

Deciphering the non-ribosomal code: the language of antibiotics and other natural products

Barbara Terlouw



Propositions

1. Protein sequences are more optimal structure embeddings than human-designed featurisations.
(this thesis)
2. Only scientists using AI will remain relevant in the field of bioinformatics.
(this thesis)
3. Knowledge-driven research leads to greater scientific breakthroughs than application-driven research.
4. Extraordinary, human-like intelligence is the consequence of language, not the other way around.
5. Monetary compensation and transparent review practices are necessary incentives for the success of the peer review process.
6. Only industry-driven initiatives can solve climate change.
7. No healthy human individual needs to eat animals.

Propositions belonging to the thesis, entitled

Deciphering the non-ribosomal code: the language of antibiotics and other natural products

Barbara Terlouw
Wageningen, 26 September
2023

Deciphering the non-ribosomal code: the language of antibiotics and other natural products

Barbara Terlouw

Thesis committee

Promotor

Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Prof. Dr M.H. Medema
Personal chair, Bioinformatics Group
Wageningen University & Research

Other members

Prof. Dr R. da Silva Torrez, Wageningen University & Research
Prof. Dr V. van Noort, Leiden University
Prof. Dr G. van Westen, Leiden University
Dr J. Masschelein, Catholic University Leuven, Belgium

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Deciphering the non-ribosomal code: the language of antibiotics and other natural products

Barbara Terlouw

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 26 September 2023
at 11 a.m. in the Omnia Auditorium.

Barbara Terlouw

Deciphering the non-ribosomal code: the language of antibiotics and other natural products,
210 pages

PhD Thesis, Wageningen University, Wageningen, The Netherlands (2023)

With references, with summary in English

ISBN: 978-94-6447-817-4

DOI: <https://doi.org/10.18174/635495>

Table of contents

Chapter 1	<i>General introduction</i>	3
Chapter 2	<i>Ecology and genomics of Actinobacteria: new concepts for natural product discovery</i>	15
Chapter 3	<i>MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters</i>	35
Chapter 4	<i>Revealing determinants of translation efficiency via whole-gene codon randomisation and machine learning</i>	45
Chapter 5	<i>Predicting NRPS A-domain specificity with PARAS and PARASECT: two structure-informed machine learning algorithms</i>	69
Chapter 6	<i>The Turterra web portal for protein family data visualisation and analysis</i>	99
Chapter 7	<i>PIKACHU: a Python-based Informatics Kit for Analysing CHemical Units</i>	111
Chapter 8	<i>RAIChU: automating the visualisation of natural product biosynthesis</i>	133
Chapter 9	<i>General discussion</i>	159
	<i>References</i>	173
	<i>Summary</i>	205
	<i>Acknowledgements</i>	207

Chapter 1

General introduction

1.1. Natural products – nature's molecular technology

We often associate nature with serenity. A forest or an ocean floor, a field of wheat or a cow grazing in a green pasture; to the human eye they are scenes of tranquillity. But look closer at the narrow spaces between the roots of pines and grass and wheat plants, at a pinch of earth stuck to the bottom of a child's shoe, at minute dents and hollows in sponges and corals, at the microscopic landscape of a cow's intestines, and we find that this apparent peace is a façade. At best they are bustling living communities with rivalling factions, where microorganisms talk to each other through the language of chemistry; at worst they are warzones: hostile environments where microbes battle for survival, fighting over scarce resources like iron and nitrogen with animals, plants, and each other. Microbes often have to withstand highly challenging conditions: we find their communities from the arctics¹ to the deep jungle², in the guts of animals³ and the glands of insects⁴, in the coldest oceans⁵ and hottest deserts⁶, exposing them to cold, drought or unforgiving animal and plant immune systems while they compete and collaborate among themselves for space and nutrients.

In any war, success depends on innovative technology. The faction that is quicker to develop novel weaponry, tactics, defensive resources, and methods of communication has a better chance of survival. At the scale of microbes, these technologies are molecular. An important subset of these constitutes natural products, or specialised metabolites: small molecules with a wide variety of functions that give the microbe advantages over other species in the habitat. Some of these natural products are largely aimed at survival of environmental challenges: in iron-deficient conditions, many microbes produce natural products called siderophores, literally translated from Greek as 'iron carriers', which they send out into the environment to collect iron or other scarce metals^{7,8}; in arid and saline environments, bacteria often produce ectoines, which protect their cells from osmotic stress caused by high salt concentrations and limited water availability⁹⁻¹¹; and some microbes even produce sunscreens to protect their machinery from damaging UV light¹². Other natural products serve a more aggressive purpose by either inhibiting the growth of competing microbes through hijacking their regulatory pathways or by outright killing them. These are natural antibiotics, such as the famous fungal molecule penicillin which kills bacteria by inhibiting the production of peptidoglycan, a key component in their cell walls^{13,14}. Some microbes make antibiotics not to protect themselves but rather their symbiotic host; for instance, endophytic bacteria from the *Bacillus* genus which live within the tissues of maize plants produce a range of lipopeptides with antifungal activity, protecting the maize plant from fungi in exchange for tapping into its nutrients¹⁵. Also, specialised metabolites may mediate communication between microbes, or microbes and other organisms. Some specialised metabolites, especially those produced by plants, may act as a rallying cry, attracting beneficial microbes to their direct environment with the promise of nutrients in exchange for defensive aid¹⁶; others may serve as a signal to allied microbes to produce weaponry against a common threat¹⁷, or inform other microbes of a change in environment¹⁸. Finally, identical natural products may occupy different roles depending on their concentration: for instance, while antibiotics are lethal at high concentrations, they have been postulated to mediate communication at lower concentrations which are often observed in nature, changing the behaviour of microbes in their environment without inhibiting their proliferation^{19,20}.

1.2. Exploiting microbial armouries

The discovery of penicillin, one of the most important molecules in human history, was a happy accident. In 1928, Alexander Fleming unwittingly contaminated a petri dish of *Staphylococcus aureus*²¹, a bacterial pathogen which still causes 1-1.5 million deaths annually²². He observed that

wherever the contaminating fungus *Penicillium notatum* grew, the bacterium did not. He later discovered that the antibacterial properties came from a molecule which he named penicillin, a natural product secreted by the fungus to kill competing bacteria in its environment¹³. Since, bacterial infection has stopped being a leading cause of death, and complex surgeries that carried risk of bacterial contamination have become possible. Fleming became one of the first to leverage an isolated natural product to benefit humanity, and many would follow suit. Now, natural products and their derivatives constitute 75% of the antibiotics used in the clinic²³, such as vancomycin²⁴ and daptomycin^{25–27}, and have also found applications as anticancer drugs (bleomycin)^{28,29}, antifungals (pyrrolnitrin)³⁰, fungicides (UK-2a)³¹, pesticides (Spinosad)³², antimalarials (quinine)³³ and herbicides (glufosinate)³⁴.

While Fleming's fortuitous contamination changed the world, progress in the natural product discovery field would be slow if it relied solely on accidents to drive research forward. Also, rediscovery of known compounds is a real issue with undirected research, costing a fortune in time and resources without yielding the desired societal impact. Fortunately, the discovery of the double helical structure of DNA in the 1950s³⁵ and the subsequent elucidation of the genetic code³⁶ helped gain fundamental understanding into the molecular flow of information in organisms, which meant that researchers no longer had to depend on such lucky coincidences. In brief, sections of DNA, the blueprint of the cell, are copied (transcribed) into mRNA: instruction manuals for building proteins. Proteins are the molecular workhorses of the cell: they transport ions and molecules into, out of and within the cell; they build, modify, and destroy other proteins; copy, repair, and proof-read DNA; and carry messages to communicate within and between cells. In short, if something needs to be done in the cell, there is usually a protein involved. Relevant to this thesis, proteins control reactions that produce natural products with a wide variety of functions as described above. Therefore, it is possible to find evidence of the production of natural products in an organism's DNA.

1.3. Sniffing out natural products in hiding

While it is possible to directly measure which specialised metabolites are produced by a microbe with mass spectrometry and other metabolomics approaches, this does not always give a complete overview of the full specialised metabolic capacity of a microbe. Producing natural products is a laborious process for a cell: it demands energy, as well as nucleic acids for mRNA, amino acids for multiple (large) proteins, and often rare building blocks for assembling the specialised metabolite itself. As such, it is beneficial for the microbe to only produce a natural product when it is strictly necessary for the microbe's survival. Typically, a competition-devoid petri dish filled with ample nutrients does not drive microbes to produce specialised metabolites, whose ecological function is often tightly linked to environmental and developmental conditions. Also, some specialised metabolites are only produced in the presence of certain building blocks or produced at such low quantities that they are not detectable with current methods. As a result, many natural products 'in hiding' are missed when monitoring their production directly.

There are ways to pre-select potent natural product producing genera, or to 'motivate' microbes to activate their specialised metabolism by manipulating their environment in the lab^{37,38}. However, as the production of all specialised metabolites is ultimately hardcoded in the microbe's DNA, it can be more lucrative to infer natural product production by analysing DNA sequence. Especially now that the cost of DNA sequencing is at an all-time low, DNA analysis has become a staple in natural product discovery. The prediction, characterisation, and prioritisation of natural products from DNA sequence is the main topic of this thesis.

1.4. From DNA to bioactivity: a multifarious odyssey

There are many biochemical steps that lie between the DNA sequence underlying specialised metabolism and the biological function of the produced specialised metabolite. The DNA sequence is first transcribed to mRNA, which is then translated into protein. The mRNA sequence does not only determine the amino acid sequence of the biosynthetic machinery, but also the efficiency of translation and therefore the quantity of enzyme that is produced. This process relies heavily on how readily the mRNA folds into three-dimensional structures^{39–43}, which in turn depends on its underlying DNA and mRNA sequence.

While biosynthetic enzymes are produced as linear strings of amino acid building blocks, they fold into three-dimensional structures that ultimately dictate their role in natural product production. In the same way that mechanical arms in factories move semi-manufactured products from one machine to the next, there exist protein arms that move molecular intermediates between different enzymes in a biosynthetic pathway. Like their engineering equivalents, the dimensions of these arms and all other mechanical components of the assembly line must be precisely adjusted to correctly interact with each other and the raw materials and building blocks; a process carefully supervised by evolution. Thus, protein sequence governs protein structure, which in turn determines the natural product building blocks selected and the order in which they are assembled, therefore dictating the chemical structure and properties of the natural product. Finally, the chemical structure of the natural product and its concentration ultimately decide the molecules it can interact with, and thus its biological function.

Alongside classical wet-lab experiments, considerably cheaper bioinformatics methods are becoming increasingly important in natural products research. Researchers are developing computational pipelines and machine learning algorithms that take DNA sequences as input and predict if the encoded proteins produce a molecule with interesting properties³⁸. Such pipelines can give insight into important questions: is the produced compound structurally different from known natural products; is the metabolite dangerous for human consumption; does the molecule kill multi-resistant bacteria; and where in nature do we look to find the next treasure trove of potent natural products? While the results of such predictive algorithms are not yet reliable enough to completely replace experimental work, they can greatly aid in accelerating research by narrowing the search space to a manageable range of molecules. To build such bioinformatics tools, understanding the biochemical journey from DNA to bioactivity is key. And the first step to understanding natural product assembly in microbes is to discuss Biosynthetic Gene Clusters (BGCs).

1.5. Biosynthetic Gene Clusters: concise instruction manuals for biosynthetic LEGO

A biosynthetic Gene Cluster (BGC) is a stretch of DNA containing a set of genes that jointly encode a biosynthetic assembly line. Each gene corresponds to one protein: one machine along the conveyor belt that builds, extends, or alters the natural product using basic molecular building blocks and reactions. It is efficient for microbes to have all the genes encoding a biosynthetic pathway in the same place: this way, the entire production line can be activated by transcribing DNA from a single genomic locus.

There exist many different types of BGC, each encoding for the synthesis of different classes of natural products. Examples include BGCs for production of non-ribosomal peptides (NRPs),

polyketides, alkaloids, terpenes, ribosomally synthesised and post-translationally modified peptides (RiPPs). Polyketides are briefly discussed in chapter 3; their chemistry is summarised in the introduction of that chapter. Central in this thesis are NRPs: peptides that are assembled by modular mega-enzymes in an mRNA-independent fashion.

1.5.1. Who needs mRNA? A deep dive into the chemistry of non-ribosomal peptide synthetases

Like proteins, non-ribosomal peptides (NRPs) are largely built up from amino acid building blocks, but that is the only thing they have in common. Their size, composition, chemical structure, biological function, and biosynthesis are vastly different. NRPs are small, with a typical length of 3-15 amino acids, unlike proteins which usually require hundreds of amino acids for a functional scaffold. Also, they are synthesised without the use of an mRNA template. Rather, NRPs are assembled by non-ribosomal peptide synthetases (NRPSs): massive enzymes that are often organised as modular assembly lines, with each module attaching an amino acid building block to the NRP scaffold.

An NRPS module minimally consists of three domains: a condensation (C) domain, which functions as a biosynthetic glue gun, connecting the amino acid subunits by catalysing a condensation reaction; an adenylation (A) domain, a three-dimensional mould which selects which amino acid can be incorporated into the NRP based on how well it fits into its active site pocket⁴⁴; and a peptidyl carrier protein (PCP) domain, a mechanical arm and attachment point which holds the selected amino acid building block in place and moves them from the shape-sorting A domain to the C domain in the next module. As such, each module is responsible for linearly extending the NRP scaffold by exactly one amino acid building block. The size of the resulting NRP is therefore dependent on the number of modules, and the sequence of building blocks is determined by the order of the A domains in the NRPS^{45,46} (Figure 1.1).

1.5.2. NRPS collaboration: fruitful endeavour and logistic headache

Often, multiple NRPSs work together to create a single NRP. In these cases, the PCP domains have to carry partially synthesised scaffolds and amino acid building blocks between different enzymes. To make sure that the subunits are still connected in the right order, the NRPSs can make use of a couple of biochemical tricks. In bacteria, they can be transcribed from a single mRNA molecule^{27,47}; this way, the domains that need to interact with each other are already positioned closely together in space from the moment they are created. An alternative is to use docking domains: short, compatible protein domains at the end and start of two interacting NRPSs, which bring the enzymes together through a structure-governed secret handshake^{48–50}. Also, it is possible to split a module or even a single domain across two different NRPSs⁵¹, such that interaction is a prerequisite for NRP assembly. Additionally, it has been shown that some C domains only tolerate certain building blocks and scaffolds in their active sites, providing a ‘gatekeeping’ function in addition to their catalytic activity^{52–54}. In NRPS engineering efforts, where researchers try to design novel NRPSs by swapping out domains and modules and combining NRPSs from different origins, this property is more troublesome than it is useful, greatly limiting the number of combinations that are possible. However, this dual activity of the C domain may have a very important biological function in multi-enzyme NRPS systems: by starting an NRPS with a C domain that is only tolerant of the substrate of the preceding module (located in a different NRPS), the C domain effectively coordinates correct NRP assembly⁵⁵. Also, it may prevent unwanted cross-reactions between NRPS machinery of different BGCs, which would more often than not lead to a lot of resource investment into the production of non-functional or

even toxic NRPs. Finally, the architectures of the first and the last NRPS in the assembly line can be adapted such that they cannot function at any other position in the production line. The A domains of starter modules often recognise non-amino acid building blocks such as benzoic acid and anthranilic acid⁵⁶, or the first amino acid in the sequence is directly attached to a fatty acid through a specialised C starter domain⁵⁷. As C domains cannot catalyse the formation of an upstream peptide bond with these acids, they can only ever function as the first NRPS in the assembly line. Analogously, terminal domains usually contain thioesterase (TE)⁵⁸ or terminal reductase (TD) domains⁵⁹, biosynthetic ‘scissors’ that cut the covalent bond between the NRP and the PCP domain. This releases the NRP, making it unavailable for further extension by other modules, therefore forcing NRPSs containing such domains to the end of the assembly line.

1.5.3. The varied amino acid diet of non-ribosomal peptide synthetases

At first glance, it seems strange that nature has evolved such a complex and metabolically expensive method for linear peptide generation when a perfectly good, energy-efficient alternative exists in mRNA template-guided ribosomal peptide assembly. In fact, there exists a class of small bioactive peptides called Ribosomally synthesised and Post-translationally modified Peptides (RiPPs) whose scaffolds are ‘classically’ synthesised by mRNA-dependent ribosomal machinery⁶⁰. However, like proteins, RiPPs are limited to the 20 canonical amino acids that mRNA can encode. Because NRPSs use A domains for amino acid selection, which depend on a lock-and-key interaction between enzyme and substrate, they enjoy much more freedom in building block choice. So far, over 300 A domain substrates have been discovered, most of them amino acids⁶¹. As a result, NRPs have a much greater potential for chemical and functional diversity than their ribosomal equivalents. This is evident from non-canonical NRPs that are currently used for a wide variety of applications, such as the kynurenine-containing antibiotic daptomycin, which was the last antibiotic to enter the clinic²⁵; the herbicide Bialaphos⁶², which contains the non-canonical phosphinothricin; and the antifungal UK-2A³¹, which uses a 3'-hydroxy-4'-methoxypicolinic acid building block.

The genes encoding NRPSs usually form the core of NRP-producing biosynthetic gene clusters (BGCs), but there are often many other genes that are also involved in NRP biosynthesis. Such genes may encode proteins that synthesise non-canonical amino acid building blocks, or enzymes that chemically modify the NRP scaffold during or after scaffold assembly. Also, there exist hybrid NRPS BGCs, which combine NRPS chemistry with other biosynthetic machinery, such as polyketide synthases^{63–65}. This all further enhances the potential chemical and functional space of NRPs.

1.6. Leveraging bioinformatics to find molecules of interest

There are various possible approaches to computationally predict the structure and biological function of a natural product from DNA sequence, each using as a starting point one or more of the waypoints described in section 1.4 (Figure 1.2). A complete prediction pipeline includes three steps: BGC detection, prediction of the chemical structure of the natural product from BGC sequence, and bioactivity prediction from natural product structure. The algorithms employed vary widely, and include rule-based methods, classical machine learning approaches, and deep learning.

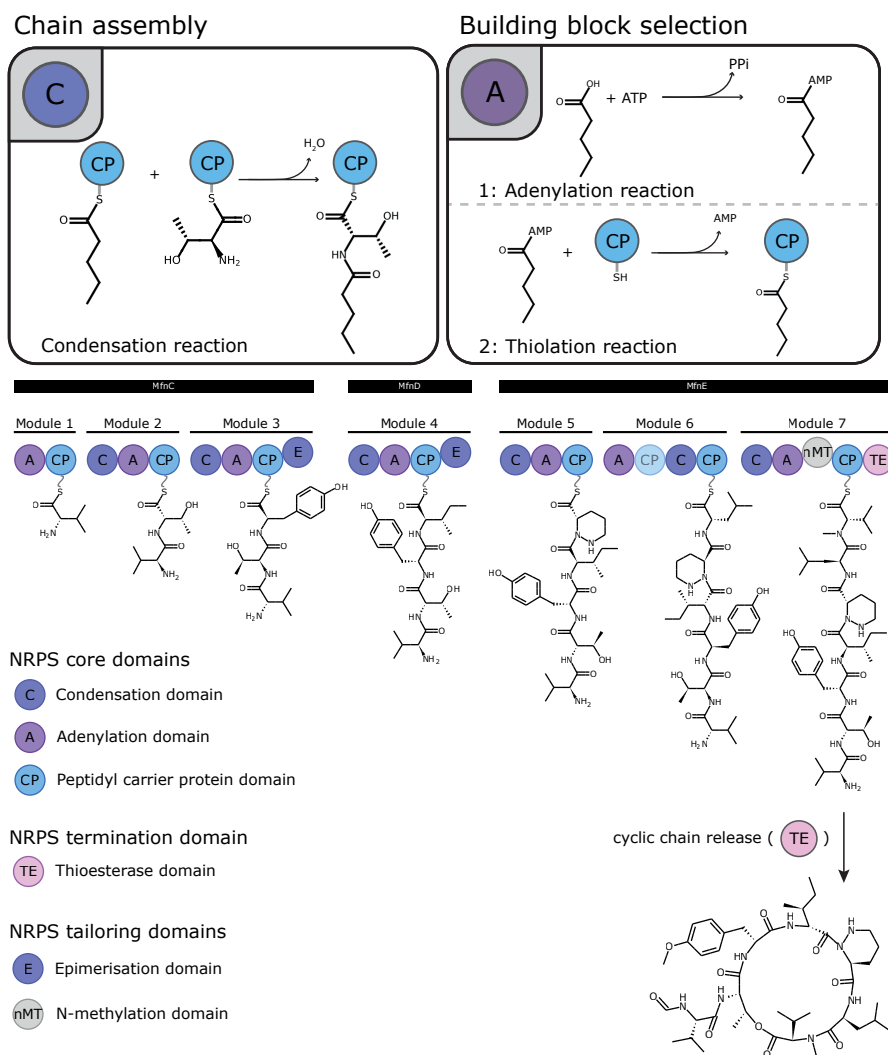


Figure 1.1. Example of a NRPS assembly line. Black boxes indicate NRPS enzymes, lines represent modules, and circles represent domains. Adenylation domains select the building blocks and condensation domains attach them end-to-end to assemble the natural product. Peptidyl carrier protein domains act as molecular scaffolds which building blocks and chain intermediates can be attached to between reactions. The thioesterase domain releases the natural product from the last NRPS enzyme, MfnE, and cyclises it. The example shown is the biosynthetic gene cluster for marformycin A⁶⁶.

1.6.1. BGC detection

Often, an analysis starts with the detection of BGCs in microbial genomes. The tool most widely used for this purpose is antiSMASH⁶⁷, which takes the DNA sequence of a bacterial genome as input and creates an interactive web page where the user can explore the different BGCs found in their bacterial genome or metagenome. For fungal and plant genomes, the tools fungiSMASH and

plantiSMASH respectively produce similar output with different underlying gene finding algorithms and/or detection rules^{68,69}. It is then possible to compare detected BGCs to BGCs with known products to get a quick overview of which known compounds might be produced by an organism. AntiSMASH already supports the option to do a 'KnownClusterBLAST', which automatically compares detected BGCs to BGCs of known function in the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database⁷⁰⁻⁷². There are also tools like BiG-SCAPE, which uses a clustering algorithm to group detected and known BGCs into families which are expected to possess similar biological functions, and CORASON, which can analyse these BGC families phylogenetically and highlight families that are expected to encode novel chemistry⁷³. For BGCs that are very similar to previously characterised BGCs, these resources make it possible to directly get an approximation of the molecular structure and function of the produced compound from BGC sequence. For more novel BGCs, structure and function prediction often require additional steps.

It is possible to directly predict bioactivity from BGC sequence. This was attempted by Walker and Clardy⁷⁴, who trained various classifiers on the MIBiG database and achieved average prediction accuracies of 70-80%. Alternatively, we can first predict the structure of the compound from BGC sequence, and then predict bioactivity from the proposed structure.

1.6.2. Structure prediction of non-ribosomal peptides

Structure prediction is highly dependent on the biosynthetic class that a BGC belongs to, as it is determined by very class-specific biosynthetic steps that lead to natural product formation. In NRPS clusters, biosynthetic enzymes belong broadly to three categories: precursor biosynthesis enzymes, which synthesise building blocks for the NRP; scaffold assembly enzymes, which are mainly the NRPS enzymes that build the main peptide scaffold of the NRP as discussed in section 1.5.1; and tailoring enzymes, which chemically modify the peptide scaffold during or after assembly to add molecular moieties, create covalent cross-links between different building blocks of the peptide, or remove atoms. In theory, if we can predict the catalysed reaction, reaction order, and site of action of each of these enzyme types, we should be able to determine the chemical structure of NRPs. Currently, NRP structure prediction efforts are mainly aimed at scaffold assembly enzymes, as building block precursors can usually be inferred from attributes of the scaffold assembly enzymes themselves⁴⁴, and the site of action of tailoring enzymes is notoriously difficult to predict.

1.6.2.1. A chronicle of A domain selectivity prediction

Most NRP structure prediction methods involve the prediction of building block selection by A domains from the protein sequence of NRPSs. In the late 1990s, two groups independently worked on identifying which regions of A domains were essential for building block recognition. Based on crystal structures of a phenylalanine-recognising A domain from the NRPS enzyme GrsA, they agreed on a core of nine amino acid residues in the binding pocket^{44,75}, with Stachelhaus *et al.* reporting an additional tenth⁴⁴. Importantly, they were able to correlate these 9-10 residues with the building block that could fit into the A domain binding pocket. Since, these ten residues have been commonly referred to as the 'Stachelhaus code'. Generally, it is accepted that if the Stachelhaus codes of two A domains are highly similar, then so are their building blocks. Therefore, it became possible to extract the Stachelhaus code from each A domain by aligning its sequence to the A domain from GrsA, and to predict its selectivity by comparing it to the Stachelhaus codes of A domains of known specificity. This rule-based approach is still used by many experts to quickly gauge what the structure of a NRP scaffold looks like.

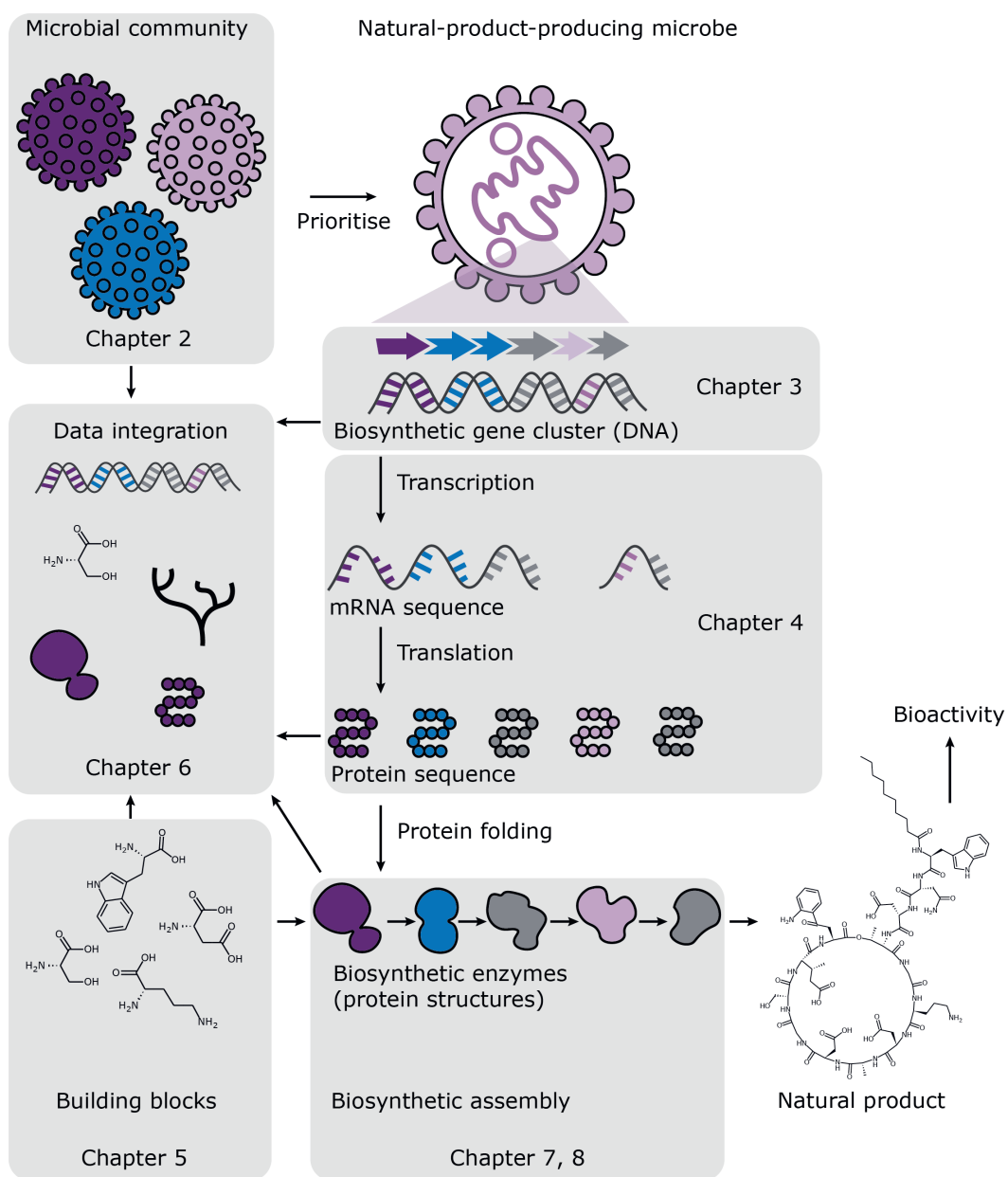


Figure 1.2. The journey from microbial community to natural product. Grey boxes indicate topics addressed by the chapters in this thesis.

In 2005, Rausch *et al.* proposed a different, more generalisable approach: they argued that interactions may happen with any residues located within 8 Angstrom (Å) of the active site⁷⁶. In the case of the GrsA A domain, this translates to 34 residues that may be involved in building block coordination. They used transductive support vector machines (tSVMs) to develop NRPSPredictor, a semi-supervised machine learning method, to automatically learn patterns in the 34 amino acid sequences from bacterial A domains. They also devised a method for featurising the amino acids as lists of numbers that represent their physicochemical properties like size and hydrophobicity, making it possible for the model to recognise the similarity between amino acids with comparable chemical properties, such as isoleucine and leucine. Röttig *et al.* later released an improved version of this tool, NRPSPredictor2⁷⁷, which was trained on more data and includes a predictor for fungal A domains as well. NRPSPredictor2 is still used in antiSMASH as one of the methods to predict A domain selectivity, alongside Stachelhaus code comparison.

There are other software packages that apply NRP structure prediction directly following BGC detection. The software package PRISM 4⁷⁸, which detects BGCs and subsequently predicts the structure of many different natural product classes, uses profile Hidden Markov Models (pHMMs) for this purpose. pHMMs can be thought of as flow charts that are built from a library of sequences to capture probabilistic patterns that are characteristic of those sequences. pHMMs can be used to query new sequences to determine if they likely contain these conserved patterns and belong to the same sequence group. In the example of A domain selectivity prediction, PRISM 4 employs a different pHMM for each A domain substrate and assigns the selectivity of the best matching pHMM to the queried A domain. A similar method was previously employed by Khayatt *et al.*, who used pHMMs for a dedicated A domain predictor⁷⁹.

Chevrette *et al.* aimed to leverage all these different methods by building the ensemble algorithm SANDPUMA, which feeds the output from six different prediction methods into a decision tree to generate the best consensus selectivity⁸⁰. These prediction methods include percent identity to the best match in the training set, Stachelhaus code similarity, the SVM method used by NRPSPredictor2, the pHMMs used by Khayatt *et al.*, and two phylogeny-based prediction methods (prediCAT) created by the authors. They also expanded their training set compared to previous methods and achieved the best accuracy.

A domains play a role in NRPS BGCs beyond recognising amino acid building blocks. Many recognise fatty acid tails and other molecular moieties that are fused to NRP scaffolds during or after scaffold assembly. An example are the lipid tails of lipopeptides, as in the NRP antibiotics daptomycin and CDA^{25,81}. These lipids are often selected by the A domains of Co-enzyme A ligases and fatty acid AMP ligases. Robinson *et al.* developed a random forest classifier that also aims to predict the substrates of these non-NRPS A domains to improve the structure prediction of compounds that incorporate such moieties⁸².

Despite this plethora of A domain selectivity prediction methods available, each suffers from shortcomings. Stachelhaus matching assumes that if the 9 binding pocket residues are identical, then so are the selected substrates, something that has been proven incorrect by various examples of A domains with identical Stachelhaus codes and different selectivities (examples listed in Chapter 5). Also, many A domains recognise multiple substrates. The few methods that even use such A domains in their training set group them under a unique label rather than using a multi-label classification system^{77,80}, which makes it impossible for an algorithm to base its prediction task on underlying biology. Furthermore, in training various methods the datasets were not properly balanced and/or

stratified into training and test sets^{77,80}. Consequently, the algorithms often seem to work much better than they actually do, performing very well on overrepresented substrates such as serine or threonine, but performing poorly on underrepresented classes. In addition, some methods, while (seemingly) highly accurate, are too slow to perform well in the context of mass analyses. For example, SANDPUMA was added to⁸³ and subsequently removed from⁸⁴ the widely accessed tool antiSMASH, as each A domain selectivity prediction took over 2 minutes to compute. Also, the best prediction method only uses a dataset of around 1,000 datapoints⁸⁰, which is small for machine learning. Finally, none of the tools directly address the biological basis for substrate selectivity: the 3D shape of the active site pocket.

1.6.2.2. From building blocks to compound: simulating chemical reactions

While predicting the selectivity of NRPS A domains may be the most important step in estimating the structure of the produced NRP, many additional computational operations are needed to go from predicting raw building blocks to inferring the final product. First, we need to express our building blocks as a machine-readable string of characters, such as SMILES strings⁸⁵. Then we require a software package that can interpret those building block representations and perform operations on them, such as connecting them in the right order. While there are currently no tools yet to accurately predict the site of action of tailoring reactions such as cyclisations, the field progresses quickly, and as such it makes sense to utilise computational chemistry toolkits that can simulate such chemical reactions. Finally, it is important to automatically visualise predicted molecules and reaction pathways, such that the human eye can interpret the predicted chemistry too. A software package widely used for such operations is RDKit⁸⁶, a cheminformatics toolkit which boasts an enormous library of functionalities that can read, visualise, and manipulate molecular representations.

We mentioned antiSMASH before in the context of BGC detection, but it is also important to address it in the context of structure prediction. AntiSMASH occupies a dominant role in natural product research, with an incredibly large user base of over 400,000 unique users having submitted around 1.5 million jobs since its inception. As such, if we could expand the BGC detection capabilities of antiSMASH to include good structure prediction and rendering, this would likely greatly accelerate natural product research and compound discovery worldwide. To facilitate this, antiSMASH requires a solid chemistry toolkit at its base, preferably implemented in Python, the programming language that antiSMASH was written in. It was attempted to add chemistry to antiSMASH using RDKit before; however, due to its C++ backend and large number of dependencies, antiSMASH ran into compatibility issues with other software packages that it relies on. Therefore, antiSMASH requires a Python-based informatics kit with a small number of dependencies.

1.6.3. Predicting bioactivity: conquering the central dogma of specialised metabolism

The central dogma of molecular biology is a well-known concept that describes the flow of information in all living organisms: DNA is transcribed into mRNA, which is translated into proteins. Similarly, we can define a central dogma of specialised metabolism: BGC sequence dictates natural product structure, which in turn determines its bioactivity. Therefore, as the last step in a prediction pipeline, we can attempt to predict a compound's bioactivity from its (predicted) molecular structure.

There are many aspects of bioactivity that are interesting to predict: what is a compound's ecological function and how does that ecological function translate to societally relevant applications; which molecular targets does a compound interact with and what is its mechanism of action; how soluble

is the compound; and at what quantities does a compound become toxic. Such complicated biological questions warrant high-quality, standardised databases, which brings me to a major challenge in the field: there are not many good ontologies for bioactivities and mechanisms of action. Many are reported as non-standardised textual descriptions, and databases that do enforce standardised vocabularies are often not cross-compatible, making it difficult to integrate data from different sources into single datasets that can be used for machine learning. As an added challenge, bioactivity of a compound is directly tied to its three-dimensional structure, while most bioactivity databases such as ChEMBL⁸⁷ only store compound structures as 1D SMILES or InChI strings, which can be converted to 2D graphical representations at best.

1.7. A three-dimensional journey through the world of natural products

In this thesis, I invite you to join me on a computational adventure that takes us from raw microbial DNA sequence to bioactive compounds that could be leveraged to better the world. We will start our journey in the world of natural products. There, a particularly talented bacterial phylum of natural product producers is waiting for us in chapter 2, where we review state-of-the-art experimental and computational methods to discover natural products with desirable bioactivities. In chapter 3, our path will take us to the largest library of BGC sequence information in the world, MIBiG, which is visited by hundreds of people every year to find data that will help them in their scientific endeavours, or to add their own. As a part of the annotation marathon for the third release of this database, we describe how we mobilised a community of natural product researchers to annotate the substrate selectivity of 2,000 additional A domains. While in the library quarter, we will make a brief detour to a transcription specialist called MEW in chapter 4. This explainable machine learning algorithm will teach us about the impact of mRNA structure on protein expression. Next, I implore you to don a pair of 3D glasses as we arrive in chapter 5, where we dive deep into the heart of adenylation domain structures for a three-dimensional tour. We use what we learn there to build two A domain selectivity predictors, PARAS and PARASECT, two structure-informed tools which aim to address the shortcomings of current methods as discussed in section 1.6.2.1. With our pockets full of sequence, structure, and substrate data, we will continue to chapter 6, where we seek out the data organiser Turterra, who collates enzyme and domain data into comprehensive interactive websites. Turterra's automatically generated web portals are designed to intuitively analyse multifaceted datasets for any enzyme or domain family. We applied Turterra to two natural product enzyme/domain families, including NRPS A domains. Then, we will cross over into the realm of cheminformatics in chapters 7 and 8, where we explore the potential of raw Python code for computational chemistry and automatic visualisation of natural product chemistry. In chapter 7, we discuss the development of the light-weight software package PIKACHU, a Python-based Informatics Kit for Analysing Chemical Units, which, with only a single dependency, is suitable for integration into antiSMASH. In chapter 8 we describe how we leveraged PIKACHU to build RAICHU, an antiSMASH-compatible software suite that can automatically create publication-quality visualisations of NRP and polyketide assembly lines from antiSMASH results. Finally, we conclude our scientific trip in chapter 10, where we discuss the impressions and insights we have gathered along the way and come to the conclusion that we stand only at the foothills of the vast mountain range that encompasses the field of natural product prediction.

Chapter 2

Ecology and genomics of Actinobacteria: new concepts for natural product discovery

Doris A. van Bergeijk*, Barbara R. Terlouw*, Marnix H. Medema, and Gilles P. van Wezel

* These authors contributed equally to this work

This chapter has been published as

van Bergeijk, D.A., Terlouw, B.R., Medema, M.H., van Wezel, G.P., Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nature Reviews Microbiology* 18, 546–558 (2020). <https://doi.org/10.1038/s41579-020-0379-y>

Abstract

Actinobacteria constitute a highly diverse bacterial phylum with an unrivalled metabolic versatility. They produce most of the clinically used antibiotics and a plethora of other natural products with medical or agricultural applications. Modern ‘omics’-based technologies have revealed that the genomic potential of Actinobacteria greatly outmatches the known chemical space. In this Review, we argue that combining insights into actinobacterial ecology with state-of-the-art computational approaches holds great promise to unlock this unexplored reservoir of actinobacterial metabolism. This enables the identification of small molecules and other stimuli that elicit the induction of poorly expressed biosynthetic gene clusters, which should help reinvigorate screening efforts for their precious bioactive natural products.

2.1. Introduction

The phylum Actinobacteria represents one of the most diverse groups of microorganisms in nature. These Gram-positive bacteria have a high GC content and show a remarkable range of morphologies, including unicellular cocci or rods (for example, members of the genera *Micrococcus* and *Mycobacterium*), and morphologically complex multicellular bacteria (for example, members of the genera *Ammycolatopsis*, *Frankia* and *Streptomyces*)⁸⁸. Actinobacteria are widely distributed across both terrestrial and aquatic ecosystems, as well as in the microbiomes of higher eukaryotes⁸⁹. This ecological diversity is reflected in their metabolic potential as Actinobacteria are extremely versatile producers of bioactive natural products. Notably, Actinobacteria produce two thirds of all known antibiotics used in the clinic today, but also a vast array of anticancer compounds, immunosuppressants, anthelmintics, herbicides and antiviral compounds, in addition to extracellular enzymes^{90–93}. Therefore, these bacteria are attractive sources for clinical drugs^{91,94}.

The introduction of antibiotics in the twentieth century greatly contributed to the extension of the human lifespan and has saved millions of lives worldwide. However, with the increase in antibiotic resistance, we now face a huge challenge in treating infections by multidrug-resistant bacteria⁹⁵. This coincides with a dramatic decrease in the success of traditional drug development through high-throughput screening^{96,97}. Indeed, the chance of finding new antibiotics via traditional methods in randomly chosen Actinobacteria has been estimated at less than one per million^{98,99}. However, advances in genome sequencing have unveiled a vast reservoir of biosynthetic gene clusters (BGCs) for natural products in microbial genomes, even in those that had been studied extensively for decades^{100–102}. The OSMAC strategy (one strain many compounds)¹⁰³ for extensive mining of individual strains still yields promising new molecules^{104,105}. However, the rate of success has decreased dramatically since the golden years of drug discovery, primarily due to replication⁹⁹. The apparent failure to uncover the full potential of natural product-producing microorganisms is likely due to the fact that we lack the understanding that is required to activate the expression of their BGCs in the laboratory. During industrial screening, bacteria and fungi are typically grown in isolation with ample nutrients and resources, which is in sharp contrast to their natural complex and rapidly changing habitat. Antibiotics are believed to be important for survival as mediators of resource competition in a competitive environment^{106,107}, and microbial interactions have a key role in their activation^{108–110}. Hence, to increase the success of natural product-based drug discovery, we need to elucidate the triggers and cues that activate the expression of BGCs^{37,111}.

Bioactive metabolites mediate important ecological functions, which are as diverse as their chemical structures. Siderophores enhance iron uptake in environments where the bioavailability of iron is limited⁸, pigments provide protection against ultraviolet radiation and have antioxidant activity¹¹² and compatible solutes protect against osmotic stress¹¹. However, the most obvious ecological purpose is biological weaponry to outcompete other organisms for resource acquisition^{106,108}. As producers of various bioactive molecules, Actinobacteria are attractive to eukaryotic hosts as symbionts. For instance, *Streptomyces* spp. that live in the antennal glands of beewolf digger wasps produce antibiotics that protect the wasp larvae from various pathogenic fungi and bacteria¹¹³, and endophytic Actinobacteria protect their host against phytopathogens^{89,114}.

