : X \to \mathbb{R}$ are two scoring functions on the set X, we can use any function $g : \mathbb{R}^2 \to \mathbb{R}$ to define a new scoring function $\psi : X \to \mathbb{R}$ by setting $\psi(x) = g(\sigma_x(x), \sigma_y(x))$, and use that to define a new order on the set X.

To make sure that the standardised strain correlation score $\bar{\sigma}_{\text{corr}}$ and the IOKR score σ_{IOKR} are comparable, we define the *standardised IOKR score* $\bar{\sigma}_{\text{IOKR}}$ as

$$\bar{\sigma}_{\text{IOKR}} = \frac{\sigma_{\text{IOKR}} - E[\sigma_{\text{IOKR}}]}{Var[\sigma_{\text{IOKR}}]}$$

with the variance and expected value estimated using the scores for all potential links.

We now turn to the problem of defining a combined scoring function based on both $\bar{\sigma}_{\text{corr}}$ and $\bar{\sigma}_{\text{IOKR}}$.

An intuitive way to define a combination of scoring functions in $\mathbb{R}^2$ is considering concentric rings centered at the origin, and assigning the points a score equal to the distance from the origin. This applies particularly in quadrant I, where both scores are positive.

Looking at the joint plots for the score distributions, in particular Figure 6.1, however, a couple of things become clear:

- As we would hope, the validated links are all in quadrant I, i.e. both scores are positive for all of the validated links.
- The distributions of the scores of the potential links are in fact not independent. Links with a high $\bar{\sigma}_{\text{IOKR}}$ are over-represented among the links with $\bar{\sigma}_{\text{corr}}$ close to 0, and links with high $\bar{\sigma}_{\text{corr}}$ are over-represented among the links with $\bar{\sigma}_{\text{IOKR}}$ close to 0.

Distances in $\mathbb{R}^2$ are usually measured in the Euclidean norm, also called the ℓ_2-norm. The ℓ_2-norm on $\mathbb{R}^2$, with $\mathbf{x} = (x_1, x_2)$, is defined as $l_2(\mathbf{x}) = \sqrt{x_1^2 + x_2^2}$, and generalises to the ℓ_p-norm as

$$\ell_p(\mathbf{x}) = (|x_1|^p + |x_2|^p)^{\frac{1}{p}} \tag{6.3}$$

for a real number $p \geq 1$. While this only fulfills the conditions for being a norm for $p \geq 1$, we can still define the function ℓ_p for $0 < p < 1$, sacrificing the triangle inequality. Figure 6.8 shows the set of points where $x = 1$ for $p = \frac{1}{2}$, $p = 1$ and $p = 2$.

As we are mainly interested in using the score to order links, rather than evaluate the difference in scores, and the power function $f(x) = x^p$ is monotonic for all p, we can simplify Eq. 6.3 to

$$\ell_p(\mathbf{x}) = |x_1|^p + |x_2|^p \tag{6.4}$$

without affecting the resulting order.

In the positive quadrant of the coordinate system (quadrant I) we can use Eq. 6.4 directly, with x_1 as $\bar{\sigma}_{\text{corr}}$ and x_2 as $\bar{\sigma}_{\text{IOKR}}$. As ℓ_p is calculated using the absolute value of the scores, and we still want to penalise points coming from quadrants other than I, in particular from quadrant III, where both scores are negative, we can multiply the scores with their signs after the power function. Using $\text{sgn}(x)$ as the function sending x to 1 if $x \geq 0$ and to -1 otherwise, we can combine the standardised strain correlation score $\bar{\sigma}_{\text{corr}}$ and the standardised IOKR score $\bar{\sigma}_{\text{IOKR}}$ into a new score ℓ_p, by setting

$$\ell_p = \text{sgn}(\bar{\sigma}_{\text{corr}})|\bar{\sigma}_{\text{corr}}|^p + \text{sgn}(\bar{\sigma}_{\text{IOKR}})|\bar{\sigma}_{\text{IOKR}}|^p \tag{6.5}$$

Table 6.3 shows the ranking of the validated links in the Crüsemann data set, starting from the validated BGC, using the ranking inherited from the various scoring functions, as well as the product order. Columns five through nine show the rank of the links using the order inherited from σ_{IOKR}, $\bar{\sigma}_{\text{corr}}$, and the ℓ_p scores using $p = \frac{1}{2}$, $p = 1$ and $p = 2$. The last column shows the rank of the link using the product order. The best rank for each link is indicated in bold, with lower numbers being better. This excludes the product order, since as discussed before it is not particularly useful to prioritise links due to the large number of incomparable elements.

While none of the rankings consistently places the validated link at the top of the list of links, the $\ell_{\frac{1}{2}}$ score gives the best ranking in 10 out of 15 cases, including in three out of the five cases where the link is unambiguous. For instance, the $\ell_{\frac{1}{2}}$-ranking for BGC0001228 (retimycin A) is 253, and for BGC0000241 (lomaiviticin A) is 5, both of which are considerably better than for the individual scoring functions, as well as for other values of p explored. While other values of p, or other approaches to combining the scoring functions, may very well yield better results, we have clearly demonstrated the utility of combining the scoring functions.

		Gross	Leão	Crüsemann	All
All	Verified	5	8	15	28
	Total	782160	9900	1404676	2196736
	Ratio	6.39×10^{-6}	0.00081	1.07×10^{-5}	1.27×10^{-5}
$\geq 95\%$ σ_{IOKR}	Verified	1	3	4	8
	Total	39108	495	70234	109837
	Ratio	2.56×10^{-5}	0.00606	5.70×10^{-5}	7.28×10^{-5}
$\geq 95\%$ $\bar{\sigma}_{\text{corr}}$	Verified	2	5	10	17
	Total	56634	1493	68560	126687
	Ratio	3.53×10^{-5}	0.00335	0.00015	0.00013
$\geq 95\%$ both	Verified	1	1	3	5
	Total	3101	72	4895	8068
	Ratio	0.00032	0.01389	0.00061	0.00062

Table 6.4: The number of links (potential and verified) scoring above the 95th percentile for the microbial data sets using the combined IOKR model.

6.6 Combining scores using the combined IOKR model

While the experiments previously detailed in this Section were performed for the MS2-MIBiG IOKR model described in Section 5.10, similar results hold for the combined IOKR model described in Section 5.12. In particular, Tables 6.4 and 6.5 show the enrichment of validated links above the 95th and 90th percentiles respectively, for the combined IOKR score, the standardised strain correlation score and both scores. Similar to Tables 6.1 and 6.2, the higher percentiles are significantly enriched for validated links.

Similar to Figures 6.1 – 6.3, Figures 6.9 – 6.11 show the joint distribution of the combined IOKR score and correlation score, demonstrating the concentration of the validated links in quadrant I of the coordinate system, i.e. where both scores are positive.

Finally, Table 6.6 shows that similarly to Table 6.3, the $\ell_{\frac{1}{2}}$ achieves the best rank of validated links starting from BGCs for the combined IOKR model as well as the MS2-MIBiG IOKR model described in Section 5.10, although not by as large a margin as for the MS2-MIBiG IOKR model.

6.7 Summary

In this section, we have conclusively demonstrated the complementarity of the feature-based and correlation-based approaches to scoring BGC-MS2 links. While this complementarity has been implicitly assumed, it has previously neither been demonstrated nor quantified.

		Gross	Leão	Crüsemann	All
all	Verified	5	8	15	28
	Total	782160	9900	1404676	2196736
	Ratio	6.39×10^{-6}	0.00081	1.07×10^{-5}	1.27×10^{-5}
$> 95\%$ σ_{IOKR}	Verified	1	3	7	11
	Total	78216	990	140468	219674
	Ratio	1.28×10^{-5}	0.00303	4.98×10^{-5}	5.01×10^{-5}
$> 95\%$ $\bar{\sigma}_{\text{corr}}$	Verified	4	5	13	22
	Total	82622	1493	123829	207944
	Ratio	4.84×10^{-5}	0.00335	0.00010	0.00011
$> 95\%$ both	Verified	1	1	7	9
	Total	9566	172	18538	28276
	Ratio	0.00010	0.00581	0.00038	0.00032

Table 6.5: The number of links (potential and verified) scoring above the 90th percentile for the microbial data sets using the combined IOKR model.

Furthermore, we have proposed a way to combine the two scoring methods proposed in Chapters 4 and 5 to yield a single scoring function that can be used to prioritise links for further verification, based on observations of the correlation of the scores.

While further optimisation of the combination of functions is surely possible, we believe that a principled approach such as the one described here present a sensible way to prioritise potential links.

MIBiG ID	BGC	GCF	MF	Rank of validated link					
				σ_{IOKR}	$\bar{\sigma}_{\text{corr}}$	$\ell_{\frac{1}{2}}$	ℓ_1	ℓ_2	σ_{IOKR} and $\bar{\sigma}_{\text{corr}}$
BGC0001228	990	161	492	1774	**1152**	1680	1699	1726	629
BGC0000241	2688	377	381	642	**16**	19	17	17	9
BGC0000333	739	132	489	1069	25	23	**11**	**11**	10
BGC0000827	1642	232	206	74	**1**	**1**	**1**	**1**	1
BGC0001830	2104	295	309	149	7	**3**	**3**	7	3
BGC0000137	394	71	353	130	16	7	**6**	9	4
BGC0000137	394	333	353	130	189	**41**	51	169	17
BGC0000137	394	71	358	475	1968	889	**853**	904	293
BGC0000137	394	333	358	473	612	**301**	524	774	117
BGC0000137	711	48	353	**79**	287	129	154	166	15
BGC0000137	711	123	353	79	126	**47**	88	115	8
BGC0000137	711	367	353	80	216	**25**	37	99	9
BGC0000137	711	48	358	628	282	**198**	451	701	94
BGC0000137	711	123	358	614	101	**73**	97	108	28
BGC0000137	711	367	358	526	90	**71**	92	138	20
Best ranks				1	3	**9**	5	2	

Table 6.6: **Scoring function performance.** Columns one through four show MIBiG ID and BGC, GCF and MF IDs of the validated links. Columns five and six show the rank of the validated link using σ_{IOKR} with the combined IOKR model, and $\bar{\sigma}_{\text{corr}}$. Columns seven through nine show the rank of the validated link using the ℓ_p scoring functions, and column eight using the product order. Lower ranking is better, and the best rank for each link (excluding the product order) is indicated in bold.

