Exploring the chemical space of natural products from Streptomyces using multi-omics approaches

Nuñez Santiago, I.

01

CHAPTER 1

Exploring the Chemical Space of Microbial Natural Products: Priorization, Strategies and Challenges

Isabel Nuñez Santiago, Gilles P. van Wezel

Introduction

The discovery and exploitation of natural products (NPs) as source and inspiration for our medicines has been a cornerstone of human health, spanning a wide range of sources from plants to fungi. Historically, these natural substances have played a pivotal role in the treatment and prevention of diseases, forming the basis of many therapeutic practices across cultures (Coy-Barrera *et al.*, 2023). In 1928, Alexander Fleming's landmark discovery of penicillin, produced by the fungus *Penicillium notatum*, caused a paradigm shift, underscoring the immense potential of microbial-derived compounds in medicine (Cragg *et al.*, 1997, Newman & Cragg, 2016). Since then, microbial natural products (NPs) have emerged as an invaluable resource for drug discovery, leading to the identification and development of a vast array of bioactive compounds with diverse structures and functions.

Natural products have significantly contributed to the advancement of modern medicine and other fields. Notable examples include antibiotics, antifungals, anticancer agents and immunosuppressants, many of which have become indispensable in the clinic (Hutchings *et al.*, 2019). Moreover, microbial natural products have found application beyond human health, including in agriculture, where they play roles as biopesticides and growth promoters (Ruiu, 2018), as well as in biotechnology, for example as enzymes, biofuels and cosmetics (Priya *et al.*, 2023).

Among the microbes that produce the majority of known natural products, filamentous fungi, such as those from *Penicillium*, and bacteria from the order of Actinomycetales, are among the richest sources (Bérdy, 2005, Newman & Cragg, 2007). Actinobacteria are renowned for their exceptional ability to generate a vast array of bioactive compounds. For example, approximately two-thirds of all known antibiotics are derived from Actinobacteria, with the majority produced by the genus *Streptomyces (Alam et al., 2022)*. They are Gram-positive bacteria which have a mycelial lifestyle that are found in soil and aquatic environments.

A major issue caused by the over-use of antibiotics is the sharp rise in antimicrobial resistance (AMR). In 1940 the first case of penicillin resistance was published (Abraham & Chain, 1988), and since then AMR has grown exponentially (Murray *et al.*, 2022). Alexander Fleming already warned that misusing antibiotics could lead to the emergence of drug-resistant bacteria (Rosenblatt-Farrell, 2009). We are now in the situation where if nothing changes radically, the number of human deaths worldwide due to AMR will rise from the current 700,000 per year to

over 10 million by 2050 (O'Neill, 2014). Thus, we are at the huge risk that the era of antibiotics might end a mere century after their first clinical use.

During the "golden age" of antibiotic discovery, industry was largely able to keep pace with the rise of resistance by rapidly developing new antimicrobials. However, this period of abundant discovery eventually came to a near complete stand-still, leading to what is now known as the "discovery void" (Aminov, 2010). Several factors have contributed to this stagnation, including the exhaustion of easily accessible "low-hanging fruit" and the frequent rediscovery of already known antibiotics rather than the sought-after novel compounds (Aminov, 2010). When it seemed that the discovery of new antimicrobials from microbes had reached its end, the sequencing revolution at the start of the 21st century offered new avenues. Genome sequencing revealed a hitherto unsuspected biosynthetic potential of microorganisms (Bentley *et al.*, 2002, Cruz-Morales *et al.*, 2013, Ikeda *et al.*, 2003). Natural products display a wide variety of chemical structures, including β-lactams, aminoglycosides, polyketides (PKs), nonribosomal peptides (NRPs) and ribosomally synthesized and post-translationally modified peptides (RiPPs). In prokaryotes, the genes involved in the biosynthesis of a particular molecule are often clustered together in what are known as biosynthetic gene clusters (BGCs) (Liu *et al.*, 2021). These clusters can span up to 100 kilobases and typically include genes for core enzymes, tailoring enzymes, resistance, regulation and export. Regulatory proteins ensure the correct timing of their production, which is important given that NP production is energetically costly for the organism (Weber *et al.*, 2015). The combination of genetics, bioinformatics and metabolomics has enabled scientists to link each BGC to its corresponding NP and unravel the underlying biosynthetic logic (Weber *et al.*, 2015, Medema *et al.*, 2011). This has not only expanded our understanding of microbial metabolism but has also opened new avenues for the discovery and development of novel natural products.

These advancements led to the discovery of previously hidden gems within microbial genomes. A notable example is the model organism *Streptomyces coelicolor*. In the 1950s, David Hopwood selected *S. coelicolor* as a model organism due to its production of distinctive blue and red pigments, which were later identified as the antibiotics actinorhodin and prodigionines, respectively (Hopwood, 2007). When the genome of *S. coelicolor* was fully sequenced in 2002 (Bentley *et al.*, 2002), many previously unsuspected BGCs were discovered, including one for the biosynthesis of a polyketide antibiotic, now known as coelimycin (Gomez-Escribano *et al.*, 2012). This led to a revolution in microbial drug discovery, whereby scientists sought to discover new chemical space based on such 'cryptic' or 'silent' BGCs (Hoskisson & Seipke, 2020).

Following the genome sequencing revolution, new high-throughput (HT) analytical methods emerged (Ayon, 2023). Besides the well-known HT screening regimes of compound and strain libraries, new methodologies were developed in bioinformatics for large-scale genome and metagenome analysis (Kalkreuter *et al.*, 2020), providing insights into genetic variation and microbial community structures (Kang *et al.*, 2024, Carrión *et al.*, 2019); genome-wide transcriptomics to profile gene expression across whole cells or communities (Lee *et al.*, 2022); mass spectrometry-based quantitative proteomics for detailed protein identification, quantification, and interaction mapping (Gubbens *et al.*, 2014); and metabolomics based on mass spectrometry (MS) or nuclear magnetic resonance (NMR) to uncover and quantify metabolites, revealing metabolic pathways and systemic changes (Kersten & Dorrestein, 2009).