In this Review, we discuss why Actinobacteria excel as natural product producers and how their specialized metabolism and the regulatory mechanisms governing this metabolism have evolved in the context of ecology and genomic structure. Finally, we explore how ecological insights can be translated into approaches for computational and experimental genome mining strategies that yield novel bioactive molecules, in particular antibiotics.

2.2. A mycelial lifestyle

The propensity to produce bioactive molecules and the richness of bacterial genomes in terms of BGC diversity have been correlated with key organismal features, such as multicellularity, endospore formation and genome size^{115,116}. Bacteria can be divided broadly into two groups based on their adaptability: specialists and generalists, each with their own environmental niches. Specialists are dedicated to life in specific environments and therefore require less extensive metabolism and, accordingly, smaller genomes^{117,118}. *Mycoplasma genitalium* is a well-known example of an organism with a small genome of around 580 kb and fewer than 500 genes¹¹⁹. The smallest genomes known to date belong to parasites and symbionts, with the beetle symbiont *Stammera* spp. as a remarkable example of an organism with a small genome of around 270 kb and 250 genes¹²⁰.

By contrast, generalists usually have larger genomes and a complex morphology, such as multicellularity and the formation of endospores^{117,118}. Their ability to use multiple nutrient sources enables them to adapt to diverse environments and growth conditions, which requires complex metabolic regulation. Therefore, it is not surprising that this group of bacteria includes the most important producers of natural products such as Actinobacteria, Cyanobacteria and Myxobacteria^{118,121}. Hallmark features of multicellularity and development include intraspecies communication, morphological differentiation and programmed cell death (PCD)¹²².

The linkage between morphological and chemical differentiation is best explained using *Streptomyces* as an example. Streptomycetes are mycelial organisms that reproduce by sporulation, with a life cycle similar to that of filamentous fungi. When the conditions are favourable, a single and uninucleoid spore will germinate and the hyphae grow out by a combination of apical growth and branching, which results in a complex mycelial network^{90,123}. Exo-enzymes are released to break down natural polymers, such as cellulose, mannan and chitin, thereby providing nutrients. The vegetative hyphae are compartmentalized by occasional semi-permeable cross-walls to form large multinucleoid cells. The next step in the developmental programme is the formation of new sporogenic aerial hyphae, which eventually differentiate into chains of uninucleoid spores^{90,123}. To fuel the onset of the developmental programme, old vegetative mycelia are autolytically degraded through PCD to liberate the necessary nutrients for the new biomass¹²⁴ (Figure 2.1). Eventually,

reproductive aerial hyphae differentiate into long chains of spores. The onset of development is controlled by the *bld* (bald) genes, so-called because mutants fail to produce the fluffy aerial mycelium¹²⁵. Genes that control the distinct steps leading towards the maturation of aerial hyphae and subsequent sporulation are called *whi* (white) genes, referring to the white appearance of mutants due to their failure to produce grey-pigmented spores¹²⁶.

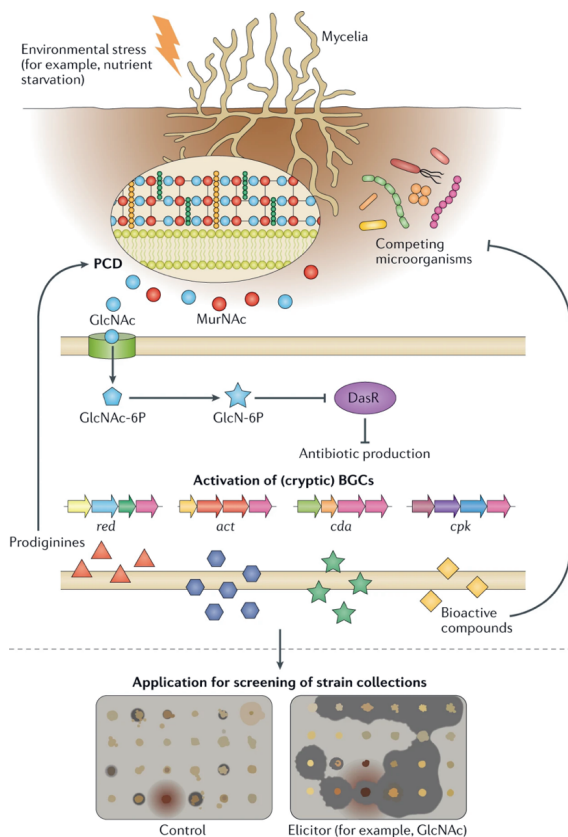


Figure 2.1. From biological understanding to elicitation of BGCs. Upon environmental stress, such as nutrient starvation, programmed cell death (PCD) leads to the autolytic degradation of the old mycelia, which liberates the necessary nutrients to fuel the onset of development and the new biomass. The onset of morphological development correlates temporally with the production of antibiotics. Specifically, cell-wall peptidoglycan is recycled to release the amino sugars N-acetylmuramic acid (MurNAc) and N-acetylglucosamine (GlcNAc). GlcNAc is internalized as GlcNAc-6P and converted to GlcN-6P, which functions as a ligand for the global regulator DasR. Binding of GlcN-6P inactivates the repressor DasR, which results in the derepression of pathway-specific activators of antibiotic biosynthetic gene clusters (BGCs)¹²⁷. The resulting production of antibiotics (such as calcium-dependent antibiotic (Cda), coelimycin P1 (Cpk) and actinorhodin (Act)) is likely to provide a line of defence to protect the nutrients that have been released by PCD against motile saprophytic bacteria, whereas the synthesis of DNA-degrading prodiginines (undecylprodigiosin (Red)) promotes PCD. GlcNAc elicits the transcription of BGCs that are not expressed under standard laboratory conditions and can thus be used in screening regimes for drug discovery.

The onset of morphological development correlates temporally with the production of antibiotics^{128,129}. The production of antibiotics is likely to provide a line of defence to protect the PCD-released nutrients against motile saprophytic bacteria, while at the same time, antibiotic-mediated lysis of these competitors may serve as an alternative food source^{90,122,130,131}. Interestingly, prodiginines — which have DNA-damaging properties and are not secreted, and hence damage the DNA of the producer — apparently facilitate PCD in *Streptomyces coelicolor*¹³². This suggests that some antibiotics may even drive development via the initiation of PCD.

2.3. Missing signals and cryptic antibiotics

The genome sequences of extant Actinobacteria ‘document’ their present and (recent) past ecology. They specify various functionalities — including the natural products they produce — used to cope with challenges that these bacteria face in their specific niches. Not surprisingly, genome sequencing has had a huge impact on natural product discovery. This is exemplified by the specialized metabolism of *S. coelicolor* A3(2), the model organism for antibiotic biosynthesis. For a long time, this organism was known to produce four antibiotics, namely actinorhodin (Act), calcium-dependent antibiotic (Cda), undecylprodigiosin (Red) and the plasmid-encoded methylenomycin (Mmy). When the genome of *S. coelicolor* was sequenced, many more BGCs were uncovered than originally anticipated, and the same was true for other model streptomycetes^{100,101,133}. Renewed efforts for drug discovery led to the surprising discovery of yet a fifth antibiotic in *S. coelicolor*, called coelimycin P1¹²⁷, in a model organism that had been studied by thousands of scientists in hundreds of laboratories around the world. Moreover, a novel branch of the biosynthetic pathway of the model polyketide actinorhodin was shown to be activated by co-cultivation with the fungus *Aspergillus niger*¹³⁴, and mass spectral imaging revealed substantial changes in the secreted metabolome during interaction of *S. coelicolor* with five other Actinobacteria¹¹⁰. These examples illustrate the concept of cryptic biosynthetic pathways, namely those that have been identified in the genome but the cognate natural products of which are not synthesized under laboratory conditions (Box 1). This concept has revolutionized the field of antibiotic research¹³⁵, leading to the current era of genomics-based drug discovery^{90,91}.

Nowadays, next-generation genome sequencing enables scientists to explore large numbers of micro-organisms in search of novel BGCs. However, there are several challenges. A first challenge lies in prioritizing (cryptic) BGCs in terms of their potential for chemical novelty and/or clinical relevance to optimally exploit the wealth of available biological, genomic and metabolomics data. Second, the bulk of the microorganisms in soil and marine environments resist cultivation under laboratory conditions and thus represent a huge ‘white space’ of biochemical diversity^{136,137}. Last, many BGCs are not expressed during laboratory cultivation. To unlock this potential, we need to better understand the ecological context in which Actinobacteria live, as this will provide clues to the mechanisms that activate the biosynthetic pathways of natural products. To leverage genome information to this end, it is of major importance to understand how such ecological forces shape actinobacterial genomes and how biosynthetic diversity evolves.

Box 2.1. Silent biosynthetic gene clusters

What we have learned from genome sequencing efforts it is that many of the biosynthetic gene clusters (BGCs) are not accounted for in the corresponding metabolomes. These BGCs are referred to as 'silent' or 'cryptic'. Although these terms are sometimes interchanged, they have different meanings. A BGC is 'cryptic' when it has been identified but cannot yet be linked to a product, activity, or phenotype. To establish whether a cluster is 'silent', experimental validation is required, at the level of gene expression or metabolomics (in case the compound is known). Silence of a BGC may reflect the fact that we do not yet understand the environmental conditions that are required for their expression. Alternatively, a BGC may have been silenced as a first step towards loss of the complete cluster. A transcriptome study of *Salinispora* strains revealed that more than half of their BGCs were expressed at levels that should facilitate discovery of the compounds they produce¹³⁸. As transcriptome data of many BGCs are lacking, this suggests that most of the unexplored BGCs may be cryptic — that is, not linked to their products — rather than silent. For these BGCs, our failure to detect the corresponding metabolites is linked to other factors, such as translation efficiency and extraction methods.

Although the importance of 'omics'-based natural product discovery is clear¹³⁹, a proportion of BGCs remain transcriptionally silent in the laboratory. Transcriptome analysis revealed that biosynthetic potential is more complicated than the presence or absence of a BGC alone, as a substantial proportion of the BGCs were differentially expressed between strains¹³⁸. An argument often heard during discussions between scientists in the field is that if a few different isolates of the same species are analysed, only one of these may express the BGC in question. True transcriptional inactivity of BGCs may be due to a mutation in a structural or regulatory gene, although the former is less logical as the entire biosynthetic machinery would be produced in vain. An intriguing concept is that regulatory genes may sustain a single (frameshift) mutation, rendering the gene inactive, which may be easily restored by a compensatory mutation as a strategy for bet-hedging in a community, as seen for isolates of *Streptomyces lunaelactis*¹⁴⁰. Such 'light-switch silencing' may be a primary mechanism by which strains are able to maintain large numbers of BGCs, as natural product biosynthesis is an energy-intensive process. Still, it is reasonable to assume for many of the silent BGCs that the right conditions for their transcriptional activation are lacking when the strain is grown in the laboratory.

2.4. Evolution of biosynthetic repertoires**2.4.1. Vertical inheritance and horizontal gene transfer**

The diversity of the environments that Actinobacteria inhabit is extraordinary^{89,141}. Specialized metabolites have been isolated from marine organisms such as *Salinispora* spp.¹⁴², arctic *Streptomyces nitrosporeus*¹⁴³, desert-dwelling *Streptomyces* spp.¹⁴⁴ and species that live in the microbiomes of animals, insects and plants^{4,114,144,145}. This ecological diversity has major implications for the evolution of secondary metabolic repertoires, which are driven by a combination of vertical inheritance and horizontal gene transfer (HGT). In terms of vertical inheritance, *Salinispora* spp. show a strong relationship between phylogeny and BGC diversity, with close relatives sharing nearly all BGCs¹⁴⁶. Similarly, in *Amycolatopsis*, BGCs are conserved within clades but not between them¹⁴⁷. These BGCs may be more important for adaptation than for survival. Similarly, rapid demographic expansion of *Streptomyces* into previously uninhabited niches has led to differentiation into various species clusters, distinguishable from each other by ancestral homologous recombination events¹⁴⁸.

HGT is another important evolutionary driver of chemical complexity of natural products in *Streptomyces*^{146,149}. In *Salinispora*, it has been estimated that up to 96% of the biosynthetic pathways may have been acquired through HGT¹⁴⁶. It is unclear how frequent HGT occurs in Actinobacteria. Some argue that lateral acquisition and subsequent maintenance of complete BGCs is very rare¹⁵⁰. Nevertheless, it is likely that HGT has a key role in shaping BGC repertoires. Many more ancient HGT events of complete BGCs will have occurred than those that can still be reliably inferred, because those BGCs will have diverged over time since the moment of transfer¹⁵¹. The large discordance between gene phylogenies of core biosynthetic genes and species phylogeny testifies to this^{73,152,153}. When HGT leads to acquisition of BGCs with similar functions, there may be strong evolutionary pressure for BGC loss. Indeed, in the evolutionary history of the genus *Salinispora*, genes for the biosynthesis of the siderophore desferrioxamine were lost in three strains independently as a direct consequence of HGT of a functionally similar BGC. Such events may happen at a large scale and therefore could mask a large proportion of historical HGT events¹⁵⁴. In fact, rates of HGT may vary strongly between different types of BGCs, depending on their ecological roles; many species have a conserved 'core' set of BGCs, together with a strongly varying set of 'accessory' BGCs that is acquired or exchanged through HGT^{147,155}. This is evident in *Amycolatopsis*, in which BGCs that were acquired through HGT largely localize to non-conserved genomic regions¹⁴⁷.

2.4.2. Chromosome structure

Genomic structure is intimately tied with genetic change and conservation in specialized metabolism, possibly both as cause and effect. Actinobacteria have either linear or circular chromosomes. In linear chromosomes, as seen in *Streptomyces* and *Rhodococcus*, strain-specific genes and BGCs that are incorporated through HGT tend to be localized in the unstable subtelomeric end regions of the chromosome^{149,156–158}. Selective pressure might induce migration of such BGCs towards the chromosomal core, where they are likely more stably maintained through vertical inheritance and thus display much higher levels of conservation across different actinobacterial species^{156–158}. A recent heterologous expression study also showed that the chromosomal location of BGCs has an effect on expression levels, with expression levels of a β -glucuronidase reporter gene measured highest in the central regions of the chromosomal arms¹⁵⁹. In *Streptomyces*, more centralized BGCs tend to encode molecules such as ectoines and siderophores (Figure 2.2), which are likely to be essential for survival of the genus. In the circular chromosomes of *Salinispora* and *Amycolatopsis*, species-specific BGCs are largely located on genomic islands relatively distant from the origin of replication¹⁶⁰. Conversely, conserved BGCs tend to be localized in the 'core genome'. As in *Streptomyces*, in *Amycolatopsis* these BGCs specify metabolites, such as ectoines, siderophores and terpenes, whereas strain-specific BGCs, which make up as much as 67% of all BGCs in *Amycolatopsis*, localize largely away from the core genome to the genomic islands¹⁴⁷. Although, to our knowledge, no studies have investigated BGC migration from peripheral regions to core regions, BGC migration between genomic islands has been inferred to be likely based on phylogenomic analyses, which supports the hypothesis that BGCs preferentially migrate towards genomic core regions¹⁶⁰.

One of the reasons why actinobacterial genomes are packed with BGCs is because there are well-established mechanisms in place to acquire and exchange BGCs. One common method of acquisition is through integrative and conjugative elements (ICEs). These ICEs are plasmids with the ability to integrate themselves into the chromosome. The prevalence of these ICEs seems to be partially dictated by their ecological background: Actinobacteria originating from soil, plants or aquatic

environments contain a greater number of ICEs than species from other environments¹⁶¹. In addition to ICEs, giant linear plasmids have a high density of BGCs for antibiotics, which suggests an important role in the acquisition of bioactive metabolites throughout evolution^{162–164}.

2.5. Control of antibiotic production

Just as ecological forces shape the organization of BGCs across actinobacterial chromosomes, they also shape how the clusters are regulated. A range of external signals influence production either directly, through activation of pathway-specific activators, or indirectly, via an interactive network of pleiotropic regulators and intracellular signalling molecules^{129,165,166}. Additionally, cross-regulation can exist between different BGCs, as was recently shown for the two chemically unrelated specialized metabolites antimycin and candicidin produced by *Streptomyces albus* S4. Their BGCs are separated by 9 kb on the chromosome, but the pathway-specific regulator of candicidin, FscR1, can bind directly upstream of the genes encoding antimycin biosynthesis and is essential for the activation of this cluster¹⁶⁷. This regulatory complexity is highlighted by the strong emphasis on regulation in the genome of the model organism *S. coelicolor*, which encodes close to 800 regulatory proteins, representing >10% of the total proteome¹⁰⁰. Higher-level control is likely to be tied to the ecological conditions in which these adaptive responses have evolved. Environmental and physiological signals have been integrated into the regulation of specialized metabolism to ensure that these costly molecules are only produced when required^{89,165}. The involvement of environmental signals in the control and complexity of specialized metabolism is illustrated by the control of the BGC for ferroverdins and bagremycins in *Streptomyces lunaelactis*¹⁴⁰. The metabolite produced from the BGC depends on iron availability. In iron-limiting conditions, the antimicrobial bagremycins are produced. When there is an excess of iron, the amino group of bagremycin is replaced by a nitroso group to generate ferroverdin, a siderophore that is used as an anticholesterol drug and that, in nature, likely functions to limit iron-mediated oxidative damage¹⁴⁰.

In a laboratory setting, the absence of the triggers and cues that would activate antibiotic production in the original habitat offers a possible explanation for why so many BGCs remain poorly expressed or silent under laboratory growth conditions (Box 2.1). However, the identity of environmental ligands and/or signals perceived by both pleiotropic and pathway-specific regulatory proteins is a major area of investigation and if resolved could lead to the activation of silent BGCs and thus drug discovery¹²⁹. The key lies in understanding the biology of the producing bacteria and translating these insights into solutions to activate antibiotic production (see below).

Here, we provide some background on the complex transcriptional control of BGCs and then look into the regulatory networks that control the well-established connection between the onset of development and antibiotic production. For detailed reviews, we refer the reader elsewhere^{128,129,168}.

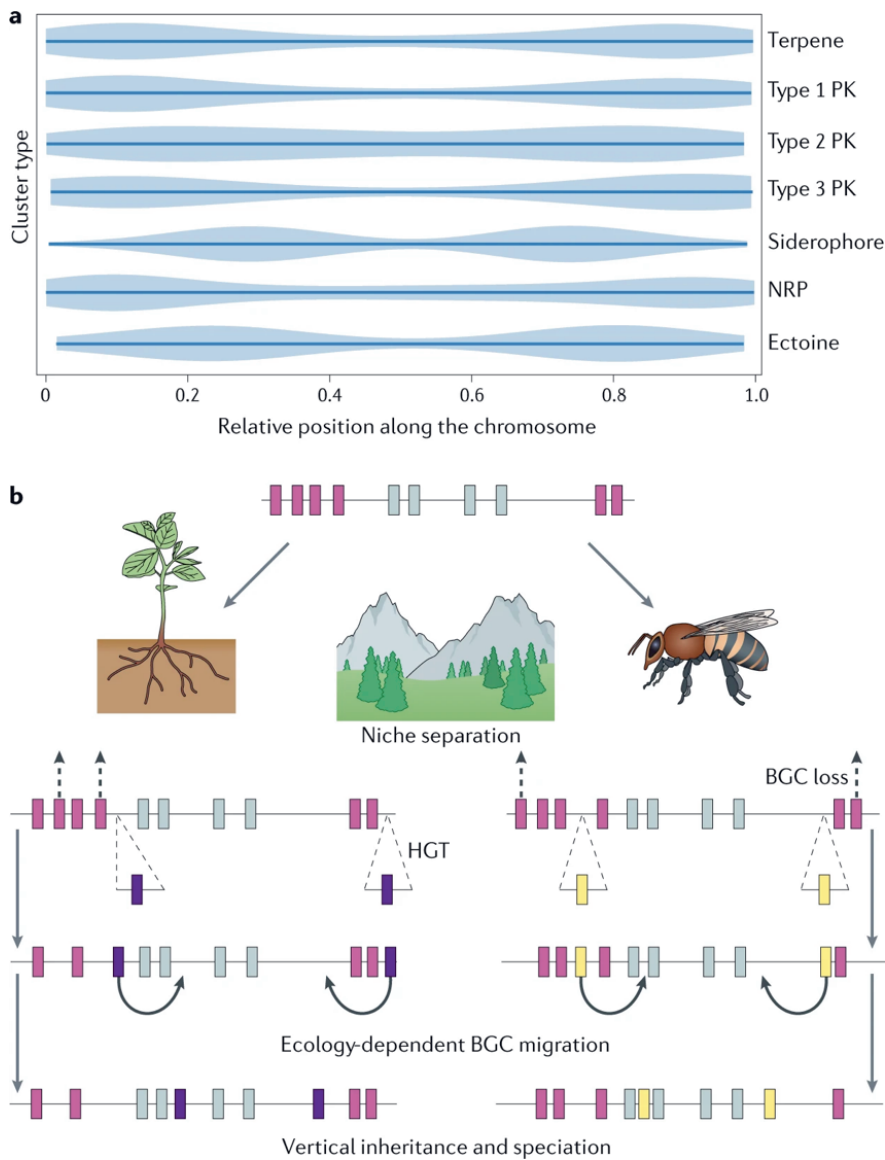


Figure 2.2. BGC distribution and evolution in *Streptomyces* chromosomes. A. Distribution of biosynthetic gene clusters (BGCs) encoding terpenes, non-ribosomal peptides (NRPs), polyketides (PKs), siderophores and ectoines in 100 linear *Streptomyces* chromosomes. Clusters encoding type 1 PKs, type 3 PKs and NRPs localize to the chromosome ends, whereas clusters encoding type 2 PKs, terpenes, siderophores and ectoines localize more towards the chromosome core. B. Proposed movement of BGCs throughout evolution. Upon niche separation, genomes in different environments take up different genetic material through horizontal gene transfer (HGT) and/or BGCs are lost. In the long term, BGCs that are important for survival in that niche migrate towards the chromosome core and have a greater chance of being maintained through vertical inheritance.

2.5.1. Principles of the control of antibiotic production.

To coordinate the metabolic responses to specific ecological challenges, Actinobacteria have evolved complex multilevel regulatory networks. These networks are composed of multilevel transcriptional and translational control. This control is required for the correct interpretation of the signals that reach the colony and to translate them into appropriate responses. Much of our knowledge of the control of antibiotic production has been obtained from the study of the BGCs for Act, Cda and Red in *S. coelicolor* and for streptomycin in *Streptomyces griseus*^{128,129,168}. These clusters are controlled by the pathway-specific activators ActII-ORF4, CdaR and RedD, which belong to the SARP family of *Streptomyces* antibiotic regulatory proteins¹⁶⁹, and StrR (ParB-Spo0J family¹⁷⁰), respectively. These cluster-situated regulators directly control the level of transcription of the BGC, which in turn dictates the production level of the cognate natural product^{171,172}. Interestingly, *actII-ORF4*, *cdaR* and *redD* are all subject to translational control by the tRNA that recognizes the rare UUA codon for leucine. This tRNA, encoded by the *bldA* gene, is also required for the proper translation of many developmental genes and thus links morphological to chemical differentiation^{173,174}.

Multiple cluster-situated regulators may control a single BGC, and in addition the BGCs are subject to global control. This enables the cell to coordinate specialized metabolism with growth and development, the balance in carbon, nitrogen and phosphorus metabolism, and other major cellular pathways, thereby generating a complex hierarchy of regulatory networks. To enable efficient responses to external stimuli, the activity of many regulatory proteins is determined by small molecules. This includes the hormone-like γ -butyrolactones^{175,176}, feedback through biosynthetic intermediates^{177–179} and sugar-based ligands^{180,181}. The identification of such external signals, often referred to as elicitors, is of key importance for the rational activation (elicitation) of natural product biosynthesis and thus for the revitalization of drug discovery.

2.5.2. PCD and the DasR regulatory network.

As described above, PCD mediates the provision of nutrients at the onset of morphological and chemical differentiation. The signalling pathway from PCD to differentiation revolves around the global nutrient sensory GntR-family regulator DasR (Figure 2.1). Autolytic degradation of cell-wall peptidoglycan releases the amino sugars *N*-acetylglucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc) around the colonies. Under nutrient-limiting conditions (famine), the accumulation of GlcNAc around colonies triggers development and antibiotic production, whereas under rich growth (feast) conditions, GlcNAc blocks both processes¹⁸¹. The rationale behind this is that under feast conditions GlcNAc is seen as derived from chitin and signals nutrient abundance and promotes growth, whereas under famine conditions it signals hydrolysis of the bacterial cell wall and thus the need for development. GlcNAc-derived glucosamine-6-phosphate (GlcN-6P) and other phospho-sugars act as ligands for DasR, and thereby inactivate the repressor, which results in the derepression of BGCs¹⁸¹. DasR directly controls the pathway-specific activators of BGCs for all antibiotics and siderophores in *S. coelicolor*^{182–184}. Addition of GlcNAc under nutrient-limiting conditions activates the transcription of antibiotic BGCs, including *cpk* which specifies the cryptic polyketide coelimycin; this principle is now also being applied in industrial screening regimes. Interestingly, there is direct competition between the DasR regulon and the regulatory networks governed by the transcription factors AtrA and Rok7B7: DasR represses *actII-ORF4* and *nagE2*, encoding the activator of Act biosynthesis and the GlcNAc transporter, respectively, whereas these genes are activated by (and depend on) Rok7B7 and AtrA101. This system highlights the complex control of bioactive molecules

in response to environmental changes. It also illustrates the value of discovering the ecological rationale behind antibiotic production and using such knowledge to activate antibiotic production (Figure 2.1).

2.5.3. Identifying the elicitors that activate cryptic BGCs.

Various other approaches have been developed to identify the environmental signals involved in the regulation of actinobacterial specialized metabolism. The signals that trigger the expression of BGCs act through a transcriptional regulatory network, governed via *cis*-regulatory elements targeted by transcription factors. Many transcription factors will respond to ligands, but how do we uncover what these ligands are? Genomic context is one major pointer. For example, if a regulatory gene lies next to a metabolic operon (such as for sugar metabolism), this may be an important clue. Additionally, transcription factors are often autoregulatory, and the *cis*-regulatory element is therefore typically found in the upstream region of the gene. With the *cis*-regulatory element in hand, computational approaches can then be used to predict the regulatory network in silico. In the case of DasR, it was immediately obvious that the best hits in those predictions all related to GlcNAc metabolism or transport, and identifying glucosamine-6P as the ligand was then fairly straightforward¹⁸¹. Methods directed at single bacterial producer strains include varying the composition of growth media^{103,185}, inducing antibiotic resistance^{186,187} and microbial co-cultivation^{188–190}. Screening for new chemical elicitors of antibiotic production is a promising approach, as this can enhance the chance of success in high-throughput screening of bacterial strain collections. Logical elicitors to include in such screens are those with a proven pleiotropic activity, such as GlcNAc¹⁸¹, γ -butyrolactones^{176,191} and histone deacetylase inhibitors¹⁹². Screening compound libraries for small molecules that perturb antibiotic production was shown to be an effective strategy to identify novel elicitors of antibiotic production¹⁹³. Another example is bioactivity high-throughput elicitor screening technology, in which a wild-type microorganism is subjected to a library of small molecules, and the resulting induced metabolomes are screened for bioactivity against a chosen indicator strain^{194,195}. Use of this method led to the identification of various cryptic antibiotics, including the novel lanthipeptide cebulantin¹⁹⁴ and the novel naphthoquinone epoxide hiroshidine¹⁹⁵. The method also identified atenolol, a β -blocker clinically used to treat hypertension, as a global elicitor¹⁹⁵.

2.5.4. Chemical ecological relationships as elicitors of antibiotic production.

Within their natural environment, Actinobacteria are part of diverse microbial communities that include archaea, bacteria, fungi, protists, and viruses. Within these communities, specific interactions have evolved, and small molecules, such as specialized metabolites, facilitate interactions between different microbial species (symbionts or competitors), including the activation of antibiotic production^{7,196}. By mimicking these naturally occurring chemical-ecological relationships in so-called co-culture experiments, cryptic BGCs might be activated in the laboratory.

Indeed, co-cultivation of Actinobacteria with other bacteria or fungi changes their specialized metabolite production profile. Examples include the production of alchivemycin A by a *Streptomyces* strain following co-cultivation with the mycolic acid-producing *Tsukamurella pulmonis*¹⁹⁷, biosynthesis of a range of metabolites during co-culture of *Aspergillus nidulans* and various streptomycetes¹⁹⁸, and the activation of a silent pathway of actinorhodin in *S. coelicolor* upon co-cultivation with *A. niger*¹³⁴. Co-culture of marine-derived *Streptomyces* spp. with different human pathogens, including methicillin-resistant *Staphylococcus aureus*, resulted in increased production of

different antibiotics and enhanced biological activity¹⁹⁹. Co-culture with multidrug-resistant bacteria might emerge as an effective, targeted approach to find novel bioactive compounds with activity against the pathogens for which new antibiotics are desperately needed.

The signals and cues that mediate the observed changes in specialized metabolite production are diverse and include physical cell-cell interactions^{197,200}, a higher rate of nutrient depletion²⁰¹, enzymatic conversion of precursors to active metabolites²⁰², HGT²⁰³ and microbial small molecules^{110,204,205}. However, for many interactions, the signals – and specifically the molecular mechanisms – are as yet unknown. The development of analytical techniques such as nanospray desorption electrospray ionization and MALDI-TOF imaging mass spectrometry enables the direct visualization of molecules exchanged during the chemical communication between microorganisms²⁰⁶. These methods might help with the elucidation of the signals involved in above-described interactions.

Actinobacteria are also found in close association with various eukaryotic hosts (Figure 2.3), and there are multiple examples of defensive symbioses between Actinobacteria and the host, in particular for plants (suppressive soils) and insects (fungus-growing ants)^{141,207,208}. An interesting example is provided by leafcutter ants, which live in symbiosis with the fungus *Leucoagaricus* and with *Pseudonocardia* bacteria²⁰⁸. The *Pseudonocardia* produce bioactive compounds to protect the fungal cultivar against infection by other fungi. Recent work showed that, in return, the presence of *Pseudonocardia* elicits the production of antimicrobials by pathogenic *Escovopsis* fungi during infection of the cultivar²⁰⁹. This exemplifies an evolutionary arms race between the Actinobacterium and the fungus.

Interestingly, many plant growth-promoting bacteria produce phytohormones, such as auxin and gibberellic acid²¹⁰. This suggests that the host and the microorganism communicate (or hijack each other's communication channels) through the production of such metabolites. This may indeed work both ways, as plant hormones influence growth and specialized metabolism by endophytic Actinobacteria²¹¹ (A. van der Meij, J.M. Raaijmakers and G.P.v.W., unpublished observations), which reveals that host-specific signals can also affect specialized metabolism (Box 2.2).

2.6. Genome mining strategies

The biosynthetic diversity found in Actinobacteria is enormous. Genomic data sets have become increasingly larger, with massive strain collections, pan-genomes and metagenomes sometimes containing genetic information for thousands of Actinobacteria at once. A range of computational tools (for example, antiSMASH⁶⁷ or PRISM⁷⁸) have been developed in the past decade that automate the identification of BGCs in these genomes and, to a certain extent, facilitate prediction of the structures of their products (Box 2.3).

Given the increasing size of genomic or metagenomic data sets, studying BGCs on a case-by-case basis is often no longer feasible. To address this problem, sequence similarity networking approaches have been developed to automatically relate predicted BGCs to gene clusters of known function (from, for example, the MIBiG database^{70–72}) and to group them into gene cluster families (GCFs)^{116,146,212}. The members of a GCF are then predicted to produce the same or highly similar molecules. Using their streamlined computational framework for BGC similarity networking, researchers recently studied 3,080 actinobacterial genomes and found that they contained around

~18,000 distinct GCFs, the vast majority of which have no known products⁷³. This constitutes an enormous potential for discovery and to some extent enables us to differentiate between BGCs that are likely to produce compounds we have seen before and BGCs that may encode novel chemistry. However, the number of GCFs to which no known functions or chemistries can be linked is so great that it is difficult to know which of the BGCs belonging to them encode the production of the pharmaceutically most interesting molecules.

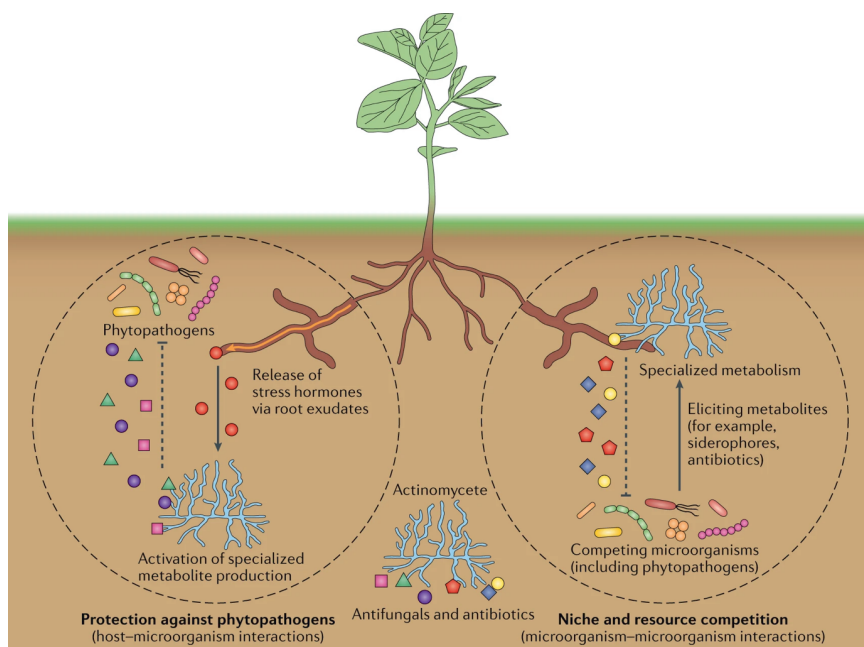


Figure 2.3. Natural products of Actinobacteria in host-microorganism and microorganism-microorganism interactions. Within the natural habitat of Actinobacteria, environmental signals are likely to have a key role in the activation of specialized metabolism. Chemical ecological interactions can, for example, be found in the rhizosphere of plants. Upon stress, plants release stress hormones (for example, jasmonic acid) via their root exudates. These hormones can activate antibiotic production by Actinobacteria, which can protect the plant against phytopathogens ('cry for help'). Additionally, competitive interactions occur between different members of the microbial soil community (both symbionts and pathogens). Metabolites, such as siderophores and antibiotics, from the competing microbial community can elicit changes in the metabolic profile of Actinobacteria, including antibiotic production. These molecules have an important role in shaping the rhizosphere microbiome and inhibiting (opportunistic) pathogens, thus also protecting the host.

Box 2.2. Host-derived signals as elicitors.

The high abundance of Actinobacteria in plant microbiomes, specifically members of the genus *Streptomyces*, suggests that these bacteria are actively recruited by plants, possibly via root exudates^{213–215}. Plants exude up to 30–60% of their photosynthate into the rhizosphere, which contains large quantities of sugars and amino acids that can influence the growth of beneficial bacteria as well as their antibiotic production (for example, through catabolite repression)^{89,211,213}. These exudates contain plant metabolites, such as the phytohormone salicylic acid, which has been positively correlated with the abundance of operational taxonomic units attributed to Streptomycetaceae in the soil, as well as endophytic Actinobacteria²¹⁴. As such hormones are released under pathogenic stress, an interesting hypothesis is that exudation of such compounds may be a means to ‘cry for help’: the release of stress hormones by the host results in the recruitment of antimicrobial-producing Actinobacteria to reduce the severity of pathogenic infection^{89,216} (Figure 2.3). Interestingly, when Actinobacteria are grown in the presence of plant stress hormones, such as jasmonic acid and salicylic acid, changes in the production of specialized metabolites are observed, often leading to increased antimicrobial activity²¹¹ (A. van der Meij, J.M. Raaijmakers and G.P.v.W., unpublished observations). This could indicate that evolved regulatory networks exist between the host and members of its microbiome, opening up the intriguing possibility to design microorganisms that produce bioactive molecules specifically in response to stress induced by pathogen-derived signals. We hypothesize that similar use of Actinobacteria as ‘medicine producers’ may happen in our own microbiome, with mammalian stress hormones as activators of specialized metabolism of Actinobacteria during infection.

Some advances have been made to predict the functions of BGCs. For example, the Antibiotic Resistance Target Seeker (ARTS²¹⁷) prioritizes BGCs based on co-localization with resistance genes. The rationale behind this is that bacteria need a mechanism to protect themselves against the antibiotic molecules they are producing. Therefore, resistance markers may function as beacons to prioritize specific BGCs for antibiotic discovery. Unfortunately, only a small percentage of BGCs have distinguishable self-resistance markers. Ecological insights are needed to provide complementary strategies to know which microorganisms are most likely to encode biosynthetic pathways of interest, which BGCs among these are functionally most desirable and how they can likely be activated.

2.6.1. Where to find chemical novelty.

To effectively mine Actinobacteria for drug discovery, there is a need for guidance towards chemical novelty. First of all, this requires a better understanding of the environmental and taxonomic distributions of BGCs. Such knowledge can help estimate where to search for novel producers and whether this search should be based on phylogeny, geography or specific environmental niches^{115,147,212}. Taxonomic groups that are particularly gifted in terms of their natural product diversity include Streptomycetales, Streptosporangiales, Frankiales, Micromonosporales and Pseudonocardiales²¹². To decrease the risk of rediscovery of known molecules, known as replication, focus is directed towards rare Actinobacteria, of which many taxa have been greatly underexplored. Indeed, genera such as *Micromonospora*, *Amycolatopsis*, *Salinispora*, *Nocardia* and *Verrucosispora* are a source of chemically unique metabolites with potent antibacterial activities, such as abyssomycins and proximicins²¹⁸. Besides phylogeny, geographic location has also been proposed as an indicator of BGC distribution²¹⁹. However, comparative genomics of *Amycolatopsis* and *Salinispora*

strains both show that taxonomy is a more important indicator of BGC distribution than geographic location^{142,146,147}. Intriguingly, regardless of their geographic origin, strains of *Salinispora arenicola* all produced rifamycin, staurosporine and saliniketol¹⁴², suggesting that, independent of the niche, strains may produce the same molecules or analogues. To a certain degree, “everything is everywhere, and the environment selects”, as Dutch microbiologist Baas-Becking proposed almost a century ago. Indeed, a study of biosynthetic diversity in soils from Central Park in New York City, United States, suggests that the degree of novelty found in common areas may be similar to that in more exotic locations²²⁰.

Box 2.3. Computational tools for genome mining of biosynthetic diversity.

The most commonly used tool for predicting biosynthetic gene clusters (BGCs) in bacterial genomes is antiSMASH. Based on core genes, antiSMASH predicts the gene cluster type (non-ribosomal peptide, polyketide, terpene, siderophore and so forth). It also annotates and groups accessory genes and minimally predicts the specialized metabolite core scaffold, enabling researchers to quickly identify BGCs of interest in an uploaded genome⁶⁷. For instance, antiSMASH output was used as the starting point for a phylogenetic prioritization method leading to the discovery of the novel compound corbomycin¹⁶². Other BGC identification tools also exist, such as BAGEL and PRISM^{78,221}. The database MIBiG, which charts known BGCs and their products, provides a means of easily comparing predicted clusters with experimentally characterized ones that have known chemical products^{70–72}.

For BGC mining on a larger scale, the networking tool BiG-SCAPE can be used to cluster both full and fragmented BGCs into gene cluster families, to obtain a comprehensive overview of BGC diversity in large genome collections or metagenomes. In such large data sets, the tool CORASON can help with prioritizing BGCs of interest by providing insights into the evolutionary context of BGCs through phylogenetic analysis⁷³. Genome mining approaches can also be combined with proteomics for more efficient prioritization²²². Additionally, metabolomics data can be coupled to structure predictions yielded by genomics data or to absence–presence patterns of BGCs, in order to link molecules to BGCs^{223,224}. The power of these methods will only increase now that tools are able to predict the function of natural products from their (predicted) structures²²⁵. As structure prediction is easier on a substructure level, there are ongoing efforts to attempt linking substructure predictions to mass shifts, which can be a great aid in elicitation studies and dereplication²²⁶.

Still, exploration of extreme or unusual environments, such as hyper-arid deserts, permafrost soils, mangrove trees, caves and deep-sea sediments that are characterized by challenging conditions (aridity, high salinity, low nutrient sources and extreme temperatures), showed high diversity of BGCs^{144,227}. Between 2010 and 2018 alone, taxonomically diverse microorganisms originating from extreme environments have been the source of nearly 200 new specialized metabolites, many of which were produced by Actinobacteria¹⁴⁴. Microorganisms from the permafrost soil synthesize a broad range of chemical compounds²¹⁸. Other interesting sources of gifted Actinobacteria are the microbiomes of diverse eukaryotic hosts, including insects, sponges and humans^{4,145,228}. Within microbiomes, pathogen pressure selects for Actinobacteria that produce efficacious and relevant antimicrobials. Furthermore, this host association could potentially enrich for compounds with low toxicity to animals. This makes host microbiomes a promising source of novel molecules, with possibly a higher potential to be successfully used in the clinic. Metabolomic analysis of *Streptomyces* spp. from insect microbiomes displayed immense potential for novel chemistry in these strains⁴. The same study demonstrated how principal component analysis can be leveraged for strain

prioritization; a strain characterized as a metabolic outlier produced cyphomycin, a novel antifungal agent active against multidrug-resistant fungi⁴.

However, we should remind ourselves that even well-studied organisms such as *S. coelicolor* still harbour undiscovered biosynthetic pathways. For instance, it was shown that inactivation of the biosynthetic genes for the common antibiotics streptothricin and streptomycin resulted in the production of hidden antibiotics, such as the rare amicetin²²⁹. It is intriguing that despite the extensive research into such organisms, metabolic products of several putative BGCs have so far eluded discovery. Also, there are still many exciting questions about the regulation of these specialized metabolite pathways, the signals that can activate production and the ecological role of many of these molecules. Even now, many lessons remain to be learned from *S. coelicolor* and other well-studied model streptomyces. The challenge is to find ways to leverage this potential, and ecology can play a key role in this.