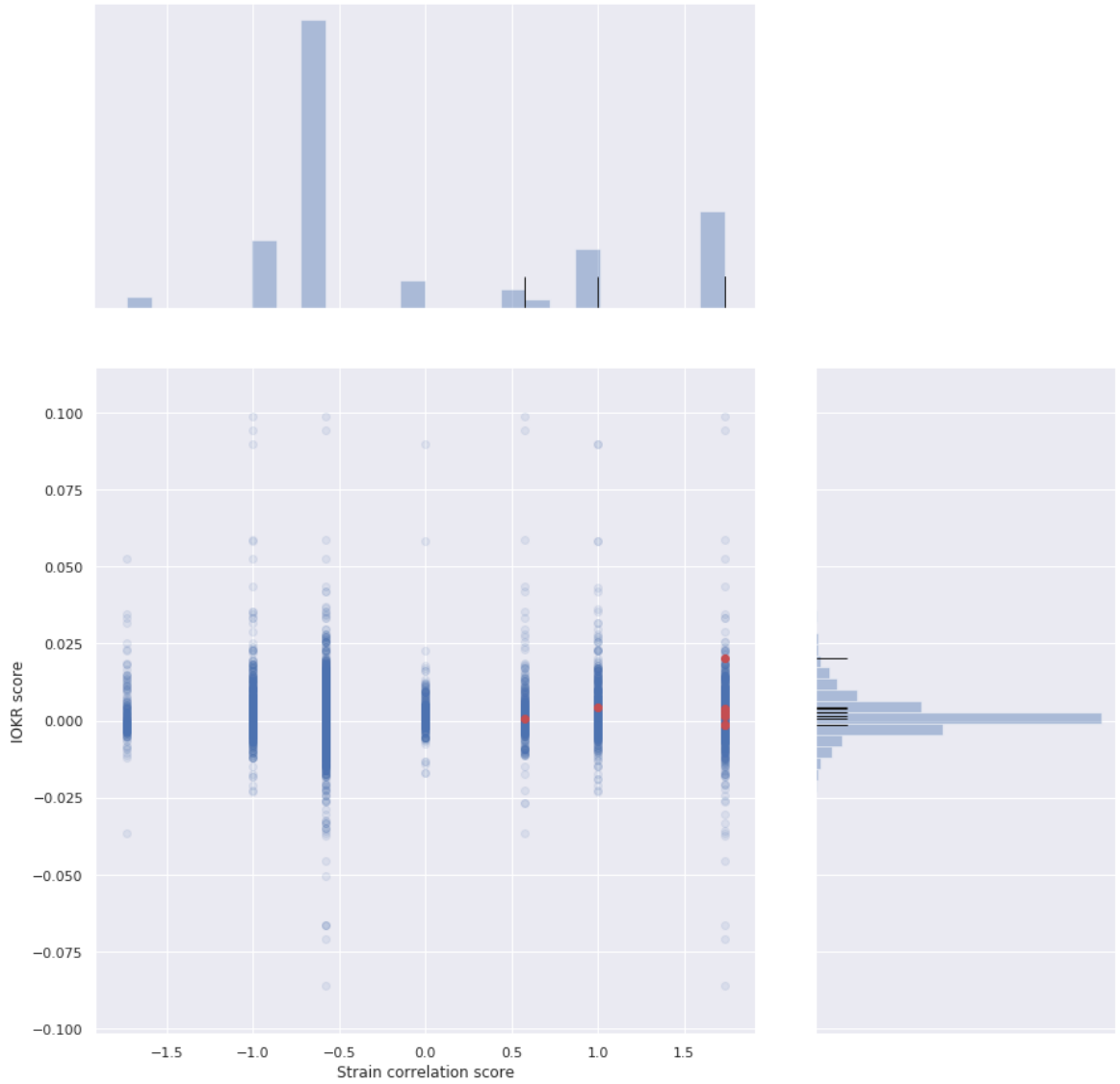


Figure 6.2: **Distribution of IOKR- and standardised strain correlation scores for the Leão data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

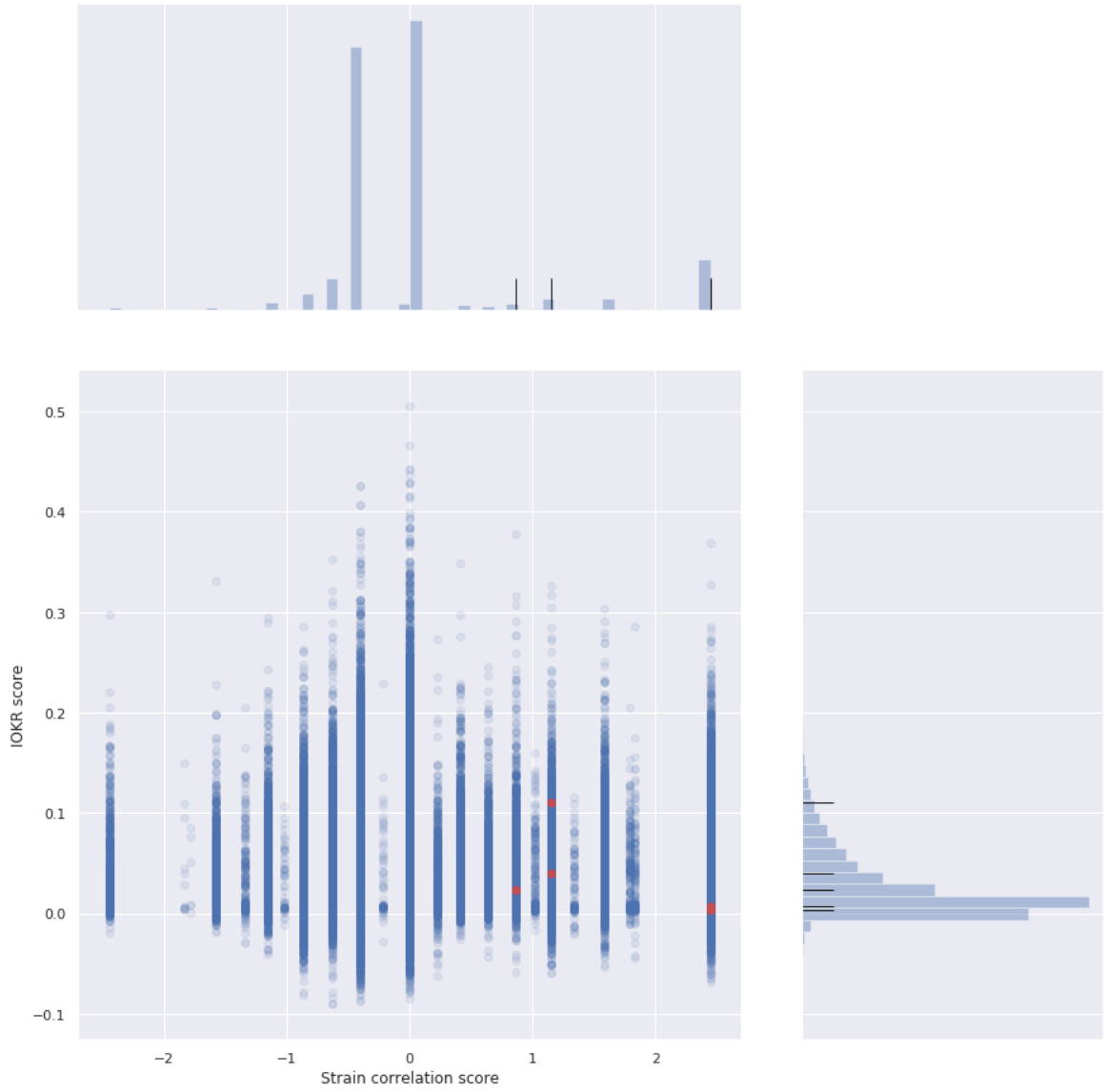


Figure 6.3: **Distribution of IOKR- and standardised strain correlation scores for the Gross data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

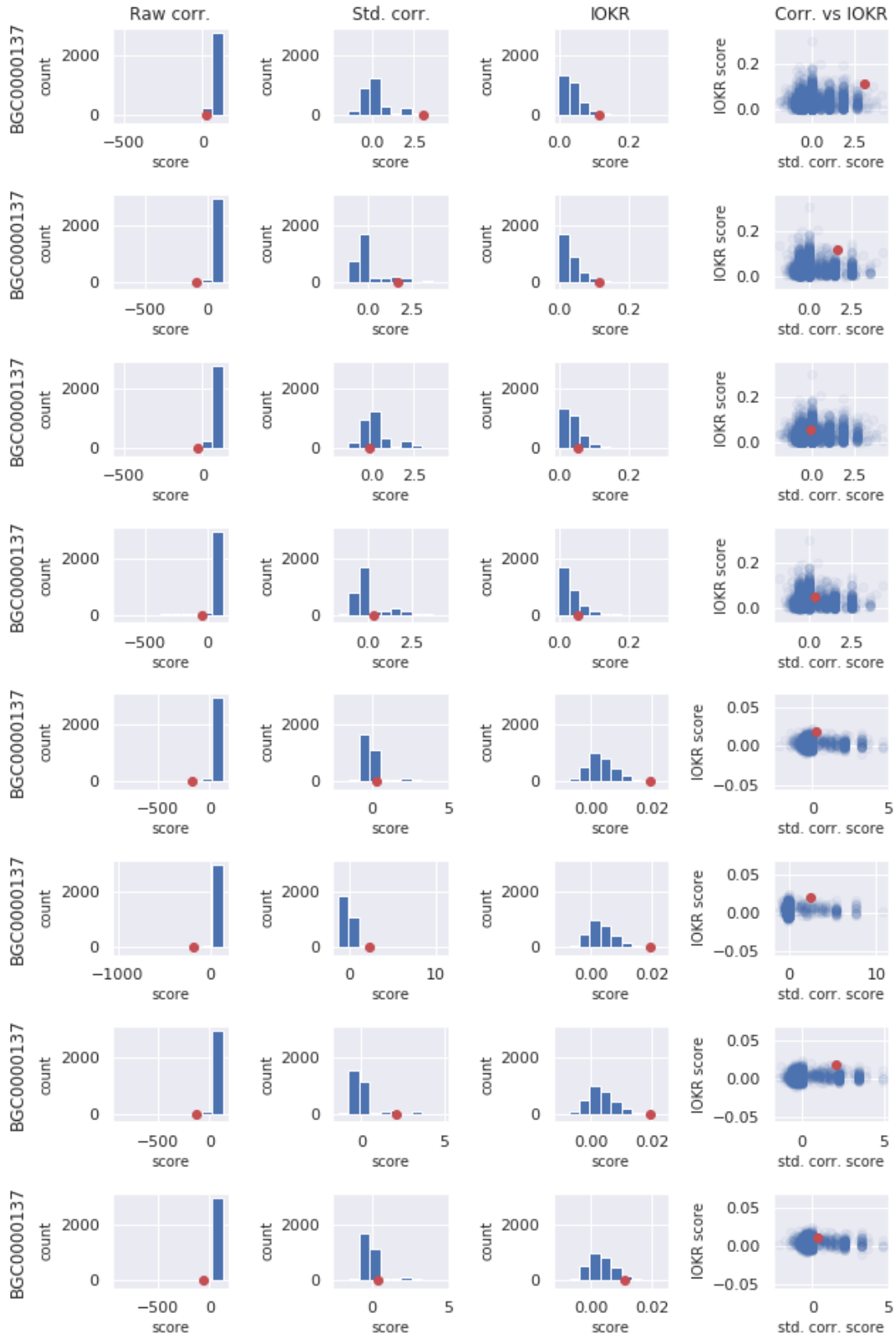


Figure 6.4: **Distribution of scores** Distribution of scores starting from verified BGCs in the Crüsemann data set.