In this review, we assess the latest methodologies used in natural product discovery, highlighting their advantages for NP discovery, while taking into account their inherent biases and limitations. We also explore how integrating the various techniques and disciplines into a more holistic approach can help scientists to unlock the full potential of the extent natural product resources (Fig. 1).

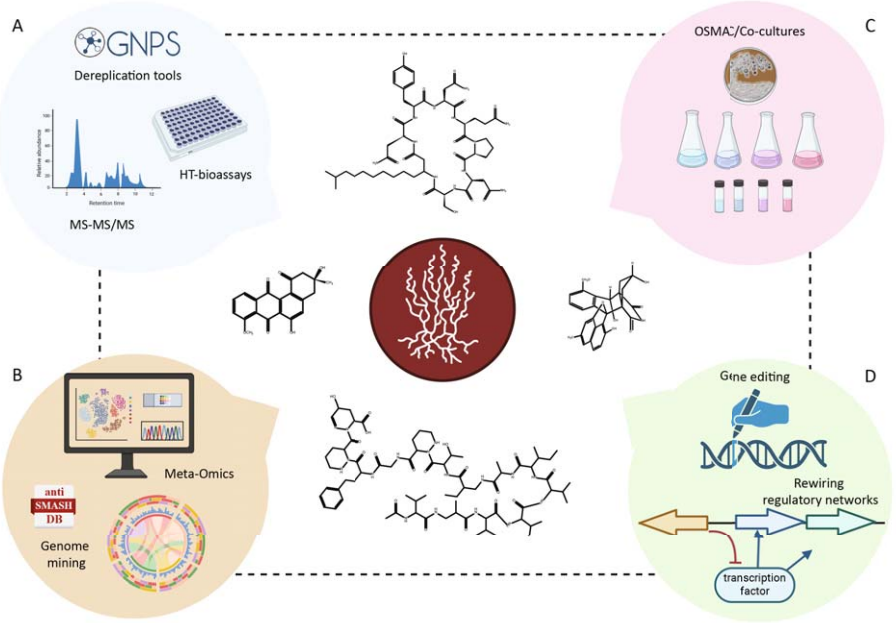

Figure 1. Integrating technologies to explore the chemical diversity of *Streptomyces*. By integrating high-throughput methods (A, B) with medium-throughput technologies (C, D), the exploration of chemical space can be conducted more efficiently. The molecules shown represent metabolites produced by *Streptomyces*, highlighting the chemical diversity of this genus, which can synthesize a broad range of bioactive compounds.

The era of mega data: exploring vast collections, high-throughput screening and meta-omics

Exploring microbial collections

Microbial collections, often built through years of dedicated sampling and isolation efforts, are rich repositories of microbial diversity. The quest for novel natural products has led researchers to explore new habitats, such as deserts, deep-sea vents and permafrost, as sources of rare and previously untapped microbes. Many collections exist at research institutes and at pharmaceutical companies. These include the Natural Product Discovery Centre, the collection of Fundación Medina (Motamedi *et al.*, 1995) (formerly part of MSD), and Naicons which are collections of Actinobacteria that are accessible to the broad scientific community (Zdouc *et al.*, 2021). However large these collections are, a major challenge lies in accessing and prioritizing their biosynthetic potential. One such challenge is that many BGCs are not expressed under routine laboratory conditions. Strains are cultivated under a number of growth conditions, and most NPs produced under such conditions have by now been identified. Elicitation approaches to activate silent BGCs and identify their compounds are discussed in more detail below. Additionally, many strains produce one or two compounds at a very high level, which overshadows minor metabolites that remain undetected or lost in the noise. These highly abundant and well-known compounds are also referred to as nuisance antibiotics (Culp *et al.*, 2019). These factors contribute to the frequent rediscovery (known as replication) of known molecules. While the sheer scale of these collections offers the potential to uncover new chemistry, the inefficiencies in current mining approaches can lead to a waste of resources and missed opportunities for breakthrough discoveries.

High-Throughput Screening

High-throughput screening (HTS) allows the large-scale processing and analysis of collections of natural product extracts, facilitating the discovery of novel molecules derived from natural sources (Genilloud, 2012, Genilloud *et al.*, 2011, Katz & Baltz, 2016, Lewis, 2013). Activity- and target-based screening thereby offers insights into the biological activities of natural products, elucidating their mechanisms of action and potential therapeutic applications as drug candidates (Henrich & Beutler, 2013). These screenings are categorized into three main approaches: classical, bioactivity-guided, and non-classical methods such as heterologous expression.

Classical HT screening of microbial strain collections involves systematically testing large collections of natural product extracts or purified compounds against a predefined panel of bioassays (Genilloud, 2012, Lewis, 2013). This approach is particularly effective in identifying bioactive compounds and serves as a foundational step in drug development pipelines. Bioactivity-guided screening refines the classical approach by linking the detection of biological activity directly to the isolation and identification of active compounds (Atanasov *et al.*, 2021). This method improves efficiency by focusing resources on compounds with proven activity, thus enhancing the likelihood of discovering potent new lead compounds. Lastly, the non-classical approach of heterologous expression involves transferring BGCs from the original microorganism to a host organism that can be more easily manipulated or optimized for production. This technique allows researchers to express and study natural products that might be difficult to detect in their native hosts, broadening the scope of discovery (Teijaro *et al.*, 2019, Wenzel & Muller, 2005, Seong *et al.*, 2021). Moreover, it can be beneficial in cases where the native producer cannot be cultured under laboratory conditions (Hover *et al.*, 2018).