2.6.2. Ecology to identify BGCs of interest.

Even when selecting Actinobacteria from under-mined taxa and from high-potential environments, one will still end up with thousands of distinct BGCs to study, many more than can realistically be targeted for experimentation with the currently available tools. Even the most high-tech synthetic biology approaches – although they are certainly game changers that allow bypassing regulation for specified BGCs – will not enable synthetic refactoring of sufficiently large numbers of BGCs to facilitate global screening of all actinobacterial biosynthetic diversity for years to come. One way to somewhat narrow down that number is to focus our attention on BGCs within non-core regions of actinobacterial genomes, which are more likely to encode compounds of chemical and functional novelty as HGT, and therefore the uptake of new, less conserved BGCs, mostly happens in these regions, as seen in *Amycolatopsis*¹⁴⁷. However, even then there are still far too many BGCs to explore. Hence, further prioritization is required. For this, genomic and meta-omic data are potentially very useful (Figure 2.4).

First of all, predicting how BGCs are regulated can shed light on both their ecological functions and which triggers or cues can be used to activate their expression in the laboratory (Figure 2.4B). Computational tools that predict regulons can help uncover the regulatory networks responsible for BGC control. One such tool is PREDetector²³⁰, which uses position weight matrices of transcription factor binding sites to predict which regulators are likely to bind DNA sequences within BGCs and thus likely regulate their expression. Often, a BGC encodes a pathway-specific regulator that is in turn regulated by a more pleiotropic (global) regulator. The identity of the global regulator can potentially be very informative about the ecological function of the BGC and therefore the function of its product. For example, in a plant microbiome setting, BGCs regulated by DasR are likely to respond to GlcNAc, which is also a break-down product of fungal cell walls; hence, this could point to a possible antifungal role of a compound produced by such a BGC. Computational searches have the potency to identify the entire regulons associated with them^{230,231}, and the gene content of these regulons may provide valuable data on the ecological functions of this regulon and by proxy of the BGC in question²³². The specific molecules eliciting the activation of these regulons would then still need to be identified; potentially, paired metabolomics and metatranscriptomics of native communities where the Actinobacterium resides may provide means to identify which molecules are specifically present when expression of the BGC is triggered. Such predicted regulatory cues can in turn feedback into tools such as PREDetector to find novel BGCs in other species that may be similarly elicited.

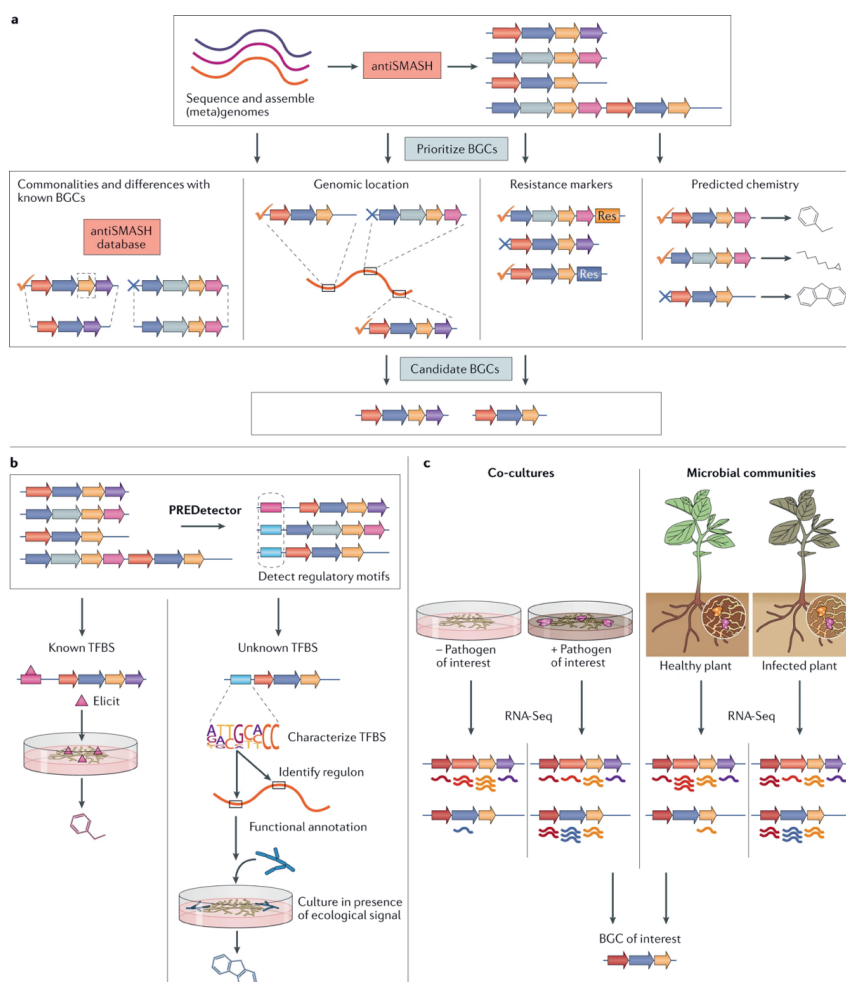


Figure 2.4. Omics strategies for BGC prioritization and elicitation. A. Genomics-based prioritization of biosynthetic gene clusters (BGCs). After identifying putative BGCs through antiSMASH, BGCs can be prioritized based on commonalities and differences with known BGCs, their location on the genome, resistance markers in the vicinity of the BGC and predicted chemistry of the compound that the BGC produces. B. Genomics-based prediction of BGC regulatory sequences and elicitors. After predicting both known and unknown transcription factor binding sites (TFBS) with PREDetector, species with known TFBS can be directly elicited. Unknown TFBS can be characterized by generating a sequence profile of similar TFBS and searching the genome for the entire regulon regulated by the corresponding transcription factor. From this, the genes in the regulon can be functionally annotated, and the ecological signal that triggers BGC expression can be inferred. C. Transcriptomics-based prioritization of BGCs. Differential gene expression of BGCs can be quantified through RNA-sequencing (RNA-Seq) of co-cultures or microbiomes with or without a pathogen of interest. BGCs that are expressed when the pathogen of interest is present are more likely to have a role in targeting this pathogen and host protection.

Furthermore, using metatranscriptomic data (or transcriptomic data of co-cultures) is likely to be a powerful technology to predict the roles of BGCs in interaction with other organisms (Figure 2.4C). Knowing under which conditions members of certain GCFs are expressed can illuminate their likely functions and hint towards how they may be regulated. For example, determining which bacterial BGCs in a plant endosphere microbiome were upregulated upon fungal infection recently led to the identification of a gene cluster essential for disease suppression²³³. The attractive concept that chitooligosaccharides produced from hydrolysis of the fungal cell wall elicit the production of the antifungal needs to be tested. Additionally, expression of BGCs can be correlated to the absence/presence/abundance of specific other organisms in the community, to identify whether they might either be triggered by their presence or effectuate their loss from the community by, for example, antibiosis. Adding metabolomics to the equation may provide further means of prioritization, as candidate products for a BGC can be identified that specifically appear when it is expressed, and tandem mass spectrometry analysis algorithms^{226,234–237} can be used to dereplicate them and predict (parts of) their structures to assess their novelty (Box 2.3).

2.6.3. Synthetic biology approaches to express gene clusters.

Although computational strategies can be used to prioritize the most novel BGCs for experimental characterization, we then face the challenge of identifying the cognate chemical products of these BGCs. This currently remains a bottleneck, as thousands of potentially interesting BGCs can be found across publicly available genome sequences. By direct sequencing of environmental DNA, metagenomics even enables us to predict the chemical space of microbial ‘dark matter’: thus far uncultivated bacteria, which represent a promising source of novel natural products²³⁸. Synthetic biology is a powerful strategy to facilitate expression of BGCs observed in genome or metagenome sequences²³⁹. This strategy is illustrated by recent work on the identification of bioactive molecules from the human microbiome²⁴⁰. The synthesis of a metagenomic BGC and subsequent heterologous expression in *S. albus* enabled the isolation and identification of new polyketides, designated metamycin A, metamycin B, metamycin C and metamycin D. Advances in synthetic biology and genome engineering (reviewed in detail here²³⁹) can become very useful in the expression of cryptic BGCs and the identification of their chemical products. However, designing a DNA sequence for a large pathway that will be functional in a model production host has remained more challenging than previously anticipated, as achieving the required stoichiometry between transcriptional units in a BGC for it to produce fully elaborated products is non-trivial²⁴¹: the required precursors or cofactors may be lacking in the heterologous host²⁴², and low production titres from synthetic BGCs may hamper chemical characterization. Although all of these bottlenecks are being addressed, understanding the regulatory mechanisms behind BGC expression for now remains of key importance for the identification of their products. Once high-throughput refactoring of complex BGCs becomes a reality, we anticipate that ecological and regulatory information will be crucial to predict BGC functions and thus prioritize them for synthesis and expression.

2.7. Conclusions and future perspectives

Despite our advances in niche exploration revealing great potential for drug discovery, the current state of knowledge regarding BGC diversity and distribution in terms of ecology and phylogeny limits our ability to guide drug discovery. It is therefore necessary to further characterize the extant microbial diversity from different ecological niches and create a global survey of niche-correlated natural product diversity. Moreover, characterizing their functions in their native microbial communities, as well as their modes of action, will be crucial to advance our understanding of the regulation of specialized metabolism and hence for our effective ability to prioritize BGCs and elicit their expression. New technologies will be required for this, and in the 'omics' area we specifically envision a larger role for transcriptomic and metatranscriptomic studies of specialized metabolism, as well as regulatory network reconstruction, targeted to the most relevant microbiomes and ecological niches. Once we better understand which cellular and ecological conditions induce the expression of BGCs, this will greatly facilitate prioritizing gene clusters that are likely to have functions of interest and to predict which molecular stimuli are likely to activate them.

Chapter 3

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters

Barbara R. Terlouw*, Kai Blin*, Jorge C. Navarro-Muñoz, Nicole E. Avalon, Marc G. Chevrette, Susan Egbert, Sanghoon Lee, David Meijer, Michael J. J. Recchia, Zachary L. Reitz, Jeffrey A. van Santen, Nelly Selem-Mojica, Thomas Tørring, Liana Zaroubi, Mohammad Alanjary, Gajender Aleti, César Aguilar, Suhad A. A. Al-Salihi, Hannah E. Augustijn, J Abraham Avelar-Rivas, Luis A. Avitia-Domínguez, Francisco Barona-Gómez, Jordan Bernaldo-Agüero, Vincent A. Bielinski, Friederike Biermann, Thomas J. Booth, Victor J. Carrion Bravo, Raquel Castelo-Branco, Fernanda O. Chagas, Pablo Cruz-Morales, Chao Du, Katherine R. Duncan, Athina Gavriilidou, Damien Gayraud, Karina Gutiérrez-García, Kristina Haslinger, Eric J. N. Helfrich, Justin J. J. van der Hooft, Afif P. Jati, Edward Kalkreuter, Nikolaos Kalyvas, Kyo Bin Kang, Satria Kautsar, Wonyong Kim, Aditya M. Kunjapur, Yong-Xin Li, Geng-Min Lin, Catarina Loureiro, Joris J R Louwen, Nico L L Louwen, George Lund, Jonathan Parra, Benjamin Philmus, Bita Pourmohsenin, Lotte J. U. Pronk, Adriana Rego, Devasahayam Arokia Balaya Rex, Serina Robinson, L. Rodrigo Rosas-Becerra, Eve T. Roxborough, Michelle A. Schorn, Darren J. Scobie, Kumar Saurabh Singh, Nika Sokolova, Xiaoyu Tang, Daniel Udway, Aruna Vigneshwari, Kristiina Vind, Sophie P. J. M. Vromans, Valentin Waschulin, Sam E. Williams, Jaclyn M. Winter, Thomas E. Witte, Huali Xie, Dong Yang, Jingwei Yu, Mitja Zdouc, Zheng Zhong, Jérôme Collemare, Roger G. Linington, Tilmann Weber, and Marnix H. Medema

* These authors contributed equally to this work

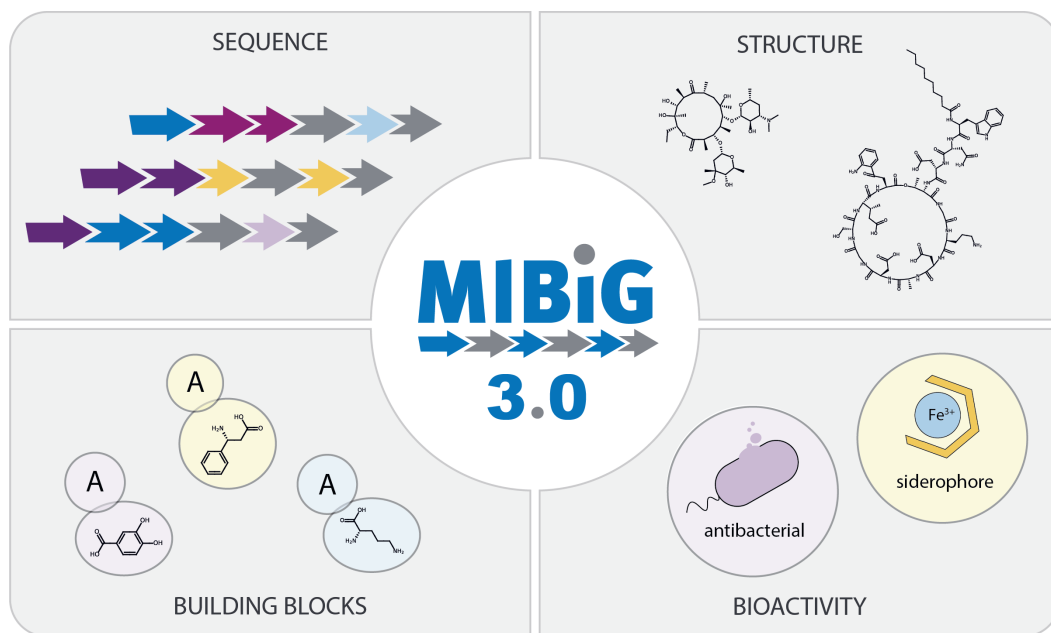
This chapter has been published as

Terlouw, B.R., Blin, K. et al., MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, 51(D1) D603–D610 (2023). <https://doi.org/10.1093/nar/gkac1049>

Abstract

With an ever-increasing amount of (meta)genomic data being deposited in sequence databases, (meta)genome mining for natural product biosynthetic pathways occupies a critical role in the discovery of novel pharmaceutical drugs, crop protection agents and biomaterials. The genes that encode these pathways are often organised into biosynthetic gene clusters (BGCs). In 2015, we defined the Minimum Information about a Biosynthetic Gene cluster (MIBiG): a standardised data format that describes the minimally required information to uniquely characterise a BGC. We simultaneously constructed an accompanying online database of BGCs, which has since been widely used by the community as a reference dataset for BGCs and was expanded to 2021 entries in 2019 (MIBiG 2.0). Here, we describe MIBiG 3.0, a database update comprising large-scale validation and re-annotation of existing entries and 668 new entries. Particular attention was paid to the annotation of compound structures and biological activities, as well as protein domain selectivities. Together, these new features keep the database up-to-date, and will provide new opportunities for the scientific community to use its freely available data, e.g., for the training of new machine learning models to predict sequence-structure-function relationships for diverse natural products. MIBiG 3.0 is accessible online at <https://mibig.secondarymetabolites.org/>.

Graphical abstract



3.1. Introduction

Across all kingdoms of life, organisms produce specialised metabolites: molecules that are produced by bacteria, fungi, and plants to gain an advantage over their competitors in challenging environments. Specialised metabolites, also referred to as secondary metabolites or natural products, exhibit a wide variety of biological activities, including many that are useful for pharmaceutical and agricultural applications, e.g., antibiotics, anti-cancer drugs, pesticides, and herbicides. The production of specialised metabolites is typically encoded by biosynthetic gene clusters (BGCs): groups of co-localised and co-regulated genes that jointly encode a biosynthetic pathway. Therefore, microbial and plant genomes can be mined for novel specialised metabolite production by detecting BGCs and predicting their encoded products and functions. Similar to how the relationship between DNA, mRNA and protein describes the flow of information in cells, we can define a ‘central dogma’ of specialised metabolism: a BGC sequence encodes a set of enzymes, which together assemble a compound structure (or a cocktail of structural analogues), which in turn dictates specialised metabolite function. Understanding how information is translated from sequence to structure to function is key to natural product discovery. To address the first stage, sequence information, various tools have been developed that automatically detect BGCs from DNA sequence, including antiSMASH and its siblings fungiSMASH and plantiSMASH^{68,69}, GECCO²⁴³, DeepBGC²⁴⁴, RiPPMiner²⁴⁵, and PRISM 4⁷⁸.

To facilitate dereplication and comparative analysis of predicted BGCs with known BGCs, and to characterise the interplay between sequence, structure and function, standardised data annotation and storage are essential. To this purpose, we developed the Minimum Information about a Biosynthetic Gene cluster (MIBiG) standard and built a database which contains standardised entries for experimentally validated BGCs of known function^{70,71}. Each entry minimally contains information about the nucleotide entry and coordinates of the genomic locus involved, the producing organism’s taxonomy, biosynthetic class, name of the produced compound(s), and literature reference(s). There are also various optional fields for non-minimal entries, including fields for gene function, product structure and bioactivity, crosslinks to chemical structure databases such as NP Atlas²⁴⁶ and PubChem²⁴⁷, and monomer identity. With MIBiG 2.0 containing over 2000 entries, the database has become an important reference for many researchers that mine genomes for natural products. For example, it has been used to estimate the potential for biosynthetic novelty in large-scale microbiome studies^{248,249}, to identify conserved amino acids playing key roles in catalytic activities across enzyme families²⁵⁰, to help guide natural product discovery efforts towards high-potential taxa²⁵¹, and to train machine-learning algorithms for natural product activity prediction⁷⁴.

Here, we present MIBiG 3.0: an update designed to increase the number of non-minimal entries in our database and adding new data entries through a large-scale community annotation effort. We focused on three features: the characterisation and cross-linking of 918 chemical structures; the annotation of 1002 bioactivities of BGC products; and the validation and annotation of 2027 protein domain substrates of nonribosomal peptide synthetases (NRPSs). In addition, we added 668 novel BGCs to the MIBiG database which were published since the last database update and removed 63 duplicate and low-quality entries (Figure 3.1). Together, these additions keep the database current, and provide unique opportunities for exploring complex sequence-structure-function relationships in diverse natural product domains.

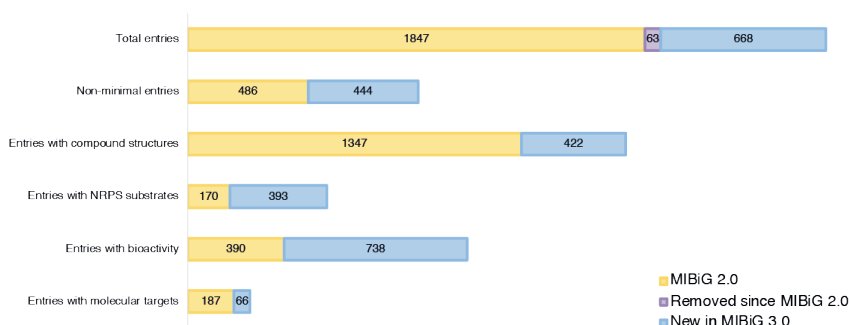
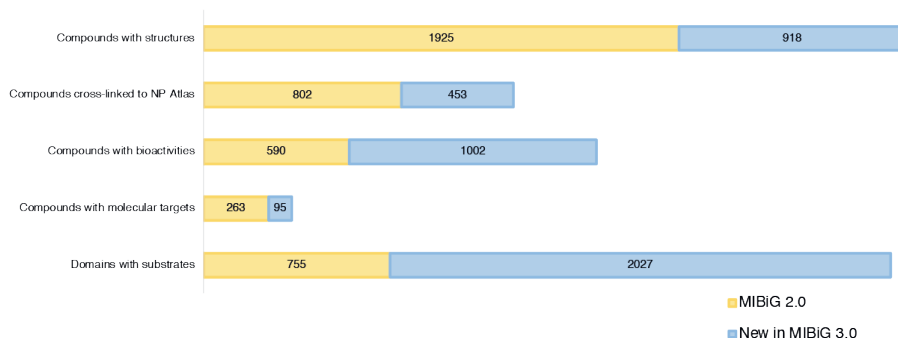
a**b**

Figure 3.1. Overview of MIBiG 3.0. A. Added, removed, and updated entries since MIBiG 2.0. B. Improvements in the annotation of compounds, bioactivities, molecular targets and NRPS domain substrates.

3.2. Methods and implementation

3.2.1. Manual curation through crowdsourcing and mass online ‘annotathons’

As authors themselves typically have the best understanding of the BGC they have studied, we greatly encourage natural product researchers to submit their BGCs to MIBiG during the process of publishing their work. To this purpose, MIBiG supplies an online form through which researchers can request a unique MIBiG identifier and submit their experimentally verified BGCs, pre- or post-publication. Since MIBiG version 2.0, this has yielded 97 manually submitted, high-quality entries which have now been incorporated into MIBiG 3.0. Still, there are far more published BGCs that are not manually submitted to MIBiG.

With an increasing number of papers describing novel BGCs being published every year, manually annotating, validating, and adding BGCs to MIBiG has become a mammoth task. Therefore, we took to social media to gauge the community’s interest in participating in an online annotation event. We received many positive responses, with 86 people from four different continents volunteering to participate in our MIBiG ‘annotathons’. We organised eight three-hour online sessions, accommodating different time-zones, with various breakout rooms dedicated to specific annotation tasks: annotating new clusters, annotating, and cross-linking compound structures, annotating

compound bioactivities, and assigning substrate selectivities to NRPS protein domains. We prepared multiple instruction videos and assigned an expert to each of the breakout rooms who could be directly approached with questions from annotators to ensure that annotation quality was consistent. In addition, one of our annotators at the CINVESTAV research institute mobilised fourteen MSc Integrative Biology students of their 2021 Bacterial Genomics class to annotate compound bioactivities under supervision. Finally, we resolved 125 database issues that were raised by users on our GitHub page, redefining BGC boundaries, correcting biosynthetic classes, adding, and removing literature references, fixing compound structures, and removing duplicate entries.

3.2.2. Annotating and cross-linking compound structures

Since version 2.0, compound structures in MIBiG have been cross-linked to the NP Atlas database: a database containing structures of natural products isolated from bacteria and fungi. During the preparations for version 3.0, we collaborated with the NP Atlas team to 1) add structures for compounds in SMILES format⁸⁵, including stereochemical information where possible, and 2) cross-link them to five databases of chemical structures: NP Atlas²⁴⁶, PubChem²⁴⁷, ChemSpider²⁵², LOTUS²⁵³, and ChEMBL⁸⁷. If compound entries were found in multiple databases, SMILES strings from NP Atlas were prioritised. SMILES strings were also collected for existing entries that were already cross-linked to a database but did not report a SMILES string. Correctness of SMILES syntax was validated with PIKACHU²⁵⁴.

3.2.3. Annotating compound bioactivities

To improve MIBiG as a resource for machine learning models predicting sequence-structure-function relationships, we added bioactivity data for 1002 compounds and chemical target data for 95 compounds. 708 of these annotations were transferred from the dataset assembled by Walker and Clardy, who designed a machine learning model to predict BGC function from sequence⁷⁴. To accommodate these annotations, we added 40 functional categories in addition to the eight categories previously described in MIBiG.

3.2.4. Annotating NRPS protein domains

To concretise the relationship between NRPS sequence and the structure of its produced nonribosomal peptide (NRP), we annotated and validated the substrate selectivities of 2782 NRPS adenylation (A) domains. A-domains dictate which monomers (predominantly amino acids) are incorporated into (hybrid) NRP scaffolds. Substrate annotation can be performed at different levels: we can define the pre-tailored substrate precursor (e.g., L-aspartic acid); the substrate as recognised by the A-domain (e.g., (3R)-3-hydroxy-L-aspartic acid); or the post-tailored integrated monomer that ends up in the final NRP scaffold (e.g., (3R)-3-hydroxy-D-aspartic acid). We chose to annotate the substrates as recognised by the A-domain, as this best reflects the biological relationship between A-domain and incorporated monomer. In addition to substrate identity, we also recorded evidence for substrate selectivity in the form of an evidence code and literature references. To this purpose, we added 13 evidence codes to the JSON schema which is used to standardise MIBiG entries (Table 3.1).

After community annotation, substrate naming was homogenised and each stereochemically ambiguous substrate was manually curated by an expert. Where stereochemistry could be inferred from structure, this is reflected in the substrate name for each stereocenter. Exceptions are amino acid names, which are assumed to be in their L-configuration. To avoid any ambiguity in substrate naming, we also linked each of our 274 unique substrate names to an isomeric SMILES string representing the substrate structure (Figure 3.2). SMILES validation and deduplication were handled using PIKACHU²⁵⁴.

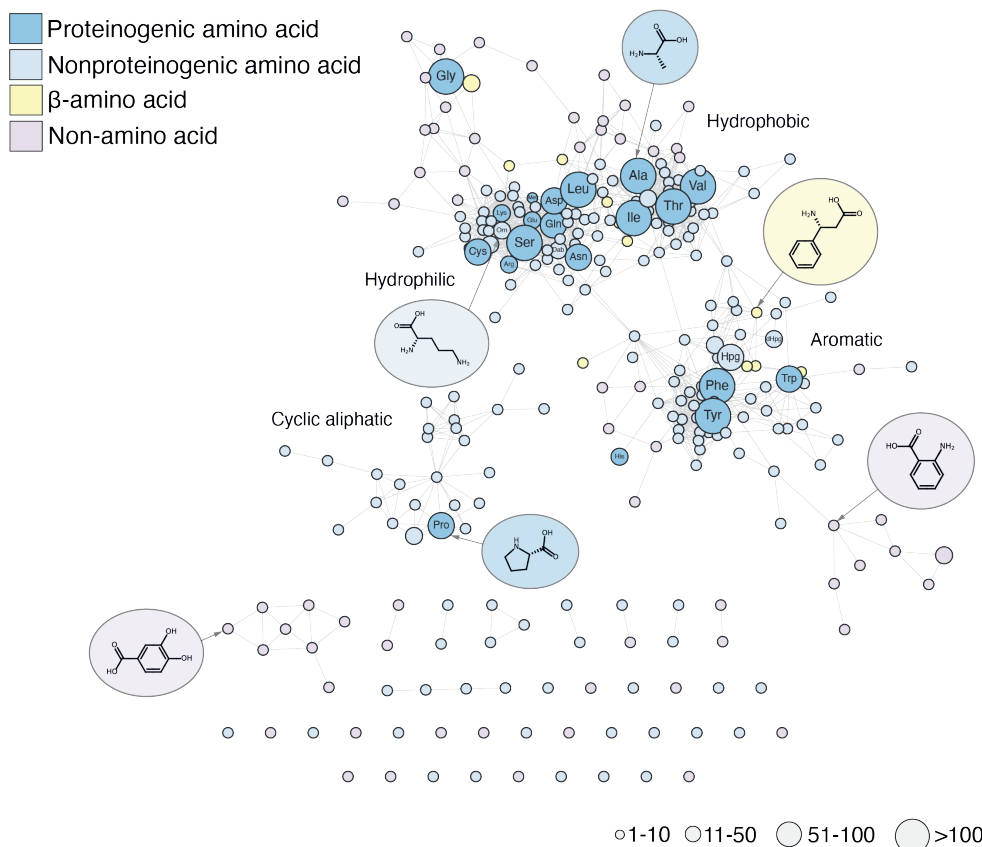


Figure 3.2. Similarity network of annotated NRPS substrates. Each node represents one of 274 unique NRPS substrate structures in MIBiG 3.0. Colours indicate substrate categories, and node size correlates with the number of annotations for that substrate in the MIBiG database. Substrates were clustered based on Tanimoto similarity of ECFP-4 molecular fingerprints²⁵⁵ (edge cut-off=0.46).

Table 3.1. Evidence codes for adenylation domain substrate annotations. As indicated, some evidence codes are only accepted as evidence for substrate specificity when combined with a second evidence code that provides further support for a data entry. 13 evidence codes were newly introduced in MIBiG 3.0.

ACVS assay: δ -(L-R-Aminoadipyl)-L-cysteiny-D-valine synthetase assay, specific for measuring penicillin production.

HPLC: High-Performance Liquid Chromatography

NMR: Nuclear Magnetic Resonance

Evidence code	Accepted as standalone evidence	New in MIBiG 3.0
Activity assay	X	
ACVS assay	X	X
ATP-PPi exchange assay	X	X
Enzyme-coupled assay	X	X
Feeding study	X	
Heterologous expression	X	X
Homology		X
HPLC	X	X
In-vitro experiments	X	X
Knock-out studies	X	X
Mass spectrometry	X	X
NMR	X	X
Radio labelling	X	X
Sequence-based prediction		
Steady-state kinetics	X	X
Structure-based inference	X	
X-ray crystallography	X	X

3.3. Results and discussion

3.3.1. Taking the ‘minimal’ out of MIBiG

While MIBiG 2.0 serves an important role in the community as a reference database to quickly identify whether a BGC is similar to any known BGCs, its utility as a resource for exploring sequence-structure-function relationships could be improved. This can mainly be explained by the high number of minimal entries in the database: entries that only contain sequence and compound information that could be augmented by adding further standardised annotations. For MIBiG 3.0, we aimed to promote as many existing and novel entries as possible to non-minimal entries by annotating compound structures (918), bioactivities (1002), and NRPS substrates (2027). In total, we added 668 novel BGCs and 4553 separate data entries to our database, increasing our number of non-minimal entries from 486 to 930 (Figure 3.1). MIBiG 3.0 now contains 2515 entries, spanning 16 phyla across 5 kingdoms of life (Table 3.2).

3.3.2. Streamlining research into the central dogma of specialised metabolism

With 905 NRPS and modular Type I PKS BGCs in MIBiG 3.0, modular BGCs constitute a substantial part of our database. Modular systems are characterised by enzyme complexes comprising repeating domain architectures, which collectively assemble a natural product scaffold. When the substrate selectivities of the recognition domains are known (acyltransferase (AT) domains for PKS and A-domains for NRPS), these consistent architectures make it possible to predict the structure of chemical scaffolds with reasonable accuracy. Most AT domains in PKS systems recognise one of two substrates, malonyl-CoA or methylmalonyl-CoA, and excellent bioinformatics tools exist to distinguish between the two²⁵⁶. However, for A-domains in NRPS systems, which recognise over 300 known substrates²⁵⁷, substrate prediction is a greater challenge, which will require substantially more data to obtain models of comparably predictive power. Therefore, we decided to make the annotation of the substrate selectivity of NRPS A-domains a major focus of MIBiG 3.0. MIBiG 3.0 now contains annotations for 2782 A-domains (compared to 755 annotations in MIBiG 2.0; Figure 3.1B), covering 274 unique substrates which are identified by stereochemically curated isomeric SMILES strings (Figure 3.2). This makes MIBiG the largest resource for A-domain substrate data, containing 3-4 times as many labelled data points as the training sets used for the A-domain selectivity predictors SANDPUMA⁸⁰ and NRPSPredictor2⁷⁷. We hope that eventually this dataset will be leveraged to train an improved A-domain substrate predictor, which can in turn be integrated into tools like antiSMASH to improve NRP scaffold structure prediction.

Since version 2.0, we have added structural identifiers of 918 compounds to our database in SMILES format¹⁶, increasing the number of BGCs with structural data from 1347 to 1769 (Figure 3.1). By pulling SMILES strings directly from cross-linked databases where possible, we avoid conflicts caused by versioning and SMILES formatting. Additionally, we linked 1002 additional compounds to 51 unique bioactivities, creating opportunities for computationally predicting compound bioactivity from structure. For a further 95 compounds, we were also able to annotate their molecular targets (Figure 3.1B).

By centering MIBiG 3.0 around the annotation of substrate building blocks, compound structures, and bioactivities, we aspired to streamline future research into all aspects of sequence-structure-function relationships that lie at the heart of natural product research. All data can be easily

downloaded and parsed in bulk from our database in JSON and GenBank format or accessed on an entry-by-entry basis through our searchable online repository (Figure 3.3). As such, we hope that MIBiG 3.0 will prove an important resource for future machine learning endeavours that aim to decode the central dogma of specialised metabolism.

Table 3.2. Entries in MIBiG 3.0 by phylum

Kingdom	Phylum	Number of BGCs in MIBiG 3.0
Bacteria	Actinobacteria	1051
	Proteobacteria	528
	Firmicutes	230
	Cyanobacteria	139
	Bacteroidetes	17
	Candidatus Tectomicrobia	6
	Chloroflexi	4
	Verrucomicrobia	3
	Planctomycetes	2
	Kiritimatiellaeota	1
	Unknown	41
Fungi	Ascomycota	416
	Basidiomycota	23
	Unknown	3
Plantae	Streptophyta	44
	Rhodophyta	2
Archaea	Euryarchaeota	3
Chromista	Bacillariophyta	1
	Dinophyceae	1

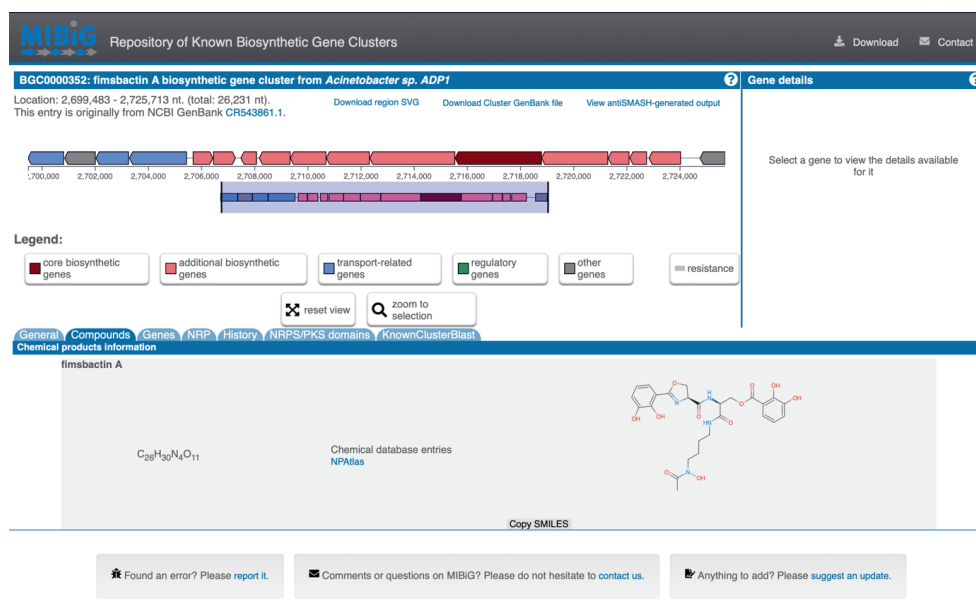


Figure 3.3. Example of a MIBiG overview page. Users can navigate to different tabs to access data such as compound structure, domain architecture and NRPS substrates.

3.4. Data availability

The MIBiG Repository is available at <https://mibig.secondarymetabolites.org/>. There is no access restriction for academic or commercial use of the repository and its data. The source code components, JSON-formatted data standard, and SQL schema for the MIBiG Repository are available on GitHub (<https://github.com/mibig-secmet>) under an OSI-approved Open Source licence.

Chapter 4

Revealing determinants of translation efficiency via whole-gene codon randomisation and machine learning

Thijs Nieuwkoop*, Barbara R. Terlouw*, Katherine G. Stevens, Richard A. Scheltema, Dick de Ridder, John van der Oost, and Nico J. Claassens

* These authors contributed equally to this work

This chapter has been published as

Nieuwkoop, T., Terlouw, B.R., et al., Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning, *Nucleic Acids Research*, 51(5), 2363–2376 (2023).
<https://doi.org/10.1093/nar/gkad035>

Abstract

It has been known for decades that codon usage contributes to translation efficiency and hence to protein production levels. However, its role in protein synthesis is still only partly understood. This lack of understanding hampers the design of synthetic genes for efficient protein production. In this study, we generated a synonymous codon-randomised library of the complete coding sequence of red fluorescent protein. Protein production levels and the full coding sequences were determined for 1459 gene variants in *Escherichia coli*. Using various machine learning approaches, these data were used to reveal correlations between codon usage and protein production. Interestingly, protein production levels can be relatively accurately predicted (Pearson correlation of 0.762) by a Random Forest model that only relies on the sequence information of the first eight codons. In this region, close to the translation initiation site, mRNA secondary structure rather than Codon Adaptation Index (CAI) is the key determinant of protein production. This study clearly demonstrates the key role of codons at the start of the coding sequence. Furthermore, these results imply that commonly used CAI-based codon optimisation of the full coding sequence is not a very effective strategy. One should rather focus on optimising protein production by reducing mRNA secondary structure formation within the first few codons.

MEW: the mRNA Expression Wizard is available at <https://github.com/BTheDragonMaster/mew>

4.1. Introduction

Due to degeneracy in the genetic code, a protein with a single amino acid sequence can be encoded by an extremely large number of different coding sequences (CDS). While different synonymous codons do not alter the amino acid sequence, they are known to influence translation efficiency and in some cases even protein folding properties^{258–262}. However, many questions about the roles of codons and their often subtle and intertwined effects are still unanswered. Several studies have revealed the importance of the first few codons on overall protein production, but there is still no complete consensus on the underlying mechanisms^{259,261,263–267}. Also, it is unclear to what extent codons further down-stream the coding sequence influence protein production²⁵⁸. Understanding codon usage is key to grasping one of the fundamental processes of life: the translation of mRNA into proteins. In addition, precise control over translation efficiency is highly desirable in both biotechnology and synthetic biology to make the process of protein production and cell engineering more predictable.

Since the early days of DNA sequencing, it has been observed that, depending on the organism, specific codons are overrepresented²⁶⁸. This led to the hypothesis that frequently occurring codons could be translated more efficiently, e.g. due to a higher abundance of corresponding tRNAs. It also led to the development of the so-called Codon Adaptation Index (CAI)²⁶⁸, which is defined as the geometric mean of the relative codon usage in a specific coding sequence, based on the average codon usage in the genome or a subset of highly expressed genes. In other words, a CDS with a high CAI primarily uses frequent codons, while a CDS with a low CAI contains more rare codons. However, the hypothesis that a high CAI is related to high protein production has been disputed in several studies in recent years^{259,261,269}. Especially the hallmark study by Kudla et al. in the bacterium *Escherichia coli* revealed that CAI does not seem a major determinant of high protein production²⁵⁹. In this study, a set of 154 codon variants of the Green Fluorescent Protein (GFP) was generated. The authors could not correlate the CAI to the protein production levels. However, the predicted folding

energy of the mRNA around the start codon did correlate with protein production. They hypothesised that protein production efficiency in *E. coli* is mostly influenced by the availability of the ribosome binding site (RBS) for translation initiation. Relatedly, already in the 1980s it had been observed that the codon identity in the 5' of the CDS has major effects on protein production efficiency of some recombinantly expressed genes in *E. coli*^{264,270}. Several studies have since followed up on this and suggest a key role of mRNA secondary structure around the RBS and start codon^{261,263,271,272}. The codon usage at the 5' CDS has also been hypothesised to be involved in a so-called codon ramp, as these regions generally contain more rare codons^{266,273}. This ramp supposedly results in a slow initial translation elongation speed, reducing the risk of detrimental ribosomal collisions along the length of the CDS. Finally, codon usage has also been associated with mRNA stability^{258,260,274}. Especially in eukaryotes, slow-moving ribosomes can initiate RNA decay, thereby linking translation elongation efficiency to mRNA decay²⁷⁵.

Despite a range of studies in this field, there are still many open questions about the complex role of codon usage on protein production: what are the key determinants of protein production, do weaker determinants play a role, and to what extent do these determinants change depending on the producing organism or the cell's environment? Consequently, reliable models for predicting protein production based on codon usage are not available. Many of the algorithms used for 'codon optimisation' are based on the CAI score or variations thereof, which in practice often fail to yield optimal results²⁷⁶.

To further contribute to the understanding and predictability of protein production based on codon usage, specifically aiming to understand influences throughout the whole coding sequence, we decided to generate a large, gene-wide synonymous codon library of the gene encoding the monomeric Red Fluorescent Protein (mRFP) and express this library in *E. coli*. This reporter protein has rarely been used in studies that focus on codon usage, as opposed to GFP, which is predominantly used in this field. We chose a different reporter protein to examine if a different gene candidate would lead to new findings on the determinants of codon usage. To improve our fundamental understanding of the impact of codon usage, and to improve the predictability of optimal codon usage for protein production, we employed various interpretable machine learning approaches. Very recently, some studies have successfully utilised machine learning methods to predict protein production efficiency based on randomised sequence libraries for non-coding gene regions, such as promoters and 5' untranslated regions (5' UTR) in *E. coli* and *Saccharomyces cerevisiae*^{277,278}, as well as for a factorially designed 5' end of the CDS²⁷⁹.

In this study, we constructed mRFP gene libraries for which the codon usage throughout almost the whole gene was fully randomised. To this end, we used an assembly approach based on Type IIS restriction and ligation, which has not been used previously for whole-gene codon randomisation. After library assembly, high-quality curated protein production levels were measured for 1459 individual clones, and their specific coding sequences were accurately determined. We then used these pairs of CDS sequences and corresponding expression values as training data for our machine learning algorithm MEW (mRNA Expression Wizard), to establish an algorithm that can predict protein production from the CDS. Remarkably, we show that only a window covering codons 2-8 is required to accurately predict mRFP production, based on sequence information only. This further strengthens conclusions from previous studies and demonstrates that, in this study, codons further downstream in the CDS are not predictive of protein production. This also underlines that future

studies aiming to optimise protein production should focus on codon usage of the first 2-8 codons of the coding sequence, rather than the current practice of full CDS optimisation.