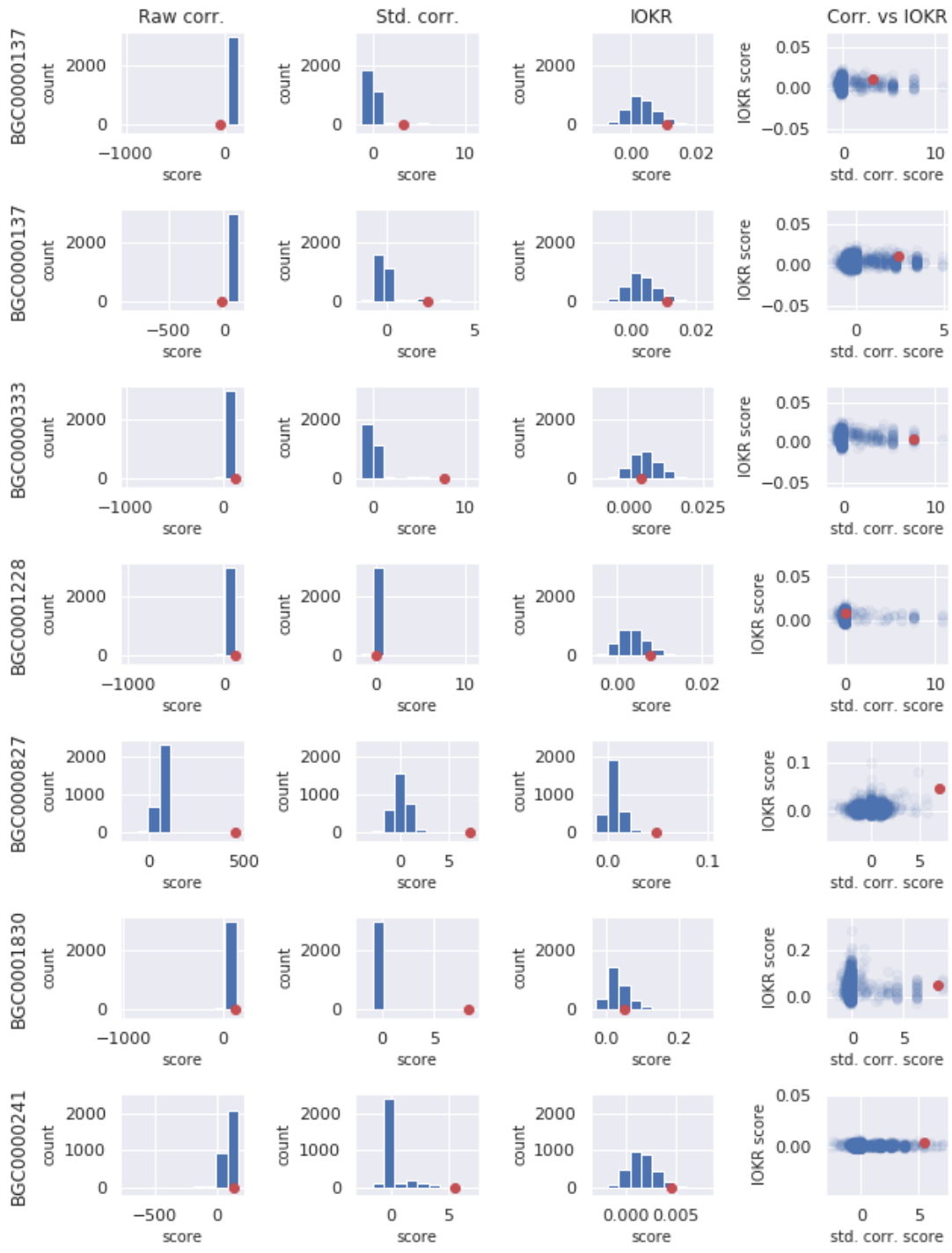


Figure 6.5: **Distribution of scores** Distribution of scores starting from verified BGCs in the Crüsemann data set (continued).

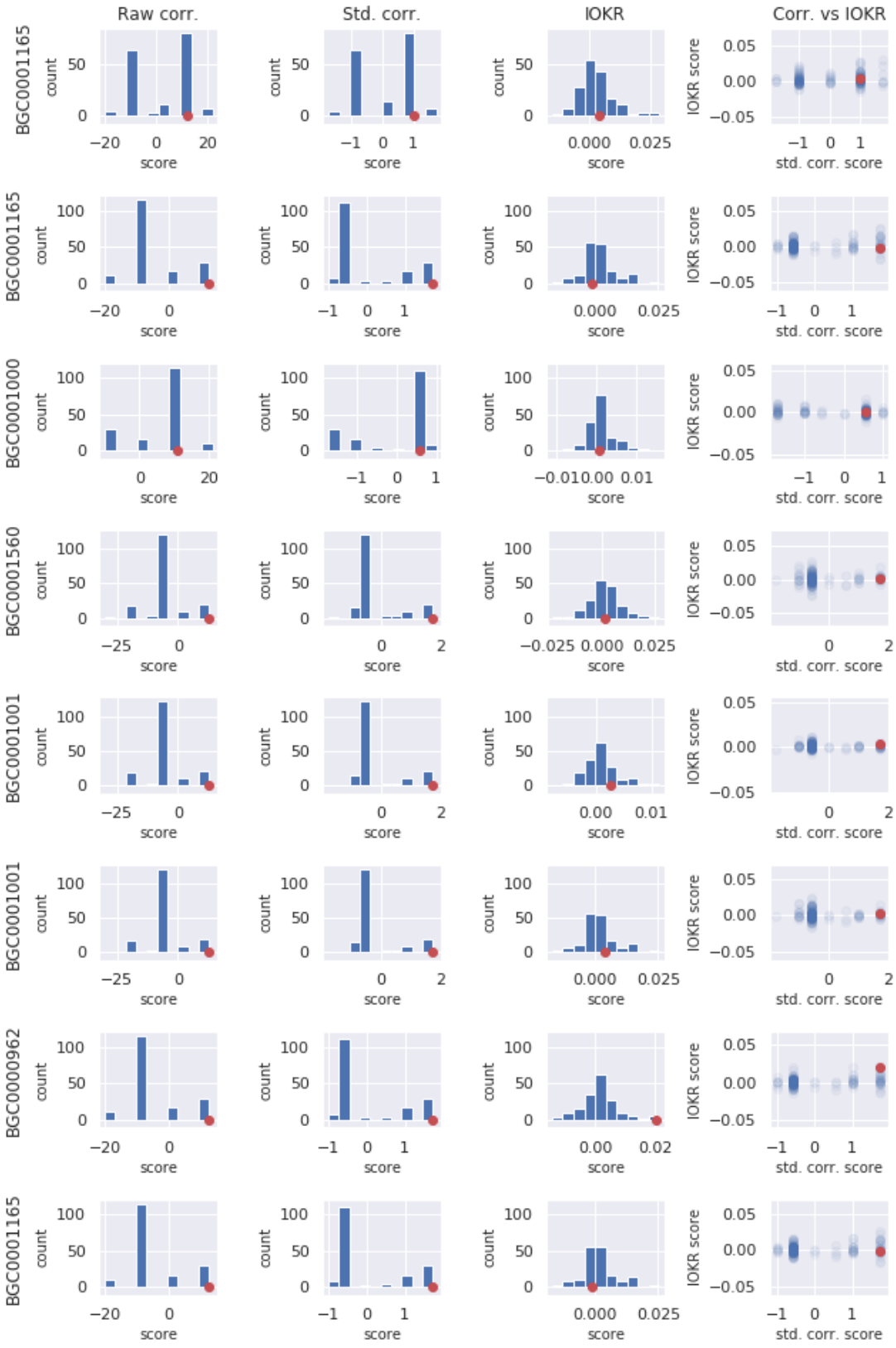


Figure 6.6: **Distribution of scores** Distribution of scores starting from verified BGCs in the Leão data set.

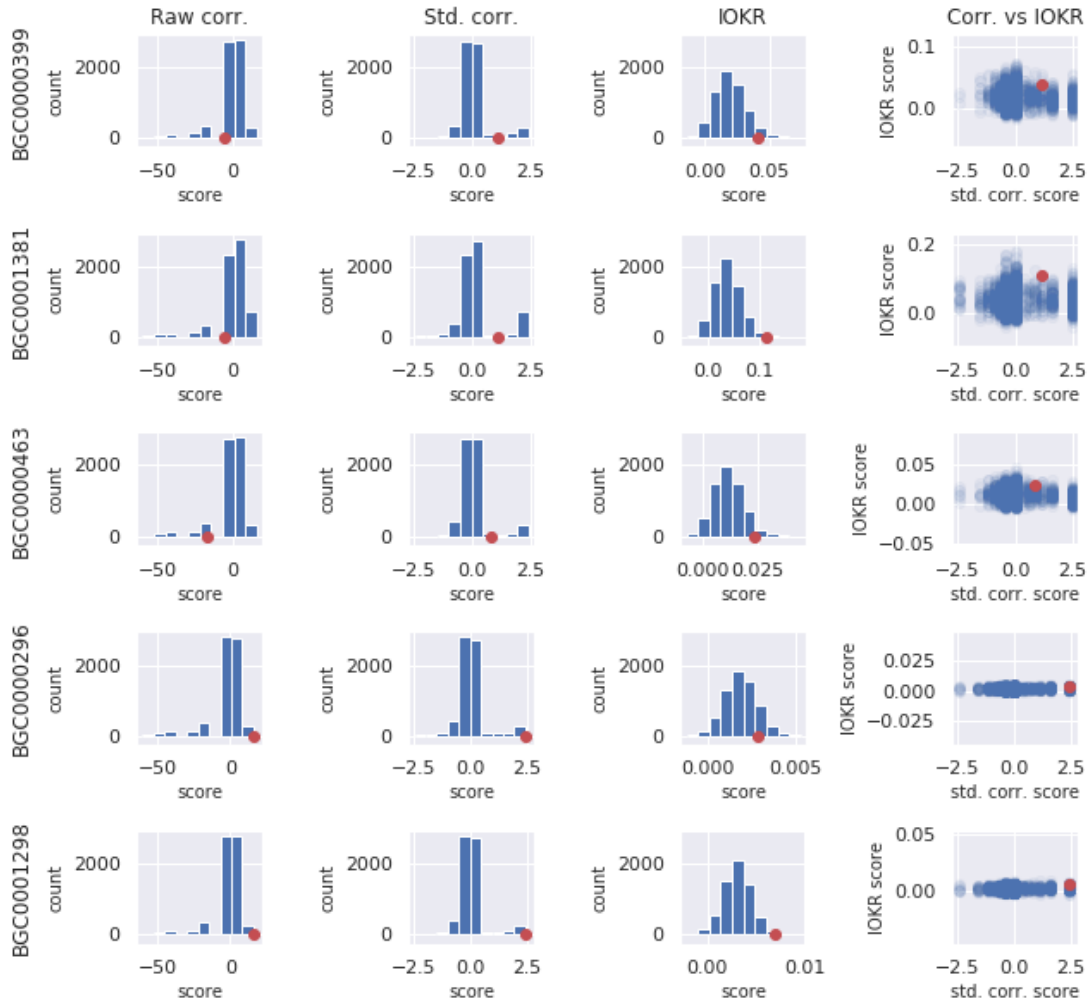


Figure 6.7: **Distribution of scores** Distribution of scores starting from verified BGCs in the Gross data set.