HTS also has its limitations (Shen, 2015, Kolter & van Wezel, 2016). One major drawback is that the success of HTS relies on the alignment between the bioassay and the target activity. If the chosen bioassay does not accurately reflect the desired activity, valuable compounds or strains may be overlooked, leading to missed opportunities in drug discovery (Henrich & Beutler, 2013). Additionally, HT methods may fail to detect less abundant or less active compounds, particularly when dominant metabolites overshadow these minor substances (Atanasov *et al.*, 2021). These limitations can lead to an incomplete understanding of the natural product diversity within a microbial community. It is crucial to consider these when using data from high-throughput screenings to prioritize strains (Kolter & van Wezel, 2016). This issue is closely tied to the challenge of dereplication—the repeated rediscovery of known compounds, which are often ubiquitous in natural product screens (Atanasov *et al.*, 2021, Avery *et al.*, 2010). Such replication hinders efforts to identify new, often minor, compounds within the natural product pool. To address this, effective dereplication tools are essential in HT screenings. Implementing these tools requires adherence to F.A.I.R. (Findable, Accessible, Interoperable, Reusable) (Dierkes & Rücknagel, 2017) data principles and ensuring open access to well-characterized compounds. By doing so, researchers can streamline the

discovery process globally, focusing on genuinely novel compounds rather than revisiting known ones.

Strategies for strain prioritization in natural product discovery

As microbial strain collections expand, there is a clear need for new approaches to effectively prioritize strains. Efficient strain prioritization will clearly aid in the discovery of novel microbial natural products. Traditionally, methods for selecting bacterial strains for drug discovery have relied on phenotypic traits or genetic variations, rather than on the laboratory production of secondary metabolites (Hou *et al.*, 2012). To address this, untargeted LC/MS-based secondary metabolomics was assessed as a rapid and efficient approach for analyzing bacterial natural products, where the metabolites are identified using techniques like liquid chromatography coupled with mass spectrometry (LC–MS) and/or NMR spectroscopy. Statistical approaches, including Pearson correlation, partial least squares (PLS), discriminant analysis (such as PCA-DA, PLS-DA, and OPLS-DA), and hierarchical cluster analysis (HCA), are subsequently applied to correlate metabolite fingerprints with bioactivity data.

In the priorization workflow, efficient dereplication is crucial to prevent the redundancy of rediscovering known compounds, saving both time and resources. In this context, the Global Natural Products Social (GNPS) molecular networking platform, developed in the Dorrestein laboratory, represents a valuable tool in the process (Aron *et al.*, 2020). GNPS serves as an open-access knowledge base for the community-wide organization and sharing of raw, processed, or identified tandem mass spectrometry (MS/MS) data. It enables crowd-sourced curation of reference MS libraries, improving compound annotations and facilitating better data interpretation (Aron *et al.*, 2020). The platform's data-driven, social-networking approach fosters collaboration and supports the identification of spectra. Additionally, GNPS introduced the concept of "living data," with continuous reanalysis of deposited data, ensuring that the knowledge base remains dynamic and up to date.

Building on this, several platforms have been developed to support high-throughput screening guided by bioassays, such as FERMO (Zdouc *et al.*, 2022), which relies on bioinformatic predictions to link molecules and bioactivity; nanoRAPIDS, which uses experimental testing via nanofractionation and HT-bioassays (Nunez Santiago *et al.*, 2024). These platforms complement the capabilities of GNPS, further enhancing the ability to rapidly identify and validate bioactive compounds.

Meta-omics and culture-independent methods

While the screening of culturable bacteria and filamentous fungi has historically been highly effective in the search for novel antibiotics, there is a clear need for novel approaches. Culture-dependent methods have several handicaps. Firstly, most microorganisms are resistant to cultivation as monocultures under standard laboratory conditions and they are therefore known primarily through their DNA sequences (Ledford, 2015). Uncultured bacteria represent around 99% of all species in natural environments and serve as an untapped reservoir for novel antibiotics (Ling *et al.*, 2015). Second, as previously mentioned, even when microorganisms can be successfully cultured in the lab, they typically express only a fraction of their biosynthetic potential under the conditions typically used in standard screening programs. Recent advances in '-omics' methodologies have facilitated reliable, high-throughput, and cultivation-independent approaches for studying the secondary metabolism of microorganisms directly within their natural environments (Geers *et al.*, 2022).

Metagenomics has unlocked vast amounts of information about microbiome composition and diversity in natural environments (Chen *et al.*, 2024). Large-scale metagenome analyses have been instrumental in, for example, understanding the composition and the role of microbial communities in soil health, helping to explain why some soils are prone to diseases while others offer protection (Carrión *et al.*, 2019). Metagenome analyses also provide further insights into how environmental factors influence microbial communities (Mackelprang *et al.*, 2011). Additionally, sequencing efforts focused on specific functional genes have provided valuable insights into the distribution and diversity of various microbial functional groups in nature (Imhoff, 2016). One of the primary limitations of shotgun metagenomics is the fact that the technology does not allow scientists to differentiate between active and inactive members of a microbiome. As a result, it remains unclear which microorganisms are actively contributing to functional traits in the ecosystem and which are simply present, likely waiting for more favorable conditions to thrive (Shakya *et al.*, 2019).

Metatranscriptomics and metaproteomics represent significant advances in next-generation sequencing technologies. These approaches not only reveal the taxonomic structure of soil microorganisms but also provide detailed insights into their functional attributes and diversity (Shakya *et al.*, 2019). Gene expression under various environmental conditions can be studied using techniques like polymerase chain reactions and microarrays, while metatranscriptomics

enables genome-wide analysis of gene expression, offering new perspectives on the ecological roles of microbial processes (Dubey *et al.*, 2020). The vast datasets generated by meta-omics require significant computational resources for analysis. This is a new form of 'brute force', the true power of which has still to be unlocked. Fortunately, advancements in computing power and the adoption of F.A.I.R. data management principles (Dierkes & Rücknagel, 2017) have made it possible to process and share this information effectively, benefiting the scientific community. However, there remains a concern that while vast amounts of data are being generated, much of it is simply described and archived rather than utilized to the full potential. Despite the sheer volume of data collected, the efficiency in translating these findings into new molecules remains low. How to effectively prioritize BGCs from the vast pool of millions of potential BGCs remains a key challenge. Moreover, a significant bottleneck of culture-independent methods persists in the laboratory, where molecular engineering and biotechnology are still necessary to clone and express the chosen BGC (Hover *et al.*, 2018), techniques that are still low-throughput and expensive. Nevertheless, similar to the rapid advances in DNA sequencing over the past few years, it is expected that DNA synthesis will soon experience significant improvements in speed and a dramatic reduction in cost. This will make it possible to synthesize, clone, and express thousands, or even millions, of biosynthetic gene clusters in heterologous hosts, followed by screening for antibiotic activity (Kolter & van Wezel, 2016).