4.2. Materials and Methods

4.2.1. *mRFP codon randomisation*

The amino acid sequence of the monomeric Red Fluorescent Protein (mRFP) was used to generate three degenerate DNA sequences representing our libraries (CAI_L, CAI_M and CAI_H). Libraries were designed such that they could be assembled from DNA oligos with a Type IIS restriction enzyme-based assembly method. Each degenerate library sequence was split into blocks of roughly equal sizes (80-90 nucleotides) in such a way that each block has a unique 4-bp overhang with neighbouring blocks. Overhangs were selected from a set that is optimised for high ligation fidelity via Type IIS assembly²⁸⁰. To create each required fixed overhang sequence, we fixed degenerate codons in such a way that the separate blocks were roughly equal in size, and that loss of degeneracy was limited. For example, fixing the degenerate sequence ARAT to AAAT would result in the loss of 1 codon possibility, while fixing the degenerate sequence YGCN to CGCC would result in the loss of seven codon possibilities. 5' and 3' flanking sequences containing recognition sites for the Type IIS restriction enzyme BsaI-HF[®]v2 (NEB, R3733) were added to each DNA block to generate unique, single-stranded overhangs after digestion. The 5' end of the first block and the 3' end of the last block contained SapI (NEB, R0569) recognition sites instead. Each block was ordered as a DNA oligo (Ultrasaver[®] DNA Oligonucleotides, IDT) and using a strand-displacing Taq polymerase (NEB, M0482), the ssDNA was converted to double-stranded DNA via PCR. PCR reactions containing the dsDNA block were cleaned and concentrated to 20 µl mQ using the DNA Clean & ConcentratorTM-5 kit (Zymo, D4004). 4 µl Gel Loading Dye, Purple (6x) (NEB, B7024) was added to each block, after which they were loaded on a 1% agarose gel and ran for 30 min at 100 V. The dsDNA blocks were excised from the gel and purified to 20 µl mQ using the ZymocleanTM Gel DNA Recovery Kit (Zymo, D4002). 5 µl of dsDNA was used to quantify the DNA concentration with the Qubit assay (Invitrogen, Q32853) according to the manufacturer's protocol.

The dsDNA blocks were mixed in an equal molar ratio to a total volume of 41 µl, with 5 µl T4 Ligase Buffer (NEB, B0202), 400 units T4 Ligase (NEB, M0202) and 60 units BsaI-HF[®]v2 (NEB, R3733). The assembly reaction was done overnight at 37°C for 18 h, followed by 5 min at 60°C and a holding step at 12°C. The assembly was cleaned and concentrated to 15 µl mQ using the DNA Clean & ConcentratorTM-5 kit (Zymo, D4004). 3 µl Gel Loading Dye, Purple (6x) (NEB, B7024) was added, and the assembly mixture was loaded on a 1% agarose gel and was run for 40 min at 100 V. The full-length assembled product was excised from the gel and purified to 44 µl mQ using the ZymocleanTM Gel DNA Recovery Kit by Zymo (D4002). 10 units of SapI (NEB, R0569) were added with 5 µl CutSmart Buffer (NEB, B7204) and digested for 2 h at 37°C. The digested codon random mRFP library with single-stranded overhangs was cleaned and concentrated to 15 µl mQ using the DNA Clean & ConcentratorTM-5 kit by Zymo (D4004). The complete 15 µl containing the codon random mRFP library was used in a ligation reaction to generate the plasmid library.

4.2.2. *Plasmid preparation and library generation*

The pFAB3909 plasmid²⁸¹ (Addgene #47812) with a P15A origin, kanamycin resistance gene and bicistronic design element was modified to be able to accept the codon randomised mRFP library and

include a constitutively expressed GFPuv gene. The relatively weak bla promoter was used to drive the mRFP expression, keeping the total protein yield relatively low to prevent overburdening of the protein production machinery and negative growth effects on production for high producing mRFP codon variants. A strong terminator was used for efficient transcription termination and to enhance mRNA stability²⁸². The open reading frame was replaced by Sapl recognition sites to generate the sticky overhangs that accept the mRFP library, and a large fragment of nonsense DNA was inserted between the Sapl sites to be able to easily separate the double Sapl digested plasmid from linear product based on size on a gel. A GFPuv gene, driven by the P4 promoter²⁸¹, was added to the plasmid as an internal standard for gene expression. Expression of GFPuv by this promoter is weak as to not interfere with the mRFP expression efficiency, but strong enough for detection with flow cytometry to allow for data normalisation.

About 3 µg plasmid was digested with 20 units Sapl (NEB, R0569) and dephosphorylated with 3 units rSAP (NEB, M0371) with 6 µl CutSmart Buffer (NEB, B7204) in a total volume of 60 µl for 3 h at 37°C, followed by an inactivation step at 65°C for 20 min. The linear plasmid was excised from the gel and purified to 30 µl mQ using the Zymoclean™ Gel DNA Recovery Kit by Zymo (D4002). The codon random mRFP library (15 µl) was ligated into 30 ng linear plasmid with 400 units of T4 ligase (NEB, M0202) and 2 µl T4 Ligase Buffer (NEB, B0202) in a total volume of 30 µl for 18 h at 16°C. The ligation mixture was cleaned and concentrated to 10 µl mQ using the DNA Clean & Concentrator™-5 kit by Zymo (D4004). 1 µl of the codon randomised mRFP library was transformed into electrocompetent DH10B cells (20 µl competent cells, 2 mm cuvette, voltage: 2500 V, resistor: 200 Ω, capacitor 25 µF, BTX® ECM630). Cells were recovered in 1 ml NEB® 10-beta/Stable Outgrowth Medium (NEB, B9035) at 37°C for 1 h. The cells were transferred to a 50 ml tube and 9 ml LB (10 g/l Peptone (OXOID, LP0037), 10 g/l NaCl (ACROS, 207790010) and 5 g/l yeast extract (BD, 211929)) were added with 50 µg/l kanamycin (ACROS, 450810500) and incubated for 18 h at 37°C.

4.2.3. Expression range enrichment and selection

A Fluorescence Activated Cell Sorter (FACS) (Sony, SH800S Cell Sorter; GFPuv excitation at 488 nm, emission at 525/50 nm; mRFP excitation at 561 nm, emission at 617/30 nm) was used to sort 50,000 cells of the overnight cell culture into three groups based on protein production levels. The left and right tail of the normal distribution and a fraction of the middle peak were sorted to create 3 groups of low, medium, and high production. The three cell groups were put on individual agar plates (10 g/l Peptone (OXOID, LP0037), 10 g/l NaCl (ACROS, 207790010), 5 g/l yeast extract (BD, 211929), 15 g/l agar (OXOID, LP0011), 50 mg/l kanamycin (ACROS, 450810500)) and grown overnight at 37°C. From these plates, individual colonies were picked and grown in 2 ml 96-well plates with 200 µl LB with kanamycin (10 g/l peptone (OXOID, LP0037), 10 g/l NaCl (ACROS, 207790010), 5 g/l yeast extract (BD, 211929), 15 g/l agar (OXOID, LP0011), 50 mg/l kanamycin (ACROS, 450810500)) for 18 h at 37°C.

4.2.4. Measurements and sequencing

The cell cultures were diluted 100× in PBS (8 g/l NaCl (ACROS, 207790010), 200 mg/l KCl (ACROS, 196770010), 144 mg/l Na₂HPO₄ (ACROS, 12499010), 240 mg/l KH₂PO₄ (ACROS, 447670010)). mRFP expression was measured using a flow cytometer (Thermo, Attune NxT Flow Cytometer; GFPuv excitation at 405 nm, emission at 512/25 nm; mRFP excitation at 561 nm, emission at 620/15 nm; stop option 200,000 single cells). A gate was used to exclude GFPuv outliers (±10% of the total population) aiming to reduce unrelated biological variance as GFPuv expression levels are expected

to stay constant. From the overnight cultures, 1 µl of cells were used in a PCR reaction to amplify the DNA for Sanger sequencing using Q5 (NEB, M0492). The PCR reaction was sent to MacroGen Europe B.V. for sample clean-up and Sanger sequencing.

All cell cultures were also measured using a microplate reader (BioTek, Synergy Mx). 50 µl overnight cell cultures were diluted in 50 µl PBS (8 g/l NaCl (ACROS, 207790010), 200 mg/l KCl (ACROS, 196770010), 144 mg/l Na₂HPO₄ (ACROS, 12499010), 240 mg/l KH₂PO₄ (ACROS, 447670010)). The plates were incubated at room temperature for 1 h before measuring (cell density measured at 600 nm; GFPuv excitation at 395/9 nm, emission at 508/9 nm; mRFP excitation at 584/9, emission at 607/9 nm). The microplate reader fluorescent readings were normalised with the OD₆₀₀ for both the GFPuv and mRFP readings.

4.2.5. Data validation

Before using the measurements in our machine learning approach, we set a few criteria that the data had to meet to exclude artefacts in our dataset. First, the sequencing data should be of sufficient quality and the encoded amino acid sequence should be correct. The raw sequence data was validated by extracting the open reading frame sequence using in-house scripts. All bases in the open reading frame needed a Phred quality score >20 (which translates to a base call accuracy of at least 99%) and the translated sequence should match the mRFP amino acid sequence in order to pass. Second, any double populations or clear changes in cell morphology or culture density were excluded. Double populations were detected in the flow cytometry data but were already automatically excluded during sequence quality-based filtering, as a double population will result in poor Sanger sequencing data. Very rarely, we observed an unexplained shift in cell morphology, observed as increased forward and side scatter values obtained during flow cytometry. Finally, rarely, a difference was observed in fluorescence between flow cytometry measurements and microplate reader measurements. Based on a set threshold of 25% deviation from the average relationship between the two measurement methods we excluded these deviating cell cultures. All exclusions were made prior to our analysis to generate high-quality data to feed into the machine learning algorithm. For the remaining data points, we assessed dataset-wide biases and correlations, such as assembly bias and the correlation between protein production level and GC content to ensure the dataset on the whole was appropriate for machine learning. All raw data and validated data are made available in the Supplementary material online at <https://doi.org/10.1093/nar/gkad035>.

4.2.6. Proteomics sample preparation and digestion

For 10 strains with a wide linear range of fluorescence levels, mRFP levels were also verified using proteomics. For each, 10 ml cell culture was grown overnight at 37°C. The cell pellets were resuspended in 1 ml lysis buffer containing 20 mM HEPES, 150 mM NaCl, 1.5 mM MgCl₂, 0.5 mM dithiothreitol, 20 units/ml DNase and cOmplete protease inhibitor cocktail (Roche) (pH 7.8). 400 µl of lysate was sonicated using a Qsonica Q500 sonicator (Qsonica LLC) operating at 80% amplitude with on/off interval of 2 and 8 s, respectively, then pelleted at 13,000 RPM for 15 min. Protein concentrations for supernatants (soluble fractions) were measured using the Qubit Protein Assay Kit (Invitrogen) and concentration for the pellet fraction were estimated by assuming 1.5 mg total protein in a 10 ml culture and subtracting the measured amount of soluble protein. Pellets were resuspended in 400 µl lysis buffer (insoluble fractions). Urea was added to all soluble and insoluble fraction samples to a final concentration of 8 M. 100 µl (~12.5 µg protein) of each sample was

reduced and alkylated by incubation with 8.4 μ M dithiothreitol at 60°C for 1 h, followed by incubation with 19 μ M iodoacetamide in the dark at room temperature for 30 min. Samples were incubated overnight at 37°C with trypsin (Sigma-Aldrich) and LysC (Wako) with enzyme-to-substrate ratios of 1:26 and 1:44, respectively. Samples were acidified with 10% trifluoroacetic acid to pH < 3 to quench the digestion, then desalted via SPE using an Oasis PRiME HLB 96 well plate (Waters) and stored at –20°C until further analysis.

4.2.7. Liquid chromatography–mass spectrometric proteomics analysis

Proteomics samples were reconstituted in 10 μ l 2% formic acid and analysed using an Ultimate 3000 HPLC system coupled on-line to an Orbitrap Fusion mass spectrometer (both Thermo Fisher Scientific). Trapping was performed on a 300 μ m \times 5 mm PepMap Neo trap cartridge (Thermo Fisher) in 100% solvent A (0.1% formic acid) at 300 nl/min for 5 min, prior to separation on a 75 μ m \times 500 mm column packed in-house with ReproSil-Pur 120 C18-AQ 2.4 μ m resin (Dr Maisch) at 300 nl/min using a 90 min gradient as follows: 9% B (80% acetonitrile, 0.1% formic acid) for 1 min, 9–13% B for 1 min, 13–44% B for 70 min, 44–99% B for 3 min, 99% B for 4 min, 99–9% B for 1 min, 9% B for 10 min. Peptides were ionised using a 2.0 kV spray voltage. MS scans were acquired within a 375–1500 m/z range with a maximum injection time of 50 ms at a mass resolution of 60,000 and an automatic gain control (AGC) target value of 4×10^5 in the Orbitrap mass analyser. Dynamic exclusion was set to 12 s for an exclusion window of 10 ppm with a cycle time of 1 s. MS/MS scans were performed for precursors with 2+ to 8+ charge states and intensities above 5×10^4 at a constant normalised collision energy of 30%. MS/MS scans were acquired within a 100–2300 m/z range with a maximum injection time of 22 ms at a mass resolution of 15,000 at AGC target of 5×10^4 in the Orbitrap mass analyser.

4.2.8. Proteomics data analysis

Raw files were processed using MaxQuant version 2.0.3.1. Proteins and peptides were identified using a target-decoy approach with a reversed database, using the Andromeda search engine integrated into the MaxQuant environment. The database search was performed against the *E. coli* (strain K12) Swiss-Prot database (version October 2022) supplemented with the full protein sequence for mRFP, and against the common contaminants database integrated in MaxQuant. Default search settings were used, including methionine oxidation and protein N-term acetylation as variable modifications, enzyme specificity set to trypsin with maximum two missed cleavages, a minimum peptide length of seven amino acids, a maximum peptide mass of 4600 Da and 1% false discovery rates. Label-free quantification via MaxLFQ algorithm²⁸³ was performed, and ‘match between runs’ was enabled. Microsoft Excel 2016 and Graph-Pad Prism 9 were used for further data analysis and graph plotting. Adjusted mRFP LFQ intensities were calculated to estimate the truly insoluble mRFP abundance, as part of the pellet contained unlysed cells with soluble protein. The adjusted insoluble mRFP LFQ intensity was calculated by multiplying the LFQs of mRFP for the peptide fraction by the ratio of pellet-to-soluble LFQ intensities for peptide deformylase (DEF, UniProtKB P0A6K3, a highly soluble cytosolic protein) in the same fraction, and then subtracting this value from the original insoluble LFQ intensity:

$$\text{Adjusted Insoluble mRFP} = \text{LFQ}_{\text{mRFP}(\text{pellet})} - (\text{LFQ}_{\text{mRFP}(\text{soluble})}) \times \frac{\text{LFQ}_{\text{DEF}(\text{pellet})}}{\text{LFQ}_{\text{DEF}(\text{soluble})}}$$

4.2.9. Training machine learning regressors

To assess if protein production levels could be predicted from sequence, we employed two different machine learning approaches: Random Forest (RF) regression and LASSO. We implemented RF and LASSO using the scikit-learn package (v0.23.0, ref) in python (v3.7.6), with the `sklearn.ensemble.RandomForestRegressor` and `sklearn.linear_model.Lasso` modules respectively. For RF, default settings were used, while for LASSO, various values for alpha were assessed (0.05, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0 and 100.0 for all regressors, in addition to 500.0 and 1000.0 for regressors trained on full-length sequences; max iterations = 10,000). An alpha of 100.0 gave rise to the best predictions for full-length sequences; for windows, different alphas performed better for different featurisations, but differences were marginal. Separate regressors were constructed for full-length featurised mRNA sequences and for featurised sliding windows of 10, 20, 30, or 40 bases. Prior to training, an independent test set comprising 10% of the data was set aside. Regressor accuracies were then evaluated on the remaining training data through 10-fold cross-validation, where 90% of the training data were used to predict the translation efficiency of the other 10%. This was done for each 10% of the data, such that we obtained a predicted translation efficiency, measured with flow cytometry, for each data point. From these predictions, Pearson and Spearman correlations were computed for actual flow versus predicted flow and used as measures for model accuracy. Feature importances were extracted from all ten regressors built in cross-validation, averaged, and plotted and visualised with matplotlib (v3.2.1). Finally, for regressors trained on full-length sequences and the best-performing sequence windows, new regressors were trained using all training data, and model accuracies were re-evaluated on the independent test-set. Code and regressors are made available at <https://github.com/BTheDragonMaster/mew>.

4.3. Results and discussion

4.3.1. Type IIS assembly method allows for generation of codon-randomised libraries at different CAI levels

To randomise synonymous codon usage throughout the whole mRFP CDS, we developed a randomisation-assembly method based on Type IIS restriction and ligation. We used this approach to generate three codon-randomised mRFP libraries with either fully randomised codon usage or with a focus on more frequent or rare codons. The first CAI library ('medium', CAI_M) is fully randomised and uses an equal distribution of all synonymous codons for each amino acid. Only for the 6-codon amino acids arginine, leucine, and serine, as well as the stop codons, the full codon space could not be covered due to the sequence limitations in degenerate oligos. For each of the three 6-codon amino acids, four of the six possible codons were included, while the stop codon was kept constant at TGA.

Theoretically, the maximum number of CDSs coding for the mRFP protein is 3.19×10^{104} . By limiting the aforementioned amino acids, the stop codon and a few fixed codons needed for assembly purposes the library still contains at most 3.68×10^{93} variants. Obviously, this is an astronomically large number, and generated libraries and experimental efforts can only cover a very small fraction of this diversity.

The randomised design approach results in a uniform codon bias distribution across the gene, with an overall medium CAI of 0.67. To see if codon randomisation with an overrepresentation of frequent

or rare codons affects translation differently, we generated two additional libraries. By restricting the allowed relative adaptiveness (the usage ratio of a codon to that of the most abundant synonymous codon), we generated libraries that use more frequent or rare codons. The library limited to rare codons used only synonymous codons with a relative adaptiveness <0.60 , or the lowest relative adaptiveness if the synonymous codons are used in a close to equal ratio. This rare codon library (CAI_L) had an average CAI of 0.41. The library using frequent codons (CAI_H) used only synonymous codons with a relative adaptiveness >0.50 , resulting in a library with an average CAI of 0.83.

To create the three libraries, we divided the complete degenerate CDS into eight blocks of ~ 85 bases, which could be assembled with unique overhangs between each adjacent part. In order to generate complementary four base pair overhangs between the parts, some codons with multiple synonymous options needed to be fixed to a single codon. The eight DNA parts were ordered as single-stranded oligos with additional Type IIS restriction sites flanking the blocks. The oligos were converted to double-stranded DNA using PCR and consequently assembled using Type IIS restriction and ligation (Figure 4.1A, B). Only a small fraction of the total DNA parts assembled into the full product of 707 bp (Figure 4.1B, indicated with the 'mRFP' label). Seven intermediate products were observed that did not further assemble into the full gene. The assembly limitations might have originated from synthesis errors in the initial oligos, preventing Type IIS restriction or resulting in incorrect overhangs. The band corresponding to the fully assembled product was purified and ligated into an expression vector (Figure 4.1C).

The expression vector contains a native, relatively weak, beta-lactamase promoter (Pbla). A weak promoter poses less burden on transcriptional and/or translational processes, reducing the risk of reaching an upper limit in the protein production process and thus preserving the full expression range as determined by the coding sequences. For the 5' UTR, a medium strength bicistronic design (BCD) was chosen based on the work of Mutalik et al.^{281,284}. This BCD element (BCD5) was previously reported to reduce the influence of mRNA secondary structures on expression. Including this element allows us to study the more nuanced features associated with codon usage and should reduce strong effects on translation caused by mRNA structure formation between the coding sequence and the constant 5' UTR.

4.3.2. Expression of libraries in *E. coli* results in a wide expression range and allows for high-quality data collection

The library containing codon randomised mRFP was transformed into *E. coli* DH10B. A single transformation of the libraries in *E. coli* yielded between 150,000 and 320,000 colonies. After 18-h cultivation on liquid, roughly 70% of the cells gave a detectable level of red fluorescence (measured using flow cytometry, Figure 4.2). The remaining 30%, for which no or very little fluorescence was measured, was later confirmed via sequencing to mainly comprise constructs that had a frameshift in the ORF. This is not unexpected, as some sequence blocks are likely missing one or multiple nucleotides since the coupling efficiency of oligos is not 100%. These errors eventually lead to frameshifts and thus protein truncations or mutations.

The flow cytometric evaluation of the three library populations showed that the average expression of the CAI_L , CAI_M and CAI_H libraries increases with average library CAI (Figure 4.2). This suggests that an overall higher CAI leads to higher expressing constructs on average. However, for all three

libraries, expression could be observed at the highest end of the expression spectrum, suggesting that a high CAI is not the most important requirement for high protein production.

To obtain high-quality expression and sequence data for the downstream machine learning analyses, we decided to collect expression levels and related sequences of individual clones. We favoured this method over previously used FlowSeq methods, which perform sequence analysis on large mixes of clones obtained in certain bins during fluorescence activated cell sorting (FACS). FlowSeq typically employs short-read sequencing (Illumina), which will not cover the full CDS length in a single read and due to the whole-gene codon degeneracy in this study, it would be difficult to assemble reads into contigs. Alternative long-read single-molecule methods (e.g. PacBio) would offer a solution, but it was questionable whether a sufficiently high coverage could be achieved to reach meaningful conclusions. FlowSeq has another limitation, as the fluorescence level detected from cells with the same genotype can already cover a relatively wide range²⁸⁵. This increases the likelihood that individual cells are binned incorrectly, and that the resulting dataset is too noisy to be analysed meaningfully through statistical analyses and machine learning. This can potentially be solved by sequencing with very high coverage; however, this is hard to achieve for high-quality long reads as mentioned above. We chose to select a limited number of individual clones for which mean expression values could be accurately determined, as well as their full gene sequences using Sanger sequencing.

To allow the selected clones to cover a wide range of low, medium and high expressing constructs, and exclude non-expressing (e.g. frameshifted) constructs, three expression- level groups were preselected for each CAI library using FACS (Figure 4.1D). After sorting, we picked colonies from each group. These clones were all inoculated in liquid culture (for a total of 480, 1440 and 480 individual cultures for CAI_L, CAI_M and CAI_H respectively). The fluorescence of the cell cultures was measured using both flow cytometry and a microplate reader. mRFP expression was normalised through comparison with the constant constitutively expressed GFP (Figure 4.1C). The mRFP coding sequence (and untranslated regions) was amplified using colony PCR and amplicons were analysed by Sanger sequencing. Next, the data were evaluated to exclude low-quality sequencing reads, amino acid mutations, mixed populations, and rare deviations in cell morphology or culture density. After this filtering step, 1459 sequences which showed high sequence diversity for each of the variable bases in the CDS were selected for further analysis. The exclusion criteria are further described in section 9.2.5. This yielded 1459 high-quality data points that we could use in our machine learning approach (Figure 4.1E).

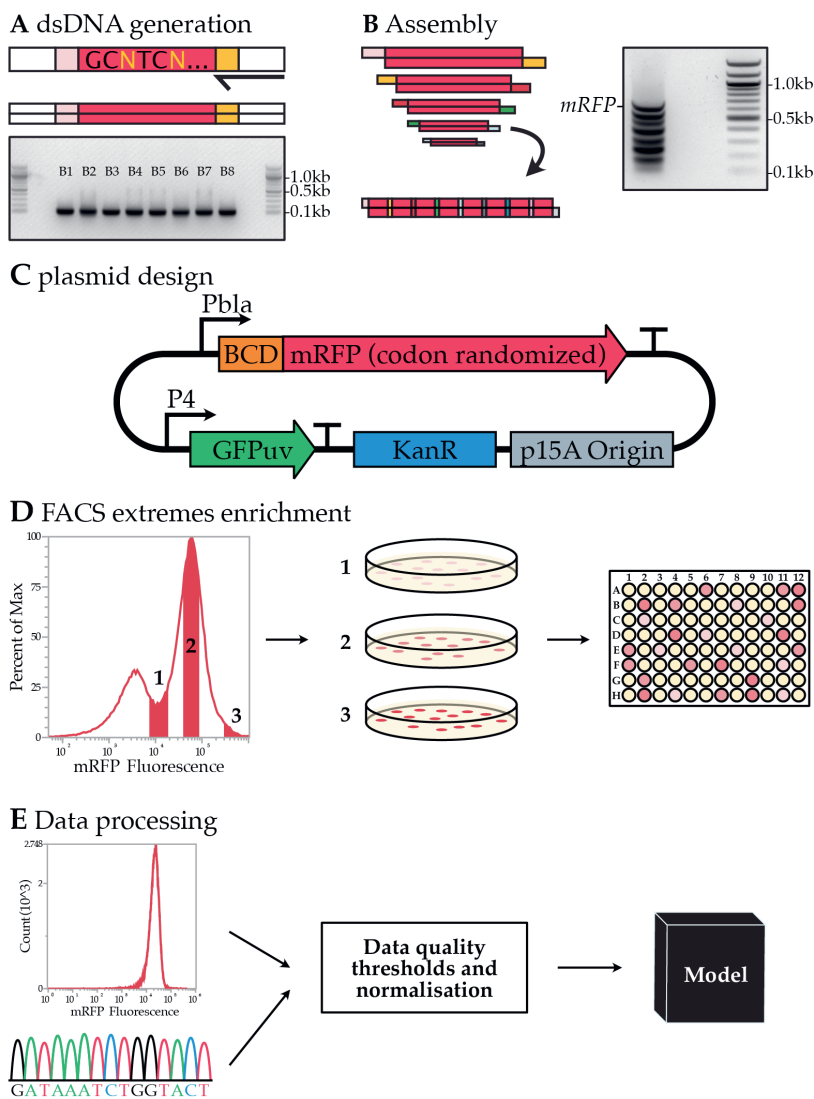


Figure 4.1. Codon-randomised library generation and analysis. A. PCR is used to generate dsDNA from oligos, and an electrophoresis gel yields the eight dsDNA blocks used to build codon-randomised mRFP. B. The assembly reaction and the electrophoresis gel result of the assembly. The complete assembly of all eight blocks is indicated with the mRFP tag. The seven bands below it are intermediate products. C. The expression vector used to express codon-randomised mRFP. D. FACS enrichment for a wide expression range within the library is used to obtain a higher representation of the high and low expressing codon variants. E. Flow cytometry analysis of cultures and Sanger sequencing data are quality filtered and used in machine learning models.

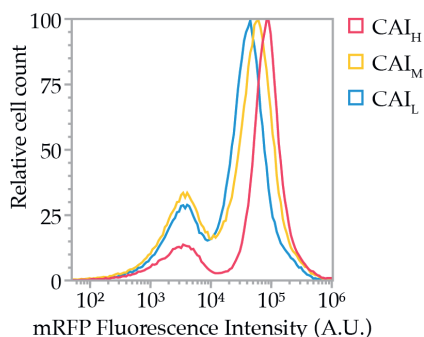


Figure 4.2. Normalised flow cytometry overlay of the mRFP fluorescence intensity from the CAI_L , CAI_M and CAI_H libraries. The left peak is part of the population showing no fluorescence, mainly due to assembly errors in the CDS. The right peak shows the mRFP expression of each library. The average expression of the CAI_L , CAI_M and CAI_H increases with average library CAI, but high expressing variants are found in all libraries (right tail). The ratio between the left and right peaks is a measure of library fidelity, as the left peak consists of autofluorescence of non-expressing or non-functional variants.

4.3.3. LC–MS/MS-based proteomics demonstrates fluorescence intensities correlate well with mRFP levels

It was previously reported that codon changes can also influence the folding properties, which can result in divergent protein functionalities or misfolding for some proteins^{286,287}. Misfolded or differently folded mRFP could in theory influence fluorescence levels. So far, this potential influence of codon variation on fluorescence levels has been typically ignored in studies using fluorescent proteins (mostly GFP) as reporter protein. In this study, we verified that the measured fluorescence levels correlate to the mRFP protein levels by cross-checking abundances with quantitative LC–MS/MS-based proteomics. We selected 10 mRFP gene variants that cover the complete observed fluorescence range in our libraries. For strains harbouring these variants, the soluble and insoluble protein fractions were quantified from mass spectrometry data. This analysis revealed that fluorescence levels correlated well with the determined abundances (Pearson correlation 0.901) and that only a minor part of mRFP ends up in the insoluble fraction (Figure 4.3). Hence, we conclude that the fluorescence level is an effective approximation of mRFP protein abundance in the cell and that fluorescence intensity data can be used for machine learning.

4.3.4. Different machine learning approaches can predict mRFP production levels

To identify the determinants of protein production levels and to assess if the protein production levels could be predicted from gene sequence, we employed two different machine learning approaches: Random Forest Regressor (RFR) and LASSO (Least Absolute Shrinkage and Selection Operator). For this purpose, we developed MEW: the mRNA Expression Wizard, which can train and test a variety of machine learning models to predict protein production levels from mRNA sequence, using different types of featurisations. These featurisations include methods that focus on base pair composition of the coding sequence, to observe the effects of factors like translation elongation efficiency, and featurisations that reflect the probability that a base is paired in the context of mRNA secondary structure.

Our rationale to use both LASSO and RFR is that due to their stepwise decision making, RFRs can model non-linear interdependencies between bases, while LASSO is better suited to straight-forward linear regression and feature selection. Importantly, for each regressor we extracted feature importances as this could help identify determinants of translation efficiency. We trained separate regressors for both full-length featurised mRNA sequences and for sliding windows of varying sizes along the entirety of the mRNA, to assess if certain windows are more predictive of translation efficiency than others.

As performance of machine learning algorithms depends greatly on their input data, featurising our mRFP data in a way that captures most information was key. We used three featurisation methods: one based on one-hot encoding of base identity, another based on predicted base pairing probabilities for each individual base by calculating mRNA secondary structure probabilities of the full transcript using ViennaRNA²⁸⁸, and a third featurisation method which combines these two. In theory, one-hot encoding should also capture base pairing probability. However, we decided to include a featurisation method specific to mRNA secondary structure as this can to some extent 'isolate' this feature from all other features linked to specific bases and codons. We will call the three types of featurisations one-hot, BPP (base pairing probability), and one-hot + BPP respectively.

We trained and validated the RFR and LASSO regressors with flow cytometry data and Sanger sequencing data, using 10-fold cross-validation, yielding a value of predicted mRFP production level for each data point. Depending on the method used (LASSO or RFR) and the featurisation method, prediction accuracy somewhat varied. However, all methods can predict protein levels reasonably well, with Pearson correlation coefficients ranging from 0.546 to 0.776 (Figure 4.4).

The predictive strength of one-hot encoded featurisation is stronger than that of BPP featurisation. This is not unexpected, as BPP featurisation assumes that mRNA secondary structures are the only cause of expression variance. Also, base pairing featurisation is done based on calculated mRNA secondary structure probabilities (ViennaRNA), which likely cannot perfectly predict exact base pairing. Still, BPP featurisation yields reasonable performance, which suggests that mRNA structures affect protein production levels. One-hot encoding captures all information in the sequence and expectedly performs considerably better. Combining one-hot encoding with BPP featurisation does not substantially improve predictions, in line with our expectation that all information on base pairing probability should already be captured by one-hot encoding. No large differences in performance between the two regression methods, LASSO and RFR, are observed. When only BPP features are used, RFR performs slightly better than LASSO; when one-hot or one-hot + BPP are used as features, LASSO slightly outperforms RFR.

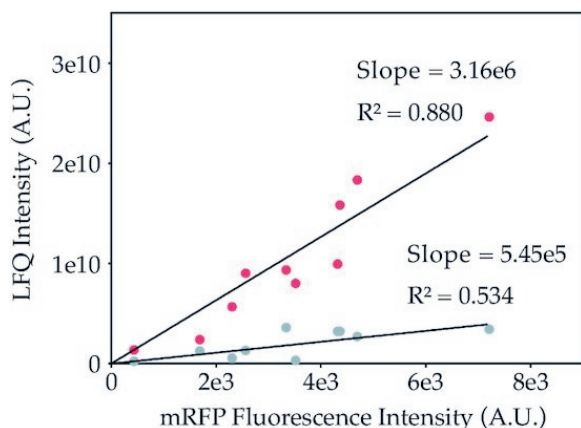


Figure 4.3. Relation between LC-MS/MS based quantification of mRFP and fluorescent intensity for the soluble (red) and insoluble (grey) protein fractions of 10 mRFP codon variants. LC-MS/MS based quantification of mRFP was performed to determine relative, label-free quantities (LFQ) of both soluble and insoluble mRFP. Insoluble LFQ intensities are adjusted for soluble mRFP ending up in the pellet fraction (in non-lysed cells) by correcting these values based on LFQ intensities for a highly soluble *E. coli* protein in the pellet fraction.

4.3.5. Bases surrounding the start codon and the RBS are most predictive of translation efficiency

Next, we assessed which features, and by extension which bases, are most predictive for translation efficiency. We did this by extracting the coefficients for LASSO and the feature importances for RFR and plotting them against sequence position (Figure 4.5). For BPP featurisation, information can be obtained for every nucleotide, including constant nucleotides in the CDS and the constant UTRs, as they are still involved in overall secondary structure formation predictions. However, for one-hot encoding, information is limited to nucleotide identity and can therefore be plotted per codon as only every third base varies across gene variants.

Overall, we found that independent of the algorithms used, the most predictive bases were always close to the start of the CDS, including the 5–10 bases before the start codon for BPP featurisations and the first 25 bases following the start codon for all featurisations. In comparison, the remaining codons play a minimal role in predicting translation efficiency. This strongly suggests that mRNA secondary structure or other factors around the start of the coding sequence play a dominant role in determining translation efficiency. This is in agreement with previous studies that found the same effect for GFP and largely attributed this observation to the necessity for an accessible RBS for translation initiation^{259,261}. However, in this study we used a BCD design in the 5' UTR, which should in theory improve RBS accessibility, independent of sequence context. The BCD design encodes RBS1, which initiates the translation of a short leader peptide. The presence of RBS1 should lead to efficient ribosome recruitment for the translation of mRFP, as the ribosomes translating the leader peptide can unfold mRNA secondary structures around RBS2, which is the translation initiation site for mRFP. Therefore, this 'unfolded' mRNA region should be better available for translation initiation and mRNA structures should have less influence on translation initiation. Still, we observe that the first codons and their base pairing probabilities explain a large part of variation in protein production level. This could be explained by the fact that the BCD design may be unable to completely resolve inhibitory

secondary structure effects, or other factors at the start of the coding sequence may still play a dominant role.

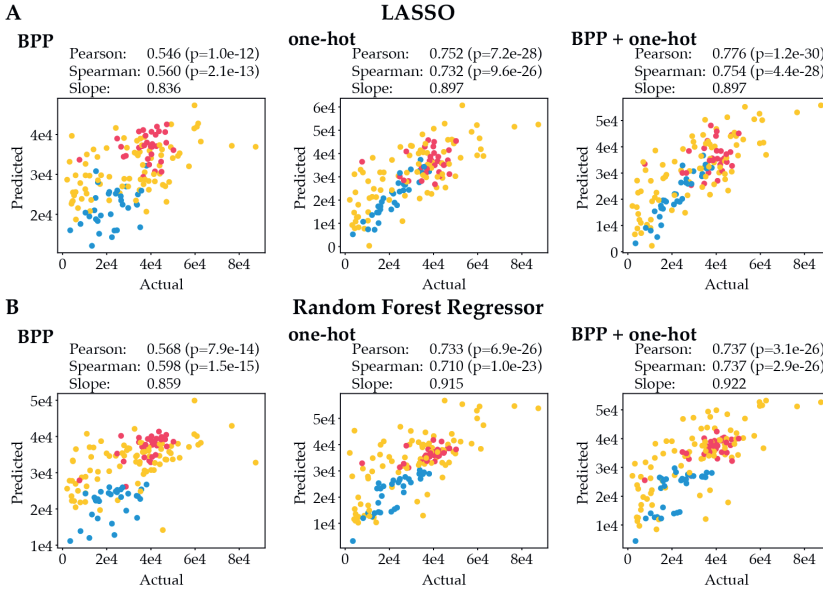


Figure 4.4. Actual expression data vs predicted expression using various machine learning algorithms and featurisations. Blue, yellow, and red points indicate data points from the leave-out test sets for the CAI_L , CAI_M and CAI_H libraries respectively. (A) Actual expression data vs predicted expression using LASSO. (B) Actual expression data vs predicted expression using the Random Forest Regressor (RFR). Regressor accuracies were evaluated through 10-fold cross-validation.

Interestingly, the LASSO regressor using BPP for featurisation assigns positive coefficients to the bases immediately trailing RBS2 in the BCD region (Figure 4.5A, first panel). A positive coefficient indicates that involvement in mRNA secondary structures at this position is positively correlated with gene expression levels. In contrast, the bases in the RBS2 itself and the 5' of the coding region are overwhelmingly assigned negative coefficients, which supports the hypothesis that minimal secondary structure involvement of bases in the RBS is beneficial for high protein production levels. In concordance, the presence of A or T bases in the 5' of the coding region, particularly 'A', is strongly positively correlated with protein production levels, while the presence of G or C bases tends to be negatively correlated with protein production levels in this region (Figure 4.5B, first panel). As A-T base pairs, and their A-U equivalents in mRNA, only form two hydrogen bonds versus three in G-C base pairs, the resulting secondary structures are weaker, and as a result, the RBS may be more accessible.

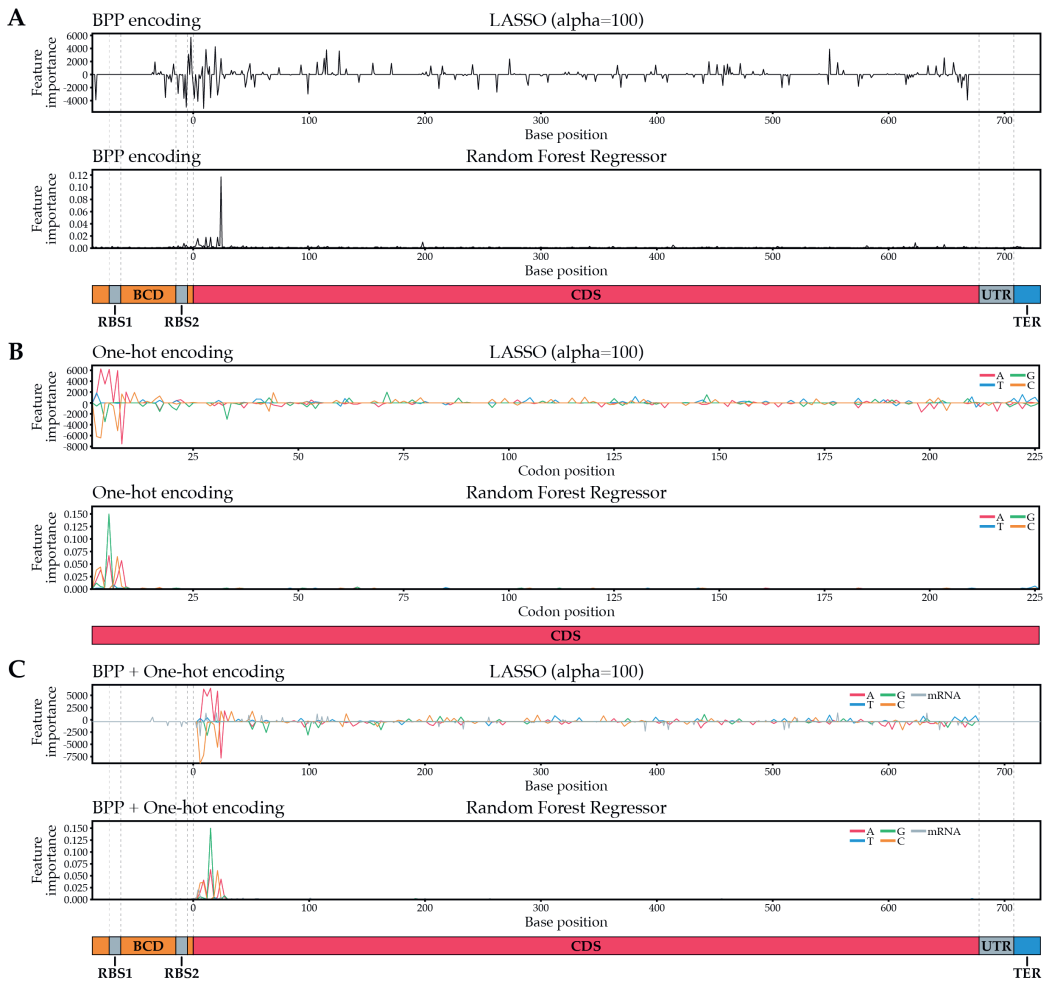


Figure 4.5. Feature importances for various machine learning algorithms and featurisations. LASSO feature importances are coefficients: a positive coefficient indicates a positive correlation between a base and translation efficiency, a negative coefficient indicates a negative correlation. In RFR, feature importances are always positive and therefore not indicative of the directionality of the correlation. A. Feature importances for algorithms using BPP featurisation. B. Feature importances for algorithms using one-hot encoding. Since only every third one-hot encoded base of the coding sequence varies, only every third base of the coding sequence was plotted. C. Feature importances for algorithms using BPP + one-hot featurisation. BCD = bicistronic design 5' untranslated region; RBS = ribosome binding site; CDS = coding sequence; TER = terminator.

4.3.6. A sequence window covering first eight codons can predict protein production

To further substantiate our finding that bases surrounding the start codon dictate translation efficiency, we trained regressors on sliding windows of 10, 20, 30 or 40 bases to visualise which regions of the mRNA were most predictive of translation efficiency. For each sliding window, we performed a 10-fold cross-validation and plotted the correlation between actual expression data and

the predicted expression data as a function of the position of the sliding window (Figure 4.6). Clear peaks of increased predictive power can be observed around the start codon, which corroborates our earlier finding that this region primarily dictates translation efficiency. This is especially apparent in models trained with one-hot encoded features and BPP + one-hot encoded features. Specifically, the 20 nucleotides surrounding base 15 (bases 6-25) lead to high prediction accuracy (Figure 4.6B, C, window size 20). This window covers codons 2-8 and logically does not cover the start codon or the first two nucleotides of the second codon, as these are constant in our design and thus cannot have any predictive power using one-hot featurisation.