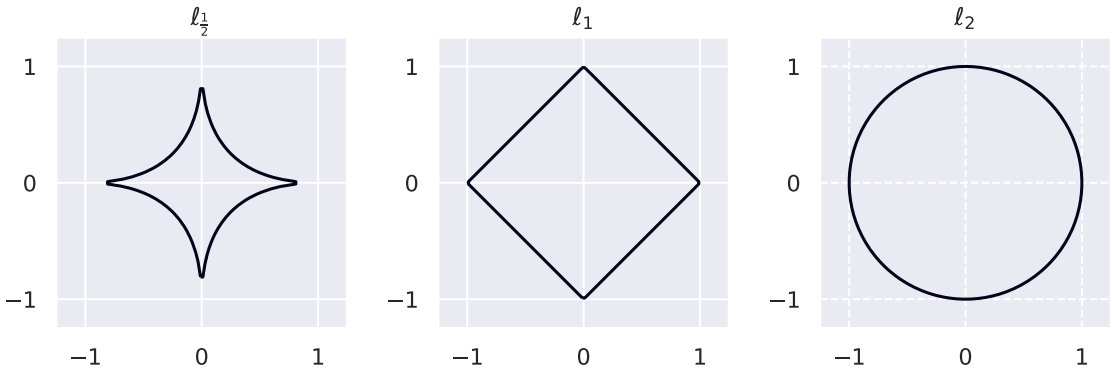


Figure 6.8: ℓ_p **iso-lines in** $\mathbb{R}^2$**.** The set of points (x, y) such that $\ell_p(x, y) = (x^p + y^p)^{\frac{1}{p}} = 1$ for $p = \frac{1}{2}$, $p = 1$ and $p = 2$. ℓ_2 is the usual Euclidean norm. This demonstrates the form of the iso-lines of scores using the ℓ_p-function for different values of p to combine $\bar{\sigma}_{\text{corr}}$ and $\bar{\sigma}_{\text{IOKR}}$, i.e. $\ell_p(\bar{\sigma}_{\text{corr}}, \bar{\sigma}_{\text{IOKR}})$.

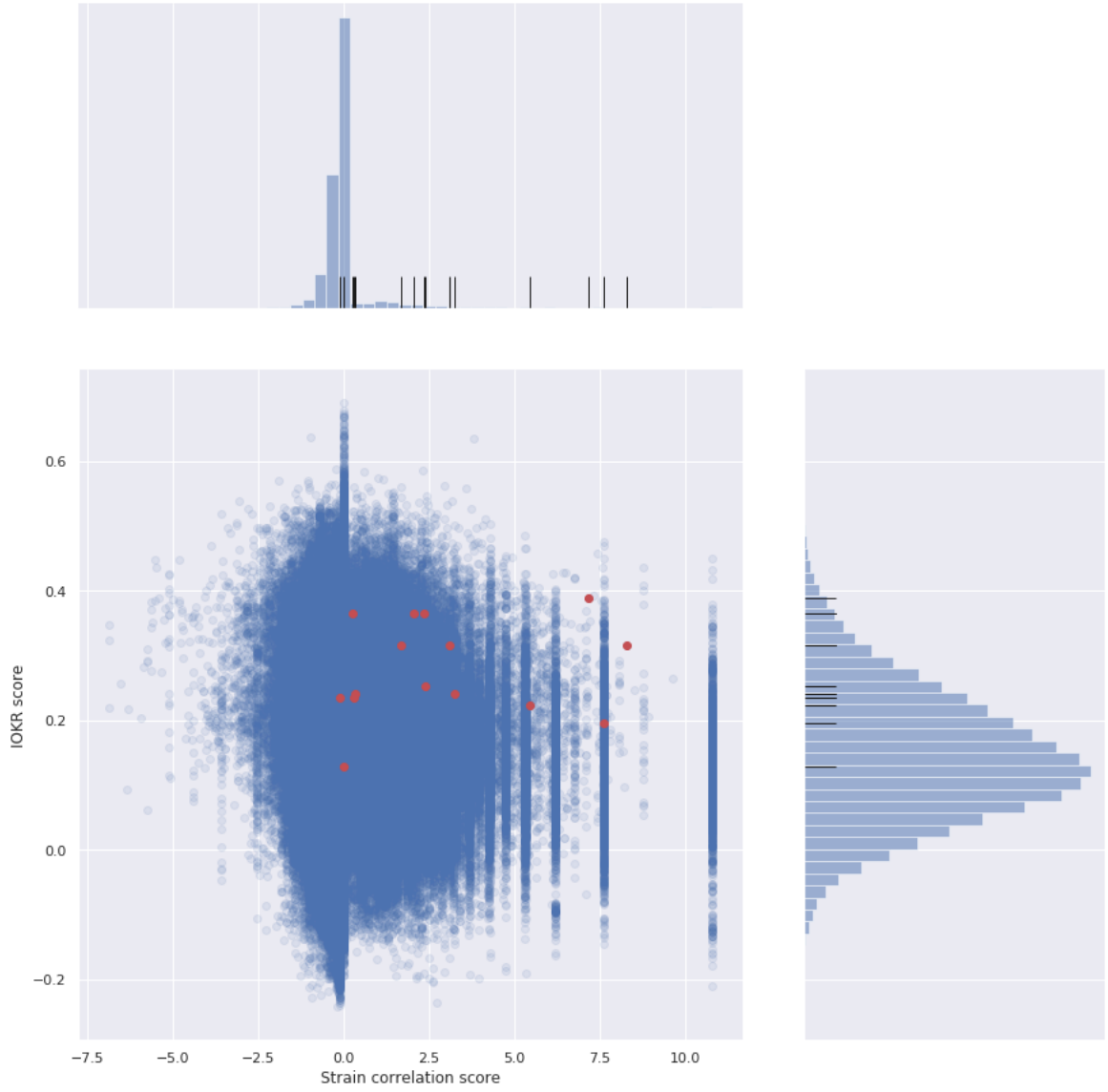


Figure 6.9: **Distribution of IOKR- and standardised strain correlation scores for the Crüsemann data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

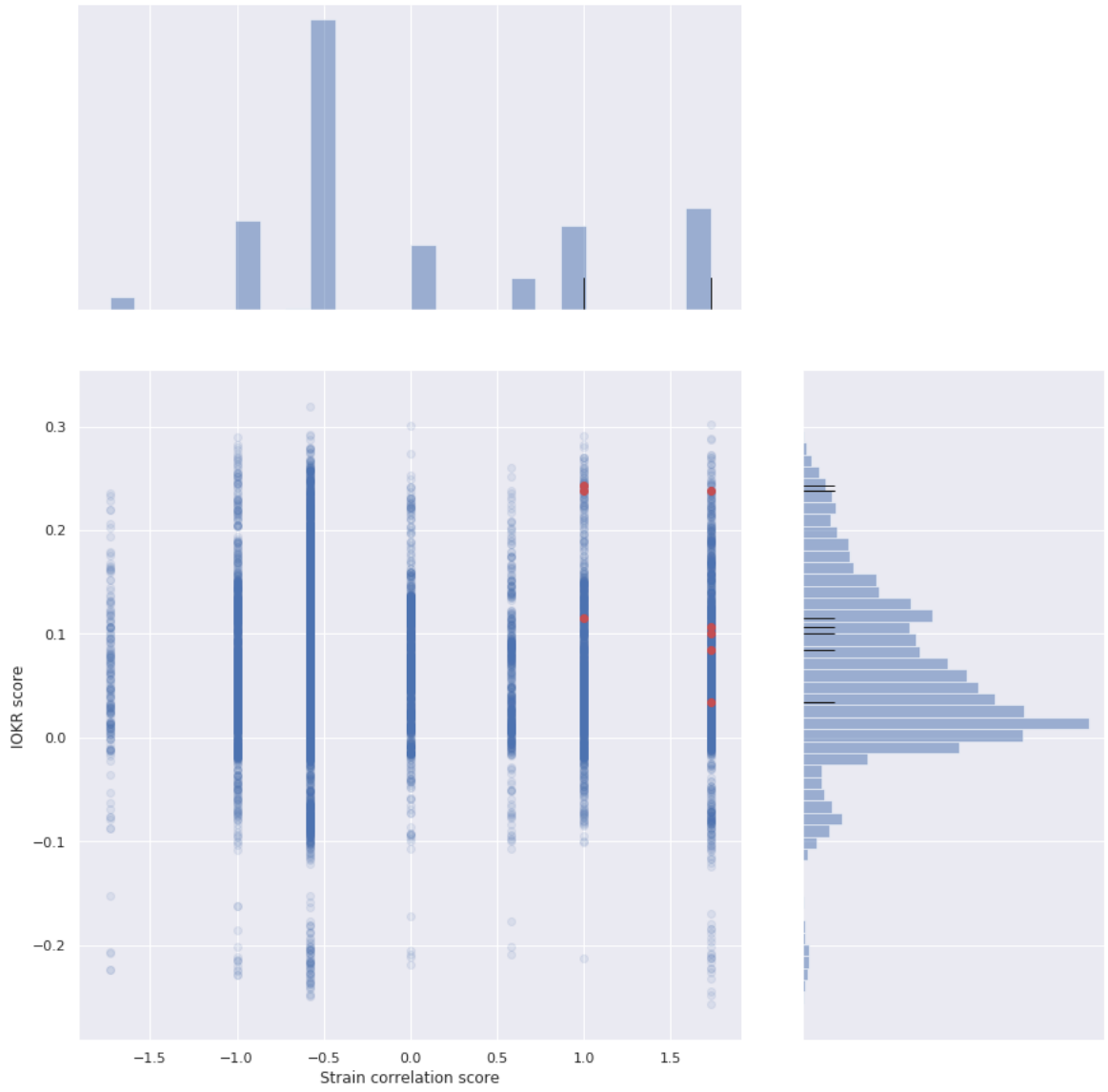


Figure 6.10: **Distribution of IOKR- and standardised strain correlation scores for the Leão data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