One-Strain, Many Compounds approaches for NP discovery

In natural product discovery, especially in the academic setting, single-strain approaches are commonly employed. These methods focus on thoroughly exploring the potential of one single strain, aiming to extract as much information and as many compounds as possible from it. After all, less than 10% of the BGCs are estimated to be expressed at a sufficient level for their associated natural product to be detected under laboratory culture conditions (Katz & Baltz, 2016). While such analyses do not allow the discovery of many different compounds, the knowledge gained from such detailed studies can serve as template for research on other strains. *Streptomyces coelicolor* serves as a prime example; as a model organism, it is extensively studied, providing a wealth of knowledge that aids in further exploration.

The “One Strain MAny Compounds” (OSMAC) strategy is an effective method for natural product discovery in a single strain (Scherlach & Hertweck, 2009, Zhang *et al.*, 2024). By modifying culture conditions, the OSMAC approach can trigger the expression of microbial genes in a non-targeted manner that may influence multiple biosynthetic gene clusters (BGCs) (Bode *et al.*, 2002). In the fungus *Stachybotrys chartarum*, OSMAC allowed the discovery of the distachydrimanes, a unique class of dimeric phenylspirodrimanes with an unusual C-18–C-23′ linkage (Alam *et al.*, 2022). The OSMAC also allowed the discovery of two analogues of actinomycin, the first antibiotic ever discovered in an actinomycete. Actinomycin L produced by *Streptomyces* sp. MBT27 incorporates anthranilamide, which only occurs under specific nutrient conditions (Machushynets *et al.*, 2022). This is an example of how new molecules can still be discovered even in well studied families of molecules. Importantly, OSMAC also led to the discovery of lugdunomycin, an angucycline-derived compound produced by *Streptomyces* sp. QL37 that features a novel carbon skeleton (Wu *et al.*, 2019b). The biosynthesis of lugdunomycin is a major topic in this PhD thesis (see Chapter 3).

Co-cultivation to elicit BGCs

Microorganisms are increasingly understood to communicate and influence one another through a complex chemical signaling system. This “chemical language” allows them to coordinate behavior, share resources, and respond to environmental cues, playing a pivotal role in shaping microbial communities and ecosystem functions (Krespach *et al.*, 2023). Microbial co-cultivation is a promising technique for activating silent biosynthetic pathways (van Bergeijk *et al.*, 2020, Bertrand *et al.*, 2014). In such experiments, microbial communication occurs through either volatile compounds or direct in-situ signaling, leading to the regulation of specialized metabolite production (Krespach *et al.*, 2023, Selegato & Castro-Gamboa, 2023, Rateb *et al.*, 2013, Wakefield *et al.*, 2017, Arora *et al.*, 2020). This regulation of biosynthetic pathways can be mediated by exogenous metabolites or autoregulatory molecules (Schroeckh *et al.*, 2009, Rateb *et al.*, 2013) resulting in pleiotropic metabolic induction. Co-cultivation is effective both between closely related *Streptomyces* species (Westhoff *et al.*, 2021) as well as between streptomycetes and other microorganisms (Seyedsayamdost *et al.*, 2012, Krespach *et al.*, 2023), such as between *Streptomyces* and filamentous fungi like *Aspergillus (Wu et al., 2015b)*. This

approach is advantageous as it does not require prior genomic knowledge or specialized equipment for cultivation and data interpretation. For more details on co-cultivation dynamics for NP discovery the readers are referred to recent reviews (Selegato & Castro-Gamboa, 2023).

Rewiring the regulatory networks

Antibiotic production in *Streptomyces* is tightly controlled by hierarchical transcriptional regulatory cascades. The regulatory complexity is underscored by the significant focus on regulation in the genome of the model organism *S. coelicolor*, which encodes over 800 regulatory proteins, accounting for over 10% of its total proteome (Bentley *et al.*, 2002). In recent years, several reviews have been published on the regulatory networks of *Streptomyces* (Romero-Rodríguez *et al.*, 2018, Van Der Heul *et al.*, 2018, Palazzotto *et al.*, 2019, Rigali *et al.*, 2018, Augustijn *et al.*, 2024b). A significant portion of our understanding of antibiotic production regulation comes from studying the BGCs for Act, Cda, and Red in *S. coelicolor* (Van Der Heul *et al.*, 2018, Bibb, 2005, van Bergeijk *et al.*, 2020). These clusters are regulated by pathway-specific activators: ActII-ORF4, CdaR, and RedD, which belong to the SARP family of *Streptomyces* antibiotic regulatory proteins (Wietzorrek & Bibb, 1997), and StrR, a member of the ParB-Spo0J family (Autret *et al.*, 2001). These regulators, located within the clusters, directly influence the transcription levels of the BGCs, thereby determining the production of their respective natural products (Gramajo *et al.*, 1993, Tomono *et al.*, 2005). Notably, ActII-ORF4, CdaR, and RedD are subject to translational regulation by a tRNA that recognizes the rare UUA codon for leucine, which is encoded by the *bldA* gene. This tRNA is also essential for the translation of numerous developmental genes, thus linking morphological and chemical differentiation in *Streptomyces* (Lawlor *et al.*, 1987, Fernandez-Moreno *et al.*, 1991).