It should be noted that the remainder of the CDS, while a lot less predictive than its 5' region, still holds some predictive power (Pearson correlation ~ 0.4). We ascribe this to the inclusion of the CAI_H and CAI_L libraries in our dataset. These datasets use completely different sets of codons, meaning that the machine learning approaches can infer from small sequence windows the identity of the library from which the sequence originated. Since data points from the CAI_H library display higher expression levels than data points from the CAI_L library on average (Figure 4.2), we attribute the non-zero Pearson correlations observed for windows downstream of the 5' end of the CDS mostly to the algorithm's ability to detect the library of origin of CAI_H and CAI_L data points. Inspection of scatter plots for windows in these regions confirmed this (Figure 4.7).

We also observed some 'dips' in predictability performance in the sliding window analysis. One such dip can be seen for small window sizes in the 3' UTR for the LASSO regressor trained on BPP features (Figure 4.6A, C). This region is very invariable both in terms of sequence and secondary structure: since the terminator almost always forms a strong secondary structure, the bases directly before it are less likely to be involved in secondary structures. As a result, the BPP features representing this region hold practically no information. The effect is exacerbated for small windows, as they are less likely to capture predictive residues upstream or downstream of an information-devoid region. In contrast, the secondary structure of the terminator itself does appear to be slightly informative. An unlikely but possible explanation could be that certain codon sequences interfere with the terminator stem formation and thereby influence mRNA stability. However, it is important to keep in mind that correlations between actual and predicted expression data for regressors trained on this region are still extremely low. Therefore, while the 3' UTR region holds some information, it is unlikely to be very influential.

A second dip is located around base 165 and 166 for regressors using one-hot-encoded featurisations (Figure 4.6B). This information valley is caused by an unusually constant region in the mRFP gene, particularly in the CAI_H library, due to low codon variability of local amino acids and a fixed boundary region of two assembly blocks. This is an artefact of our method, and hence not a biologically relevant observation. This dip is not observed for featurisation methods that also include base pairing probabilities, as base pairing interactions of constant regions with other bases can still be informative.

To better understand which sequence elements in the 20-base window surrounding base 15 affect translation efficiency, we plotted feature importances for each regressor trained on this window (Figure 4.8). From this, we inferred that especially at position 15 (codon 5), low probabilities of involvement in mRNA secondary structure are predictive of high expression. This is in line with the current consensus that minimal mRNA secondary structure surrounding the 5' end of the coding

region is conducive to efficient translation. In the case of mRFP, this low base pairing probability seems to be primarily achieved by placing an 'A' at position 15 (Figure 4.8B, C).

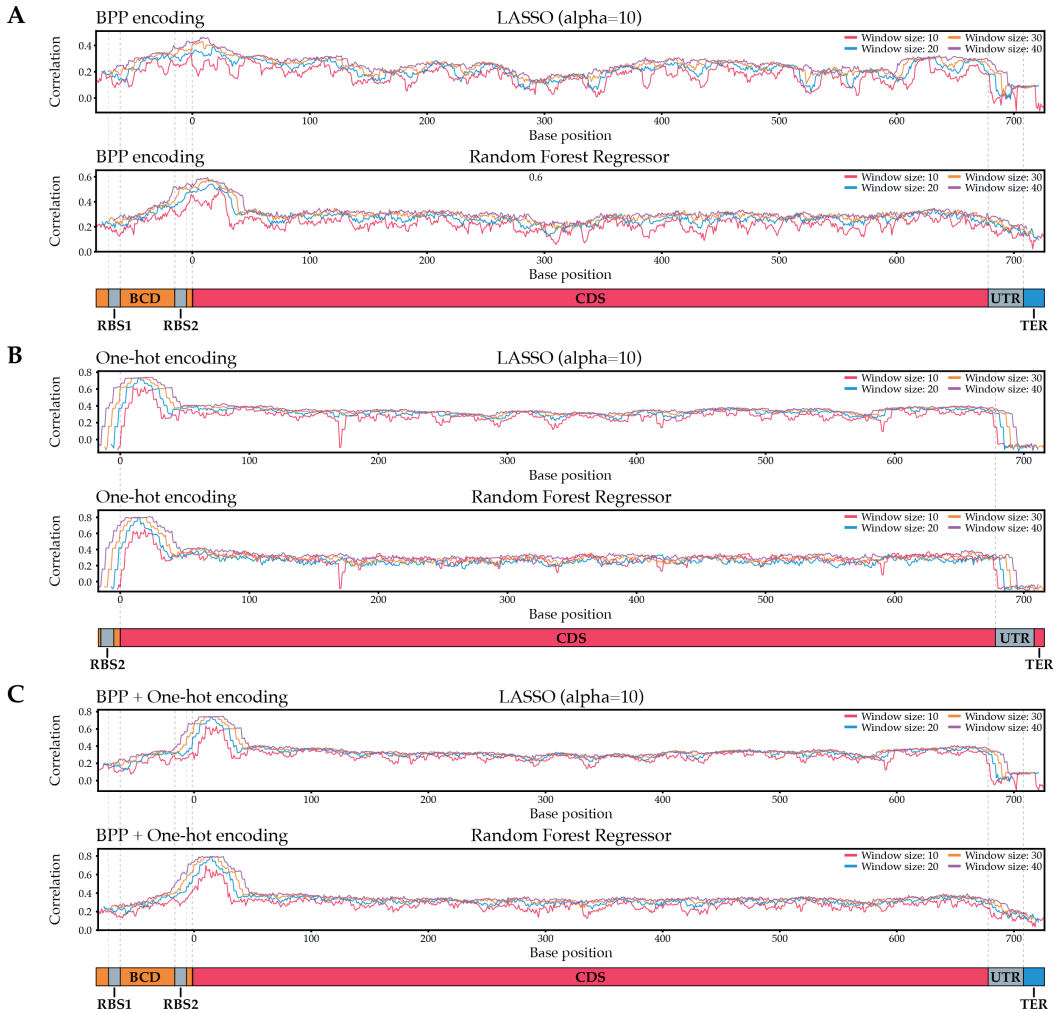


Figure 4.6. Predictive regions of translation efficiency in mRFP mRNA. The x-axis represents the central base of a sliding window of indicated lengths, the y-axis the correlation between actual expression data and the expression as predicted by a machine learning algorithm trained on solely that sliding window. A. Predictive regions found by algorithms trained with BPP featurisation. B. Predictive regions found by algorithms trained with one-hot encoding. As the one-hot encoded features for the UTRs are constant and thus contain no predictive information, windows that only contain residues in the UTRs were omitted. C. Predictive regions found by algorithms trained with BPP + one-hot featurisation.

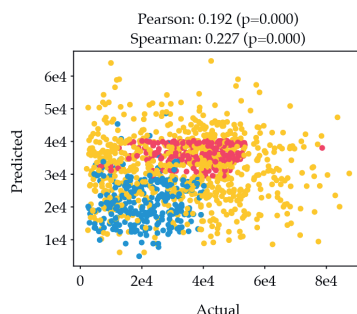


Figure 4.7. Models trained on downstream sequence windows make predictions based on the sequence library of origin. This plot represents the actual flow vs flow as predicted by a random forest regressor (one-hot encoded) through cross-validation on window 251 (window size 20). The clear divide between the CAI_L (blue) and CAI_H (red) libraries shows that the model's (limited) predictive power relies on its ability to distinguish between these two libraries. Since these libraries use completely different codon sets, and since data points from the CAI_H library display higher expression than data points from the CAI_L library on average, it makes sense that the model is able to infer from just a small sequence window which library the sequence originated from. This accounts for the non-zero Pearson correlation we observe for most sequence windows.

Of all our regressors, six random forest regressors outperformed the rest. As these six regressors were comparable in performance, we selected the model that used the fewest features and retrained the model using our full training set. We then plotted actual expression data against predicted expression for each data point in our leave-out test set. This revealed a strong correlation (Pearson $r = 0.762$) for all three libraries (Figure 4.9), which demonstrates that mRFP protein production can be correlated extremely well to sequence by just looking at bases 6–25 of the entire coding sequence. The bases 6–25 correspond to codons 2–8. The nucleotides in the third position of these codons that contribute to high expression are mostly A, as well as T in codon 2 (as shown earlier by the feature importances, Figure 4.5B and Figure 4.8B, C). An important question in the field is if the benefit of these codons at the start of the coding sequence is related to mRNA secondary structures and/or the efficient translation of these codons. To approximate which codons can be translated efficiently, indices such as the CAI and tRNA adaptation index (tAI)²⁸⁹ are commonly used. Interestingly, the beneficial codons ending in 'A' have a lower CAI and tAI index than other synonymous codons in most cases (Figure 4.10; Table 4.1), but still are highly important for high expression. This indicates that it is advantageous to use these codons for a reason that is unrelated to translation elongation efficiency. The most plausible mechanistic explanation is the lower secondary structure propensity of A/T-ending codons due to weaker A-T/U binding in general. This is further supported by our observations for the second codon: the T-ending variant is also related to high expression but is the 'least efficient' codon based on CAI and tAI. However, for the eighth (and last) codon of the important codon window, we observe the reverse: a C is favoured over A for high expression. The A-ending codon has very low tAI and CAI values, lower than any available codon in the first 8 codons. For this codon, the importance of 'optimal translation' seems to be dominant over the influence of secondary structure binding propensity of C/G-ending codons. Since this codon within our codon window is furthest away from the translation initiation region it is less likely to form detrimental RBS-obscuring secondary structures.

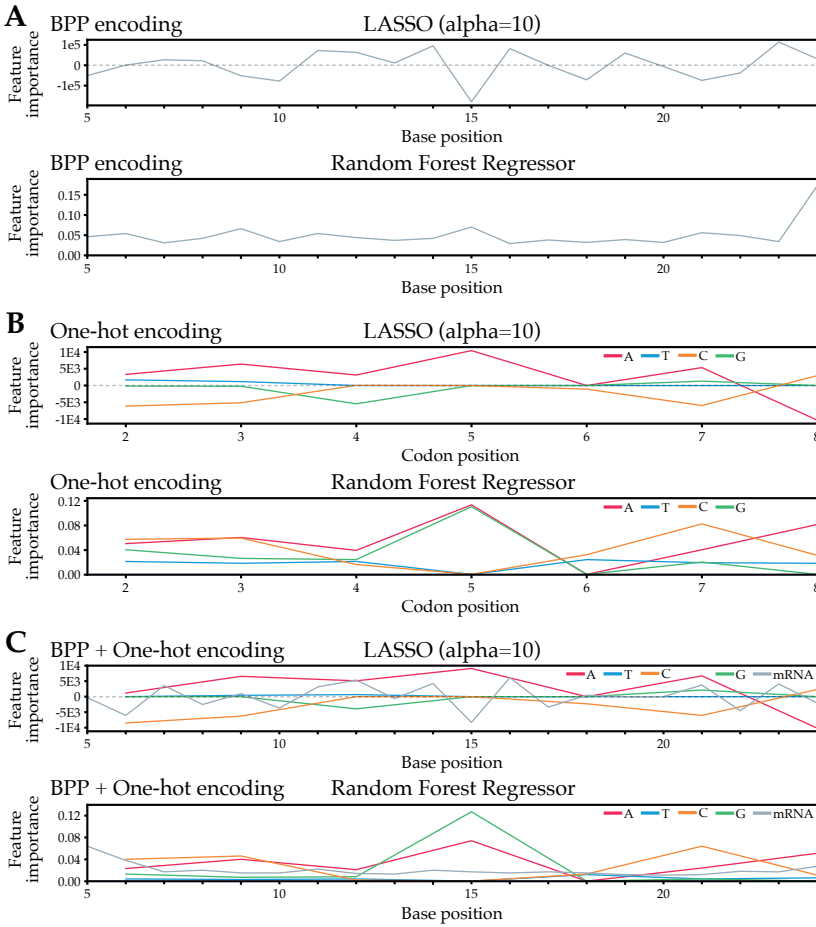


Figure 4.8. Feature importances for various machine learning algorithms and featurisations trained on a 20-base window around base 15. LASSO feature importances are coefficients: a positive coefficient indicates a positive correlation between a base and translation efficiency, a negative coefficient indicates a negative correlation. In RFR, feature importances are always positive and therefore it contain no information about the directionality of the correlation. A. Feature importances for algorithms using BPP featurisation. B. Feature importances for algorithms using one-hot encoding. Since only every third one-hot encoded base of the coding sequence varies, only every third base of the coding sequence was plotted. C. Feature importances for algorithms using BPP + one-hot featurisation.

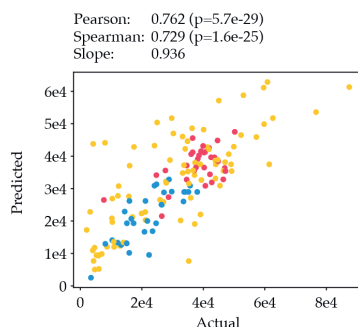


Figure 4.9. Actual vs predicted expression for one of our six best performing regressors. The RFR regressor trained on one-hot featurisation and a window size of 20 located at base 15 showed a high correlation between actual and predicted expression (Pearson correlation 0.762).

Because translation initiation is the major rate-limiting factor, the effects of codon usage throughout the gene seem less apparent. This is also exemplified by our finding that the highest expressing variants originate from our CAI_M library. This library displays greater codon variety than the CAI_H library, which is particularly important for the 5' of the CDS. Altogether, our results show that while the CAI has some influence on gene expression (Figure 4.2), the majority of translation regulation arises from codon usage in the 5' of the CDS. This matches the conclusion previously drawn by Kudla et al., who predicted a window of bases -4 to +37 (spanning codon 1–13) as a key determinant for GFP production. In our analysis, a slightly smaller window of codon 2–8 is sufficient to predict protein production. The same 8-codon window was also found in another systematic effort parallel to our study, using an alternative non-fluorescent reporter system (Bxb1 recombinase)⁴³. This suggests that the relevance of this window may be a general phenomenon, at least for gene expression in *E. coli*.

Our study also suggests that the 5' UTR may influence expression based on the observed BPP feature importance for the bases in that region. However, this region was kept constant in this work on purpose. The parallel study by Höllner et al. randomised bases -25 to -1 in the 5' UTR alongside codons 2-16 of their reporter protein Bxb1. Indeed, this study confirmed a large contribution to expression variance from both the 5' UTR (50%) and the CDS (20%)⁴³.

Due to the partial black box nature of machine learning, design rules for the 5' CDS are not fully apparent. However, our analysis suggests that secondary structure is likely a key determinant of translation efficiency for these codons. It is clear that if high protein production is desired, the focus should be on optimising the start of the coding sequence and the 5' UTR in *E. coli*, and possibly in other bacterial hosts. Typically, codon optimisation algorithms and approaches optimise for parameters such as CAI over the full CDS but ignore the 5' UTR sequence, and therefore may introduce detrimental secondary structures. We suggest a shift in these approaches to specifically tackle optimisation of the first ~8 codons. Some previous studies have proposed randomisation approaches in the 5' UTR and/or the first codons^{270,290}, but these approaches are currently barely applied. Alternatively, existing or new in silico design tools considering secondary structures in the 5' UTR-CDS start region (such as RBS Calculator²⁹¹) could be considered to improve gene expression. This systematic study provides a clear rationale for adopting these methods, rather than commonly used whole-gene codon optimisation algorithms, to improve protein production.

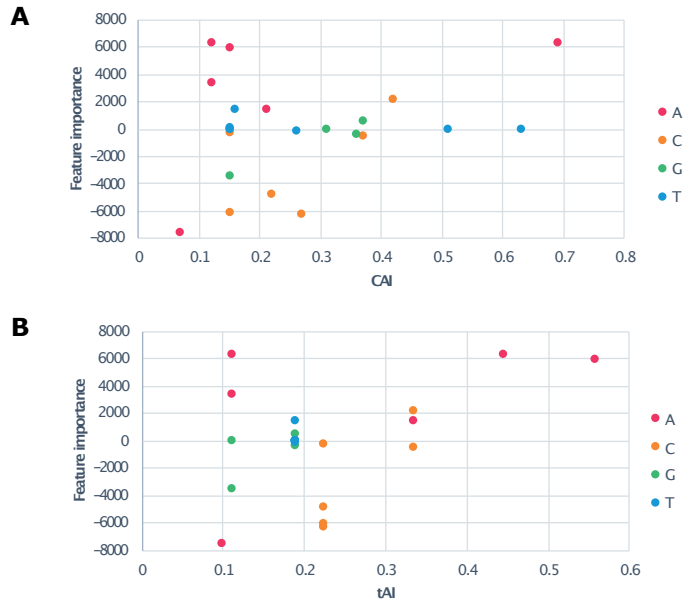


Figure 4.10. The relationship between feature importance and CAI and tAI. The feature importance of the nucleotide identity at the third position of codon 2-8 compared to the CAI of these codons. B. The feature importance of the nucleotide identity at the third position of codon 2-8 compared to the tAI of these codons. CAIs were calculated based on all *E. coli* K-12 coding sequences (RefSeq 015291845.1). tAI scores were obtained from <http://tau-tai.azurewebsites.net> for *E. coli* K-12. Feature coefficients came from our best-performing one-hot encoded LASSO model.

Table 4.1. Feature importance, CAI, and tAI score for each synonymous codon at codon positions 2-8. Green and red indicate a positive and negative feature importance respectively. The red/green shading for the CAI and tAI indicate the position between the full range score of 0 and 1. The CAI scores per codon were calculated based on all *E. coli* K-12 coding sequences (RefSeq 015291845.1). The tAI scores for each codon were obtained from <http://tau-tai.azurewebsites.net> for *E. coli* K-12. NA indicates this synonymous codon variant does not encode for this amino acid and is left out of the library. Feature coefficients came from our best-performing one-hot encoded LASSO model.

Feature importance							
Codon Position	2	3	4	5	6	7	8
Codon	GCN	TCN	TCN	GAR	GAY	GTN	ATH
Amino Acid	A	S	S	E	D	V	I
A	1461	6347	3409	6394	NA	6001	-7523
C	-6249	-6037	-191	NA	-455	-4775	2166
G	-333	0	-3441	0	NA	549	NA
T	1421	0	78	NA	0	-149	0
CAI							
A	0.21	0.12	0.12	0.69	NA	0.15	0.07
C	0.27	0.15	0.15	NA	0.37	0.22	0.42
G	0.36	0.15	0.15	0.31	NA	0.37	NA
T	0.16	0.15	0.15	NA	0.63	0.26	0.51
tAI							
A	0.333	0.111	0.111	0.444	NA	0.556	0.097
C	0.222	0.222	0.222	NA	0.333	0.222	0.333
G	0.188	0.111	0.111	0.188	NA	0.188	NA
T	0.188	0.188	0.188	NA	0.188	0.188	0.188

Chapter 5

Predicting NRPS A-domain specificity with PARAS and PARASECT: two structure-informed machine learning algorithms

Barbara R. Terlouw, José D.D Cediél-Becerra, Shanshan Zhou, David Meijer, Chris Fage, Lona Alkhalaf, Matthew Jenner, Serina Robinson, Gregory Challis, Marc G. Chevrette, and Marnix H. Medema

Abstract

Non-ribosomal peptides (NRPs) are a chemically and functionally diverse class of tailored peptide natural products. Their structures are largely governed by the substrate specificity of the adenylation (A) domains of their synthesising enzymes. Therefore, we can predict the core structure of NRPs by predicting these substrate specificities. Until now, machine learning algorithms that predict A-domain specificity have been trained using limited training data and have used sequence alignments to extract active site features for model training. However, it is ultimately the 3D structure of the active site that governs selectivity. Here, we demonstrate that two phylogenetically distinct A domains recognise their substrate tryptophan in different orientations, underlining a need for methods trained on sufficient training data that make use of both sequence and structural data to capture this natural diversity. Then, we describe the development of two structure-informed A domain predictors, PARAS and PARASECT. With three times more training data, a performance increase of 27.4%, and a runtime two orders of magnitude faster compared to state-of-the-art tools, PARAS and PARASECT pave the way towards better NRP scaffold predictions, which will help accelerate natural product research.

PARAS and PARASECT are available at <https://github.com/BTheDragonMaster/parasect>.

5.1. Introduction

Bacteria and fungi synthesise a multitude of natural products, including the highly chemically and functionally diverse class of non-ribosomal peptides (NRPs). The most well-known example of a NRP is penicillin, a modified Ala-Cys-Val tripeptide which was the first antibiotic to be discovered¹³. Since, many more naturally occurring NRPs have been leveraged for pharmaceutical and societal applications, such as the antibiotic daptomycin, the immunosuppressant cyclosporine²⁹², the fungicide UK-2A³¹, and the herbicide Bialaphos⁶².

As their name suggests, NRPs are not synthesised by ribosomes. Instead, they are assembled by non-ribosomal peptide synthetases (NRPSs): large, multi-modular enzymes that build peptides in an mRNA-independent manner. To incorporate an amino acid building block, an NRPS module minimally requires three domains: a condensation (C) domain, which catalyses the condensation reaction between the carboxylic acid group of the peptide scaffold and the amino group of the to-be-incorporated amino acid; an adenylation (A) domain, which selects the building block through a lock-and-key interaction with the amino acid substrate; and a non-catalytic peptidyl carrier protein (PCP) domain, which functions as a covalent attachment point for the peptide scaffold or amino acid building blocks⁴⁶. This mRNA-independent assembly method has as a consequence that, while ribosomally-synthesised peptides are limited to the standard 20 proteinogenic amino acids, NRPs can incorporate over 300 different building blocks⁶¹, including proteinogenic and non-proteinogenic amino acids, N-terminal acids and C-terminal amines. This, combined with a large array of tailoring enzymes that can modify the peptide scaffold, yields a vast chemical space with immense functional potential.

Now that genome sequencing costs are at an all-time low, genome mining has assumed a key role in natural product discovery. With tools like antiSMASH⁶⁷, researchers can explore sequenced genomes for biosynthetic gene clusters (BGCs): genetic units that encode biosynthetic pathways for specialised metabolites. Then, they can examine individual genes in the BGC for clues about the structure of the

produced metabolite. In the case of NRPS BGCs, the peptide scaffold structure is usually approximated by predicting amino acid building blocks based on A domain sequence.

In the late 1990s, two groups simultaneously worked on identifying residues in the A domain active site that guide substrate selectivity. Based on crystal structures of a phenylalanine-recognising A domain from the NRPS enzyme gramicidine S synthetase A (GrsA), they agreed on a core of nine amino acid residues in the binding pocket⁷⁵, with Stachelhaus *et al.* reporting an additional tenth⁴⁴. The 10 residues reported by Stachelhaus *et al.* are commonly referred to as the Stachelhaus code. Generally, it is accepted that if two A domains share identical Stachelhaus codes, they likely recognise identical substrates. Therefore, it has become a common strategy to extract the Stachelhaus code from an A domain by aligning its sequence to the A domain from GrsA, and to predict its selectivity by comparing it to the Stachelhaus codes of A domains of known specificity. Such one-to-one Stachelhaus code comparison is still used by many researchers today to predict the selectivity of A domains.

It is important to note that the Stachelhaus code includes the highly conserved D235 and K517 residues, which stabilise the amino and carboxyl groups of the amino acid substrate backbone respectively. As some NRPS A domains also recognise other acids such as benzoic acid, which lack the amino group, substitutions of the D235 residue occur, and therefore the residue at this position can still be informative. Substitution of the K517 residue is much rarer. Also, K517 is the only residue located in the small C-terminal subdomain of the A domain, while the other 9 are located in the larger N-terminal subdomain which maps to a separate Pfam profile Hidden Markov Model (pHMM). For these reasons, K517 is often ignored in Stachelhaus code comparisons.

In 2005, Rausch *et al.* chose an A domain prediction approach that was more broadly applicable. They extended the Stachelhaus code to a 34-residue signature, which represents all residues within 8 Angstrom (Å) of the GrsA A domain active site, and were the first to apply a machine learning framework to this prediction task. Importantly, they featurised the residues of the active site as vectors of 15 physicochemical properties relating to hydrophobicity, charge, and size among other features. Because of this, their transductive support vector machine NRPSPredictor, a semi-supervised machine learning model that is able to use unlabelled datapoints to better characterise sequence space, is able to take into account the similarity between active site residues while training the model. Röttig *et al.* later released an improved version of this tool, NRPSPredictor2⁷⁷, which was trained on an extended dataset and includes a predictor for fungal A domains as well. Robinson *et al.* used a similar approach to train a random forest model, AdenylPred, that makes predictions for non-NRPS A domains as well⁸².

Khayatt *et al.* employed a different sequence-based method for A domain selectivity prediction: it uses profile HMMs (pHMMs) for various possible A domain substrates, which an A domain of interest is queried against⁷⁹. The substrate corresponding to the best-matching pHMM is reported. The PRISM 4 software suite, which detects BGCs and predicts the subsequently predicts product scaffold of produced natural products, employs a similar approach⁷⁸.

Chevrette *et al.* leveraged all these different methods by building the ensemble algorithm SANDPUMA⁸⁰, which combines the outputs from six different prediction methods in a decision tree to generate a consensus selectivity. These prediction methods include various of the previously mentioned approaches, as well as two phylogeny-based prediction methods (prediCAT) created by

the authors. They also greatly expanded their training set compared to previous models and achieved the best accuracy thus far.

While there are many A domain prediction methods available, each has their shortcomings. Stachelhaus comparison is limited to A domains which share a Stachelhaus code with a known A domain, as partial Stachelhaus code matches are often a poor indicator of substrate selectivity. Also, there are examples of A domains that share identical Stachelhaus codes and 34 residue signatures, which recognize different substrates. Quite often, A domains with identical binding pockets recognise different but similar substrates. For example, the first two A domains of ATY37590.1, an NRPS enzyme encoded by the bogorol BGC, have identical active sites but recognise isoleucine and valine, respectively. More substantial substrate differences are rarer, but do exist: for instance, the first A domain of the protein ALV86867.1 in the telomycin BGC recognises alanine, while four A domains with identical 34 residue signatures in the NRPS enzymes Omn6 and Omn7 in two omnipeptin BGCs (GenBank accessions QEO74981.1, QEO74982.1, QEO75073.1, and QEO75075.1) select serine. A comprehensive list of examples is displayed in Table 5.1. It is thought that other factors may influence the selectivity of such A domains, such as residues outside the active site, substrate availability in the case of rare substrates, and C domain proofreading activity.

NRPSPredictor2 and AdenylPred also struggle with such A domains, as their underlying support vector machines and random forests respectively can only differentiate between substrates based on the 34-residue signature. SANDPUMA's ensemble architecture is less constrained by these limitations, as it also employs phylogeny to make its decisions; however, it is comparatively slow, with a single prediction taking up to two minutes to complete. As a standalone method, the pHMM approach employed by Khayatt *et al.* and PRISM 4 was shown to perform quite poorly, likely because the 9-34 active site residues only represent a small portion of the entire A domain. As such, pHMMs can mispredict the selectivity of A domains with high sequence identity to a datapoint in the training set but have a different active site architecture, as frequently occurs between A domains within a single BGC.

So far, all A domain selectivity predictors rely heavily on sequence alignments: they are required for the development of pHMMs^{78,79}, the construction of phylogenetic trees⁸⁰, and the extraction of the A domain active site^{77,80,82}. This could be an issue, as sequence alignments typically do not work very well for the alignment of highly diverse sequences such as the A domain active site. Specifically, sequence-based active site extraction of highly divergent sequences can lead to the introduction of gaps that do not and cannot actually exist in the active site. Structure-based alignments would be better, as these align residues based on the (predicted) position of residues in 3D space. Especially now that AlphaFold2 provides a quick and highly accurate method for obtaining predicted protein structures²⁹³, structure-guided alignments could provide a realistic solution.

Still, the greatest limitation of all above methods is a lack of training data, with SANDPUMA boasting the largest dataset of only ~1000 datapoints⁸⁰, some of which are duplicates. This is especially problematic as A domains with very different active site residues can recognise identical substrates, highlighting a need for a large and diverse dataset that covers as many A domain clades as possible.

Finally, none of the A domain prediction methods produce multi-label outputs, which would be very beneficial, as many A domains are promiscuous and select multiple different substrates^{294,295}. At best, such A domains are currently only labelled with one of their substrates or not included at all; at worst,

they are given their own label class. Finally, current machine learning algorithms only consider features of the A domain and not the substrate^{77,79,80,82}, meaning that the algorithms cannot learn from similarity and dissimilarity between substrates.

In this work, we will first discuss experimental work on a tryptophan-recognising adenylation domain that showcases that different A domains may recognise the same substrate in different orientations, suggesting that some A domain specificities may have evolved multiple times and that structural features should be considered alongside sequence features for training A domain selectivity predictors. Then, we will describe the development of two fast and accurate structure-informed machine learning frameworks: PARAS (Predictive Algorithm for Resolving A domain Selectivity) and PARASECT (Predictive Algorithm for Resolving A domain Selectivity by featurising Enzyme and Compound in Tandem), which aim to overcome the challenges described above by being trained on an expanded and carefully curated dataset and by using structure-guided alignments to extract the A domain active site. PARASECT is the first multi-label A domain selectivity predictor, which is trained on both the sequence and structural features of the A domain and the molecular features of the putative substrate and predicts whether the two interact. Finally, we benchmark PARAS and PARASECT against the current state-of-the-art adenylation domain predictor SANDPUMA and show a substantial increase in performance (27.4%).

5.2. Results and Discussion

5.2.1. A comparison of *TtpB* and *TycB*: two tryptophan-selecting adenylation domains

When assessing the performance of current state-of-the-art adenylation domain predictors, we noticed that they disproportionally made errors in the classification of large proteinogenic amino acids such as phenylalanine, tryptophan, and lysine⁷⁷ (F1-score of 0.688, 0.320 and 0.400 in NRPSPredictor2, respectively). We hypothesised that these substrate selectivities possibly evolved multiple times, and that due to the size and degrees of rotational freedom of their side chains, convergent evolution could lead to active site pockets that select identical substrates but have active site pockets with entirely different orientations. To test this hypothesis, we decided to computationally compare the active sites of two tryptophan-recognising adenylation domains from two phylogenetically distinct branches using AlphaFold models and molecular docking. The two domains we chose were the third A domain from *TycB* (TycB-A3), an NRPS enzyme involved in tyrocidine biosynthesis in *Brevibacillus brevis*⁴⁷, and the third A domain from the NRPS enzyme *TtpB* (TtpB-A3), which is involved in the biosynthesis of tryptopeptin in *Streptomyces sparsogenes*²⁹⁶. We chose these domains as their Stachelhaus codes are highly divergent: DAWTIAGVCK for TycB-A3, and DASVVGCVTK for TtpB-A3, which only overlap by four amino acids, two of which are the highly conserved aspartic acid and lysine that stabilise the amino acid backbone in almost all NRPS A domains. While assays have been done to assess the selectivity of TycB-A3, which recognises predominantly tryptophan and phenylalanine to a lesser extent⁴⁷, we still had to verify the selectivity of TtpB-A3.

Table 5.1. A domains with the same amino acid signature but different substrates. Protein IDs are GenBank, RefSeq or UniProt accessions. BGC IDs are given for proteins which occur in BGCs in the MIBiG database⁷².

Protein ID	Domain number	Selectivity	BGC ID	Produced compound
Signature: LGLAFDASVQQVDCLVGGEYNVYGPTTECTVDTTTC				
AJF34463.1	4	glutamine	BGC0001207	teixobactin
WP_047197778.1	2	glutamic acid	BGC0001608	glidobactin
Signature: RWMFTDVSFVWEWHFICSGEHNLYGPTAESVDVTY				
ALV86867.1	1	alanine	BGC0001406	telomycin
QED88054.1	2	serine	BGC0001967	ADEP1
QEO74982.1	2	serine	BGC0002078	omnipeptin
QEO74981.1	3	serine	BGC0002078	omnipeptin
QEO75073.1	3	serine	BGC0002079	omnipeptin
QEO75075.1	2	serine	BGC0002079	omnipeptin
Signature: SGSAFPSSPGGALQVGGAAQQVFGMAEGLVNYTR				
AAO07759.1	1	salicylic acid, 2,3-dihydroxybenzoic acid	BGC0000460	vulnibactin
QRK05499.1	1	benzoic acid, nicotinic acid	BGC0002324	myxochelin
Signature: TNNSFDGSTFDSFMVFGGEVNGYGPTETTVFATT				
AAV29580.1	2	valine		BT peptide
AAV29580.1	3	valine		BT peptide
ATY37590.1	3	valine	BGC0001532	bogorol
ATY37591.1	1	valine	BGC0001532	bogorol
ATY37591.1	2	leucine	BGC0001532	bogorol
Signature: TNNSFDGSTFEGFMLFGGEINVYGPTTESTVYATY				
ATY37590.1	1	isoleucine	BGC0001532	bogorol
ATY37590.1	2	valine	BGC0001532	bogorol
Signature: YRASFDLTVTASKVVIGGEINAYGPTEATVNCAE				
ARU08074.1	2	3-hydroxyaspartic acid	BGC0001448	malacidin
ARU08074.1	3	aspartic acid	BGC0001448	malacidin
Signature: LNDHFDVSVWEGNQIFGGEINMYGITETTVHVTY				
ACA97576.1	2	threonine	BGC0000408	polymyxin
ACA97580.1, AEZ51520.1, B9TTF7	2	threonine	BGC0000408, BGC0001153	polymyxin
ACA97577.1	1	threonine	BGC0000408	polymyxin
ABQ96384.2	1	threonine	BGC0001152	fusaricidin
ABQ96384.2	4	threonine	BGC0001152	fusaricidin
AEZ51517.1	1	threonine	BGC0001153	polymyxin
AJM89734.1	1	threonine	BGC0001192	colistin
AJM89738.1	2	threonine	BGC0001192	colistin
ATY37588.1	1	threonine	BGC0001532	bogorol
AAV29578.1	1	4R-E-butenyl-4R-methylthreonine		BT peptide
A9LJA2	1	threonine		fusaricidin
A9LJA2	4	threonine		fusaricidin

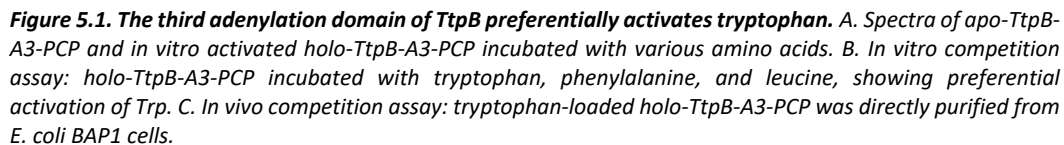
TtpB-A3 preferentially selects tryptophan

Usually, ATP-Pi exchange assays are used to test the selectivity of A domains: as ATP is converted to AMP during substrate loading, one can measure the release of radio-labelled P_i to infer the efficiency of the adenylation reaction²⁹⁷. However, these assays only measure the first reaction step; in the second step, thiolation, the adenylylated substrate is transferred to the phosphopantetheine arm that is post-translationally attached to a serine residue on the PCP domain, releasing AMP. As we wanted to ensure that both reaction steps occur during substrate activation by TtpB-A3, we decided to study the A-PCP didomain of the third module of TtpB (TtpB-A3-PCP) rather than the standalone A domain. To this purpose, we heterologously expressed TtpB-A3-PCP in two *E. coli* strains: TOP10, which is suitable for protein overexpression in general; and BAP1, a strain that is specifically designed for natural product biosynthesis²⁹⁸. Relevant to this work, *E. coli* BAP1 expresses the gene encoding the phosphopantetheinyl transferase Sfp, which posttranslationally modifies PCP domains with a phosphopantetheine arm. Therefore, TOP10 cells should be able to produce *apo*-TtpB-A3, while BAP1 cells produce *holo*-TtpB-A3 that can already activate and load substrate onto the PCP domain *in vivo*, given that the preferred substrate is available to the organism.

After overexpressing and purifying *apo*-TtpB-A3 from *E. coli* TOP10, we confirmed the mass of the didomain through mass spectrometry (Figure 5.1A, top row). Then, we converted *apo*-TtpB-A3 to its *holo* form (Figure 5.1A, second row) and used the *holo*-didomain to test the activation of five aromatic and hydrophobic substrates: tryptophan (Trp), phenylalanine (Phe), leucine (Leu), valine (Val), and histidine (His; Figure 5.1A, rows 3-7). When *holo*-TtpB-A3 was incubated with the substrates in isolation, it displayed excellent activation of Trp, and Phe, some activation of Leu, very little activation of His and no activation of Val. However, when we incubated *holo*-TtpB-A3 with Trp, Phe, and Leu at the same time, it exclusively activated Trp (Figure 5.1B), suggesting that this is the substrate that is recognised *in vivo*. We confirmed this by purifying pre-loaded *holo*-TtpB-A3 from BAP1 cells, which out of all available amino acid substrates in *E. coli* preferentially activated Trp (Figure 5.1C). This provides sufficient evidence that TtpB-A3 indeed catalyses both reactions that lead to the activation of Trp.

5.2.1.1. TycB-A3 and TtpB-A3 recognise tryptophan in different orientations

Having confirmed that TtpB-A3 also selects tryptophan, we then set out to compare the 3D architectures of the active sites of TtpB-A3 and TycB-A3 to determine how two highly divergent active site sequences could lead to the recognition of the same substrate. To this purpose, we built AlphaFold models²⁹³ for both domains and subsequently performed molecular docking using either tryptophanyl-adenylate (TYM) or AMP and Trp as ligand(s). While the docking of ligands onto static structures will remain an estimation of ligand-protein interaction, our results clearly showed that it is improbable that the two domains recognise Trp in the same orientation (Figure 5.2): the TYM ligand docked to TtpB (Figure 5.2B) is rotated with respect to its α -carbon and β -carbon compared to TycB-A3-docked TYM (Figure 5.2A), shifting the aromatic rings of the side chain slightly towards the adenylyl moiety of TYM and rotating it 180° such that the aromatic nitrogen is positioned 7.1 Å away from its corresponding position in TycB-A3 (Figure 5.2D). When looking at the domains themselves, clear differences can be seen in the lobes of the active site pocket. In contrast, the channel that binds the adenylyl moiety of the substrate are far more constant, as also reflected by the similar way in which this portion of the TYM substrate is docked in both domains. This lends credibility to our hypothesis that there may be different orientations in which A domains interact with large substrates and underlines why Stachelhaus code comparison is not a reliable method of determining substrate selectivity.



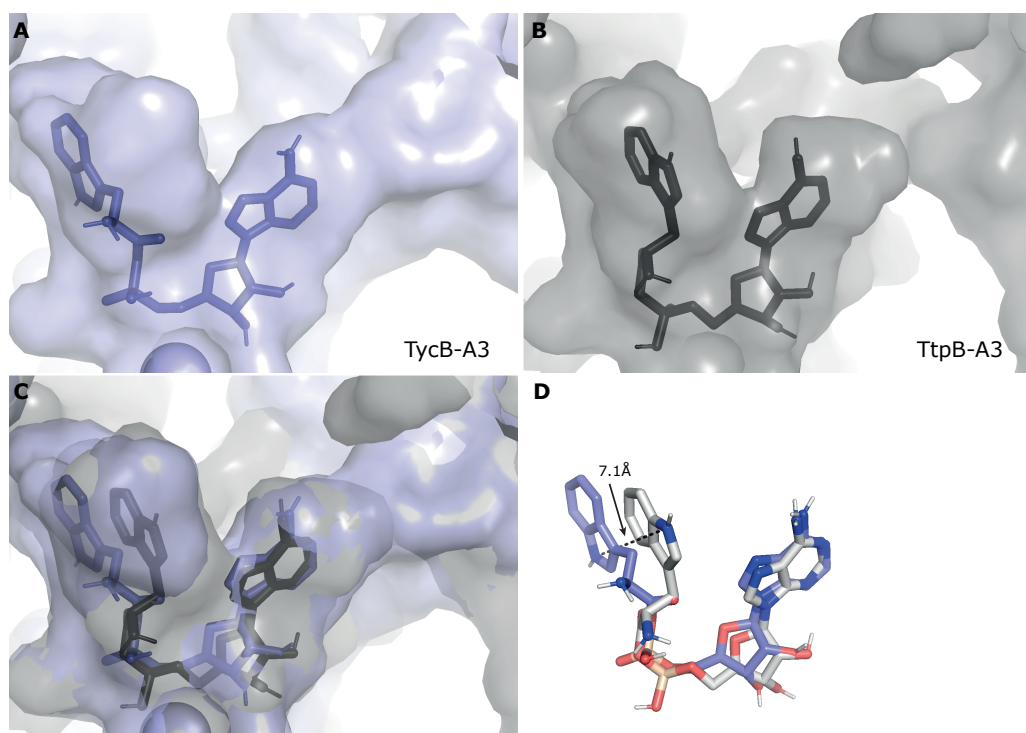


Figure 5.2. TycB-A3 and TtpB-A3 recognise Trp in different orientations. A. Tryptophan-adenylate (TYM) docked to TycB-A3. B. TYM docked to TtpB-A3. C. Comparison of TYM docked to TycB-A3 (blue) and TtpB-A3 (grey). D. The tryptophan moiety of TYM docked to TtpB-A3 (white) is shifted slightly towards the adenylate moiety and has undergone a rotation of its aromatic rings of 180° compared to TycB-A3-docked TYM (blue), leading to a positional shift of the aromatic nitrogen of 7.1\AA (yellow dashed line).

5.2.2. A structure-based approach to predicting A domain selectivity

As our analysis of the TtpB-A3 and TycB-A3 domains showed, the three-dimensional architecture of the active site determines not only which substrate is activated, but also in which orientation. Therefore, we wanted to explore various structure-based approaches to predicting A domain selectivity.

5.2.2.1. Data collection

While the chosen featurisation method and machine learning approach can affect model quality, data quality and quantity typically have an even greater impact on model performance. For this reason, we collated a dataset from three sources: MIBiG 3.0⁷², the SANDPUMA training set⁸⁰, and the NRSPredictor2 training set⁷⁷. The resulting dataset counted 3227 datapoints, which we extensively curated and validated prior to model training (see Methods). Of these, 381 domains exclusively recognised substrates that we had too few examples of ($n < 11$) in our dataset to validate model performance on these substrates. Therefore, we only used the remaining 2846 in model training, which still accounts for 2.6 and 5.3 times as many labelled datapoints as were used for training

SANDPUMA (1089 datapoints) and NRPSPredictor2 (534 datapoints), respectively. During curation, we also discovered errors in the SANDPUMA and NRPSPredictor2 datasets (137 [12.6%] in SANDPUMA, 31 [5.8%] in NRPSPredictor2), which may have impacted the performance of these models^{77,80}.