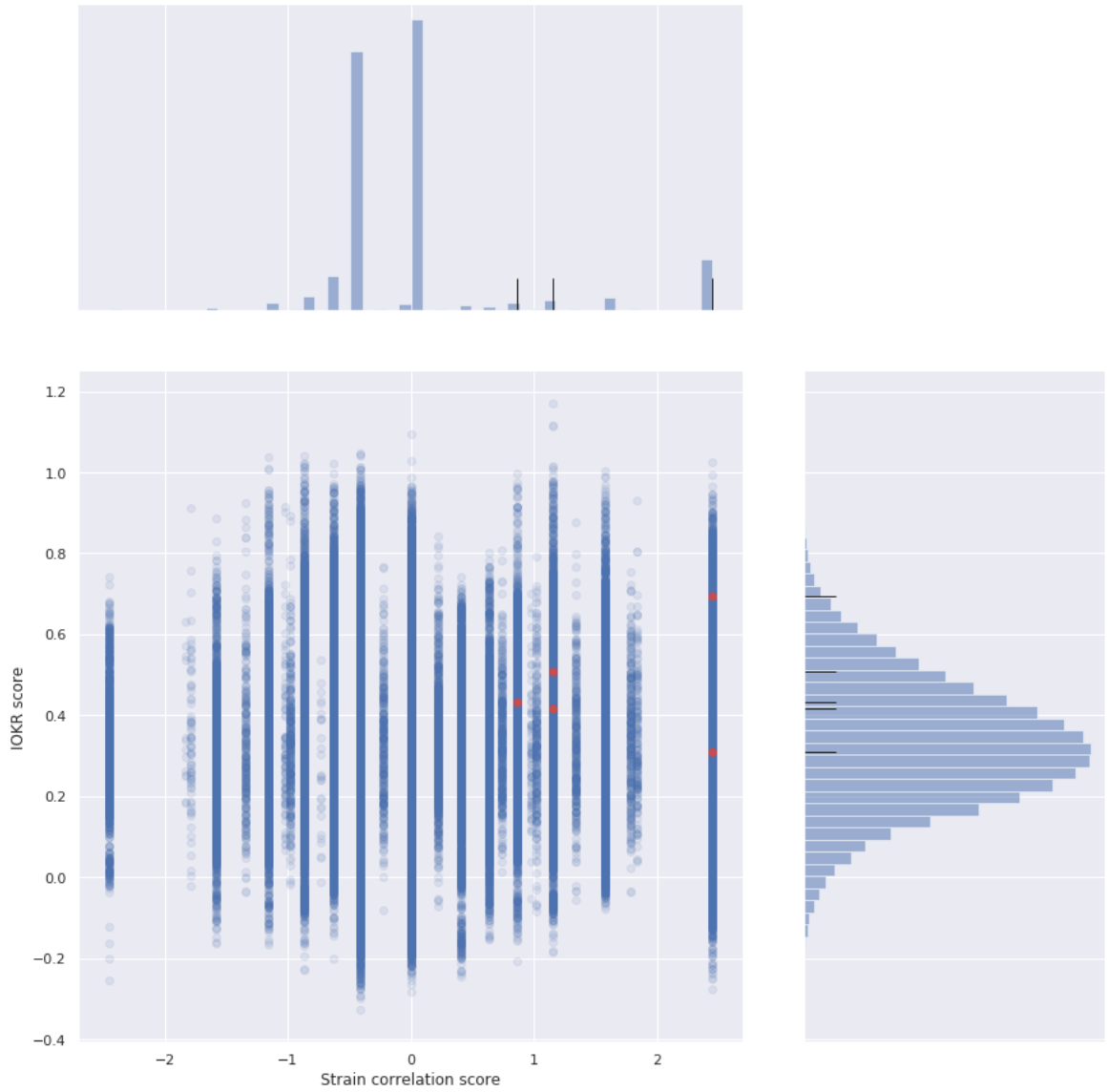


Figure 6.11: **Distribution of IOKR- and standardised strain correlation scores for the Gross data set.** IOKR score is on y-axis and standardised strain correlation score on x-axis. The histograms of the individual scores are above (standardised strain correlation score) and to the right (IOKR score) of the correlation graph. Validated links are marked in red on the joint graph, and with black lines on the histograms.

Chapter 7

Conclusion

7.1 Retrospective

The resurgent interest in microbial natural products in drug discovery makes the development of efficient ways of linking microbial BGCs to their corresponding MS2 spectra ever more important. As laid out in Section 1.1, this thesis addresses four important aspects of this development.

To begin with, in Chapter 3 we addressed the problem of quantifying the performance of link scoring methods. In the absence of benchmark data sets we merged the industry-standard MIBiG [66] and GNPS [64] databases of BGCs and MS2 spectra, respectively, to obtain a benchmark database of verified BGC-MS2 links. We also described how the Paired omics Data Platform [108] facilitates access to paired data sets of microbial genomic and metabolomic data, and validated BGC-MS2 links in those data sets, which can be used for performance evaluation.

Second, starting with the correlation-based approach to link BGCs and MS2 spectra, Chapter 4 demonstrated a serious flaw in the commonly-used strain correlation score and then proceeded to mitigate it by standardising the score to make comparison between different potential links easier. We then demonstrated how the strain correlation score can penalise BGCs shared by HGT and further showed how the phylogenetic relation between the strains in the population can be taken into account to prioritise those BGCs.

Third, turning to the feature-based approach to link BGCs and MS2 spectra, Chapter 5 introduced the a novel feature-based scoring method which unlike most such methods is not dependent on natural product class.

Finally, Chapter 6 demonstrated the complementarity of the two approaches, which has until now only been implicitly assumed, and suggested a principled way of combining multiple scoring functions to rank potential links.

Each of these presents potential avenues of inquiry.

7.2 Future directions

7.2.1 Data sets and tools

As described earlier the lack of standardised data sets makes comparison of methods difficult. Linking the MIBiG and GNPS databases using structural annotations, as described in Section 3.6, can hopefully be a step in the direction of establishing such data sets, while the Paired omics Data Platform [108] stands to vastly improve the access to microbial data sets of paired MS2 and genomic data. The success of the latter, however, is very much dependent on the adoption of the platform. In general, improvements in the sharing of data, adherence to data standards, and the richness of metadata are important factors to encourage further development of computational methods in the field.

The importance of data sharing can hardly be overstated. Although scientific journals and funding bodies increasingly require data to be shared, in the absence of commonly accepted benchmark data sets researchers in the field must do more than the bare minimum to make their data accessible. The free sharing of data sets encourages innovation by levelling the playing field and making comparison between competing approaches easier.

It is the responsibility of the researchers to make sure their data is accessible to other researchers, whether for replication of their own results or development of novel methods. This applies not only to newly created data, but also to data sets derived from existing data, where the downstream analysis required to produce the data set used must at least be properly documented if the derived data sets themselves are not made available.

Facilitating the reuse of data, as well as making analysis of multi-strain microbial data sets easier, the recently-published NPLinker platform [170] aims to make linking BGCs and MS2 spectra easier. The framework uses user-provided data, but is also integrated with antiSMASH DB [171], BiG-SCAPE [82] and PoDP [108], and allows researchers to implement their own scoring functions in addition to the built-in ones. Improved tools for data analysis will help drive further research in the area as well as facilitate development of new methods of analysis.

To increase the discriminative power of the strain correlation based approaches, the ability to pool results from different experiments would be highly useful. While genomic data is standardised and comparable, the same cannot always be said for the metabolomic data. The development of methods to evaluate similarity between MS2 spectra resulting from experiments with different instrumentation or instrument configuration would allow researchers to more efficiently use the data that already exists.

While the genomic data is comparable between experiments, the assignment of BGCs into GCFs can give different results depending on the tools used for the assignment.

As discussed in Section 3.8.1, given a known BGC in MiBIG, BiG-SCAPE and the KCB function of antiSMASH can be considered as two facets of the same thing, i.e. the actual metabolomic potential of the microorganisms. Ideally, for BGCs homologous to a MIBiG entry, the GCF containing the MIBiG entry and the set of BGCs identified as close matches to the MIBiG entry should be identical.

Cursory experiments demostrate that this is not always the case, as demonstrated by the case of BGC0000872 in the Crüsemann data set discussed in Section 3.8.1. While this may to some extent be explained by runtime parameters affecting clustering, the behaviour is consistent across all values of BiG-SCAPE cutoff parameters tested, and in any case the stability of the BiG-SCAPE clustering with regards to parameter selection would benefit from careful investigation.

A potential reason for this discrepancy is that the similarity metric on the space of BGCs is in some way incomplete or incompatible with the antiSMASH KCB results: the results can be interpreted as BiG-SCAPE and antiSMASH disagreeing on how similar the BGCs are in fact to the MIBiG BGC.

7.2.2 Similarity metrics and kernels

The underlying assumption is that similar BGCs will produce similar metabolites. While this may seem a reasonable assumption, in fact it is dependent on the similarity metric on both the space of BGCs, and of the metabolites.

The similarity metric on the space of BGCs impacts both the IOKR model and the strain correlation model. While the impact of the similarity metric is obvious in the case of the IOKR model, as it directly impacts the calculations, in the case of the strain correlation score, the BGC similarity metric impacts the clustering of the BGCs into GCFs. The word vector metric proposed on the space of BGCs in Section 5.11.1 has known limitations, because small changes in Pfam domain composition can represent significant structural changes [82]. The fact that a single Pfam domain represents all adenylation domains, irrespective of specificity, serves to emphasise the point that theoretically, two BGCs can encode products with fundamental differences — namely composed of different amino acids — but still be identical when considered in terms of Pfam domains.

To circumvent this, BiG-SCAPE relies heavily on sequence similarity when comparing modular product classes, but on the whole, this similarity measure shows worse performance in the IOKR model than the word vector metric, even with the shortcomings mentioned above.

While the weights used for the component metrics in BiG-SCAPE are optimised using available, validated data sets, they are nevertheless only based on a limited selection of component metrics. Further development of metrics on the space of BGCs represents a promising avenue of inquiry with repercussions for the whole field. While the pfam2vec model proposed in [77] still does not provide granularity beyond the Pfam domain level, vector embedding models represent a promising avenue of inquiry as BGC similarity metrics, particularly given their success in text mining tasks [172].