Global regulators influence multiple metabolic pathways and may not directly impact specific biosynthetic gene clusters. They also respond to various chemical or physiological signals, such as nutrient limitation, cell wall damage, heat shock, or changes in growth conditions (van Bergeijk *et al.*, 2020). The first example of a complete signaling pathway from external sugars to the control specific BGCs is that of DasR, a highly global regulator that controls the expression of many BGCs and responds to the intracellular concentration of amino sugars. *N*-acetylglucosamine-6P and glucosamine-6-phosphate act as

ligands for DasR and thereby inactivate the repressor, which frequently leads to de-repression of BGCs (Rigali *et al.*, 2008). DasR controls the pathway-specific activators of BGCs for all antibiotics and siderophores in *S. coelicolor* (Nazari *et al.*, 2012, Craig *et al.*, 2012, Świątek-Połatyńska *et al.*, 2015). Addition of GlcNAc under nutrient-limiting conditions activates the transcription of antibiotic BGCs, which is applied in industrial screening regimes.

Understanding the global regulatory elements that govern each BGC is useful for developing targeted approaches to activate specific BGCs. This is facilitated by bioinformatic tools such as PREDetector (Hiard *et al.*, 2007) and LogoMotif (Augustijn *et al.*, 2024a). Both programs use position weight matrices to predict and analyze regulatory interactions in bacterial genomes, focusing on identifying binding sites and target genes. In PREDetector, experimentally validated transcription factor binding sites (TFBS) are thereby identified, helping researchers to search genomes for TFBSs. LogoMotif adds a TF database focused on bacterial regulatory information, particularly in Actinobacteria (Augustijn *et al.*, 2024a). LogoMotif also includes TFBS prediction tools like MiniMotif, and this is integrated into version 7 of antiSMASH (Blin *et al.*, 2023), making this the first version of this important software that includes the prediction of regulatory elements within BGCs. By utilizing this database, researchers can better predict regulatory networks and target specific BGCs for activation, facilitating more efficient discovery of bioactive compounds.

Identifying the signals or conditions that trigger the activation of transcriptional regulators can be challenging. When activation proves elusive, an alternative approach has been applied in organisms that are susceptible to molecular genetic manipulation. This method is known as HITES. Here, the promoter of the transcription factor is replaced with an inducible promoter, allowing for artificial induction of the transcriptional regulator. This, in turn, may then lead to the production of the associated natural products (Chiang *et al.*, 2011, Bergmann *et al.*, 2007, Chiang *et al.*, 2009). A challenge arises when biosynthetic genes are controlled by multiple transcriptional regulators from different classes, as is often the case in large BGCs. Nevertheless, the use of inducible promoters remains a promising tool for advancing research in this area.

In addition, removing or deactivating these genes that encode repressors of BGCs is time-consuming, but could potentially activate silent BGCs, leading to the production of natural compounds. Transcriptional regulators from the

TetR families, which often act as repressors, are particularly promising targets for this strategy. An example of TetR-mediated repression was demonstrated in *Streptomyces* sp. PGA64 and *S. ambofaciens*, where null mutations of the repressors within two silent gene clusters led to the activation of these clusters (Bunet *et al.*, 2011, Metsä-Ketelä *et al.*, 2004). This resulted in the production of the corresponding natural products: a novel angucyclinone in *S. PGA64* and kinamycins in *S. ambofaciens*.

Awakening silent BGCs via elicitation approaches

Small molecule elicitor screening is a technique used to awaken cryptic biosynthetic gene clusters by introducing specific molecules that trigger their activation. This method has gained significant traction over the past five years due to its potential in high-throughput screening to discover novel, previously undetected natural products. Elicitors, typically small molecules, act as chemical signals that mimic environmental or physiological triggers, stimulating silent BGCs into producing secondary metabolites. Several notable studies have highlighted the success of this approach in uncovering cryptic compounds (Seyedsayamdost, 2014, Okada & Seyedsayamdost, 2017, Mao et al., 2018, Xu et al., 2017, Zhang & Seyedsayamdost, 2020).

A variety of strategies have been developed to pinpoint the environmental signals that regulate the specialized metabolism of actinobacteria. The activation of BGCs is mediated by a transcriptional regulatory network influenced by cis-regulatory elements (CREs) and transcription factors (TFs) (Augustijn *et al.*, 2024b). While many TFs can respond to specific ligands, determining the nature of these ligands poses a challenge. The genomic context serves as a key indicator; for instance, a regulatory gene located adjacent to a metabolic operon, such as one involved in sugar metabolism, can offer valuable insights. Furthermore, TFs frequently exhibit autoregulation, typically having their CRE situated in the upstream region of the corresponding gene. By identifying the CRE, researchers can employ computational methods to predict the regulatory network in a virtual environment. For example, in the case of DasR, predictions showed that the most significant associations were linked to N-acetylglucosamine metabolism or transport, making it easier to identify glucosamine-6P as the ligand (Rigali *et al.*, 2008). Methods directed at single bacterial producer strains include varying the composition of growth media, inducing antibiotic resistance, and microbial cocultivation.

Strategies focused on individual bacterial producer strains include altering growth media composition (Bode *et al.*, 2002, Zhu *et al.*, 2014), inducing antibiotic resistance (Hosaka *et al.*, 2009, Tanaka *et al.*, 2009), and conducting microbial co-cultivation (Bertrand *et al.*, 2014, Hoshino *et al.*, 2015, Sugiyama *et al.*, 2015). Exploring new chemical elicitors for antibiotic production holds promise, as it increases the likelihood of successful high-throughput screenings of bacterial strain libraries. These screenings have tested various substances beyond GlcNAc through interference with DasR (Rigali *et al.*, 2008), γ-butyrolactones (Willey & Gaskell, 2011, Hsiao *et al.*, 2009), and histone deacetylase inhibitors (Albright *et al.*, 2015). One notable approach for discovering new chemical elicitors is the high-throughput elicitor screening technology (HiTES) (Moon *et al.*, 2019a, Moon *et al.*, 2019b), where a wild-type microorganism is exposed to a library of small molecules, followed by the analysis of the resultant metabolomes for bioactivity against a specific indicator strain. Utilizing this method has resulted in the identification of several cryptic antibiotics, such as the novel lanthipeptide cebulantin (Moon *et al.*, 2019a) and the new naphtoquinone epoxide hiroshidine. It also revealed atenolol, a β-blocker used in hypertension treatment, as a global elicitor (Moon *et al.*, 2019b).