5.2.2.2. *Structure-based extraction of the A domain active site improves the performance of predictive models*

As structural modelling is time-intensive, we wanted to examine structure-informed methods that do not require the modelling of domains that the user wants to query. For this reason, we first focussed our attention to structure-based extraction of the 34 amino acid active site signatures as also employed by NRPSPredictor2⁷⁷. Current active site extraction methods rely on profile alignments, which use a sequence alignment as guide and append a query sequence to the alignment such that residues of interest can be identified and extracted. It can be challenging to obtain a sequence alignment for A domains that adequately aligns the active site: this region of the domain is highly variable, especially the residues of interest, and it can therefore be difficult to determine from sequence alone which residues truly belong to the active site. In particular, we observed that active site extractions made using sequence-guided alignments contained relatively many gaps, while typically one would not expect gaps to occur in an active site. We reasoned that a structure alignment might resolve this, as it uses the three-dimensional position of residues to align sequences rather than residue identity to construct alignments. Indeed, active sites extracted using a structure-based alignment contained ten times fewer gaps, with an average of 0.06 gaps per active site signature for structure-guided extraction compared to 0.58 gaps per active site signature for sequence-guided extraction. For the Stachelhaus code this difference was much smaller: 0.03 gaps per active site for structure-guided extraction compared to 0.04 gaps for sequence-guided extraction. We also investigated mismatches, and found that on average, 1.01 residues were different for each active site signature (1.65 including gaps), and 0.12 for each Stachelhaus code (0.20 including gaps). Most of these mismatches were accounted for by alignments to tryptophan (Figure 5.3A, C), which were often aligned against gaps or different residues. This makes sense, considering that the BLOSUM62 matrix, which is typically used for local sequence alignments, highly penalises mismatches with tryptophan. We also examined how often each residue of the Stachelhaus code and active site signature contained a mismatch. We observed that residues 278 and 331 (indexed according to the reference structure 1AMU) accounted for most mismatches in the Stachelhaus code, and residues 213 and 214 contributed most towards mismatches in the active site signatures (Figure 5.3B, D).

To see if these differences led to differences in predictive power, we built homology models for a random subset of our dataset – a subset as profile alignment scales poorly with the number of sequences – and built ten different structure alignments for cross-validation purposes, each using 90% of the data. Then we trained basic predictive models to assess if there was a difference in performance when the active sites were extracted using sequence alignments or structure alignments as guide. On average, models trained on active sites that were extracted using structure-based guide alignments outperformed those that were trained on residues extracted from sequence-based guide alignments by about 1.7% (paired t-test: $p=0.03$; Table 5.2). This demonstrates that it is still possible to leverage structural information to improve model performance, even when structural data of the queried domain is not directly used. We opted to use structure-based guide alignments for identifying active site residues going forward.

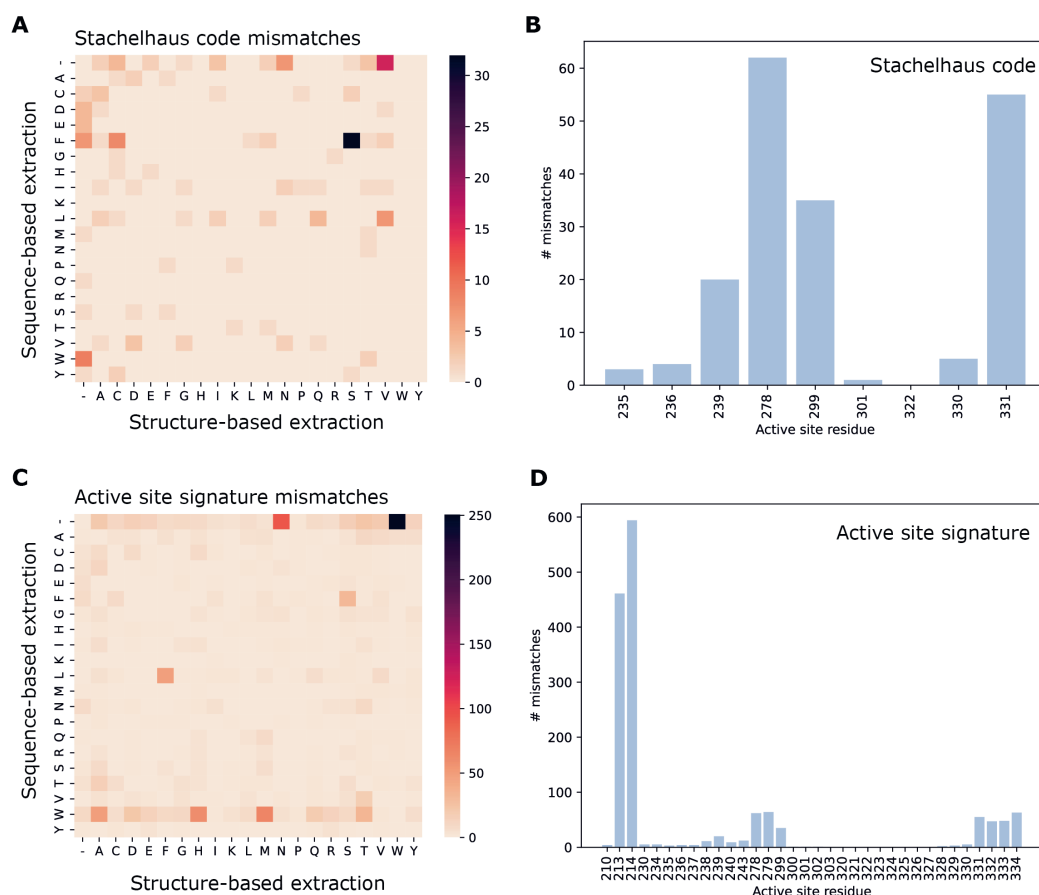


Figure 5.3. Differences between active sites extracted using sequence-based and structure-based approaches. A. Stachelhaus mismatches between sequence-based and structure-based extraction per residue type. B. Stachelhaus mismatches per active site position. C. Active site signature mismatches between sequence-based and structure-based extraction per residue type. D. Active site signature mismatches per active site position. Active site positions are labelled according to positions in the reference structure 1AMU.

5.2.2.3. PARAS and PARASECT: two structure-informed adenylation domain predictors

To provide the natural products community with improved methods to predict the substrate selectivity of A domains, we developed two novel, structure-informed random forest algorithms, PARAS and PARASECT. PARAS is a single-label classifier that always outputs one of 34 different substrates (Figure 5.4), which uses only domain features as input. In contrast, PARASECT predicts the probability that an A domain and a substrate interact. PARASECT can be run in two modes. In automatic mode, PARASECT automatically queries an A domain against each of the 34 substrates that it was trained on and returns the substrate which has the greatest probability of interacting with the A domain. In this mode, it is also possible to request multiple top predictions. In manual mode, the user can provide a list of substrate names or SMILES strings and specifically query whether a domain is likely to interact with those substrates.

Table 5.2. Structure-guided active site extraction boosts model performance. We assessed the performance of ten random forest models following structure-guided active site extraction or sequence-guided active site extraction. Per substrate class, datapoints were randomised and subsequently distributed across the validation sets in order. As a result, validation sets with low indices contain a disproportionate number of datapoints for which there were few or no examples in the training set. This explains the drop in model performance in these sets.

Validation set	Structure-guided alignment	Sequence-guided alignment
1	0.656	0.640
2	0.785	0.710
3	0.790	0.800
4	0.763	0.742
5	0.833	0.811
6	0.874	0.862
7	0.894	0.859
8	0.841	0.854
9	0.924	0.911
10	0.821	0.821
Average	0.818 (standard deviation = 0.076)	0.801 (standard deviation = 0.0818)

For both PARAS and PARASECT, we explored three different approaches for featurising the A domain: the first uses the sequence features of the 34 residues of the active site, extracted using a structure-guided alignment; the second uses a set of 68 principal components, which we used to represent a much larger feature set of 56000 derived from our homology models; and the third combines both approaches. Prior to training, we divided our dataset into train and test sets in two different ways: by stratifying on substrate class, and by stratifying on the phylogeny of the adenylation domains by clustering our domains on bitscore prior to data stratification. This enables us to assess how well our algorithms will perform on sequences that are phylogenetically distinct from datapoints in the training set. The average % identity of a test datapoint to its closest relative in the training set was $75.3\pm 17.14\%$ for the class-based split and $58.2\pm 16.3\%$ for the phylogeny-based split. In both cases, the ratio of training datapoints to test datapoints was approximately 3:1.

On our class-stratified dataset, our single-label predictor PARAS performed slightly better than PARASECT in automatic mode (Figure 5.5A, B), with a peak performance of 89.4% compared to PARASECT's 86.4%. This makes sense when we measure accuracy based on the first returned label, as PARAS was specifically trained to return a single label while PARASECT has to infer the best label from a set of interaction probabilities. However, unlike PARASECT, PARAS cannot determine if there are any secondary substrates that the A domain may also select.

As expected, both PARAS and PARASECT performed better on class-based data splits (Figure 5.5A) than on phylogeny-based data splits (Figure 5.5B), demonstrating the importance of stratifying based on phylogeny when training models on sequence data. On these phylogenetic splits, PARAS and PARASECT still achieved average accuracies of 77.4% and 73.7%, respectively.




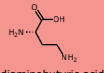









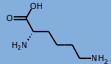
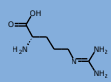
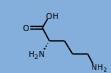
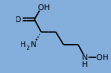
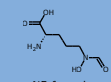
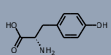
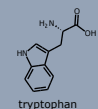
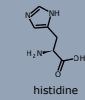
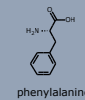
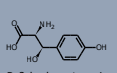



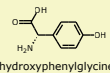
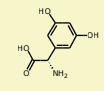






Small polar  cysteine  threonine  serine  2,4-diaminobutyric acid	Small  glycine  alanine  D-alanine  2-aminoisobutyric acid	Asx/Glx  aspartic acid  asparagine  glutamic acid  glutamine  2-aminoadipic acid	N-containing  lysine  arginine  ornithine  N5-hydroxyornithine  N5-formyl-N5-hydroxyornithine	Aromatic  tyrosine  tryptophan  histidine  phenylalanine  R-β-hydroxytyrosine	Acids  anthranilic acid  2,3-dihydroxybenzoic acid  salicylic acid Phenylglycines  4-hydroxyphenylglycine  3,5-dihydroxyphenylglycine
Cyclic aliphatic  proline  pipecolic acid	Small hydrophobic  β-alanine  valine  isoleucine  leucine				

Figure 5.4. An overview of the substrates that PARAS and PARASECT can predict. While over 200 substrates are represented in our raw dataset, we only have sufficient examples for 34 substrates (inclusion limit: 11). Coloured rectangles indicate the substrate groups that we used for PCA.

We additionally investigated if the average prediction confidence/interaction probability correlated with datapoint correctness, and we found that it did: on average, PARAS is 51.9-52.9% more confident about true positives than about false positives, and PARASECT predicts a 22.9-37.7% greater interaction probability for true positives with respect to false positives. To set up recommendations for users that indicate when predictions are trustworthy and when they are not, we also plotted accuracy against confidence score for PARAS (Figure 5.6A), and made a precision-recall curve for PARASECT, using the F1 score to determine an optimal threshold (Figure 5.6C). For PARAS, we recommend to trust predictions with a confidence of at least 50%, as prediction accuracies beyond this range lie between 80-90% (Figure 5.6A); and for PARASECT, we recommend to trust interaction probabilities with a lower bound of 78-82% (Figure 5.6C, E). Unlike SANDPUMA and NRSPredictor2, we do not enforce these guidelines by default. While we gain an accuracy increase from ~90% to ~96% for PARAS if the threshold is set to 50%, a lot of true positives are missed, with

~24% of data not getting a prediction at all (Figure 5.6B). In addition, a mispredicted substrate is often not very different from the actual substrate, and as such these mispredictions can still be informative to the researcher. For instance, PARAS wrongly predicted three A domains in our test set to select asparagine, when their actual substrates were aspartic acid in two instances and glutamine in the other, both of which are structurally very similar. Similarly, a substantial proportion of mispredictions occur between the residues isoleucine, leucine, and valine (Figure 5.7A, B). This pattern is even more evident when we look at the interaction probabilities predicted by PARASECT: it often predicts that multiple secondary substrates can be selected by an A domain, and as such has a higher false positive rate which makes it easier to observe which kind of prediction errors are made. Often, these secondary substrates are similar to the substrate that is actually selected, even when there is no experimental evidence that these substrates are indeed selected by the domain (Figure 5.7C, D). There are a few exceptions to this: for instance, PARASECT sometimes mistakes threonine for leucine, isoleucine, or valine, while PARAS never makes these specific errors. This is likely due to the fact that threonine shares a substructure with these three amino acids, which is directly used as a feature for predicting substrate-domain interaction.

We next assessed the prediction accuracy for each of the 34 substrates that PARAS and PARASECT were trained on both our class-based split and phylogeny-based split. In particular, we were interested to see if we saw a difference in performance for large amino acids between the two stratification methods, as our experimental results indicated that the A domain active site architectures of domains that recognise large substrates may vary substantially depending on which phylogenetic clade the domain belongs to, possibly due to degrees of rotational freedom of their side chains. Indeed, between the two test sets we observed a substantial drop in performance for various large substrates, including tryptophan, arginine, and phenylalanine, with PARAS and PARASECT misclassifying over 50% of datapoints for the latter (Figure 5.8). Other affected substrates were the smaller hydrophobic amino acids valine and isoleucine, whose active sites frequently clade together, illustrating the need for a large and diverse training set to cover as many phylogenetic branches as possible.

Interestingly, all our classifiers consistently misclassified pipecolic acid as proline (Figure 5.7, Figure 5.8). This is not entirely surprising, as the two amino acids have highly similar structures, with as only difference the size of their aliphatic rings. PARASECT also does the reverse and predicts that proline-recognising A domains also select pipecolic acid (Figure 5.7C, D). As pipecolic acid is a non-proteinogenic amino acid, cells need to synthesise it prior to incorporation into NRPs. This could indicate that *in vivo*, substrate preference for one or the other is not always dictated by the A domain active site, but rather by the presence or absence of pipecolic acid.

For PARASECT, we also assessed how often the correct substrate was represented among the top 1-5 predictions. Our results show that the correct substrate is found within the top 5 results 96.5% and 90.6% of the time for our class-based and phylogeny-based splits, respectively. Even when only considering the top 2 options, the probability of the correct substrate being among them is higher than the probability that PARAS makes a correct call for the substrate, at 92.9% and 82.3% for class-based splits and phylogeny-based splits, respectively (Figure 5.6D, F). Especially for A domains from underrepresented clades for which no good prediction can be made, PARASECT could provide a means to narrow down which types of substrates might be recognised.

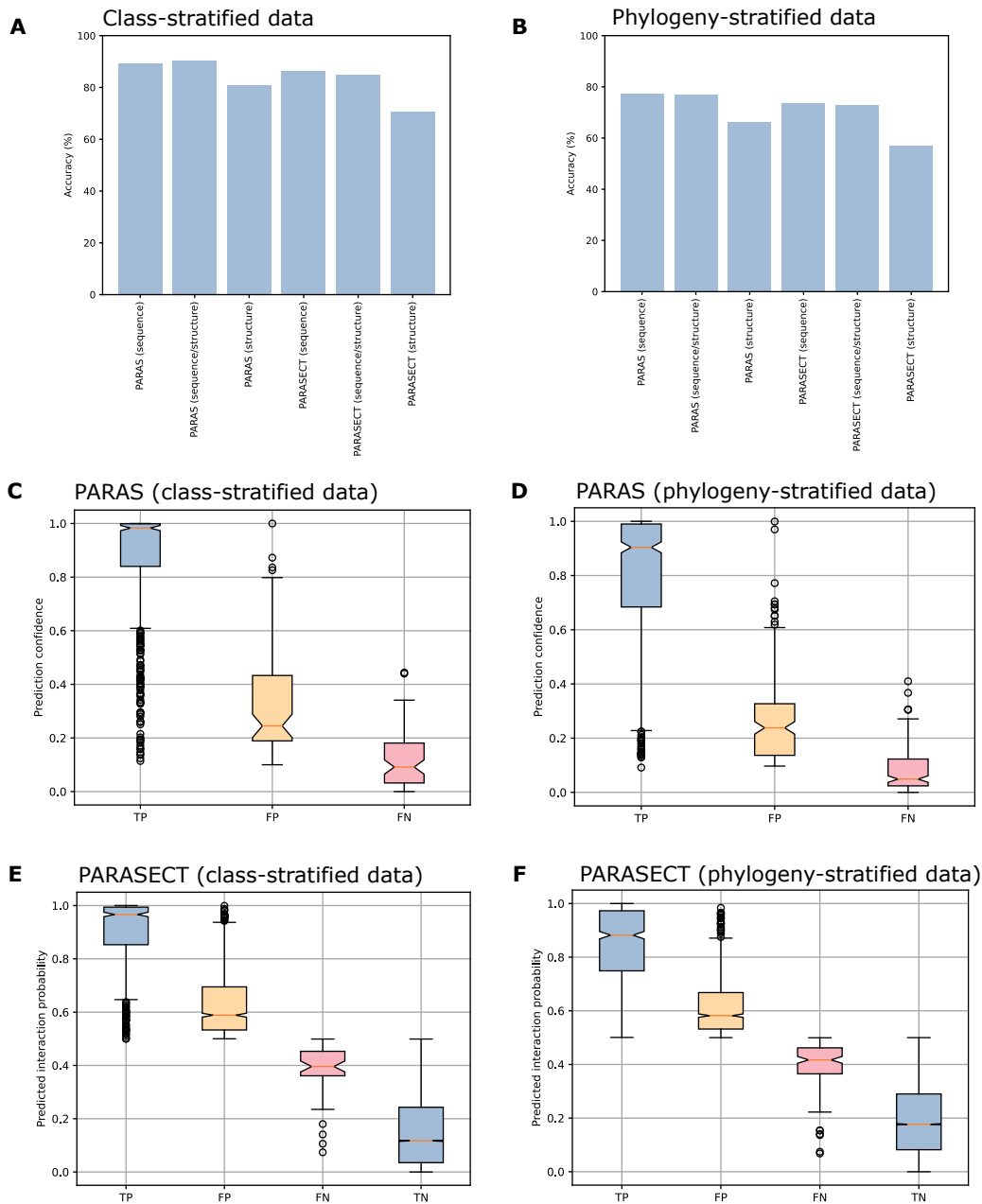


Figure 5.5. Performance of PARAS and PARASECT. Models trained on sequence features performed consistently better than models trained on structure, regardless of whether class-based stratification (A, C, E) or phylogeny-based stratification (B, D, F) was used. For both PARAS (C, D) and PARASECT (E, F), true positives have a much higher average confidence than false positives.

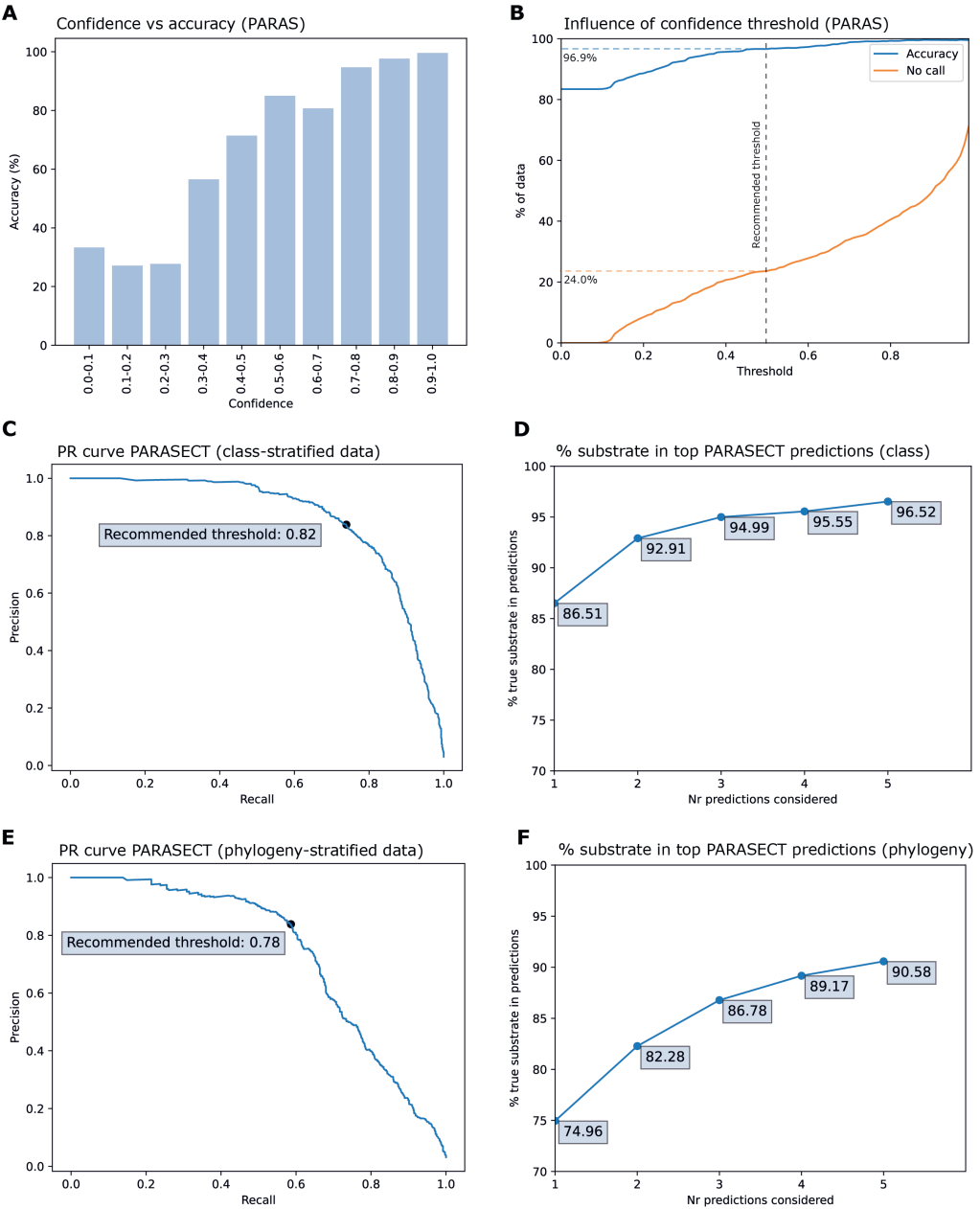


Figure 5.6. Setting thresholds for PARAS and PARASECT. Relationship between confidence score and accuracy of PARAS predictions. When PARAS is over 50% confident, prediction accuracies exceed 80%. B. Percentage of datapoints that are correctly predicted or not given a prediction for at different confidence thresholds for PARAS. C, E: precision-recall curves for PARASECT on class-stratified data and phylogeny-stratified data, respectively. D, F: the number of times a correct prediction is among the top predictions for PARASECT for class-stratified data and phylogeny-stratified data, respectively.

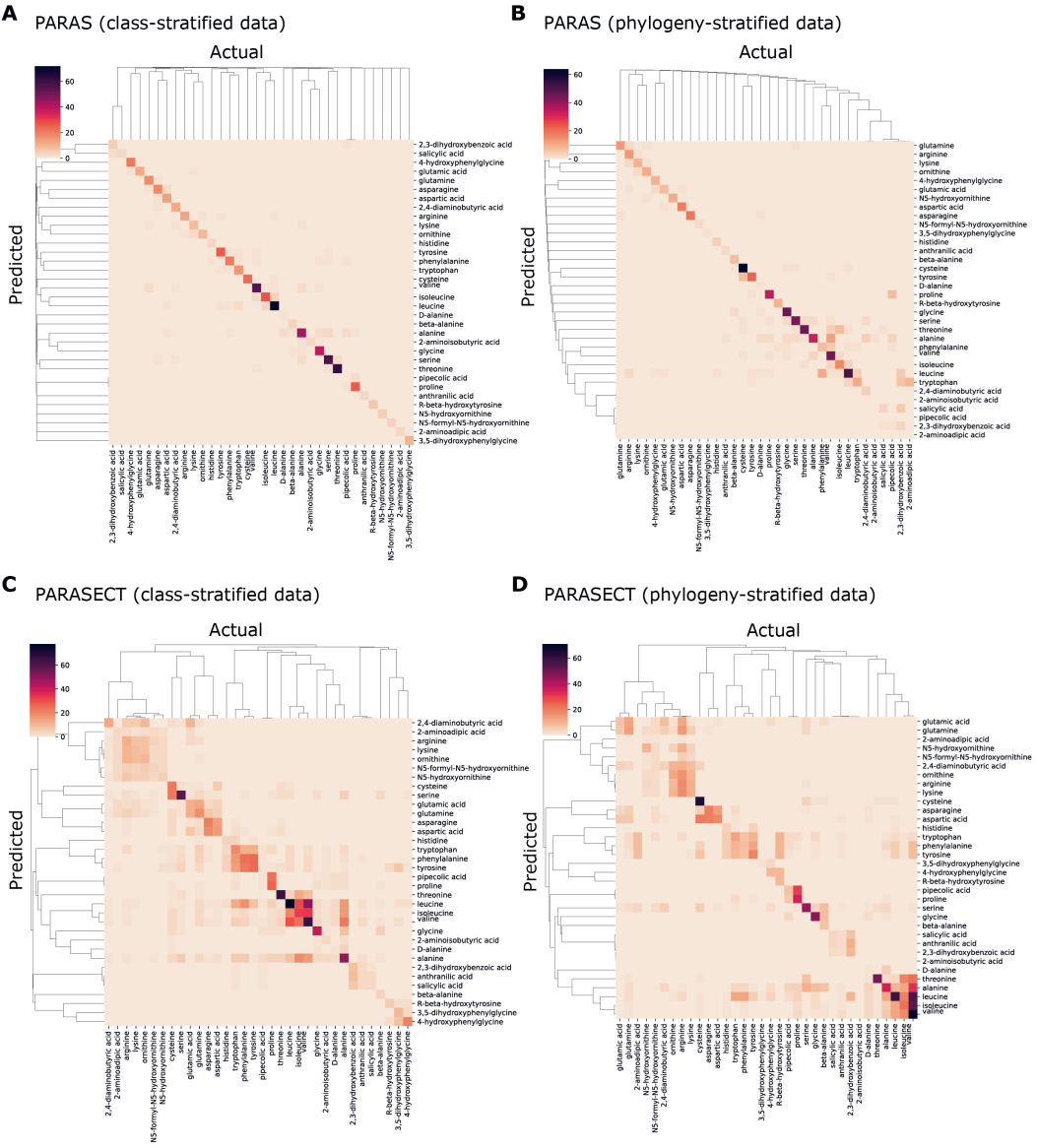


Figure 5.7. Confusion matrices showing common mispredictions by PARAS and PARASECT. A and B show true positives and false positives predicted by PARAS; C and D show true positives and false positives predicted by PARASECT. Mispredictions are slightly different for class-stratified splits (A, C) and phylogeny-stratified splits (B, D).

Interestingly, we found that sequence-based featurisation performed substantially better than structure-based featurisation across the board, independent of the stratification method used (Figure 5.5A, B). The most likely reason for this is information loss. Several predictive and reductive steps lie between an A domain sequence and the structural features that we use in our models: first,

we create structural models with AlphaFold; then, we represent the active site as a voxel grid, reducing our resolution to 1Å boxes that only contain information on which atoms it overlaps with; and third, we perform dimensionality reduction on the atom features of the resulting voxel grid such that we have a sufficiently small number of variables for machine learning. Therefore, there are three stages at which we likely lose information. Additionally, the structural models we obtain are static, while enzymes are highly dynamic. A domains in particular require large conformational changes in order to load the substrate onto the PCP domain. This flexibility is currently not captured by our structural models. In contrast, our sequence features include properties that represent which protein structures amino acids preferentially inhabit (α -helices, β -sheets, or β -turns; see methods), a characteristic that is directly tied to flexibility. These compounding factors go a long way to explaining why our sequence-based models perform consistently better. Therefore, instead of putting effort into embedding structures into sequential representations that lose the least amount of information, perhaps instead we should turn to nature, which has already provided us with high-quality structure embeddings that can be easily adapted for machine learning purposes: protein sequences themselves. That being said, when we used structural features alongside sequence features in our PARAS model, we did observe a performance increase of 0.8% (Figure 5.5A). Feature importance analysis showed that several structural features were major contributors to this model, with 4 out of the top 10 important features being structural. This suggests that structural features do contain information that is not captured by sequence features alone. Therefore, for the most accurate prediction, researchers might consider building structural models for their A domain and running PARAS with structural features enabled. However, more extensive cross-validation is required to confirm that this performance increase is significant.

While our structure-based approaches underperformed compared to sequence-based approaches, we have made them available on our GitHub such that researchers can leverage them for feature inference in future endeavours. We have also provided a script that can automatically visualise the contributions of voxels in the active site to the principal components that our algorithms were trained on. We hope that this will help researchers gain mechanistic understanding into the structural contributors to adenylation domain selectivity.

5.2.2.4. Benchmarking PARAS and PARASECT

To demonstrate the improvement of PARAS and PARASECT compared to state-of-the-art adenylation domain predictors, we benchmarked both sequence-based tools against SANDPUMA, including the 5 algorithms that it uses under the hood to make its ensemble predictions. One of these is an SVM model that is very similar to NRSPredictor2. To make as fair a comparison as possible, we used test sets comprising those sequences that did not occur in the training sets of either SANDPUMA or PARAS/ PARASECT, leaving 463 domains for our class-based split (Figure 5.9A) and 480 domains for our phylogeny-based split (Figure 5.9B). In both splits, PARAS and PARASECT substantially outperformed SANDPUMA as well as its composite algorithms, with accuracies of 90.3% for PARAS and 86.6% for PARASECT compared to 53.3% for SANDPUMA for the class-based split, when ‘no call’ is considered an incorrect prediction; and accuracies of 78.5% for PARAS and 74.2% for PARASECT compared to SANDPUMA’s 52.1% for the phylogeny-based split. The SVM models performed slightly better than SANDPUMA, with accuracies of 62.9% for the class-based split and 57.5% for the phylogeny-based split, but still performed worse than PARAS and PARASECT by 16.7%-27.4%. SANDPUMA can gauge to some extent whether it can make an accurate prediction or not. Therefore, a large proportion of mispredictions were returned as ‘no call’. When measuring its accuracy only on those datapoints it made a prediction for, it achieved an accuracy of 76.7 for the class-based split and

A PARAS (class-stratified data)

B PARAS (phylogeny-stratified data)

C PARASECT (class-stratified data)

D PARASECT (phylogeny-stratified data)

Figure 5.8. Performance of PARAS (A, B) and PARASECT (C, D) per substrate. A and C show performance on class-stratified test sets; B and D show performance on phylogeny-stratified test sets.

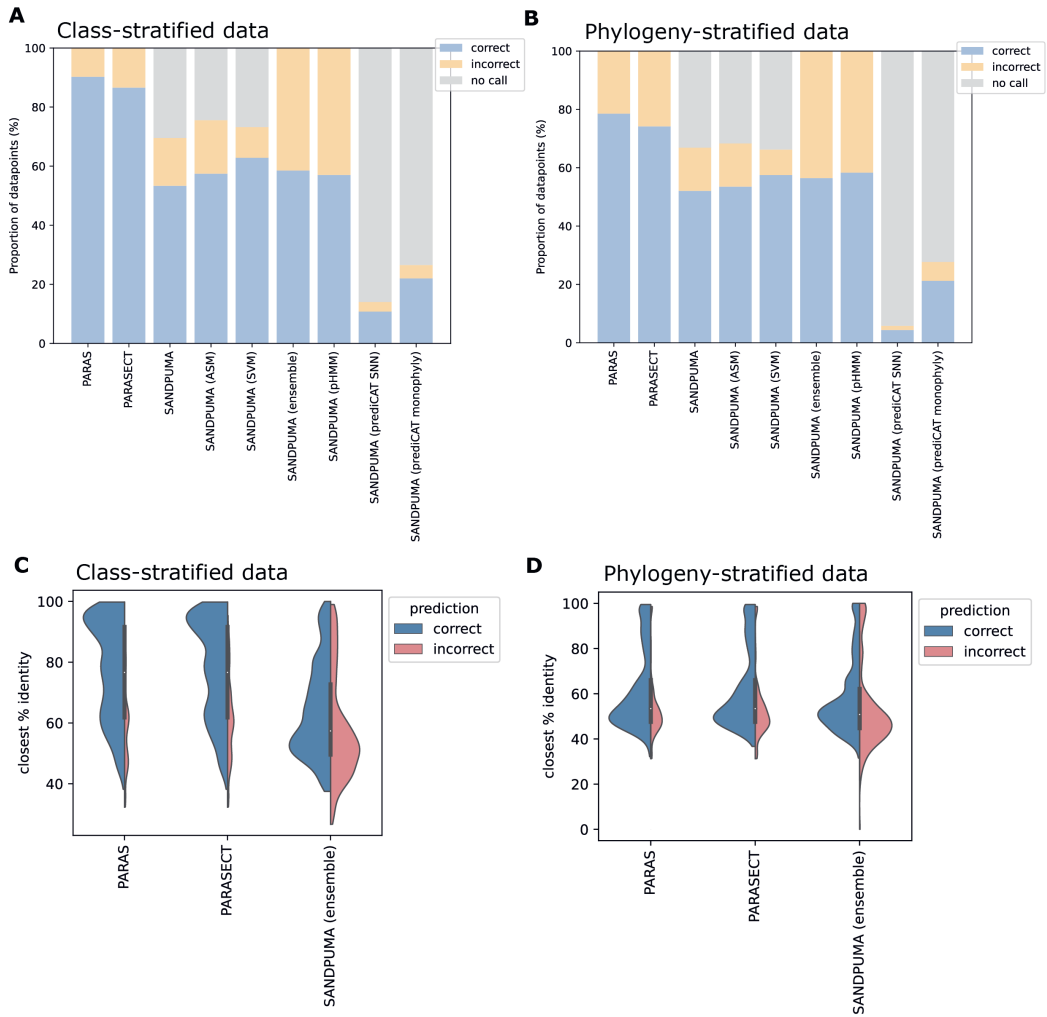


Figure 5.9. Benchmarking the performance of PARAS and PARASECT against SANDPUMA. PARAS and PARASECT both outperformed SANDPUMA as well as all sub-classifiers contributing to the ensemble algorithm by ~20-30%, regardless of whether class-based stratification (A) or phylogeny-based stratification (B) was used.

To investigate why PARAS and PARASECT outperform SANDPUMA, we also plotted the performance of PARAS, PARASECT, and SANDPUMA (when forced to make a prediction) against the % identity of the datapoint to the most similar datapoint in the training set. Looking at the distributions of test datapoints for class-based stratification (Figure 5.9C), we could clearly see that for PARAS and PARASECT, there existed a highly similar datapoint for a larger proportion of our test datapoints than for SANDPUMA. This is to be expected, considering our training dataset is almost three times as large as SANDPUMA's, and shows the importance of sufficient training data for machine learning.

Interestingly, PARAS and PARASECT still outperformed SANDPUMA after phylogeny-based stratification, where the data distribution is a lot more similar (Figure 5.9D). The 137 mistakes that we identified in the SANDPUMA dataset could contribute to this. It is possible that if SANDPUMA

were trained on the same dataset, it would perform comparably to PARAS and PARASECT. However, we decided not to explore this as SANDPUMA had a couple of other drawbacks that make it impractical to implement into natural product discovery pipelines: it was implemented in Perl, whereas the predominant programming language for BGC detection tools such as PRISM 4 and antiSMASH 7.0 is Python; and it is a couple of order of magnitudes slower than PARAS and PARASECT due to its internal phylogenetic placement algorithm PrediCAT. As a result, a single A domain prediction takes SANDPUMA several minutes to compute, compared to the 2-3 seconds that PARAS and PARASECT need for a prediction. Finally, due to the ensemble nature of the model, SANDPUMA is not as interpretable as PARAS and PARASECT, and therefore could not feasibly be leveraged for gaining mechanistic understanding into the workings of A domains.

5.3. Conclusion and Future Perspectives

In this work, we demonstrated that predicting adenylation domain selectivity is not as straightforward as previously believed. With an example from the tryptopeptin and tyrocidine BGCs, we showed how two A domains recognise the same substrate in different orientations, highlighting the need for datasets of increased size and diversity to cover the full evolutionary spectrum of adenylation domains. We then described the development of PARAS and PARASECT, two structure-informed adenylation domain predictors that offer substantial improvements compared to current state-of-the-art tools, showing performance increases of up to 27.4%. We showed that structure-based active site extraction contributed to this boosted model performance. In addition, we suspect that the increased size and quality of our dataset, which is almost three times as large as datasets that previous algorithms were trained on, had a major positive effect on model accuracy.

With the improved predictions that PARAS and PARASECT have to offer, we hope that researchers can obtain more reliable scaffold predictions from NRPS BGC sequence. To make our A domain predictor accessible to the natural products community, we are currently implementing PARAS into antiSMASH.

5.4. Materials and Methods

5.4.1. Computational methods

5.4.1.1. Data collection and curation

To train PARAS and PARASECT, we collated a dataset of 3254 unique A domains from three different sources: the NRPSPredictor2 dataset⁷⁷, the SANDPUMA dataset⁸⁰, and MIBiG 3.0⁷². We also added various novel data entries from literature. A domains were labelled by their GenBank, RefSeq^{299,300} or UniProt protein ID³⁰¹ and a number corresponding to the index of the A domain in the NRPS enzyme. For A domains from the MIBiG dataset and novel data entries, we downloaded the sequence of their protein from GenBank and extracted the sequences of N-terminal and C-terminal A subdomains with HMMer3³⁰², using the AMP-binding (PF00501.29) and AMP-binding_C (PF13193.7) Pfam pHMMs³⁰³, respectively. We were able to extract both subdomains for 2617 A domains in our dataset, and only extracted the N-terminal subdomain for the other 683. For A domains from SANDPUMA and NRPSPredictor2, we directly used the sequences from the datasets made available by the authors online. We also deduplicated our dataset through pairwise identity searches, allowing for truncations on either end of the A domain that might have occurred due to the different extraction methods used by different authors. For each A domain, we kept all protein labels.

As errors are unavoidable in human-collated datasets, we set out to curate our dataset in three steps. First, during deduplication, we flagged any domains with identical sequences but different selectivities and went back into the literature to fix their annotations. Next, we repeated the process for A domains with identical Stachelhaus codes, asserting that any differences in specificity were backed up by literature, fixing annotations where they were incorrect or omitting datapoints when we suspected experimental data to be unreliable. For instance, the Stachelhaus codes reported in literature for the first A domains in the proteins AAZ03554.1 and AAZ03552.1 were drastically different from the Stachelhaus codes we extracted³⁰⁴. Finally, we phylogenetically grouped the A domains based on the 34 amino acid signature of the active site and similarly verified the specificity of obvious outliers.

For the 113 domains for which ATP-PPi exchange assays were done, we additionally recorded the percentage activation for each of the tested substrates as compared to the best-activated substrate. To ensure annotation quality, we had two experts look at each datapoint: one to perform the initial annotation, the other to ensure its correctness. This yielded a reliable negative dataset which we used as training data for our binary predictors.

5.4.1.2. Structural modelling

Prior to structural modelling, we aligned all adenylation domains with MAFFT v7.508³⁰⁵ (default settings). We used this alignment and the reference crystal structure of the GrsA A domain (1AMU) to extract the N-terminal and C-terminal subdomains for each A domain. We then obtained predicted structures for each A subdomain domain using ColabFold³⁰⁶ in batch mode (default settings) without side chain relaxation. We only acquired structures of the C-terminal subdomain for 2617 domains in our dataset. As the C-terminal subdomain does not play a large role in substrate recognition, with only the highly conserved K517 residue involved in stabilising the amino acid backbone of the substrate^{44,77}, we decided to generate full structural models by combining the AlphaFold model of each native N-terminal subdomain with the C-terminal subdomain of 1AMU. We then aligned all full structure models to 1AMU in pymol³⁰⁷ (open-source, v2.5.0) and saved all models in identical orientations in Protein Data Bank³⁰⁸ (PDB) format. Finally, we copied the Mg²⁺ and AMP cofactors from 1AMU to the PDB files of all A domains to reflect the state of the active site pocket without the amino acid substrate bound.

5.4.1.3. Sequence-based domain feature extraction

To obtain sequence-based features for our models, we first extracted the 34 residues that correspond to the 34 residues within 8Å of the phenylalanine substrate in the A domain of GrsA, as also done by the authors of NRPSPredictor2^{76,77}. We used Muscle³⁰⁹ (v3.8.1551) profile alignments to extract these active site signatures, using the sequence of the GrsA (BAA00406.1) A domain as a reference. As a guide alignment, we used a structure-based sequence alignment of a subset of our dataset (923 A domains). We obtained this guide alignment by modelling the N-terminal A subdomains with Modeller³¹⁰ (v10.0, default: 250 models per sequence, no loop refinement) and aligning the resulting homology models with Caretta-shape³¹¹ (v1.0, default settings). As sequence extraction time scales exponentially with the size of the guide alignment, we chose to use a subset of our data rather than the full dataset. Then, each of the resulting 34 amino acid residues was featurised as a list of 15 physicochemical properties describing hydrophobicity (WOLS870101³¹², NEU1³¹³, NEU2³¹³, NEU3³¹³), size (WOLS870102³¹², TSAJ990101³¹⁴), electron state (WOLS870103³¹²), hydrogen bond donors (FAUJ880109³¹⁵), polarity (GRAR740102³¹⁶, RADA880108³¹⁷, ZIMJ680103³¹⁸), occurrence in alpha-

helices (CHOP780201³¹⁹), beta-sheets (CHOP780202³¹⁹) and beta-turns (CHOP780203³¹⁹), and isoelectric point (ZIMJ680104³¹⁸)³²⁰, as previously described by Rausch *et al.*⁷⁶ and Röttig *et al.*⁷⁷ for NRPSPredictor and NRPSPredictor2. These physicochemical features were then normalised and concatenated into feature vectors of length 510 (34 x 15; Figure 5.10B). Normalised physicochemical properties and guide alignments can be found at <https://github.com/BTheDragonMaster/paras>.