In a similar vein, the choice of kernels on the space of MS2 spectra is still an open question. In particular, after filtering — which is usually done either based on intensity or predicted fragmention tree — current similarity functions treat all peaks in the MS2 spectra as equally informative in deciding similarity, while in fact, some *m/z* values are ubiquitous and should therefore arguably be downweighted.

While Brouard and co-workers achieved their best IOKR performance using a weighted combination of kernels using MKA [113, 152, 166], the recently-proposed spec2vec algorithm [173] has demonstrated promising results in initial tests with IOKR. MS2DeepScore [174] has also demonstrated good performance predicting similarity of spectra, but remains to be tested in combination with IOKR.

Both of the latter algorithms take into account how informative each *m/z* peak is likely to be when computing the similarity between two spectra.

As the results on microbial data sets presented in this thesis are based on classical molecular networking, the choice of similarity function on the space of MS2 spectra might influence the clustering of spectra into MFs, and thus influence both the IOKR score and the strain correlation score.

Finally, the choice of molecular fingerprints influences the IOKR model. By their design, molecular fingerprints abstract a complete description of a molecule into structural features of the molecule. Designing molecular fingerprints to target chemical substructures or properties particular to microbial specialised metabolites, such as discussed in [175], can potentially increase the power of the IOKR models.

7.2.3 Clustering

A major step in the preprocessing of information for metabologenomics is clustering, both BGCs into GCFs and spectra into MFs. Both are active and rapidly developing research areas. This is particularly true for BGC clustering, where the recent creation of BiG-SCAPE and BiG-SLICE lays the groundwork for further rigorous experimentation with BGC clustering.

As mentioned in Section 3.2, data to verify these approaches is hard to come by and has until now been poorly standardised. To quantify their performance, BiG-SCAPE [82] and BiG-SLICE [90] are tested on a set of known BGCs where a number of clusters are pre-defined. Furthermore, BiG-SCAPE is tested on a microbial data set using the (non-standardised) strain correlation scoring, to demonstrate the presence of validated links in the upper end of the distribution of scores.

The performance of the clustering in both domains, BGCs and spectra, depends on two factors: the similarity function and the clustering algorithm. With multiple viable options to define similarities on both BGCs and spectra (including the metrics defined by BiG-SCAPE [82] and BiG-SLICE [90] as well as pfam2vec [86] for BGCs, and including PPKr [151], cosine similarity, modified cosine similarity [58], spec2vec [173] and MS2DeepScore [174] for MS2 spectra) the choice of metric on either space may need to take into account not only the correspondence with the similarity between metabolites, but also the metric being used on the other space. For instance, the metric defined on the space of BGCs may exhibit some properties which make it more compatible with a particular metric on the space of MS2 spectra, even if other metrics on that space show closer correspondence with the similarity metric defined on the space of metabolites.

After defining the similarity metric, a wide variety of clustering algorithms exists to choose between, depending on the properties of the data [176]. The choice of clustering algorithm depends to a large extent on the properties of the data set, but arguably the choice of parameters for the algorithms is just as important [176]. All of the clustering tools mentioned above use different clustering algorithms, with different levels of parameter tuning, both internally and user-provided. It is highly likely that further research in this area will benefit all downstream data analysis.

Other approaches to the task than traditional clustering may also be viable. For instance, using antiSMASH to create GCFs from MIBiG entries using KCB allows the same BGC to belong to more than one GCF. Approaches such as mixture models may work to assign BGCs to GCFs in a more flexible way than by clustering, where the association is probabilistic rather than absolute.

7.2.4 Structure predictions for BGCs

In Section 5.10, antiSMASH-assigned MIBiG IDs were used to assign molecular structures to the BGCs for use with IOKR. However, antiSMASH [73] and the recently-released version 4 of PRISM [102] make structural predictions for some of the metabolites encoded by the predicted BGCs. As the methods presented here are considered in relation to BGC clustering, and PRISM 4 has not yet been integrated with BiG-SCAPE or other BGC clustering tools, the structural predictions provided by PRISM are not evaluated in this thesis.

While attempts were made to use generate fingerprint vectors from the molecular structures predicted by antiSMASH, the initial results were not promising. As predicting molecular structure is a difficult problem, many of the predictions were for molecular substructures and not entire molecules. Since many molecular fingerprints represent the presence or absence of particular substructures, building such fingerprints from partial structures is difficult, as the absence of a feature from the predicted substructure does not necessarily translate into the absence of the feature from the entire molecule. Reflecting this, CDK, which was used to create the molecular fingerprints, does not handle such partial predictions.

While PRISM 4 has been demonstrated to offer a greater number of structure predictions than antiSMASH, it still only predicts structures in about a quarter of the cases [102], and their applicability to linking approaches remains to be evaluated.

7.2.5 Joint IOKR model

As demonstrated in section 5.12, two individually-trained IOKR models can be combined to score BGC-spectrum links. Each model learns a mapping from the set of BGCs or MS2 spectra into the space of molecular fingerprints. However, these learned mappings may not necessarily be optimal with respect to the shared latent space, i.e. the approximations that one model makes may be incompatible with the approximations that the other model makes. Ideally, if θ is the mapping from the space of BGCs to molecular fingerprints, and ψ from the space of MS2 spectra to molecular fingerprints, and x and y are linked BGC and spectrum, respectively, and both in the training data set, then the images $\theta(x)$ and $\psi(y)$ should be particularly close in the space of molecular fingerprints. Since the models are trained individually, and in particular make their errors independent of one another, this need not be the case.

To mitigate this problem, the IOKR models can be trained jointly, by incorporating the distance between the fingerprint vectors for the training samples into the error function and learning the mappings into the feature space with the constraint that similar metabolites map to similar points in the fingerprint space.

This complicates the training phase of the model, however, as there is no longer a known closed-form solution to the optimisation problem. Instead, the objective can be minimised using gradient descent.

Even so, using IOKR to model BGC-MS2 relations has a fundamental theoretical flaw. Recall that the BGC-MIBiG IOKR model learns a function from the space of BGCs to the space of molecular fingerprints. However, as a single BGC can encode multiple metabolites, this behaviour cannot in fact be precisely modelled by a function. As the IOKR model assumes a single output value for any input value, this will inevitably lead to inaccuracies in the out-

put space for a BGC that encodes multiple metabolites, assuming that the chosen molecular fingerprints capture the difference between the metabolites.

7.2.6 Combining multiple scores

Combining different scores, or a higher number of scores, represents a clear way forward as well. While Chapter 6.5 demonstrates the complementarity of strain correlation scoring and IOKR scoring in particular, other scoring approaches may well complement or outperform each of those.

Many of the feature-based algorithms discussed in Sections 5.2 and 2.3.1 can be converted into scoring functions and combined with others, such as Pep2path [97].

Apart from feature-based scoring, algorithms to predict natural product class from MS2 spectrum, i.e. if a spectrum belongs to a NRP, RiPP, PK or other class, can be used as a basis for a scoring function, taking the natural product class prediction from antiSMASH or other such tools as input for a GCF.

Different ways of combining individual scores represent an open question. As the ℓ_p functions are a generalisation of the concept of a sphere, they can easily be extended to any number of dimensions to accommodate any number of distinct scoring functions. However, the scoring functions could also be combined in fundamentally different ways. Such combinations could for instance include constructing polynomials from the scoring functions. While this would allow for more nuanced weighing of the individual functions, the choice of coefficients would create another optimisation problem in turn.

7.3 Summary

Microbial specialised metabolites remain one of the most promising sources of novel antibiotics, much needed to counter the spread of AMR. This is supported not only by the historic record but also by the vast chemical space represented by microbial secondary metabolites, the scale of which has only recently become apparent.

Furthermore, even though novel antibiotics are discovered, the ease with which microbes seem to develop resistance to antibiotics guarantees that there will always be a need for new antibiotics. As microbes develop resistance to current last-line antibiotics, new ones are needed to take their place. Judicious use of current antibiotics, both in humans and livestock, and careful management of pharmaceutical waste can delay the development of resistance, but not prevent it.

A key step in harnessing the metabolomic potential of microbes is establishing the correspondence between BGCs and the metabolites for which they encode. Increased throughput in genomic and metabolomic experiments enables statistical- and machine learning approaches to be brought to bear on problems related to this linking in a more effective and principled manner than before. As methodological advances in machine learning make it ever easier to process large amounts of data and detect subtle patterns therein, the application of machine learning to combined genomic and metabolomic data offers the promise of a paradigm shift in the search for new antibiotics.

Appendix A

autoMLST parameters

The *Salinispora* phylogeny tree used in Section 4.5 was created using a local installation of the autoMLST pipeline [134], commit `231601f`, run with default parameters and bootstrap replicates enabled. The local version is identical to the public web version in that the 50-genome limit in the public version is not in place.

The key processes for the tree generation are a concatenated alignment using MAAFT for every gene detected in the provided sequences from a given set of genes, as described in [134]. Following alignment, the pipeline does automated trimming using trimAL and phylogenetic inference with IQtree using gene partitions (for per-gene parameter estimation) to generate the final tree.

Bibliography

[1] Aminov R. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. Frontiers in Microbiology. 2010;1:134.

[2] Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance-the need for global solutions. The Lancet Infectious Diseases. 2013 dec;13(12):1057-98.

[3] Sengupta S, Chattopadhyay MK, Grossart HP. The multifaceted roles of antibiotics and antibiotic resistance in nature. Frontiers in Microbiology. 2013 mar;4(MAR):47.

[4] Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR. Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). Science. 1980 sep;209(4464):1532-4.

[5] 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. American Journal of Physical Anthropology. 2010 sep;143(1):151-4.

[6] Davies J. Where have all the antibiotics gone? Canadian Journal of Infectious Diseases and Medical Microbiology. 2006;17(5):287-90.

[7] Ali SM, Siddiqui R, Khan NA. Antimicrobial discovery from natural and unusual sources. Journal of Pharmacy and Pharmacology. 2018 sep;70(10):1287-300.

[8] Gould K. Antibiotics: from prehistory to the present day. Journal of Antimicrobial Chemotherapy. 2016 mar;71(3):572-5.

[9] Ventola CL. The antibiotic resistance crisis: causes and threats. P & T journal. 2015;40(4):277-83.

[10] World Health Assembly. Global action plan on antimicrobial resistance [Governing body documents]. World Health Organization; 2015.

[11] Ten threats to global health in 2019. World Health Organization; 2019. Available from: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

[12] O'Neill J. Tackling drug-resistant infections globally: final report and recommendations ; 2019. Available from: https://amr-review.org/.

[13] Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: A global multifaceted phenomenon. Pathogens and Global Health. 2015 oct;109(7):309-18.

[14] Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiology. 2018;4(3):482.