The choice of an elicitor molecule may be inspired by natural interactions between organisms and their environments. The plant stress hormones jasmonic acid (JA) and methyl jasmonate modulate the life cycle of *Streptomycetes* by promoting antibiotic production and faster development. In *S. coelicolor*, these hormones specifically stimulate the synthesis of the antibiotic actinorhodin. JA is then modified within the cell to reduce toxicity. This highlights the role of plant hormones in activating cryptic biosynthetic pathways and influencing *Streptomyces* behavior (van der Meij *et al.*, 2023). The logical behind this is that host stress hormones may acts as ‘cry for help’, to activate the biosynthetic potential of the host microbiome (van der Meij *et al.*, 2017). A striking example of the application of this logic is the application of epinephrine, a human stress hormone, to induce secondary metabolite production in Actinobacteria. Indeed, it was shown that epinephrine stimulates siderophore production, with catechol identified as the key eliciting moiety. Further exploration of the catechol-responsive strain *Streptomyces* sp. MBT84 through mass spectral networking revealed the activation of a cryptic biosynthetic gene cluster responsible for producing the angucycline glycosides aquayamycin, urdamycinone B, and galtamycin C (van Bergeijk *et al.*, 2022).

Another promising approach for discovering elicitors involves inducing the SOS response through subinhibitory concentrations of antibiotics, which can potentially activate silent BGCs. This pathway offers a mechanism to regulate gene expression on a broad scale. Although the link between antibiotics, stress responses, and cryptic metabolite production is intriguing, it remains to be fully validated and will likely be a focus of future research (Rosen & Seyedsayamdost, 2017). A critical aspect of this technique is selecting an appropriate concentration of the metabolite (s) added to the cultures. Ideally, the concentration should allow for metabolic regulation without inhibiting the inducer strain, referred to as a subinhibitory concentration. Another possible limitation is ensuring that the compounds being tested are solubilized in solvents that do not interfere with microbial metabolic production, to avoid false positives when determining subinhibitory concentrations. Notably, dimethylsulfoxide (DMSO) has a well-known effect on bacterial growth and should be used at concentrations below 2% to prevent interference (Basch & Gadebusch, 1968).

Synergism between BGCs during natural product biosynthesis

In natural product biosynthesis, typically a single BGC is responsible for producing one or multiple compounds. However, the possibility that multiple BGCs work together to produce a single compound is often overlooked. This is an important aspect to consider, for example for approaches based on heterologous expression of BGCs, which will almost certainly fail in the biosynthesis of a compound requires yet another BGC. An example is actinomycin L (figure 2), produced by *Streptomyces* sp. MBT27 (figure 2B), which was already mentioned above. This molecule features an actinomycin backbone with an anthranilamide moiety. Although the BGC responsible for actinomycin L is highly conserved compared to the BGC for actinomycin X (which lacks the anthranilamide moiety) in *Streptomyces antibioticus*, it does not account for the production of anthranilamide itself. The enzyme responsible for converting anthranilic acid to anthranilamide has yet to be identified (Machushynets *et al.*, 2022). In *Streptomyces* sp. MBT76, the biosynthesis of catecholate–hydroxamate siderophores such as qinichelins (figure 2C) showcases an intricate model of intertwined biosynthetic pathways. Here, no fewer than four distinct gene clusters collaborate to produce two types of catechol-peptide siderophores: qinichelins and griseobactin (Gubbens *et al.*, 2017). This example illustrates a complex interaction between NRPS gene clusters, akin

to the functional crosstalk observed in the assembly of erythrochelin (Lazos *et al.*, 2010) in *Saccharopolyspora erythraea* and rhodochelin (Bosello *et al.*, 2011) in *Rhodococcus jostii* RHA1. Such cross-talk enhances the structural diversity of siderophores from a limited set of biosynthetic genes, providing a competitive evolutionary advantage in iron acquisition. Further examples are found in glycosylation, a common modification in nature, where sugar moieties are added to aglycones, creating diverse natural products. *Kitasatospora* sp. MBT66 exemplifies this process by producing both endophenaside antibiotics and the plecomacrolide leucanicidin (figure 2D). Leucanicidin is derived from bafilomycin A1 via methyl-rhamnosylation, using the same sugar moiety found in endophenasides. The genes *leuA* and *leuB*, adjacent to the bafilomycin biosynthesis cluster, encode a sugar-O-methyltransferase and glycosyltransferase, respectively, and are responsible for the methyl-rhamnosylation of both phenazines and macrolides. Mutating *leuB* abolishes glycosylation of both classes of natural products, highlighting the enzymes' promiscuity in modifying distinct molecular scaffolds.

A major example that is worked out in this thesis is that of lugdunomycin, discussed in Chapter 3 of this thesis. Here, the final reaction that produces lugdunomycin requires one substrate derived from an angucycline BGC, namely elmonin. The second substrate is *iso*-maleimycin, produced from a BGC that specifies a β-lactone-like compound (Uiterweerd *et al.*, 2024). Elmonin serves as the precursor for the *iso*-benzofuran molecule, which functions as the diene in an intermolecular Diels-Alder reaction together with dienophile *iso*-maleimycin (figure 2A). For details the reader is referred to Chapter 3.

The synergy between different BGCs convey an extra challenge for NP discovery. The conventional genome mining approach and the tendency to categorize BGCs may lead us to overlook the fact that a single organism can utilize various parts of its genome in concert. In future bioinformatic predictions, incorporating the potential for multi-BGC interactions could be highly beneficial. For now, adopting a bottom-up approach—starting with the chemical characteristics of natural products and then exploring their genomic basis—may offer a practical solution for discovering novel compounds produced by multiple BGCs.