5.4.1.4. Comparing structure-based and sequence-based guide alignments

To compare structure-based and sequence-based guide alignments for feature extraction, we first split a subset of 923 domains into ten cross-validation sets, creating structure-based alignments (Caretta-shape³¹¹ v1.0) and sequence-based alignments (Muscle³⁰⁹ v3.8.1551) for each training set, stratifying on substrate class and using only the first reported substrate as response. Then, we featurised the 34 active site residues that we extracted from the structure-based and sequence-based alignments (as described above) and trained random forest models (scikit-learn³²¹, v1.1.3, default settings) on those same cross-validation sets and compared the test accuracy of the resulting two models. To ensure fair comparison, sequence-informed and structure-informed models were trained on the same datapoints using the same random seed for each cross-validation set.

5.4.1.5. Structure-based domain feature extraction

To convert our AlphaFold models into feature lists, we first placed a cubic voxel grid of 20x20x20 voxels centred around the position of the β -carbon of the phenylalanine substrate in the 1AMU reference structure. Next, for each A domain, we determined how many atoms overlapped with each voxel, recording counts for 7 atom types separately: aromatic, hydrophobic, charged, and non-charged oxygens, charged and non-charged nitrogens, and sulphurs. We then concatenated these 7 numbers for all 8000 voxels into a single vector of length 56000 representing the A domain active site. As this is too large a number of features to work with considering our dataset size, we decided to reduce dimensionality through principal component analysis (PCA). To ensure that the resulting principal components captured variation between the active sites of highly similar substrates, we first assigned each datapoint to one of 9 substrate categories: aromatic amino acids, aromatic acids, Asx/Glx, cyclic aliphatic amino acids, substrates with nitrogen-containing side chains, phenylglycines, small amino acids, small hydrophobic amino acids, and small polar amino acids (Figure 5.4). Datapoints that recognise substrates belonging to different groups were assigned to multiple groups. We only included substrates for which we have 11 or more examples in our dataset. Then, we performed PCA on each of the 9 substrate groups as well as the full dataset using scikit-learn³²¹ (v1.2.0) and stored the transformative models. For each PCA, we made a scree plot and determined the optimal number of components by using the kneedle package³²² (v0.8.3) to automatically locate the knee in the plot. This yielded between 5-8 PCs for each substrate group and 11 PCs for the full dataset, cumulating to 68 PCs in total. We then applied each of the 10 transformative models on the full dataset to obtain all 68 PCs for all datapoints. The resulting PCs can be found in the 'structure data' folder of our online GitHub repository. These 68 PCs were consequently used as features for training PARAS and PARASECT (Figure 5.10B).

5.4.1.6. Substrate feature extraction

As PARASECT predicts interaction between an A domain and a substrate based on features of both the domain and the substrate's molecular structure, we also needed a way to featurise molecules. To this purpose, we collected isomeric SMILES strings⁸⁵ for each substrate and used PIKACHU²⁵⁴ to extract ECFP-4 molecular fingerprints²⁵⁵ of length 1024 for each compound. These vectors were directly used for model training (Figure 5.10A, C).

5.4.1.7. Data stratification

Prior to model training, we stratified our data in two different ways to ensure we could properly assess how well our model will perform on unseen sequences afterwards. One split was done based on phylogeny such that each substrate was represented in both the training and test set with at least one example; and the other split was based on substrate class alone.

To split based on phylogeny, we used the PairwiseAligner module from Biopython³²³ (v1.81) to calculate bit-scores for all A domain pairs in our dataset. For each domain, we then concatenated the bit-scores for all pairs that domain was involved in (including a pair with itself) into a vector, keeping the order of domains constant for each iteration. Therefore, domains that are phylogenetically similar should have similar vectors. Next, we used the KMeans module from scikit-learn³²¹ (v1.2.0) to cluster our domains based on these bit-score vectors. We automatically detected the optimal number of clusters using kneedle³²² (v0.8.3), which came out to 4 clusters. We then distributed these clusters across our training and test sets such that both clusters had at least one example for each substrate in each – there was exactly one way to distribute the clusters that met this requirement, with 2 clusters assigned to each. We chose the larger dataset, accounting for about 75% of the data, as our training set, and used the remaining datapoints for our test set. We later used this train-test split to assess how well our models are likely to perform on A domains that are phylogenetically very different from A domains in our training set.

As many datapoints have multiple substrate labels, we used an iterative stratifying method that is suitable for multi-label datasets as proposed by Sechidis *et al.*³²⁴. This strategy assures that datapoints are divided across train and test sets such that each substrate class is proportionally represented. We used the MultilabelStratifiedShuffleSplit module from the iterstrat package³²⁵ (v0.1.6) to implement iterative stratification in python, taking into consideration only those substrates for which there were at least 11 examples in our dataset. We chose a train-test ratio of 3:1 such that the training set size would be comparable to the size of the training set that we obtained by splitting on phylogeny. Exact train-test splits can be found in the ‘train test splits’ folder on our GitHub page.

5.4.1.8. Model training and evaluation

During model development, we used the MultiLabelStratifiedKFold module from the iterstrat package³²⁵ (v0.1.6) to perform 5-fold cross-validation on our training sets to assess intermediate model performance and tune model parameters. Both PARAS and PARASECT are random forest models implemented with the RandomForestClassifier module from the scikit-learn³²¹ (v1.2.0). For each train-test split, we trained three models for both PARAS and PARASECT, each using different approaches to featurising A domains: sequence features only, structure features only, or a combination of both. As parameter tuning barely affected model performance (data not shown), we decided to train all our models with 1000 trees and default settings otherwise.

We trained PARAS as a single-label classifier, choosing the first listed label (which corresponds to the predominant substrate) as response for training purposes. PARAS output is therefore always a single substrate (Figure 5.10). For assessing model performance, we designated a prediction as correct if the predicted substrate was among the substrates listed for that domain.

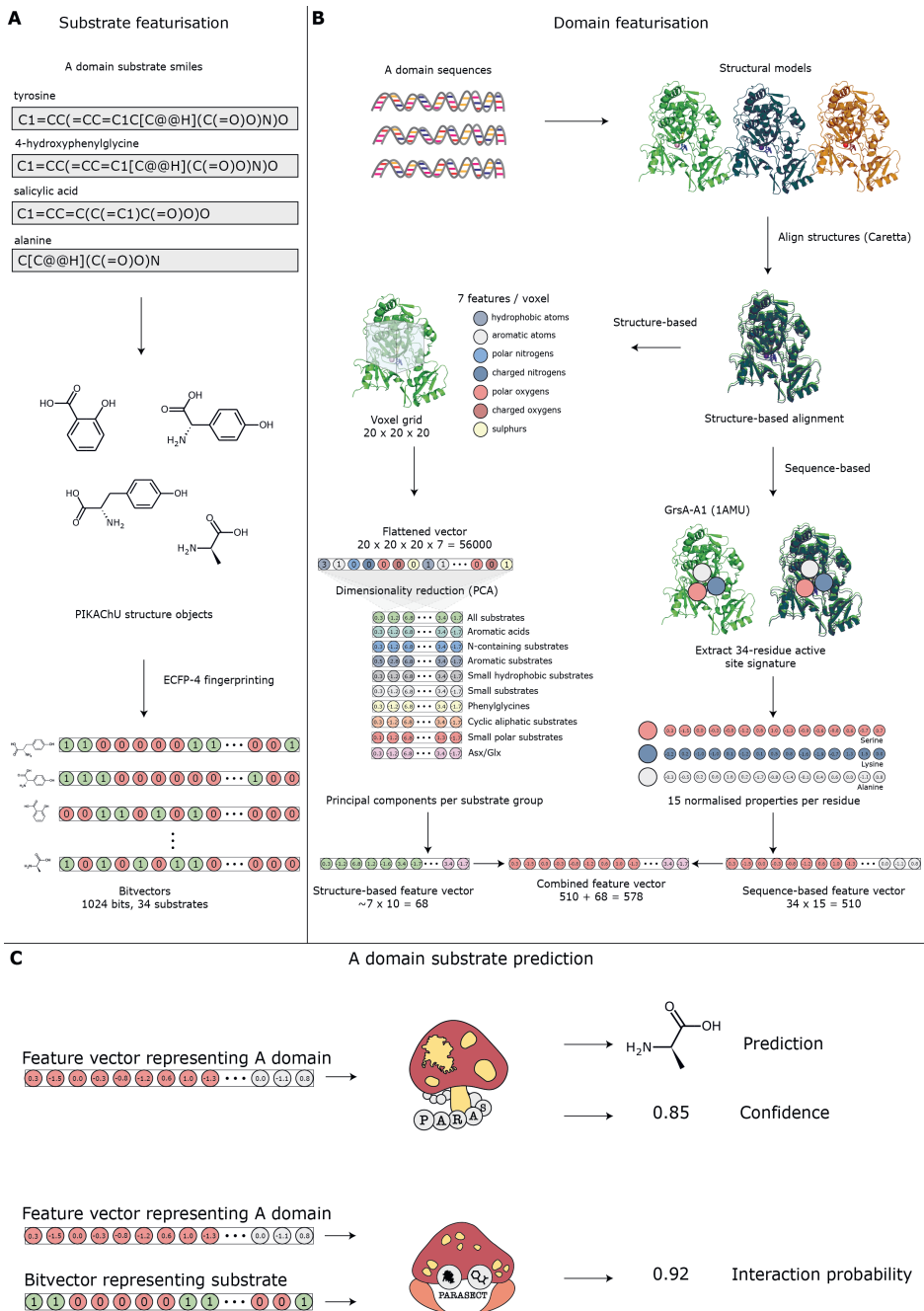


Figure 5.10. Data featurisation and model training. A. Featurisation of substrates for PARASECT. B. Sequence-based and structure-based featurisation of A domains for PARAS and PARASECT. C. Input and output of PARAS and PARASECT.

In contrast, PARASECT was trained on domain-substrate pairs, with a floating-point number as response, representing the probability of interaction between the domain and the substrate (Figure 5.10C). As this leads to highly imbalanced datasets, with a ratio of positive to negative datapoints around 1:30, we randomly under-sampled our data to achieve a 1:1 ratio using the RandomUnderSampler module from imblearn³²⁶ (v0.10.1). Model performance was subsequently assessed in two different ways: by looking at metrics that are commonly used for evaluating binary data, including precision, recall, and F1-score; and by determining if the substrate with the highest probability of interaction for a domain corresponds to one of the substrates that domain recognises. We additionally estimated a 'cut-off' interaction probability by obtaining and comparing mean and median interaction probabilities for true positives and false positives. For both PARAS and PARASECT, we also gauged model performance for each different substrate class.

5.4.1.9. Benchmarking

To benchmark PARAS and PARASECT against current state-of-the-art A domain predictors, we compared their performance against SANDPUMA⁸⁰. To this purpose, we ran SANDPUMA on all A domains that did not occur in SANDPUMA's training set. For a fair comparison, we only assessed model performance on A domains in the test sets of PARAS and PARASECT. As SANDPUMA is an ensemble predictor that internally runs other tools, including the SVM method employed by NRSPredictor2⁷⁷ and the pHMM method from Khayatt *et al.*⁷⁹, we were also able to use the SANDPUMA run to compare the performance of PARAS and PARASECT against these methods. As SANDPUMA sometimes predicts multiple substrates, a prediction was labelled as correct if there was overlap between the true substrates of an A domain and the predicted substrates.

5.4.1.10. Structure modelling and substrate docking

To obtain structural models for the tryptophan-recognising A-domains of TycB and TtpB, we separately modelled their N-terminal and C-terminal subdomains with AlphaFold with amber relaxation^{293,306} (Google Colab v1.3.0, default settings). We then aligned both subdomains to the 1AMU reference structure to obtain full A-domain models.

Next, we performed two dockings for each A domain: one with two separate ligands: tryptophan (PDB:TRP) and AMP; and the other with a reaction intermediate: tryptophanyl-adenylate (PDB:TYM). First, we prepared the ligands with the SDMolSupplier module of RDKit⁸⁶ (v2022.03.5) by moving them into the rough vicinity of the active site, and subsequently converted them to .pdbqt format with the Meeko package³²⁷ (v0.2). Then, we converted the A domains to .pdbqt format as well with MGLTools. We manually added the Mg²⁺ atom into the .pdbqt files after. Finally, we performed the docking with Autodock Vina³²⁸ (v1.2.3; exhaustiveness: 64; number of poses: 40), saving the best 20 poses. Then, we manually examined each pose to assess if the docking was reliable, specifically by determining if the position of the α -carbon of the TRP/TYM ligand was near its expected position, considering that the amino acid backbone needs to be stabilised by the conserved Asp and Lys residues in the active site. For the TycB A domain, we chose poses 12 and 5 for the TYM and AMP/TRP-dockings, respectively; and for the TtpB A domain, we chose poses 3 and 19 (Figure 5.2).

Experimental methods

To determine the substrate selectivity of the third A domain of TtpB (from now on known as TtpB-A3), we first transformed *Escherichia coli* with a plasmid containing DNA encoding the His-tagged TtpB-A3-PCP didomain, and then overexpressed and purified the protein. We opted to express the didomain rather than the standalone A-domain so that we could determine if both reactions that the adenylation domain catalyses took place: the adenylation reaction, which catalyses the conversion of ATP and the amino acid substrate to an amino-acyl-AMP intermediate; and the thiolation reaction, which transfers the amino acid to the phosphopantetheine arm on the PCP domain. After protein purification, we converted the didomain to its *holo* form *in vitro* by incubating it with CoA-SH and Sfp enzyme, a phosphopantetheinyl transferase which activates the didomain by transferring a phosphopantetheine arm onto the PCP domain from CoA. Then, we assessed the ability of *holo*-TtpB-A3-PCP to load various substrates by incubating the didomain with ATP and a tryptophan, phenylalanine, leucine, histidine and/or valine substrate, either separately for individual substrate assessment, or in tandem to observe which substrate was preferred. We also performed an *in vivo* competition assay by expressing the TtpB-A3-PCP didomain in *BAP1* cells²⁹⁸, a strain of *E. coli* that expresses Sfp, which allows *in vivo* domain activation and substrate loading. In all cases, substrate activation was measured with UHPLC-ESI-Q-TOF-MS analysis.

5.4.1.11. PCR amplification

As *Streptomyces sparsogenes* is very GC-rich, primer design with sufficiently long overlaps that still have reasonable annealing temperatures can be challenging, especially if the design requires overhangs with restriction sites for later incorporation of the DNA fragment into a plasmid. For this reason, we did two steps of PCR amplification to obtain PCR fragments containing the DNA encoding the TtpB-A3-PCP didomain: the first to amplify fragments from the GC-rich *Streptomyces sparsogenes* genomic DNA (kindly provided by the Marlene Rothe); and the second to amplify fragments with overhangs containing restriction sites (NdeI on the forward primer; EcoRI for the reverse). Detailed PCR protocols for each PCR can be found in Table 5.3. Primers were ordered from Merck. Resulting PCR reactions were run on an agarose gel (1g agarose in 100 ml 1x TBE buffer) for 70m at 130V, the bands were excised, and the DNA was extracted from the gel slices using the ThermoScientific Gel Extraction Kit.

5.4.1.12. Plasmid preparation and cloning

Next, we digested both our amplified PCR product (6 µl, ~400 ng; 0.5 µl EcoRI; 0.5 µl NdeI; 11.7 µl dH₂O; 2 µl 10X Buffer O) and the plasmid pET28A (28.3 µl, ~2 µg; 1 µl EcoRI; 1 µl NdeI; 14.7 µl dH₂O; 5 µl 10X Buffer O), which contains a T7 promoter and terminator, a kanamycin resistance cassette, an N-terminal His-tag, an N-terminal thrombin cleavage site, and EcoRI and NdeI restriction sites. Enzymes and buffers were acquired from ThermoFisher scientific. Reactions were incubated for 2h at 37°C and subsequently inactivated for 20m at 65°C.

We then set up a ligation reaction to insert the digested PCR product into the digested pET28A plasmid (2 µl ligase buffer, 1 µl T4 ligase, 4 µl 5ng/µl digested pET28A, 13 µl 5ng/µl digested insert). The reaction was incubated for 5h at 16°C and subsequently stored at 4°C. Buffer and enzyme were acquired from ThermoFisher scientific.

Next, we transformed our ligation reaction directly into chemically competent *E. coli* One Shot™ TOP10 cells (ThermoFisher Scientific) and chemically competent *E. coli* BAP1 cells, the latter of which

express the Sfp protein responsible for activating the A-PCP didomain as described above. For each strain, we added 5 μ l of our ligation reaction to one vial of competent cells and mixed gently, after which we placed the reaction on ice for 30m. Then, we heat-shocked the cells for 30s at 42°C and directly transferred them back to ice for 2m. We added 250 μ l of LB medium to each vial and shook the vials horizontally at 37°C for 1h (225 rpm). Then, we spread 20 μ l transformed cells onto LB plates with kanamycin (50 μ g/ml).

To ensure that the plasmids had incorporated the sequence encoding TtpB-A3-PCP didomain correctly, we inoculated 15 ml LB medium + kanamycin (50 μ g/ml) with the resulting colonies, incubated them overnight at 37°C, and isolated the plasmids using ThermoFisher Scientific's GeneJET Plasmid Miniprep Kit. We then checked the plasmids by restriction digestion and gel electrophoresis and sent plasmids that showed bands of the right sizes for sequencing. Cultures that harboured a correct plasmid were stored at -80°C.

5.4.1.13. Protein overexpression and purification

To obtain sufficient protein for analysis, we inoculated 1L of LB + 50 μ g/ml kanamycin with either transformed TOP10 or BAP1 cells and left them to grow at 37°C to an OD of ~1.0. Then, we added 1ml 0.5M IPTG to induce protein production. Cells were left to produce protein overnight at 15°C. Cells were spun down for 20m at 5000 rpm, 4°C. Pellets were resuspended in ~15ml Tris washing buffer (20mM Imidazole, 20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol). Next, we lysed the cells in a cell disruptor. The lysed cells were spun down for 15m at 17000 rpm, 4°C. Supernatant was transferred to a fresh tube and spun down for another 15m at 17000 rpm, 4°C. Then, the supernatant was filtered (0.45 μ l filter) and loaded onto a 1ml Cytiva HisTrap-FF column. The column was washed twice with Tris washing buffer (20mM Imidazole, 20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol), and the protein was eluted with elution buffers containing increasing concentrations of Imidazole (50 μ M – 300 μ M). Wash fractions were also collected. Fractions were denatured and run on an 8% SDS-PAGE gel for ~30m at 180V. Fractions containing protein of the expected size were concentrated to 0.5ml (5000 MWCO, 4000 rpm, 4°C), the buffer was replaced with storage buffer (20 mM Tris-HCl, 100 mM NaCl, 1-% glycerol) four times. The resulting protein solutions had a concentration of 47.673 mg/ml for *apo*-TtpB-A3-PCP (expressed from TOP10 cells) and 32.05 mg/ml for *holo*-TtpB-A3-PCP (expressed from BAP1 cells). We analysed the proteins by UHPLC-ESI-Q-TOF-MS to assert that the proteins had the correct weights and to check which substrate was loaded by *holo*-TtpB-A3-PCP in BAP1 cells.

5.4.1.14. In vitro activity assay

To assess which substrate is loaded by TtpB-A3-PCP, we first converted *apo*-TtpB-A3-PCP purified from TOP10 cells to *holo*-TtpB-A3-PCP by incubating 42 μ l 200 μ M *apo*-TtpB-A3-PCP with 5 μ l 100 mM MgCl₂ (as Mg²⁺ is required by Sfp and by the A domain itself to stabilise the negative charge of ATP in the active site), 2 μ l 20 mM CoA-SH (which provides the phosphopantetheine arm), and 1 μ l 400 μ M Sfp enzyme (isolated as described above). We incubated the reaction for 1h at room temperature, and then loaded the *holo*-enzyme with substrate by adding 50 μ l 170 μ M *holo*-TtpB-A3-PCP to 1 μ l 100 mM ATP (required for the adenylation reaction) and 1 μ l 50 μ M substrate dissolved in dH₂O. We tested the substrates tryptophan, phenylalanine, histidine, leucine, and valine. *apo*-TtpB-A3-PCP, *holo*-TtpB-A3-PCP, and loaded *holo*-TtpB-A3-PCP were analysed by UHPLC-ESI-Q-TOF-MS.

Table 5.3. PCR primers and protocols.

Primers		
Forward primer PCR 1	5'-TGTCCTTCCAGGAACAGAACGGCAGTT-3'	
Reverse Primer PCR 1	5'-TCATGCATCGTCGCCCTCTTACGG-3'	
Forward primer PCR 2	5'-AATAACATATGGCCGACGAGCGTGC-3'	
Reverse primer PCR 2	5'-AATAAGAATTCTCATGCATCGCTGCCCTC-3'	
Reaction setup PCR 1		
Reactant	Volume	Concentration
ThermoFisher Scientific HF Phusion mix	12.5 µl	2x
dH ₂ O	7.0 µl	
DMSO	2.0 µl	
<i>S. sparsogenes</i> genomic DNA	1.5 µl	73.9 ng/µl
Forward primer PCR 1	1.0 µl	10 µM
Reverse primer PCR 1	1.0 µl	10 µM
Program phase	Temperature	Time
Start-up	98°C	1m
Cycles (30x)		
Denaturation	98°C	30s
Annealing	66.3°C	30s
Extension	72°C	2m45s
Cooldown	72°C	10m
Store	4°C	∞
Reaction setup PCR 2		
Reactant	Volume	Concentration
ThermoFisher Scientific HF Phusion mix	12.5 µl	2x
dH ₂ O	7.0 µl	
DMSO	2.0 µl	
Amplified product PCR 1	1.5 µl	39.7 ng/µl
Forward primer PCR 2	1.0 µl	10 µM
Reverse primer PCR 2	1.0 µl	10 µM
Program phase	Temperature	Time
Start-up	98°C	1m
Cycles (30x)		
Denaturation	98°C	30s
Annealing	65°C	30s
Extension	60°C	2m30s
Cooldown	72°C	10m
Store	4°C	∞

5.4.1.15. UHPLC-ESI-Q-TOF-MS analysis

We analysed intact *apo*-TtpB-A3-PCP, *holo*-TtpB-A3-PCP, and loaded *holo*-TtpB-A3-PCP on a Bruker MaxIs II ESI-Q-TOF-MS connected to a Dionex 3000 RS UHPLC (equipped with an ACE C-300 RP column (100 x 2.1 mm, 5 µm, 30°C); controlled using Bruker Otof control 4.0). The column was eluted with 0.1% formic acid and 5-100% MeCN in a linear gradient for 30m. We ran the mass spectrometer in positive ion mode (scan range: 200-3000 m/z). We used the following source settings: end plate offset: -500 V; capillary: -4500 V; nebulizer gas (N₂): 1.8 bar; dry gas (N₂): 9.0 L/min; dry temperature: 200 °C. The following ion transfer conditions were used: ion funnel RF: 400 Vpp; multiple RF: 200 Vpp; quadrupole low mass: 200 m/z; collision RF: 2000 Vpp; transfer time: 110.0 µs; pre-pulse storage time: 10.0 µs. The resulting spectra were analysed with Bruker's DataAnalysis software (v4.4).

Chapter 6

The Turterra web portal for protein family data visualisation and analysis

Barbara R. Terlouw*, Janani Durairaj*, Nico Louwen*, Dick de Ridder, Marnix H. Medema, and Aalt D.J. van Dijk

* These authors contributed equally to this work

Abstract

Scientists who study individual protein families often perform a wide range of bioinformatics analyses using an assortment of different tools. There is mostly limited compatibility between tools, and without programming skills it is difficult and time-consuming to integrate and comprehensively investigate the different outputs, especially for large datasets. Here, we present Turterra: an accessible and comprehensive analysis portal for protein families. Turterra automatically constructs a web-portal from user-provided data, interactively visualising multiple sequence alignments, phylogenetic trees, protein structures and chemical substrates/products side by side. In this portal, data can be filtered by user-defined categories such as accession, species, or compound specificity, so that the user can easily visualise relevant subsets of data. Turterra also provides the option to build multiple sequence and structure alignments, phylogenetic trees, and homology models from scratch. Once the portal has been built, new sequences can be uploaded by end-users and compared to existing datasets, making Turterra a suitable engine for quick analysis of multifaceted data and for rapid deployment of protein portals. This will accelerate protein family research and facilitate collaboration between researchers working on the same families.

Turterra code and documentation are available at <https://git.wur.nl/durai001/turterra>.

6.1. Introduction

With a wealth of biological data published each day, bioinformatics analyses have become standard practice for researchers in the fields of biology and biochemistry. Novel protein sequences can be instantly compared to large databases such as GenBank^{299,300}, SWISS-PROT³²⁹, Pfam³⁰³ and PDB³⁰⁸, giving scientists direct insight into potential biological functions of their sequences. Once sufficient related proteins have been collected, a wide variety of analyses are routinely performed, including the construction of multiple sequence alignments, phylogenetic trees, and structural homology models to approximate 3D architecture. Such family-level analyses are particularly important for enzymes, which often share conserved folds as well as variable parts linked to their activity and specificity. Understanding and visualising patterns of conservation and variability is essential for grasping how different components relate to the various aspects of an enzyme's function.

Resources such as the tools published by EMBL-EBI³³⁰ have made large-scale sequence analysis more accessible to non-bioinformaticians. Still, life scientists have to employ a small army of different tools to analyse their protein data, including tree builders^{331,332}, homology modellers/structure predictors^{293,310,333}, and sequence aligners^{305,309,334}. It can be an arduous and time-consuming task to integrate different, not always compatible outputs of independent tools into comprehensive analyses that capture the essence of a research question. Also, it is often useful to look only at subsets of a dataset at a given time, or to add a new data point to an existing analysis without having to redo the entire analysis for the existing data entries. This is especially true for large datasets, which due to sheer scale are usually difficult to visualise in a digestible manner. Even for experienced bioinformaticians it can take considerable time to integrate and summarise different tool outputs into an intelligible format.

To decrease the time spent by bioinformaticians and molecular biologists on integrating different data outputs, several integrative analysis packages for protein families have already been developed. These include among others JalView 2³³⁵, a tool that specialises in multiple sequence alignment

analysis which can link to external web services like PDB³³⁶ for 3D structure visualisation and ViennaRNA²⁸⁸ for RNA secondary structure prediction; and Zebra3D³³⁷, which focuses on 3D analysis of protein homologs by 3D similarity search and structure alignment. While these tools are great at integrating and visualising the data types they were designed to handle, each has their shortcomings. For instance, JalView 2 does not provide the option to build or analyse computationally modelled structures alongside structures in the PDB database, which would be a valuable addition given the increased accuracy of computational modelling techniques. While Zebra3D does allow for the analysis of modelled structures, it lacks the interactivity that tools such as JalView 2 possess. Also, the analyses produced by these tools are not easily filterable on metadata. The inclusion of such an option would enable the targeted study or direct comparison of protein subsets based on auxiliary information such as mutation and variant studies, substrate specificity, inhibitor binding, and phenotypes at various temperature and pH conditions. Finally, the analyses performed by existing tools are not easily shareable, as most tools were designed for individual use and not targeted at communities of researchers who might work on the same protein (sub)families. A multifaceted data analysis platform that allows users to easily publish their data to the web would truly elevate integrative analyses to the next level: researchers would not have to waste time performing computationally expensive analyses that have already been done and could easily contribute to existing efforts that aim to understand their protein family of interest. Consequently, protein family information could be coherently stored in a single place for the benefit of research communities.

To provide the scientific community with a shareable platform for state-of-the-art multifaceted data analysis, we have developed Turterra: an accessible and comprehensive analysis portal for protein families. Here, we present a detailed overview of Turterra's features and demonstrate its versatility through two examples: the family of sesquiterpene synthase (STS) enzymes, which catalyse a single substrate into hundreds of 15-carbon sesquiterpene molecules, which give many plants and their fruits their smell³³⁸; and the family of non-ribosomal peptide synthetase (NRPS) adenylation domains, which govern the composition of microbial non-ribosomal peptides by specifically binding amino acid substrates in their active site^{45,339}.

6.2. Methods and implementation

Turterra contains two main executable scripts: *Turterra-build* and *Turterra*. Both scripts and their dependencies were written in Python (v3.9)³⁴⁰.

6.2.1. *Turterra-build*

Turterra-build creates the data used for visualisation and analysis by *Turterra*. If *Turterra-build* is run without arguments, it will only create the required folder architecture, which can then be manually populated by the user with the appropriate files. However, the user can also specify if they want *Turterra-build* to create homology modelled structures, a multiple sequence alignment, multiple structure alignment, profile HMM, phylogenetic tree, or any subset of the above. Multiple sequence alignments are created with MUSCLE³⁰⁹ (v3.8, default settings), structure alignments with Caretta-shape³¹¹ (v1.0, default settings), profile HMMs with HMMER³⁰² (v3.3.2, default settings), phylogenetic trees with FastTree³⁴¹ (v2.1.10, default settings), and homology models with MODELLER³¹⁰ (v10.0, default: 250 models per sequence, no loop refinement). The arguments and settings for each tool can be configured separately with a YAML-formatted configuration file. From

each set of generated homology models, the model with the lowest normalised DOPE score is selected as the best model and used for analysis and visualisation purposes.

6.2.2. *Turterra*

Turterra constructs and runs the layout of the main web portal with the Python package Dash³⁴² (v1.15.0). Apart from the built-in Dash components used for buttons, dropdown lists, and uploading files, the web portal contains five Dash components that handle data analysis: SequenceViewer, AlignmentChart, Molecule3DViewer, and Molecule2DViewer from dash-bio³⁴³ (v0.4.8), and Cytoscape from dash-cytoscape³⁴⁴ (v0.2.0). These visualise protein sequences; multiple sequence alignments (sequence-based or structure-based); (homology modelled) protein structures in 3D; substrate and/or product structures in 2D; and phylogenetic trees, respectively. A tab-separated file is provided by the user with Accession, Species, and Compounds as required columns defining each protein's accession, the species it is obtained from, and the compound (ligand/product) specificity. All other columns in the file are considered as extra data and are used in the portal to filter data. The dash-extensions package³⁴⁵ is used to allow downloading filtered subsets of data in each panel via its Download component, and to facilitate server-side filesystem storage of the data used by the application via the ServersideOutput component - this allows storing large datasets of proteins, alignments, models etc. without burdening the user's browser with intensive data transfer operations. Compound chemical structures are parsed from SMILES format using the cheminformatics kit PIKACHU²⁵⁴ (v1.0.13). ProDy³⁴⁶ (v2.0) is used to parse protein structures from PDB files, combined with the dash-bio-utils³⁴⁷ (v0.0.6) package to convert each structure into the format required for visualisation by Molecule3DViewer.

The portal also includes a component that allows the user to upload new sequences and/or associated structures. New sequences are appended to the existing sequence-based sequence alignment using MUSCLE (v3.8)¹¹. Similarly, Caretta³⁴⁸ (v1.0) rapidly aligns new structures to the consensus structure computed from the original set of 3D (homology modelled) structures, to then generate a new consensus structure. Uploaded sequences are appended to the existing phylogenetic tree using the phylogenetic placement tool epa-ng³⁴⁹ (v0.3.8, default settings, model: JTT), and the resulting JPLACE output is parsed and converted to Newick format.

The various components and panels are interconnected by means of a series of call-back functions, with different buttons and selections acting as triggers to activate changes in other entities, resulting in a highly networked application integrating the different views of data. Figure 6.1 depicts this as a flowchart, with arrows representing the relationship between triggers and their corresponding outputs.

6.2.3. *Extending Turterra*

To enable developers to easily add and interlink new custom panels, we have built Turterra in an extensible and modular fashion and provide tools and resources to quickly on-board developers to the code base. One such tool is a script to generate flowcharts using Mermaid diagram syntax (described at <https://mermaid-js.github.io>), depicting the flow of information from each component to the other. Figure 6.1 shows this chart generated for the current portal. This can be used as a reference when drafting a new component or panel, to pinpoint other components that may act as inputs, triggers, or outputs to the new one. With the predefined variable naming scheme described

in our documentation, developers can get a birds-eye view of their components labelled and styled according to their component type and the panel they are placed in, allowing for easy debugging of highly interconnected code. In addition, we have extensive documentation aimed at Python developers new to the Dash library or to GUI programming in general.

6.2.4. Data preparation

For this paper, we assembled two datasets to visualise in Turterra: a dataset of 302 STS enzymes from the characterized plant STS database (<https://bioinformatics.nl/sesquiterpene/synthasedb>)³¹, and a dataset of 1,093 NRPS adenylation domains (in-house data). For each accession, the datasets included an amino acid sequence in FASTA format, a structural (homology) model in PDB format, and the chemical structure of the enzyme's or domain's product or substrate respectively in SMILES format. Adenylation domain sequences were trimmed to only contain the N-terminal domain, as the C-terminal domain was too flexible and variable to obtain high-quality homology models. The STS homology models cover only the C-terminal domains of their sequences.

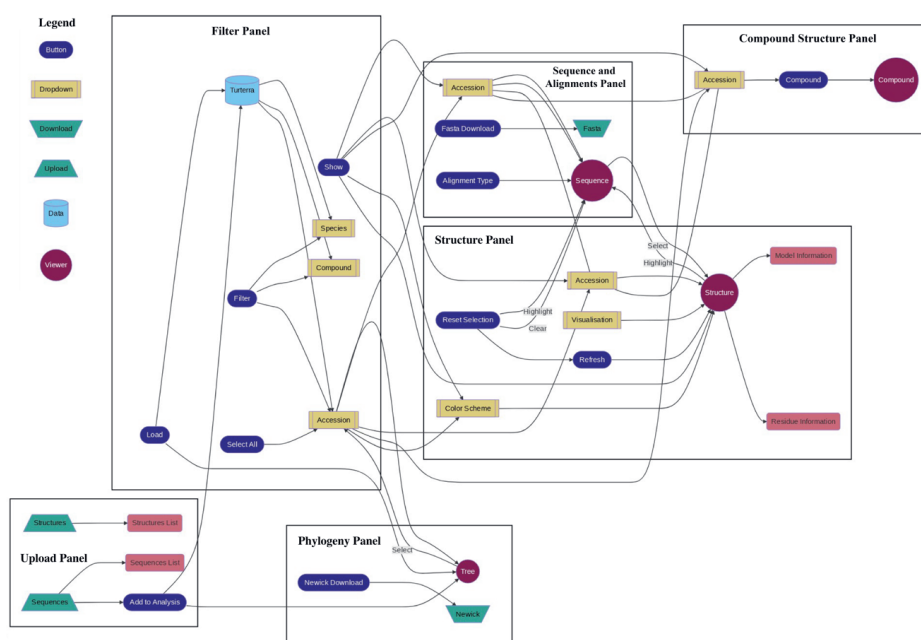


Figure 6.1. The auto-generated flowchart of call-back functions between various buttons, dropdowns, viewers, and other components on the Turterra website. An arrow from one component to another implies that a change in the first component acts as a trigger to change the second component.

6.3. Results and Discussion

6.3.1. *Turterra: an easy-to-use portal for protein family analysis*

For researchers interested in a quick initial assessment of their dataset and molecular biologists less experienced with bioinformatics tools, Turterra-build provides the option to build a multiple sequence alignment, structure alignment, phylogenetic tree and homology models, or a subset of these, from scratch. The user can provide as little as a FASTA file containing the sequences of interest, and a folder of PDB files to be used as templates for homology modelling, from which Turterra-build will create all the files that are required for the construction of the web portal. Later, any individual files can be replaced by user-provided versions. Required formats are described in-depth in the Turterra manual, making it straightforward for researchers of any discipline to provide the necessary files.

From this user-provided or Turterra-generated data, Turterra builds a comprehensive web-portal providing different views of the data. Datasets can be easily filtered according to user-defined categories and phylogeny, allowing for straightforward analysis and visualisation of relevant protein subsets. After construction of the portal, new sequences and corresponding (modelled) structures can be uploaded and compared to the pre-existing dataset. In settings where a researchers' Turterra portal is shared with others, the upload panel allows each individual viewer to independently compare their own data to the dataset shared by the portal. This unique combination of features makes Turterra a suitable engine for quick analysis of multifaceted biological data and rapid portal building for publication to the web.

To showcase Turterra's functionalities, we constructed Turterra web portals for two example enzyme families: a dataset of 302 STS enzymes, and a dataset of 1,093 NRPS adenylation domains, available at <https://bioinformatics.nl/turterra>. STSs are a large family of plant enzymes responsible for the synthesis of a large variety of sesquiterpenes: plant natural products that help give plants and fruits their distinctive smell³³⁸. Previous research has shown that sequence similarity in these enzymes is explained more by phylogeny than similarity in product specificity³⁵⁰. However, structural information has been successfully used to group these enzymes by precursor cation specificity³⁵¹, thus indicating that researchers studying STSs would benefit from an integrated appraisal of sequence, phylogeny, and structure, enabled by Turterra. Like STSs, NRPS adenylation domains are also involved in the biosynthesis of natural products. They are found in both bacteria and fungi and are core components of the much larger modular macro-enzymes called NRPSs^{45,339}. These enzymes produce peptide scaffolds, the composition of which is determined by the specificity of NRPS adenylation domains for certain amino acids. Subtle differences in sequence and structure of these otherwise highly similar domains result in the recognition of over 100 different amino acid substrates^{44,75}, making NRPS adenylation domains suitable for the multifaceted analysis Turterra provides. We use these two portals to describe Turterra's functionalities and performance below.



Figure 6.2. The panels of a Turterra web-portal built for NRPS adenylation domains. The Filter panel displays options to select proteins by various user-defined criteria. The resulting set of accessions are then available throughout the remaining panels. B. The Phylogeny panel displays a zoomable phylogenetic tree of all proteins, with selected accessions highlighted. Clicking on an unselected node in this tree adds the corresponding accession to the selection in the Filter panel. C. The Sequence and Alignments panel can show an individual searchable protein sequence, a sequence-based sequence alignment (as shown), or a structure-based sequence alignment, depending on the radio button selected. The alignments are interactive, allow zooming into specific regions, and display the level of conservation at each position on top. D. The Structure panel displays a zoomable, interactive, and customizable view of the 3D protein structure. E. The Compound Structure panel displays substrate/product chemical structures of selected accessions. F. The Upload panel allows the end-user to integrate and compare new sequences and structures with the existing dataset.

6.3.2. Turterra in action

Figure 6.2 depicts the six interconnected analysis panels in Turterra, described in detail below.

After loading the user-provided dataset, Turterra has two panels useful for filtering subsets of proteins - the Filter panel (Figure 6.2A) and the Phylogeny panel (Figure 6.2B). The former provides filtering options based on user-defined categories such as species and compound specificity. The latter depicts a phylogenetic tree of all proteins. Protein subsets can be defined, expanded, or narrowed by selecting or de-selecting clades or single enzymes in the phylogeny panel, allowing for inspecting similarities and differences between phylogenetically related proteins. These two panels are linked: selecting accessions via either panel updates the other as shown in Figure 6.3. For example, the accession outlined in red in Figure 6.3A was selected by clicking on the corresponding node in the Phylogeny panel in Figure 6.3B.

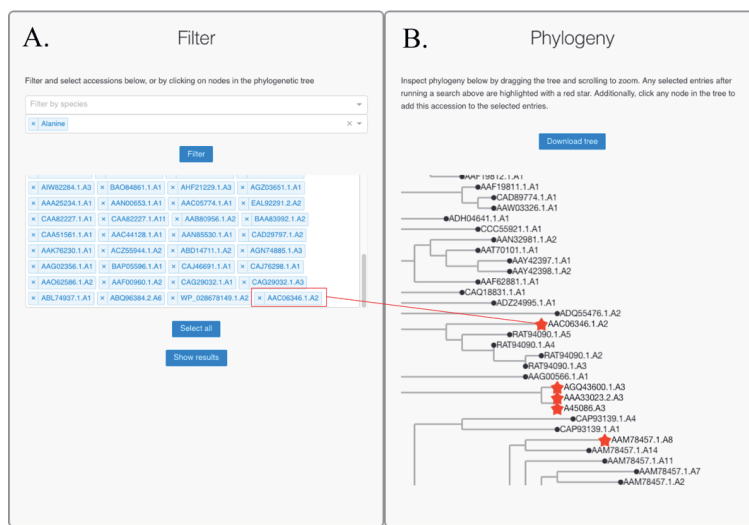


Figure 6.3. Turterra's filter and phylogeny panels. A. the Filter panel and B. the Phylogeny panel are interlinked - accessions selected in one are displayed in the other, as enzyme names in the Filter panel and as red stars in the Phylogeny panel. The accession indicated with a red line in A has been added by clicking on the corresponding node in B. The phylogeny tree in B can be exported as a Newick file using the Download button.

Once a set of accessions are selected, the next three panels allow visualisation of their sequences, 3D structures, and the compounds that they produce or modify. These are, respectively, the Sequence and Alignments panel (Figure 6.2C), the Structure panel (Figure 6.2D), and the Compound Structure panel (Figure 6.2E).

The Sequence and Alignments panel can display an individual sequence (single sequence mode), a sequence-based sequence alignment (sequence alignment mode) or a structure-based sequence alignment (structure alignment mode), controlled by a set of radio buttons. The alignments are interactive, zoomable, and scrollable, with conservation at each position represented as a bar chart above the alignment. The Structure panel displays an interactive 3D protein structure and has multiple visualisation styles and colour schemes to choose from, described in the online documentation. In addition, the Sequence panel (in single sequence mode) and the Structure panel are interlinked as shown in Figure 6.4: the residues highlighted in the sequence correspond to those highlighted and labelled in the Structure panel. These highlights can be defined in either panel through mouse selection and inter-active clicking respectively, and selected residues can be visualised differently. This is especially useful to inspect the structural context of residues deemed important from literature or from conservation analysis, and to map the sequence context of relevant residues in the structure, determined by mutational studies or by crystal structure analysis. For example, Figure 6.4 depicts the STS from tobacco producing the sesquiterpene 5-epi-aristolochene³⁵². The two stretches of sequence selected with the mouse in Figure 6.4A correspond to two known STS catalytic motifs³³⁸ which are seen to surround the active site cavity in the structure in Figure 6.4B. The residue selected by clicking in the structure in Figure 6.4B caps this active site cavity, and thus is interesting to map back to its sequence and alignment context.