[15] Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nature Reviews Genetics 2019 20:6. 2019 mar;20(6):356-70.

[16] Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nature Chemistry. 2016;8(6):531-41.

[17] Amirkia VD, Qiubao P. The Antimicrobial Index: A comprehensive literature-based antimicrobial database and reference work. Bioinformation. 2011 jan;5(8):365-6.

[18] Singh KS, Sharma R, Reddy PAN, Vonteddu P, Good M, Sundarrajan A, et al. IspH inhibitors kill Gram-negative bacteria and mobilize immune clearance. Nature. 2021 jan;589(7843):597-602.

[19] Theuretzbacher U, Outterson K, Engel A, Karlén A. The global preclinical antibacterial pipeline. Nature Reviews Microbiology. 2020 may;18(5):275-85.

[20] Butler MS, Paterson DL. Antibiotics in the clinical pipeline in October 2019. Journal of Antibiotics. 2020 jun;73(6):329-64.

[21] Davies J. Specialized microbial metabolites: Functions and origins. Journal of Antibiotics. 2013 jul;66(7):361-4.

[22] Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A. 2012 Jun;109(26):E1743-52.

[23] Baltz RH. Gifted microbes for genome mining and natural product discovery. J Ind Microbiol Biotechnol. 2017 May;44(4-5):573-88.

[24] Demain AL, Sanchez S. Microbial drug discovery: 80 Years of progress. Journal of Antibiotics. 2009 jan;62(1):5-16.

[25] Singh R, Kumar M, Mittal A, Mehta PK. Microbial metabolites in nutrition, healthcare and agriculture. 3 Biotech. 2017 may;7(1).

[26] Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020 Mar.

[27] Pham JV, Yilma MA, Feliz A, Majid MT, Maffetone N, Walker JR, et al. A review of the microbial production of bioactive natural products and biologics. Frontiers in Microbiology. 2019 jun;10(JUN):1404.

[28] Baral B, Akhgari A, Metsä-Ketelä M. Activation of microbial secondary metabolic pathways: Avenues and challenges. Synthetic and Systems Biotechnology. 2018 sep;3(3):163-78.

[29] Hoskisson PA, Seipke RF. Cryptic or silent? The known unknowns, unknown knowns, and unknown unknowns of secondary metabolism. mBio. 2020 sep;11(5):1-5.

[30] Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nature Reviews Microbiology. 2015 jun;13(8):509-23.

[31] Mao D, Okada BK, Wu Y, Xu F, Seyedsayamdost MR. Recent Advances in Activating Silent Biosynthetic Gene Clusters in Bacteria. Current opinion in microbiology. 2018 oct;45:156.

[32] Liu Z, Zhao Y, Huang C, Luo Y. Recent Advances in Silent Gene Cluster Activation in Streptomyces. Frontiers in Bioengineering and Biotechnology. 2021 feb;9.

[33] Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology. 2016 jul;17(7):451-9.

[34] de Hoffmann E, Stroobant V. Mass Spectrometry: Principles and Applications. Wiley; 2013.

[35] Nagana Gowda GA, Djukovic D. Overview of mass spectrometry-based metabolomics: Opportunities and challenges. Methods in Molecular Biology. 2014;1198:3-12.

[36] McLafferty FW. A Century of Progress in Molecular Mass Spectrometry. Annual Review of Analytical Chemistry. 2011;4(1):1-22. PMID: 21351881.

[37] Zhang J, Gonzalez E, Hestilow T, Haskins W, Huang Y. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. Current Genomics. 2009 sep;10(6):388-401.

[38] Smith R, Mathis AD, Ventura D, Prince JT. Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. BMC Bioinformatics. 2014;15(7):S9.

[39] Stricker T, Bonner R, Lisaek F, Hopfgartner G. Adduct annotation in liquid chromatography/high-resolution mass spectrometry to enhance compound identification. Analytical and Bioanalytical Chemistry. 2020 oct;413(2):503-17.

[40] Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted Metabolomics. Current Protocols in Molecular Biology. 2012;98(1):30.2.1-30.2.24.

[41] Guo J, Huan T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. Analytical Chemistry. 2020 jun;92(12):8072-80.

[42] Hilaire PBS, Rousseau K, Seyer A, Dechaumet S, Damont A, Junot C, et al. Comparative evaluation of data dependent and data independent acquisition workflows implemented on an orbitrap fusion for untargeted metabolomics. Metabolites. 2020 apr;10(4).

[43] Brodsky L, Moussaieff A, Shahaf N, Aharoni A, Rogachev I. Evaluation of peak picking quality in LC-MS metabolomics data. Analytical Chemistry. 2010;82(22):9177-87.

[44] Wandy J, Davies V, J J van der Hooft J, Weidt S, Daly R, Rogers S. In Silico Optimization of Mass Spectrometry Fragmentation Strategies in Metabolomics. Metabolites. 2019 oct;9(10):219.

[45] Andrews GL, Dean RA, Hawkridge AM, Muddiman DC. Improving Proteome Coverage on a LTQ-Orbitrap Using Design of Experiments. Journal of The American Society for Mass Spectrometry. 2011 apr;22(4):773-83.

[46] Johnson D, Boyes B, Fields T, Kopkin R, Orlando R. Optimization of data-dependent acquisition parameters for coupling high- speed separations with LC-MS/MS for protein identifications. Journal of Biomolecular Techniques. 2013 jul;24(2):62-72.

[47] Hecht ES, Scigelova M, Eliuk S, Makarov A. Fundamentals and Advances of Orbitrap Mass Spectrometry. In: Encyclopedia of Analytical Chemistry. Wiley; 2019. p. 1-40.

[48] Davies V, Wandy J, Weidt S, van der Hooft JJJ, Miller A, Daly R, et al. Rapid Development of Improved Data-Dependent Acquisition Strategies. Analytical Chemistry. 2021 apr;93(14):5676-83.

[49] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML - A community standard for mass spectrometry data. Molecular and Cellular Proteomics. 2011 jan;10(1).

[50] Sturm M, Kohlbacher O. TOPPView: An open-source viewer for mass spectrometry data. Journal of Proteome Research. 2009 jul;8(7):3760-3.

[51] Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. Nature Methods. 2020 Sep;17(9):905-8.

[52] Katajamaa M, Orešič M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2005 jul;6(1):179.

[53] Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. Journal of proteome research. 2010 aug;9(8):4152.

[54] Bazsó FL, Ozohanics O, Schlosser G, Ludányi K, Vékey K, Drahos L. Quantitative Comparison of Tandem Mass Spectra Obtained on Various Instruments. Journal of The American Society for Mass Spectrometry 2016 27:8. 2016 may;27(8):1357-65.

[55] Lu W, Su X, Klein MS, Lewis IA, Fiehn O, Rabinowitz JD. Metabolite measurement: Pitfalls to avoid and practices to follow. Annual Review of Biochemistry. 2017 jun;86:277-304.

[56] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry. 2010;45(7):703-14.

[57] Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: A Metabolite Mass Spectral Database. Therapeutic Drug Monitoring. 2005;27(6).

[58] Stein SE. Chemical substructure identification by mass spectral library searching. Journal of the American Society for Mass Spectrometry. 1995;6(8):644-55. PMID: 24214391.

[59] Nothias LF, Nothias-Esposito M, Da Silva R, Wang M, Protsyuk I, Zhang Z, et al. Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. Journal of Natural Products. 2018 apr;81(4):758-67.

[60] Quinn RA, Nothias LF, Vining O, Meehan M, Esquenazi E, Dorrestein PC. Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. Trends in Pharmacological Sciences. 2017 Feb;38(2):143-54.

[61] Atencio LA, Boya P CA, Martin H C, Mejía LC, Dorrestein PC, Gutiérrez M. Genome Mining, Microbial Interactions, and Molecular Networking Reveals New Dibromoalterochromides from Strains of Pseudoalteromonas of Coiba National Park-Panama. Marine Drugs. 2020 sep;18(9):456.

[62] Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular networking as a dereplication strategy. J Nat Prod. 2013 Sep;76(9):1686-99.

[63] Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol. 2015 Apr;22(4):460-71.

[64] Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016 Aug;34(8):828-37.

[65] Doroghazi JR, Metcalf WW. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genomics. 2013 sep;14(1):611.

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

[67] 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 Research. 2019 10;48(D1):D454-8.

[68] Ballouz S, Francis AR, Lan R, Tanaka MM. Conditions for the evolution of gene clusters in bacterial genomes. PLoS Computational Biology. 2010 feb;6(2):1000672.

[69] Johnston CW, Skinnider MA, Wyatt MA, Li X, Ranieri MRM, Yang L, et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. Nat Commun. 2015 Sep;6:8421.

[70] Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. Natural Product Reports. 2013;30(1):108-60.

[71] van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev. 2020:.

[72] Tobias NJ, Wolff H, Djahanschiri B, Grundmann F, Kronenwerth M, Shi YM, et al. Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. Nature Microbiology. 2017;2(12):1676-85.

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

[74] Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Research. 2017;45(W1):W49-54.

[75] Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genetics and Biology. 2010 sep;47(9):736-41.

[76] Coggill P, Finn RD, Bateman A. Identifying Protein Domains with the Pfam Database. Current Protocols in Bioinformatics. 2008;23(1):2.5.1-2.5.17.

[77] 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 Research. 2019 08;47(18):e110-0.

[78] Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014 Jul;158(2):412-21.

[79] Umemura M, Koike H, Nagano N, Ishii T, Kawano J, Yamane N, et al. MIDDAS-M: Motif-Independent De Novo Detection of Secondary Metabolite Gene Clusters through the Integration of Genome Sequencing and Transcriptome Data. PLoS ONE. 2013 dec;8(12). Available from: `/pmc/articles/PMC3877130//pmc/articles/PMC3877130/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3877130/`.

[80] Umemura M, Koike H, Machida M. Motif-independent de novo detection of secondary metabolite gene clusters—toward identification from filamentous fungi. Frontiers in Microbiology. 2015;6(MAY):371. Available from: `/pmc/articles/PMC4419862//pmc/articles/PMC4419862/`

?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419862/.

[81] Chavali AK, Rhee SY. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. Briefings in Bioinformatics. 2017 04;19(5):1022-34.

[82] Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020 Jan;16(1):60-8.

[83] Lin K, Zhu L, Zhang DY. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics. 2006;22(17):2081-6.

[84] Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014 Nov;10(11):963-8.

[85] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar G, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2020 10;49(D1):D412-9.

[86] 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 Oct;47(18):e110.

[87] Saitou N, Nei M, Saitou N NM. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;4(4):406-25.

[88] Daszykowski M, Walczak B. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: KDD-96 Proceedings. vol. 2; 2010. p. 635-54.

[89] Dueck D, Frey BJ. Clustering by Passing Messages Between Data Points. Science. 2007;315(5814):972-6.