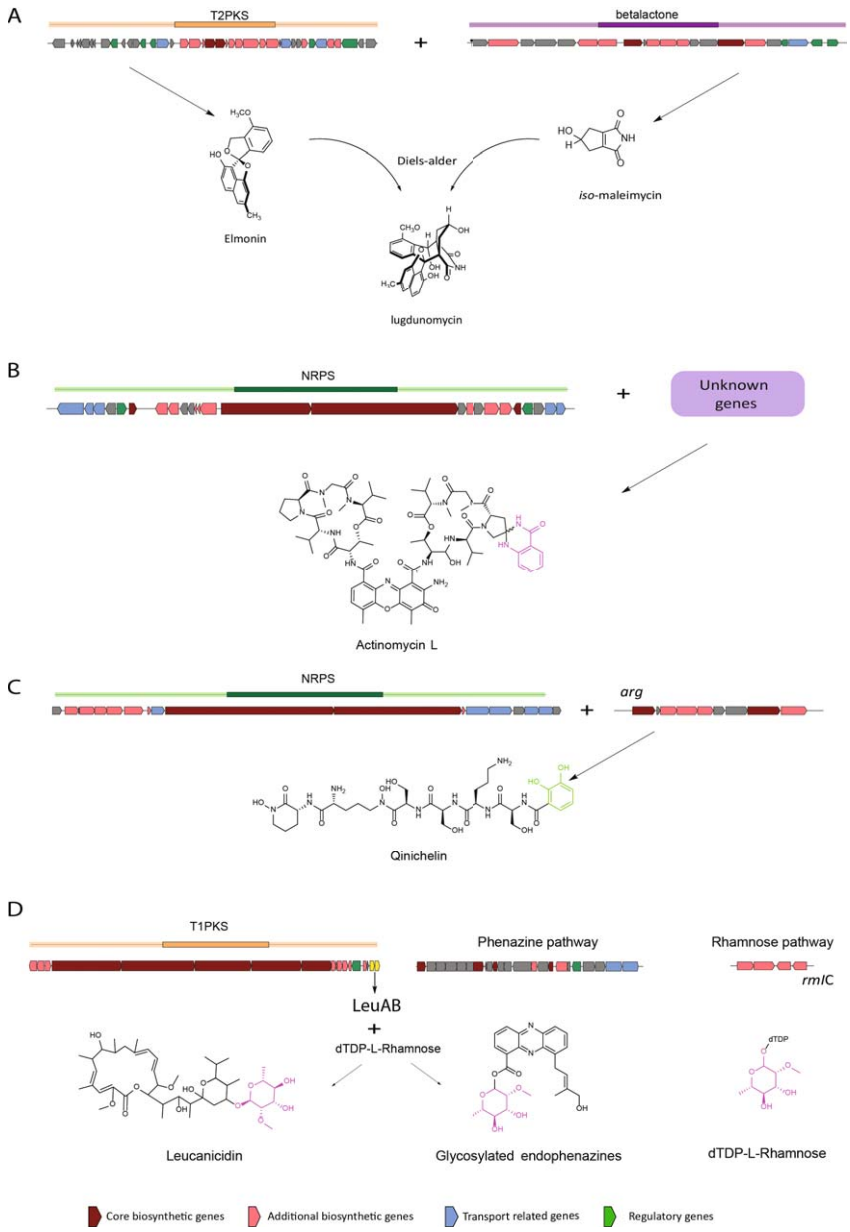

Figure 2. Overview of molecules requiring multiple BGCs for their synthesis. Notably, several classes of molecules depend on the synergy between distinct BGCs, leading to a remarkable diversity of natural products. A) Lugdunomycin pathway: The products of two different BGCs are combined through an intermolecular Diels-Alder reaction. B) Actinomycin L pathway: The addition of anthranilamide, catalyzed by an unknown enzyme not encoded by the BGC, is required for full biosynthesis. C) Simplified quinichelin pathway: A complex network of intertwined biosynthetic pathways produces catechol-peptide siderophores. D) Leucanicidin and endophenasides pathways: Both glycosylation processes are mediated by the LeuA and LeuB enzymes in the leucanicidin BGC.

A holistic approach to natural product discovery: managing data complexity and prioritizing strain potential

In natural product (NP) discovery, the sheer volume of data and resources available can be both a boon and a challenge. While the vast collections of microbial strains, extensive sequencing data, and high-throughput screening platforms provide huge opportunities, there is a risk of becoming overwhelmed by the scale of information. Integrating techniques — from exploring microbial collections and high-throughput screening to leveraging meta-omics and one-strain approaches — will aid scientists in uncovering novel natural products. In this thesis, such combination of genomics, metabolomics and screening technologies has been carried out to study the biosynthetic logic of lugdunomycin (Chapter 3) as well as for the development of novel omics technologies like nanofractionation (Chapter 2).

However, more data is not always synonymous with better results. The proliferation of data often leads to a saturation point where the novelty and utility of new findings are diminished. The reality of ever-larger datasets can detract from the meaningful exploration of individual strains and their unique biosynthetic potentials. In this context, a more balanced approach that combines the depth of single-strain investigations with the breadth of high-throughput and meta-omic techniques could prove more beneficial.

Focusing on medium-scale projects, where resources are allocated more judiciously, can lead to a more efficient use of existing data and reduce redundancy. By prioritizing detailed, targeted studies of specific strains, researchers can maximize the potential of each strain and uncover new compounds that might be missed in broader, less focused screenings. In the crucial aspect of strain prioritization for natural product (NP) discovery, the integration of artificial intelligence (AI) presents a promising frontier. Concurrently, AI approaches, particularly machine learning, have revolutionized computational drug design by predicting biological activity and facilitating de novo drug development (Mullowney *et al.*, 2023, Sahayasheela *et al.*, 2022, Saldivar-Gonzalez *et al.*, 2022). In fact, while in its infancy, AI has already helped in the discovery of novel antibiotics (Kloosterman *et al.*, 2020, Wong *et al.*, 2024). These technologies offer exciting possibilities for identifying drug candidates from the vast array of molecules produced by nature. However, it is

essential to recognize that AI, while a powerful tool, is not without limitations. It learns from the data and knowledge we provide, which means it can inherit and perpetuate existing biases within the scientific community. Thus, while AI can enhance our discovery efforts, it is important to use it with a mindful approach, acknowledging its potential for bias and the need for high-quality datasets to train algorithms effectively.

We here also want to argue that global community collaboration is key to address the challenges in natural product discovery. The chemical space explored to date is estimated to be only around 3%, with redundancy and underuse of the resources remaining significant issues. A potential solution is to create unified, global databases for characterized strains, similar to the comprehensive genomic databases that facilitate worldwide research. Such databases could also include extensive archives of natural products, enabling more efficient dereplication and reducing redundancy. Furthermore, we often overlook molecules that do not exhibit immediate bioactivity and thus are rarely published. However, this focus on current bioactivity can introduce significant bias. Molecules deemed “inactive” today may hold the key to future breakthroughs for diseases we have yet to encounter. By maintaining an open and inclusive approach to data archiving and global collaboration, we can ensure that even compounds with currently unknown activities are preserved and potentially discovered as valuable therapeutic agents in the future.

Outline of the thesis: Multi-Omics and analytical technologies for natural product exploration

Actinobacteria, and in particular members of the genus *Streptomyces*, are a major source of the known bioactive natural products, and they remain an invaluable resource for novel chemical entities, which are waiting to be fully explored and harnessed. This potential is exemplified by the gifted natural product producer *Streptomyces* sp. QL37, a central focus of this thesis. This strain, isolated from soil collected in the Qinling Mountains of the People’s Republic of China, was initially selected due to the interesting pigments it produced when cultured in the laboratory. Further investigation revealed that Streptomyces sp. QL37 produces derivatives of well-known secondary metabolites, such as angucyclinones and limamycins, as well as the unprecedented lugdunomycin. The unique molecular scaffold of lugdunomycin, never before observed in nature, prompted

exploration of its biosynthesis, with expectations that its pathway would be as unconventional as its chemical structure.

In **Chapter 2**, we present nanoRAPIDS, an analytical platform based on the nanofractionation technology (Mladic *et al.*, 2018, Mladic *et al.*, 2016), coupled to LC–MS/MS and combined with innovative mass spectral analysis tools such as featured-based molecular networking (FBMN) (Nothias *et al.*, 2020) in a single analysis. The nanoRAPIDS platform was designed for the efficient identification and prioritization of low-abundance bioactive compounds in complex extracts. Requiring as little as 10 µL of crude extract, nanoRAPIDS allows combining separation, bioassays, mass spectrometry, and GNPS molecular networking for dereplication. The platform was used to identify bioactive iturin and surfactin congeners in *Bacillus* and to track bioactive angucyclines in *Streptomyces* sp. MBT84, revealing an unusual N-acetylcysteine conjugate of saquayamycin. This technology is particularly useful for correlating changes in elicited conditions with specific molecular profiles, facilitating the discovery of novel compounds under various growth conditions.

In **Chapter 3**, we present the biosynthetic pathway for lugdunomycin in *Streptomyces* sp. QL37. The latest stages of its biosynthesis were previously unknown. Through a combination of biomimetic chemical synthesis, computational methods, genomics, and mutational analysis, we could show that lugdunomycin is formed by a rare intermolecular Diels-Alder reaction involving elmonin (as a masked diene) and *iso*-maleimycin (as the dienophile). These two substrates are encoded by separate BGCs, which demonstrates that to fully understand biosynthesis, we must consider the bacterium as a whole, as many factors—such as interactions between different biosynthetic pathways—play a critical role in shaping the final product. This work highlights the challenges of discovering new chemical space, requiring the integration of multiple biosynthetic pathways and complex reactions.

In Chapter 4, we present the analysis of angucycline BGCs in *Streptomyces* species, with a focus on comparing the angucycline BGC of the lugdunomycin producer *Streptomyces* sp. QL37 to those of other angucycline-producing strains and limamycin producers. Through a combination of phylogenomics, LC-MS/MS metabolomics, and feature-based molecular networking, we identified metabolomic profiles that linked *Streptomyces* sp. QL37 to other angucycline and limamycin producers. Their metabolic profiles revealed the presence of several bioactive compounds, including 8-O-methyltetrangomycin, fluostatin A

analogues, haloquinone, and rubiginone A2. Bioactivity assays demonstrated significant antimicrobial activity against Gram-positive bacteria, including MRSA, as well as promising anticancer activity.

In **Chapter 5**, the discovery of marushamycin is reported. Marushamycin is a piperazic acid (Piz)-containing peptide produced by *Streptomyces* sp. QL37. Using antiSMASH together with the substrate predictor PARAS, we identified two nonribosomal peptide synthetases (NRPSs) encoded by BGC20, which incorporate Piz into a 14-amino acid peptide, called marushamycin A. BGC20 was activated by culturing on MM agar with aspartic acid as the sole nitrogen source. Metabolomic analysis using GNPS molecular networking and massQL identified 14 features containing a fragment matching the y-ion of piperazic acid. Further analysis, including MS sequencing and NMR spectroscopy, confirmed the presence of multiple Piz monomers in the structure. Two additional compounds, marushamycin B and C, were identified as novel related peptides.

In **Chapter 6**, the results are summarized and reviewed in a general discussion. besides summarizing the data presented in the thesis, the importance of integrative, multi-disciplinary approaches for uncovering new chemical diversity with potential bioactivity are highlighted. The thesis provides new technologies and approaches for drug discovery, aimed at exploring the vast yet untapped chemical space of natural products produced by Actinobacteria.