[90] Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience. 2021;10(1).

[91] Männle D, McKinnie SMK, Mantri SS, Steinke K, Lu Z, Moore BS, et al. Comparative Genomics and Metabolomics in the Genus Nocardia . mSystems. 2020 jun;5(3).

[92] Soldatou S, Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Hughes AH, Rogers S, et al. Comparative Metabologenomics Analysis of Polar Actinomycetes. Marine Drugs. 2021;19(2).

[93] Soldatou S, Hjörleifsson Eldjárn G, Huerta-Uribe A, Rogers S, Duncan KR. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. FEMS Microbiol Lett. 2019 Jul;366(13).

[94] Medema MH. The year 2020 in natural product bioinformatics: An overview of the latest tools and databases. Natural Product Reports. 2021 feb;38(2):301-6.

[95] Cao L, Gurevich A, Alexander KL, Naman CB, Leão T, Glukhov E, et al. MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities. Cell Systems. 2019 dec;9(6):600-8.e4.

[96] Mohimani H, Liu WT, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA. NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. J Nat Prod. 2014;77:0.

[97] Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, et al. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. PLOS Computational Biology. 2014 09;10(9):1-7.

[98] Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics. 2017 Oct;33(20):3202-10.

[99] Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai HC, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nature Chemical Biology. 2017;13(5):470-8.

[100] Panter F, Krug D, Müller R. Novel Methoxymethacrylate Natural Products Uncovered by Statistics-Based Mining of the Myxococcus fulvus Secondary Metabolome. ACS Chemical Biology. 2019;14(1):88-98.

[101] Kersten RD, Yang YL, Xu Y, Cimermancic P, Nam SJ, Fenical W, et al. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011 Oct;7(11):794-802.

[102] 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 dec;11(1):1-9.

[103] Mohimani H, Gurevich A, Mikheenko A, Garg N, Nothias LF, Ninomiya A, et al. Dereplication of peptidic natural products through database search of mass spectra. Nat Chem Biol. 2017 Jan;13(1):30-7.

[104] Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018 Oct;9(1):4035.

[105] 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 Feb;2(2):99-108.

[106] Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 2019 aug;20(8):467-84.

[107] San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. Frontiers in Microbiology. 2020 jan;10.

[108] Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, et al. A community resource for paired genomic and metabolomic data mining. Nature Chemical Biology. 2021 apr;17(4):363-8.

[109] Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: A resource evolving in response to the needs of its scientific community. Nucleic Acids Research. 2020 jan;48(D1):D440-4.

[110] Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Central Science. 2019 nov;5(11):1824-33.

[111] Schügerl K. Extraction of primary and secondary metabolites. Advances in Biochemical Engineering/Biotechnology. 2005;92:1-48.

[112] Salem MA, Jüppner J, Bajdzienko K, Giavalisco P. Protocol: A fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. Plant Methods. 2016 nov;12(1):45.

[113] Brouard C, Szafranski M, D'alché-Buc F. Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. J Mach Learn Res. 2016;17(1):1-48.

[114] McSherry F, Najork M. Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In: Macdonald C, Ounis I, Plachouras V, Ruthven I, White RW, editors. Advances in Information Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 414-21.

[115] van Santen JA, Kautsar SA, Medema MH, Linington RG. Microbial natural product databases: moving forward in the multi-omics era. Natural Product Reports. 2020.

[116] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

[117] Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: Database boosting is needed to achieve 93% accuracy. Journal of Cheminformatics. 2017 may;9(1):32.

[118] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model. 1988 Feb;28(1):31-6.

[119] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform. 2015 May;7:23.

[120] Pletnev I, Erin A, McNaught A, Blinov K, Tchekhovskoi D, Heller S. InChIKey collision resistance: An experimental testing. Journal of Cheminformatics. 2012 dec;4(12):39.

[121] Crüsemann M, O'Neill EC, Larson CB, Melnik AV, Floros DJ, da Silva RR, et al. Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. J Nat Prod. 2017 Mar;80(3):588-97.

[122] Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Biol. 2007 Jan;14(1):53-63.

[123] Leao T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus. Proc Natl Acad Sci U S A. 2017 Mar;114(12):3198-203.

[124] Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, et al. MS/MS networking guided analysis of molecule and gene cluster families. Proceedings of the National Academy of Sciences of the United States of America. 2013 jul;110(28):E2611-20.

[125] Helfrich EJN, Vogel CM, Ueoka R, Schäfer M, Ryffel F, Müller DB, et al. Bipartite interactions, antibiotic production and biosynthetic potential of the Arabidopsis leaf microbiome. Nature Microbiology. 2018 aug;3(8):909-19.

[126] Parkinson EI, Tryon JH, Goering AW, Ju KS, McClure RA, Kemball JD, et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. ACS Chemical Biology. 2018:acschembio.7b01089.

[127] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics. 2017 jul;101(1):5-22.

[128] Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nature Reviews Genetics. 2020 jul;21(7):428-44.

[129] Gregory TR. Understanding Evolutionary Trees. Evolution: Education and Outreach. 2008 feb;1(2):121-37.

[130] Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: Building the web of life. Nature Reviews Genetics. 2015 aug;16(8):472-82.

[131] Metcalf JA, Funkhouser-Jones LJ, Brileya K, Reysenbach AL, Bordenstein SR. Antibacterial gene transfer across the tree of life. eLife. 2014 nov;3(November):1-18.

[132] Bello-López JM, Cabrero-Martínez OA, Ibáñez-Cervantes G, Hernández-Cortez C, Pelcastre-Rodríguez LI, Gonzalez-Avila LU, et al. Horizontal gene transfer and its association with antibiotic resistance in the genus aeromonas spp. Microorganisms. 2019 sep;7(9).

[133] Adamek M, Alanjary M, Ziemert N. Applied evolution: Phylogeny-based approaches in natural products research. Natural Product Reports. 2019 sep;36(9):1295-312.

[134] Alanjary M, Steinke K, Ziemert N. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. Nucleic Acids Research. 2019 04;47.

[135] Trappe K, Marschall T, Renard BY. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. Bioinformatics. 2016;32(17):i595-604.

[136] Seiler E, Trappe K, Renard BY. Where did you come from, where did you go: Refining metagenomic analysis tools for horizontal gene transfer characterisation. PLoS Computational Biology. 2019;15(7).

[137] Fields S, Kohara Y, Lockhart DJ. Functional genomics. Proceedings of the National Academy of Sciences of the United States of America. 1999 aug;96(16):8825-6.

[138] Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Research. 2017;45(W1).

[139] Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research. 2015 jul;43(W1):W237-43.

[140] Hufsky F, Böcker S. Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. Mass Spectrometry Reviews. 2017;36(5):624-33.

[141] Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. Journal of Cheminformatics. 2020 Jun;12(1):43.

[142] Klekota J, Roth FP. Chemical substructures that enrich for biological activity. Bioinformatics. 2008 nov;24(21):2518-25.

[143] Casey S, Perry JW, Publishing Corp New York Y RN, Morgan HL. Tools for Machine Literature Searching; 1964. 1.

[144] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010;50(5):742-54.

[145] Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A. 2015 Oct;112(41):12580-5.

[146] Brouard C, Shen H, Dührkop K, d'Alché Buc F, Böcker S, Rousu J. Fast metabolite identification with Input Output Kernel Regression. Bioinformatics. 2016 Jun;32(12):i28-36.

[147] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. USA: Cambridge University Press; 2004.

[148] Chen Y, Garcia EK, Gupta MR, Rahimi A, Edu LW. Similarity-based Classification: Concepts and Algorithms Luca Cazzanti. Journal of Machine Learning Research. 2009;10:747-76.

[149] Micchelli CA, Bartlett P. Learning the Kernel Function via Regularization Massimiliano Pontil. Journal of Machine Learning Research. 2005;6:1099-125.

[150] Cortes C, Mohri M, Weston J. A general regression technique for learning transductions. In: ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning; 2005. p. 153-60.

[151] Jebara T, Kondor R, Howard A. Probability Product Kernels. J Mach Learn Res. 2004;5:819-44.

[152] Brouard C, Bassé A, D'alché-Buc F, Rousu J. Improved small molecule identification through learning combinations of kernel regression models. Metabolites. 2019;9(8).

[153] Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform. 2017 Jun;9(1):33.

[154] Heinonen M, Shen H, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics. 2012;28(18):2333-41.

[155] Kalli A, Smith GT, Sweredoski MJ, Hess S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: Focus on LTQ-orbitrap mass analyzers. Journal of Proteome Research. 2013;12(7):3071-86.

[156] Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Journal of Cheminformatics. 2016 Jan;8(1):3.

[157] van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Central Science. 2019;5(11):1824-33.

[158] Bittremieux W, Chen C, Dorrestein PC, Schymanski EL, Schulze T, Neumann S, et al. Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. bioRxiv. 2020.

[159] Walther E, Boldt S, Kage H, Lauterbach T, Martin K, Roth M, et al. Zincophorin - biosynthesis in Streptomyces griseus and antibiotic properties. GMS infectious diseases. 2016 Nov;4:Doc08-8. 30671322[pmid].

[160] Soon Ong C, Mary X, Smola AJ. Learning with Non-Positive Kernels. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Canada; 2004. .

[161] Haasdonk B. Feature space interpretation of SVMs with indefinite kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005 apr;27(4):482-92.

[162] Lin HT, Lin CJ. A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods. National Taiwan University; 2003.

[163] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management. 1988 jan;24(5):513-23.

[164] Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 1972;28(1):11-21.

[165] Wu G, Chang EY, Zhang Z. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In: Proceedings of the 22nd International Conference on Machine Learning; 2005. .

[166] Cortes C, Mohri M, Rostamizadeh A. Algorithms for Learning Kernels Based on Centered Alignment. Journal of Machine Learning Research. 2012;13:795-828.

[167] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Bengio Y, LeCun Y, editors. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings; 2013. .

[168] Le Q, Mikolov T. Distributed Representations of Sentences and Documents; 2014.

[169] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics; 2013. p. 1631-42.

[170] Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLOS Computational Biology. 2021 05;17(5):1-24.

[171] Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: Increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Research. 2021 jan;49(D1):D639-43.

[172] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13. Red Hook, NY, USA: Curran Associates Inc.; 2013. p. 3111–3119.

[173] Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLOS Computational Biology. 2021 02;17(2):1-18.

[174] Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore - a novel deep learning similarity measure for mass fragmentation spectrum comparisons. bioRxiv. 2021 apr:2021.04.18.440324.

[175] Stokman S, van der Hooft J, Medema M. Creating a Natural Product Database and Generating Molecular Substructures. Wageningen University and Research; 2019.

[176] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. Clustering algorithms: A comparative approach. PLOS ONE. 2019 jan;14(1):e0210236.