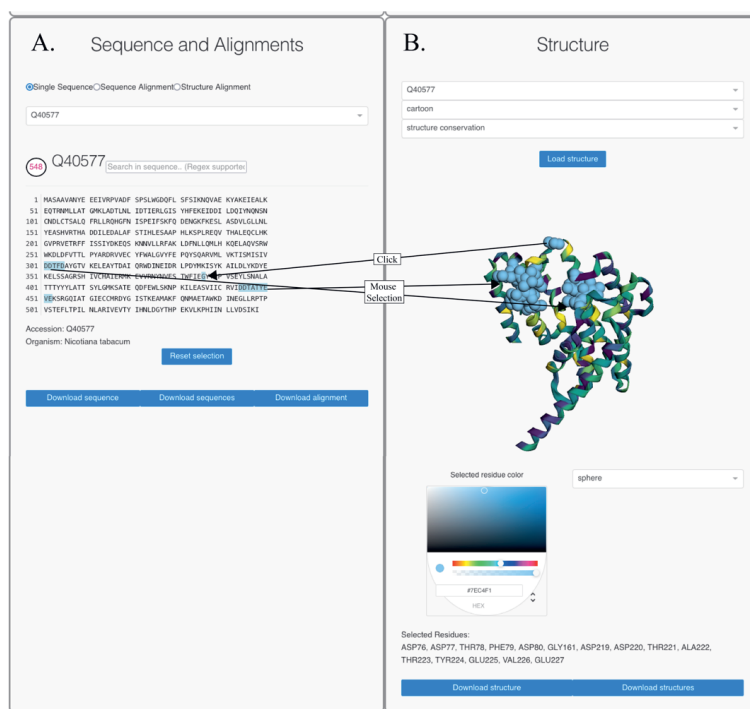


Figure 6.4. Turterra's sequence and alignments and structure panels. A. the Sequence and Alignments panel in single sequence mode and B. the Structure panel are interlinked - residues selected with the mouse in A ("Mouse Selection") are highlighted in B with residue numbers labelled in the list at the bottom, and residues clicked in B ("Click") are highlighted in A. The colour scheme and visualisation style of the structure can be changed using the provided dropdowns in B. The available colour schemes and styles are described in the documentation. Selected residues can be visualised differently from the rest of the structure using the provided colour picker and style selection dropdown, shown as light blue spheres in the figure. The three Download buttons in A allow the user to export the sequence of the selected accession, the sequences of all filtered accessions, and the aligned sequences of filtered accessions respectively as FASTA files (with radio buttons controlling whether the sequence alignment or structure alignment is exported). The two download buttons in B export the structure of the selected accession and the superposed structures of filtered accessions respectively as PDB files.

The Compound panel displays the two-dimensional structure of compounds associated with a protein, with an option to select one if multiple such compounds exist. The protein visualised across the Sequence and Alignments, Structure, and Compound Structure panels is the same and can be chosen using the accession selection dropdowns in any of these panels.

Finally, the Upload panel (Figure 6.2F) allows end-users to compare their novel proteins to the proteins in the loaded dataset. Users can upload sequences and optionally structures of their proteins and have them integrated into the existing phylogenetic tree, sequence alignment, and structure alignment without recalculation for the whole dataset. This enables comparison of putative or newly characterized proteins, and mutants or variant sequences with previously characterised proteins and their existing literature. This is especially useful in collaborative studies with researchers working on different subsets or variants of the same protein family.

Turterra tolerates missing or incomplete data: proteins without associated structures or compounds can still be analysed from a sequence and phylogeny perspective, and incomplete structural models with missing residues are still correctly mapped to the corresponding sequence. In addition, both filtered and expanded data can be exported and downloaded by end-users: the phylogeny tree in Newick format, sequences, and alignments as FASTA files, superposed structures as PDB files, and chemical compounds as SMILES strings.

6.3.3. Performance and distribution

Table 6.1 depicts the data creation, loading, response and upload times for the STS and NRPS datasets. Since the bulk of data transfer for data shared across components is performed on the server side, the Turterra portal easily scales to datasets with thousands to tens of thousands of proteins. Portals which are published and made available to users through the web can be accessed simply via a URL and do not consume any space on the end-user's filesystem unless data is explicitly downloaded. Each user receives their own session key on the portal maintainer's side which is used to store their filtering and analysis options. These keys can be used to keep track of the time since a user's analysis to inform them of the results or to eventually clear their data once enough time has passed.

Table 6.1. Portal creation, loading, response, and upload times for the STS and NRPS datasets, for which the average protein sequence length and structural model lengths are given. "Creation time" is the time taken for generating sequence and structure alignments on a single thread. We don't include the model and tree generation times here as this is highly dependent on the program and settings used and hence would differ significantly depending on the user and use case. "Loading time" is the amount of time taken upon pressing the "Load Data" button in the portal. "Response time" is the time taken to register mouse click and selection events. "Upload time" is the time taken for integrating a new sequence and structure into each portal using the Upload panel.

Dataset	No. Proteins	Av. sequence length (aa)	Av. model length (aa)	Creation time (m)	Loading time (s)	Response time (s)	Upload time (s)
STS	302	551 ± 62	265 ± 3	7 min	2s	~0.1s	4.2s
NRPS	1093	475 ± 54	370 ± 22	25 min	6s	~0.3s	37s

6.3.4. Opportunities for extension

The Turterra source code provides a good starting point for developing additional customised panels holding information specific to certain proteins, use cases, or studies. For example, a mutation panel could connect experimentally solved or modelled structure mutants with their phenotypes and allow users to visualise these mutated residues in the structure and sequence panels. Specialised predictors of compound specificity could generate their own panels giving a detailed prediction report for each protein and linking to predictive residues in other panels as well as predicted compounds in the compound panel. For multiple researchers working on a shared project, a Turterra portal could contain annotations and notes linked to the sequence, structure, compound, or phylogeny, allowing productive collaboration and easy sharing of results. With thorough documentation and the availability of helper scripts to auto-generate flowcharts of inputs and outputs, adding new panels and interlinking these with existing panels is straightforward for novice programmers as well. As usage becomes more widespread, we envision an open-source community of developers designing plug-and-play panels for Turterra users.

6.4. Conclusion

Turterra provides a comprehensive solution for highly interactive, multifaceted biological data analysis, allowing even researchers with limited bioinformatics experience to quickly analyse and filter their protein data in a single place. It is the ideal framework for in-house or web-based server publication, facilitating collaboration between researchers working on the same protein families. For more experienced users, Turterra's accessible source code makes the portal easily customisable and extendable to better fit specific protein families or to allow for integration into existing tools and databases. We hope that Turterra will drastically cut time spent on data analysis by researchers in all fields of protein biology and will organically grow to suit the needs of the scientific community.

Chapter 7

PIKACHU: a Python-based Informatics Kit for Analysing Chemical Units

Barbara R. Terlouw, Sophie P.J.M. Vromans, and Marnix H. Medema

This chapter has been published as

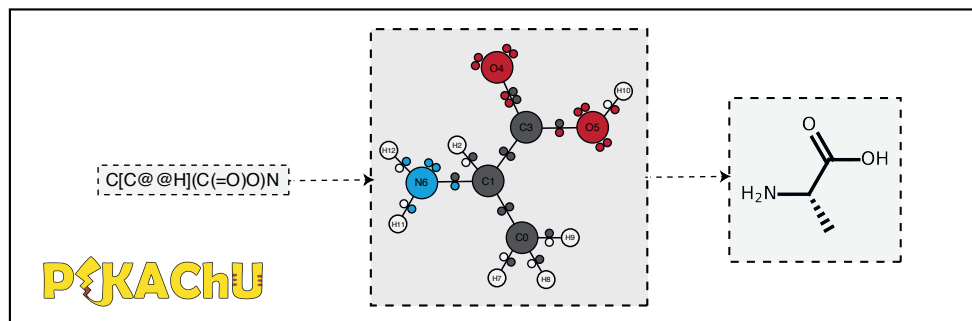
Terlouw, B.R., Vromans, S.P.J.M., Medema, M.H., PIKACHU: a Python-based informatics kit for analysing chemical units. *Journal of Cheminformatics* 14, 34 (2022). <https://doi.org/10.1186/s13321-022-00616-5>

Abstract

As efforts to computationally describe and simulate the biochemical world become more commonplace, computer programs that are capable of *in silico* chemistry play an increasingly important role in biochemical research. While such programs exist, they are often dependency-heavy, difficult to navigate, or not written in Python, the programming language of choice for bioinformaticians. Here, we introduce PIKACHU (Python-based Informatics Kit for Analysing Chemical Units): a cheminformatics toolbox with few dependencies implemented in Python. PIKACHU builds comprehensive molecular graphs from SMILES strings, which allow for easy downstream analysis and visualisation of molecules. While the molecular graphs PIKACHU generates are extensive, storing and inferring information on aromaticity, chirality, charge, hybridisation and electron orbitals, PIKACHU limits itself to applications that will be sufficient for most casual users and downstream Python-based tools and databases, such as Morgan fingerprinting, similarity scoring, substructure matching and customisable visualisation. In addition, it comes with a set of functions that assists in the easy implementation of reaction mechanisms. Its minimalistic design makes PIKACHU straightforward to use and install, in stark contrast to many existing toolkits, which are more difficult to navigate and come with a plethora of dependencies that may cause compatibility issues with downstream tools. As such, PIKACHU provides an alternative for researchers for whom basic cheminformatic processing suffices and can be easily integrated into downstream bioinformatics and cheminformatics tools.

PIKACHU is available at <https://github.com/BTheDragonMaster/pikachu>. Online supplementary material is available at <https://doi.org/10.1186/s13321-022-00616-5>.

Graphical abstract



7.1. Introduction

In a data-driven world where the discovery of novel natural and synthetic molecules is increasingly necessary, *in silico* chemical processing has become an essential part of biological and chemical research. Novel metabolites are compared or added to searchable chemical databases such as ChEBI³⁵³, PubChem²⁴⁷, NP Atlas²⁴⁶, and COCONUT³⁵⁴; molecular structures are predicted from biological pathways^{68,355}; and bioactivities and pharmaceutical properties are predicted from chemical structure^{225,356,357}. Such analyses rely on robust cheminformatics kits that can perform basic chemical processing, such as fingerprint-based similarity searches, substructure matching, molecule visualisation and chemical featurisation for machine learning purposes.

Typically, molecular processing by cheminformatics kits begins with the reading in of molecular data from chemical data formats, ranging from one-dimensional to three-dimensional molecular representations. One such format is the SMILES (Simplified Molecular-Input Line Entry System) format⁸⁵, which represents a molecule as a one-dimensional string, describing atom composition, connectivity, stereochemistry, and charge. More elaborate formats such as PDB and MOL use text files to store not just the abovementioned properties but also atom coordinates in three-dimensional space.

Depending on the application, different formats and subsequent processing are appropriate. Due to the vast number of possible chemical analyses, exhaustive cheminformatics kits have accumulated into software libraries that are so large that they can be hard to navigate and rely on so many dependencies that they can be difficult to implement in software packages. As a result, the trade-off between time spent accessing and integrating these cheminformatics kits into a codebase and time spent on actual analyses is disproportionate for users that need to perform simple *in silico* analyses such as reading in SMILES, drawing a molecule, or visualising a substructure. One popular open-source cheminformatics kit that suffers from this problem is RDKit⁸⁶. While RDKit is an incredibly fast and powerful library that supports an immense variety of possible chemical operations, its use of both Python and C++ as programming languages as well as the sheer number of dependencies it relies on frequently causes compatibility issues when integrating RDKit into other programs, and disproportionately increases the number of libraries that need to be installed. Therefore, while RDKit is great for heavy-duty *in silico* analyses such as computing 3D conformers for a compound or constructing electron density maps, it is a bit heavyweight for the basic operations that most researchers in bioinformatics and cheminformatics require.

A second widely-used cheminformatics kit is CDK³⁵⁸. Written in Java, it is well-suited for implementation in web applications, and has successfully been used for molecular processing in the COCONUT database³⁵⁴, the Cytoscape application chemViz2³⁵⁹, and the scientific workflow platform KNIME (Konstanz Information Miner)³⁶⁰. However, with Python becoming the programming language of choice for many scientists³⁶¹, especially those working in the growing field of (Deep) neural networks, CDK is not always an ideal fit.

To make basic cheminformatics processing more accessible for Python programmers, we therefore introduce PIKACHU: a Python-based Informatics Kit for Analysing Chemical Units. PIKACHU is a flexible cheminformatics tool with few dependencies. It can parse molecules from SMILES, visualise chemical structures and substructures in matplotlib, perform Extended Connectivity FingerPrinting (ECFP)²⁵⁵ and Tanimoto similarity searches, and execute basic reactions with a focus on natural product

chemistry. Therefore, we hope that PIKACHU can provide a convenient alternative for many Python-based bio- and cheminformatics tools and databases that only demand basic chemical processing.

7.2. Methods and implementation

7.2.1. Software Description

PIKACHU is implemented in Python³⁴⁰ (v3.9.7). Its only dependency is the common Python package matplotlib³⁶² (v3.4.3). PIKACHU can be run on Windows, MacOS, and Linux systems.

7.2.2. Parsing molecules from SMILES

PIKACHU takes a SMILES string as input and from it builds a graph object, in which nodes represent atoms and edges represent bonds (Figure 7.1). For each atom, PIKACHU initially stores information on chirality, aromaticity, charge, and connectivity. For each bond, it stores bond type (single, double, triple, quadruple, or aromatic), neighbouring atoms, and cis-trans stereochemistry for double bonds. Once all atoms, bonds, and their connectivities have been stored, electron shells and orbitals are constructed for each non-hydrogen atom. Next, we determine the valency for each non-hydrogen atom, taking into account atom charge. For atoms of variable valency such as sulphur (2, 4 or 6) and phosphorus (3 or 5), we select a valency that is equal to or higher than the sum of non-hydrogen bonds and explicit hydrogen bonds, prioritising smaller valencies. Double, triple, quadruple and aromatic bonds contribute proportionally to this sum. If insufficient bonding orbitals are available to achieve the desired valency, the electrons in the valence shell are excited to higher-energy orbitals, such that each orbital contains at most one electron. Implicit hydrogens are then added to the structure such that the pre-determined valencies are obeyed.

Subsequently, electrons are allocated to the p-orbitals of π bonds in double, triple and quadruple bonds, and atom hybridisation is determined from steric number. Then, all cycles in the graph are detected using an open-source Python implementation (Miles 2019) of the simple cycle detection algorithm described by D. Johnson in 1975^{363,364}. PIKACHU removes all cycles smaller than three atoms and identifies the smallest set of unique smallest rings (SSSR).

Next, the SSSR is used for aromaticity detection. This is done recursively: in each round, each cycle that has not yet been added to the set of aromatic cycles is evaluated with Hückel's $4n + 2$ rule on planar rings. We chose to assess aromatic cycles rather than systems as Hückel's rule is not always reliable for cyclic systems³⁶⁵. First, the hybridisation of all atoms in the cycle is examined. All atoms must be sp²-hybridised, or sp³-hybridised with a delocalisable lone pair that can be promoted to a p-orbital. If the cycle is planar and the sum of double bonds and lone pairs is odd, the cycle is considered aromatic. Aromatic bond stretches are locally kekulised, and double bonds are subsequently counted. When a cycle is considered aromatic, bonds and atoms in the cycle are set to aromatic, and lone pairs of sp³-hybridised atoms are promoted to p-orbitals such that the new hybridisation is sp². Recursion is needed in case double bonds in cyclic systems are defined in such a way that not all sub-cycles contain the required number of bonds to obey Hückel's rule: when adjacent bonds are updated to aromatic, they will be counted in the next round of aromaticity detection (Figure 7.2). When, after an iteration, the number of aromatic cycles no longer changes, all aromatic cycles have been detected. From these cycles, PIKACHU defines aromatic systems, where

aromatic cycles are considered part of an aromatic system if they share a bond with at least one other aromatic cycle in the system.

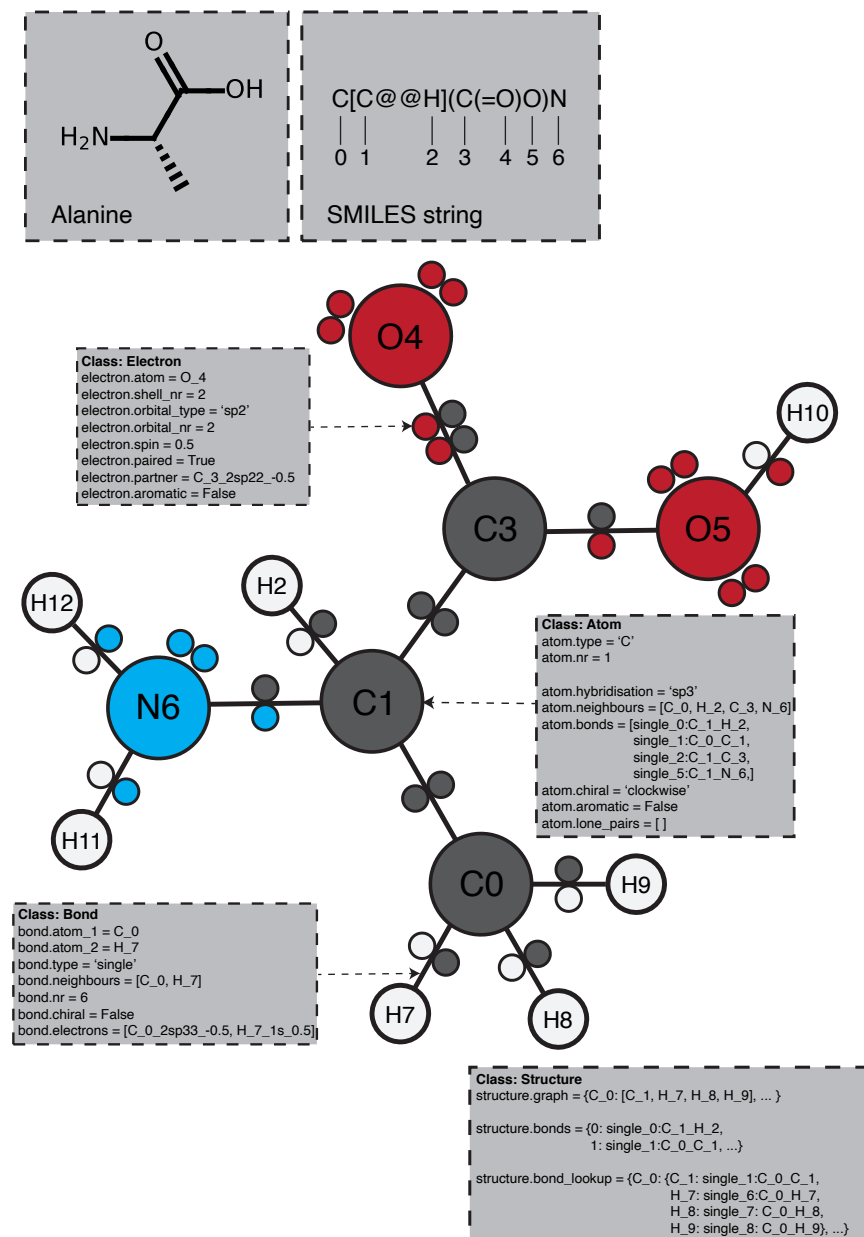


Figure 7.1. Overview of the internal structure of PIKACHU's structure graphs. This example uses L-alanine, a small amino acid. The four bottom boxes in grey indicate attributes for each of PIKACHU's major classes: Structure, Atom, Bond, and Electron.

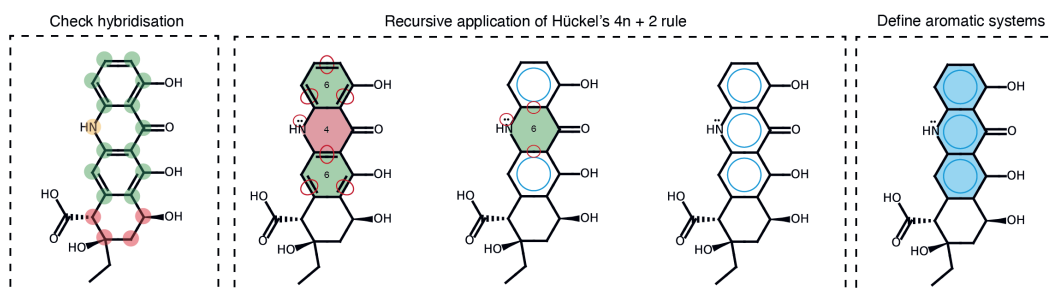


Figure 7.2. PIKACHU's recursive aromaticity detection. Aromatic cycles are individually and recursively detected and later joined into cyclic systems.

Electrons involved in σ bonds and aromatic bonds are only allocated after aromaticity detection. As electrons involved in aromatic systems are not localised to specific atoms or bonds, the p-orbitals of atoms in aromatic systems are emptied and their electrons stored in an AromaticSystem object.

Finally, any unpaired electrons are dropped back to lower-energy orbitals. A structure object is returned which can be visualised, kekulised, analysed through substructure matching and molecular fingerprinting, and altered through an assortment of built-in and custom chemical reactions.

If a SMILES string yields a structure object that is chemically incorrect due to too many or too few bonds being attached to an atom or valence shells not being filled appropriately in the case of organic atoms, PIKACHU gives a StructureError, informing the user that the parsed structure is chemically incorrect and gives a rough indication of why. Two examples of such StructureError messages are 'Error parsing "F/C(\Cl)=C(F)/Cl": Conflicting double bond stereochemistry' and 'Error parsing "CN(=O)=O": Basic bonding laws have been violated'.

7.2.3. Visualisation and kekulisation

Prior to visualisation, aromatic systems within a structure are kekulised so that aromatic systems can be represented by alternating single and double bonds. PIKACHU kekulises aromatic systems using a Python implementation³⁶⁶ of Edmonds' Blossom Algorithm for maximum matching³⁶⁷. Next, atoms are positioned using PIKACHU's drawing software. PIKACHU's python-based drawing algorithm was adapted and improved from SmilesDrawer³⁶⁸, an open-source JavaScript library for molecular visualisation. While written in different programming languages, the algorithms underlying the drawing software of PIKACHU and SmilesDrawer are largely identical. We will briefly recap this algorithm below; more detailed descriptions of the algorithm's elements can be found in the SmilesDrawer paper³⁶⁸.

First, if indicated, PIKACHU's drawing algorithm removes hydrogens from the graph. Next, it finds the smallest set of smallest rings in the structure graph. As SmilesDrawer's SSSR implementation sometimes failed to detect some rings, leading to unreadable structure renderings (Figure 7.3A, B), we implemented the SSSR algorithm ourselves. Next, like SmilesDrawer, PIKACHU classifies all rings into one of three groups: simple rings, overlapping rings, and bridged rings. Simple rings are

standalone rings that do not have any overlapping atoms with any other rings. Overlapping rings are rings that overlap with one or more other rings, where the overlap between any two rings can comprise at most two atoms, any atom in the overlap is part of at most two rings, and no atoms in the ring overlap with bridged rings. Finally, bridged rings are rings that share more than two atoms with another ring, contain atoms that are part of three or more rings, or share atoms with another bridged ring (Figure 7.4A).

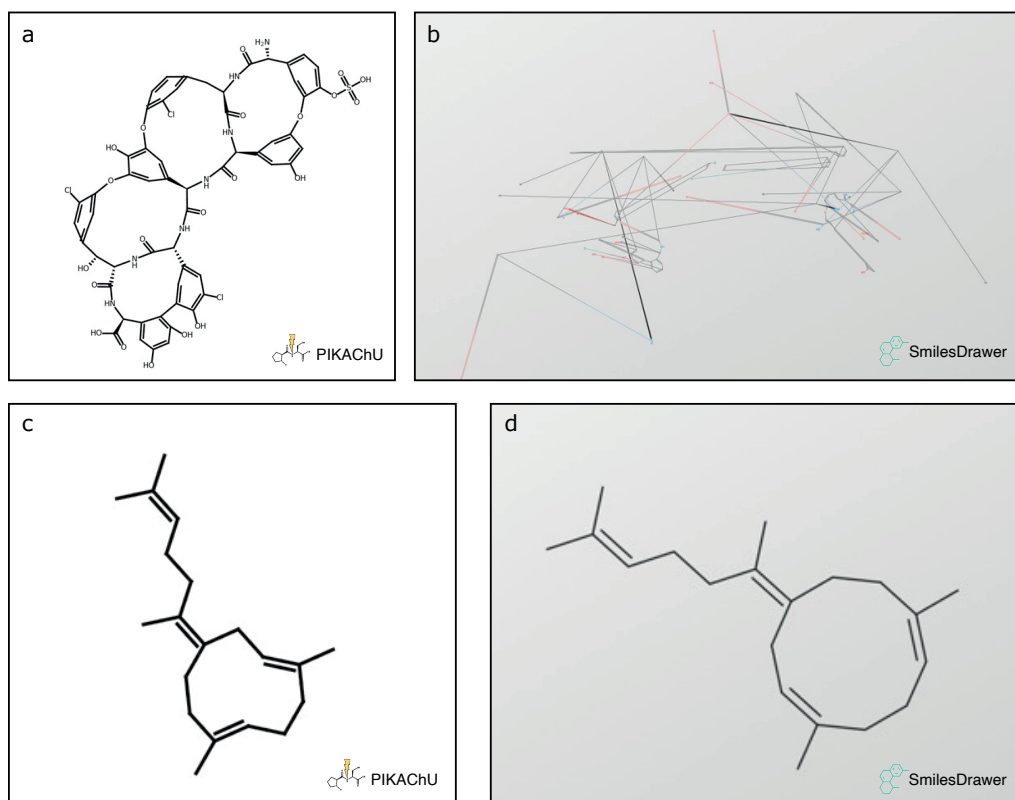


Figure 7.3. Improvements on SmilesDrawer's drawing algorithm. Macrocycle detection in PIKACHU (A) and SmilesDrawer (B), and stereobond rendering in rings by PIKACHU (C) and SmilesDrawer (D). SmilesDrawer's SSSR implementation fails to detect one of the macrocycles, leading to the unreadable structure (A). PIKACHU's SSSR implementation does recognise this ring, and therefore renders the structure correctly (B). When a stereobond occurs in a ring, PIKACHU attempts to place this bond such that the stereochemistry is visualized correctly (C) in contrast to SmilesDrawer (D).

After ring systems have been identified, atoms are placed onto a 2D coordinate system. If the molecule contains rings, positioning starts with the placement of an atom in a ring, prioritising bridged rings over simple and overlapping rings. Then, the graph is traversed one atom at a time in depth-first fashion. If an atom is part of a ring, the entire ring or ring system get placed at once. In the case of simple and overlapping rings, ring placement can be done using simple polygon geometry. For bridged rings, atoms are positioned using the force-spring model described by Kamada and Kawai³⁶⁹, where all atoms of the bridged system are initially placed in a circle, and then pulled

towards their optimal positions by minimising the difference between the desired bond length and the distance between neighbouring atoms, and maximising distances between non-neighbouring atoms. Non-ring atoms are positioned a bond length away from the previous atom, where the angle with respect to the previous atom is determined by the number of neighbours the atom has (Figure 7.4C), and the size of the molecular subtree behind each neighbouring atom (Figure 7.4D). Stereochemically restricted double bonds are always forced into the appropriate cis- or trans conformation. Unlike SmilesDrawer, which directly infers bond stereochemistry from the SMILES string, PIKACHU draws this information from bond objects stored in the molecular graph. As an improvement on SmilesDrawer, PIKACHU attempts to resolve wrongly depicted stereobonds in rings by mirroring one of the neighbouring atoms into the ring. PIKACHU always selects the atom with the smallest protruding side chain for this purpose. When multiple consecutive stereobonds are found in a ring, PIKACHU adjusts them in order, never rotating a neighbour of the same bond twice (Figure 7.3C).

Once all atoms have been assigned initial coordinates, atoms adjacent to rings are flipped outside of their ring where possible. Then, the drawing is checked for overlaps between atoms, and these overlaps are resolved by rotating branches of the molecule around single bonds. In PIKACHU, we have included an extra 'finetuning' option that is not present in SmilesDrawer. When the finetuning flag is set to True, all pairs of clashing atoms are detected. Then, the shortest path is calculated between all clashing atoms. First, PIKACHU determines which bonds are rotatable: bonds are considered unrotatable when they are a chiral bond, are adjacent to a chiral bond, or are in a cycle. As rotations around bonds located equally far away from two clashing atoms likely have the greatest impact on clash resolution, PIKACHU selects the rotatable bond that is positioned as close to the centre of the shortest path as possible. Next, PIKACHU takes the resulting set of bonds found for all clashes, and rotates each at 30° intervals, assessing and storing the number of clashes in the drawing after each iteration. The angle for which the number of steric clashes is minimised is chosen (Figure 7.4B).

Finally, some bonds adjacent to chiral centres are replaced with backward and forward wedges. They are placed such that they do not neighbour more than one chiral centre where possible, they are not part of a ring, and point in the direction of the shortest branch leading from a chiral centre, in that order of priority. The resulting image is subsequently written to a .svg or .png file or displayed directly in matplotlib.

7.2.4. Structure annotation

PIKACHU provides the option to add custom annotations to structures. Each Atom instance contains an 'annotations' attribute, which points to an AtomAnnotation instance. An AtomAnnotations instance can contain as many annotations as the user requires. Annotations can be added to all atoms in a structure at once by defining the name of the attribute with a string, and optionally providing a default value for the attribute. Subsequently, specific values can be added and retrieved for specific atoms or atom sets. A manual providing an example can be found on the PIKACHU wiki.

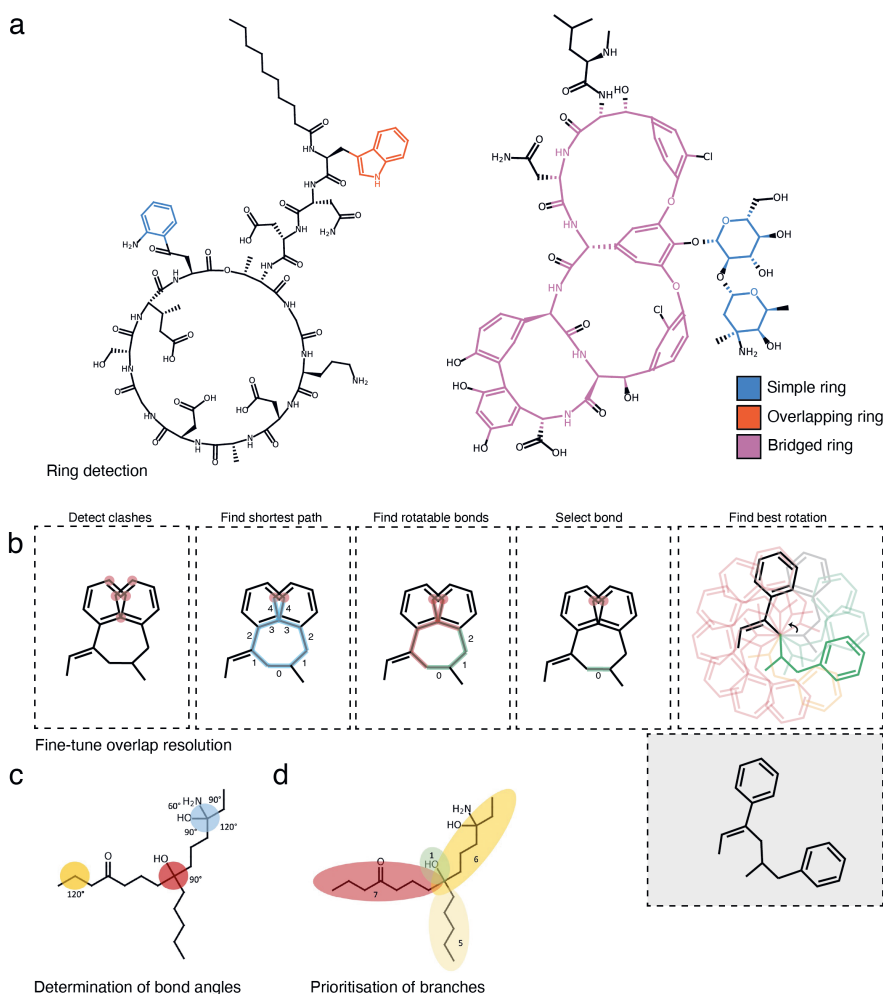


Figure 7.4. PIKACHU's drawing algorithm. A. Examples of simple (blue), overlapping (red) and bridged (pink) rings. Note that the aromatic rings in pink become part of the bridged ring system because they overlap with bridged rings. B. PIKACHU's 'fine-tune' algorithm. First, clashes are detected and the shortest path between them is found. The rotatable bond with the shortest distance to the centre of the shortest path is chosen (indicated with numbers). 12 rotations at incremental angles of 30° are evaluated for clashes. The best rotation is chosen. C. Determination of bond angles based on neighbouring atoms. If an atom has 3 or fewer non-hydrogen neighbours, the angles default to 120° (yellow). If an atom has 4 non-hydrogen neighbours, angles default to 90° if three or more of the branches have a depth more than 1, or three or four branches have a depth of exactly 1 (red). If however two of the branches have a depth of exactly 1 (blue), the angle is set to 120° between the two longest branches, 90° between any short branch and any long branch, and 60° between the shortest branches. D. Positioning of neighbouring branches depends on the depth of each branch: the two longest branches (red and dark yellow, depths 7 and 6 respectively) are always placed opposite one another.

7.2.5. Substructure matching

PIKACHU detects occurrences of a substructure in a superstructure in five steps. In all steps, hydrogens are ignored. First, PIKACHU checks for each atom type in the substructure if enough atoms of these types are accounted for in the superstructure. Second, it assesses for each atom in the substructure whether an atom exists in the superstructure with the same connectivity, looking at directly neighbouring bonds and atoms. Third, using the atom with the most diverse connectivity as a seed, it finds matches of the substructure in the superstructure using a depth-first search algorithm, ignoring stereochemistry. By first looking at atom type and atom connectivity, and by using atoms of diverse connectivity as seeds for substructure matching, the number of calls to the computationally expensive depth-first search function is minimised. Fourth, for each match, it determines if all chiral centres in the substructure have the same orientation as corresponding chiral centres in the superstructure. Fifth, PIKACHU checks if cis-trans orientation of double bonds in the substructure matches that of double bonds in the superstructure. Chiral centre and double bond stereochemistry checks can be toggled by the user independently of one another. If chirality of bonds and atoms are considered, substructures with undefined stereochemistry will still match to parent structures with defined stereochemistry. This does not apply in reverse: if a stereocentre or stereobond is defined for a substructure, it will not match to parent structures with undefined stereochemistry.

The algorithm described is somewhat similar to the Ullmann algorithm³⁷⁰, which first assesses if a candidate subgraph contains enough nodes of the correct degree prior to substructure matching and selects nodes of the most unique degree as seeds. A key difference is that PIKACHU's substructure matching algorithm also considers the identity of a node's neighbours, not just a node's degree. Substructures can be easily visualised through a range of functions in PIKACHU's 'general' library.

7.2.6. Fingerprinting

PIKACHU uses ECFP²⁵⁵, which is an improved version of the classical Morgan fingerprinting also taking into account cycle membership, to perform similarity searches and convert molecules to bit vectors for machine learning featurisation. Using Python's inbuilt hashlib library, PIKACHU initialises each atom to a 32-bit hash, derived from a tuple containing information on heavy neighbours, valence, atomic number, atomic weight, charge, hydrogen neighbours, and ring membership. Then, each atom hash is iteratively updated with hashes from its neighbours, as well as the distance from the neighbour to the atom and stereochemical information if the atom is a chiral centre. The number of iterations depends on a radius which can be set to any number (default = 2 for ECFP-4 fingerprinting). The ECFP algorithm was described in detail by Rogers and Hahn in 2010²⁵⁵. Finally, duplicate hashes are removed, as well as different hashes representing the same substructure, yielding a set of 32-bit hashes that constitutes a molecule's fingerprint.

Using ECFP fingerprinting, PIKACHU can calculate Jaccard/Tanimoto distance and/or similarity between any two molecules. Furthermore, PIKACHU can convert molecule libraries into bit vectors of varying lengths (default = 1024) and an accompanying list of substructures represented by those bit vectors that can be used in downstream machine learning algorithms.

7.2.7. Defining reaction targets

In order to facilitate implementation of reactions and reaction pathways, PIKACHU lets users define target bonds or atoms within substructures with a set of dedicated functions. These functions take a SMILES string representing a substructure, and either one or two integers that define an atom or a bond between two atoms respectively. For example, the SMILES string 'C(=O)NC', accompanied by the integers 0 (pointing to the first C atom) and 2 (pointing to the N atom), represents a peptide bond. The occurrences of these bonds/atoms are then detected within a superstructure through a substructure search and are returned as a list of bonds/atoms. Subsequently, the returned bonds/atoms can be used as reaction targets, for instance for bond hydrolysis or atom methylation, using functions in PIKACHU for breaking or creating bonds and adding or removing atoms. Reactions currently have to be encoded manually using a library of functions included in PIKACHU, which include functions for creating bonds, breaking bonds, adding and removing atoms, and splitting disconnected graphs into separate structures. We provided in-built condensation and hydrolysis functions, as well as a more elaborate ketoreductase function, as examples on our GitHub page.

7.2.8. Characterisation and visualisation of the polyketide ketoreduction reaction

We demonstrated the implementation of reactions using PIKACHU by characterising and visualising a polyketide ketoreduction reaction. We built the ketoreduction reaction by first defining a reaction target as described above, in this case a β -keto bond, and detecting its position in a polyketide chain. Next, we wrote a function that reduces the double carbonyl bond to a single bond, which identifies and removes the π -electrons in the double bond, sets the bond type to single, adjusts the hybridisation of the atoms involved and finally updates the structure object through PIKACHU's refresh functions. To finalize the reaction, two hydrogen atoms were added to the carbon and oxygen atoms of the former carbonyl bond using PIKACHU's `add_atom` function. Finally, to visualise the reaction, we highlighted the atoms and bonds of the newly formed hydroxyl group in red and drew the molecule. Detailed instructions on how to make full use of PIKACHU's range of functionalities, as well as the script used to implement the ketoreduction reaction, can be found in the online documentation.

7.2.9. Validation

To assess the correctness of PIKACHU's SMILES reading and writing software, we converted all SMILES strings from the NP Atlas database into PIKACHU Structure instances. Subsequently, we converted these structure instances back to SMILES strings. Next, we canonicalized the PIKACHU-generated SMILES and the original SMILES using RDKit⁸⁶ (v2020.09.1.0), setting the 'isomericSmiles' flag to 'True' such that correct interpretation of cis-trans bond configuration and the stereochemistry of chiral centres could also be assessed. If the two canonicalized SMILES were identical, a SMILES to structure to SMILES conversion was considered correct.

To measure PIKACHU's drawing readability, atom coordinates were computed with PIKACHU and RDKit's `rdCoordGen` module⁸⁶ (v2020.09.1.0) for the 32552 molecules in the NP Atlas database²⁴⁶ (v2021_08) and the 100,000 smallest molecules from the ChEMBL database⁸⁷ (release 30). Next, all drawings were assessed for clashes. A clash was defined as two non-neighbouring atoms sitting at less than half an average bond length distance from each other in Euclidean space. Total number of

clashes, number of structures containing clashes, and the number of structures that gave drawing errors were recorded.

To assess PIKACHU's drawing accuracy, we included a MOL file writer into PIKACHU, which stores PIKACHU-computed atom coordinates and connectivities as a MOL file. We generated such MOL files for the entire NP Atlas database and the 100,000 smallest molecules from the ChEMBL database, read the resulting MOL files with RDKit's `rdMolFiles` module (v2020.09.1.0), stored the resulting molecules as SMILES strings, and using RDKit canonicalized both the original input SMILES and the SMILES produced from the PIKACHU-generated MOL files setting the 'isomericSmiles' flag to 'True'. If the SMILES were identical, a PIKACHU-generated drawing was considered 'correct'.

7.2.10. Speed assessment

PIKACHU's speed was assessed with Python's 'time' module. As we particularly designed PIKACHU with natural product chemistry in mind, which typically involves larger and more heavily cyclised compounds than most molecules stored in small-molecule databases, we decided to test drawing speed on two different databases: the NP Atlas database²⁴⁶ and the ChEMBL database. For each database⁸⁷, we randomly selected 10,000 molecules and timed drawing speed at 10, 20, 50, 100, 200, 500, 1000, 2000, 5000 and 10,000 drawn structures using Python's 'time' module.

7.3. Results and Discussion

PIKACHU is a dependency-light cheminformatics kit implemented entirely in Python. With only matplotlib as dependency and an extensive readme, wiki, tutorials, and example scripts on its GitHub page, PIKACHU is easy to run and install, and suitable for integration into bioinformatics and cheminformatics pipelines. Below, we will first assess PIKACHU's ability to correctly interpret SMILES, draw structures accurately and readably, detect and visualize substructures, and perform ECFP fingerprinting. We also measured PIKACHU's SMILES reading and structure drawing speed. Next, we demonstrate how PIKACHU can be used to implement and visualise reactions. Finally, we compare PIKACHU to the state-of-the-art cheminformatics kits/chemical drawing libraries RDKit, ChemDraw and SmilesDrawer.

7.3.1. SMILES reading and writing

We assessed PIKACHU's ability to parse and generate correct SMILES syntax by comparing the SMILES it converts to SMILES generated by RDKit, a well-established cheminformatics package. As PIKACHU was created with natural product chemistry in mind, which typically involves large and heavily cyclised molecules the function of which depends heavily on stereochemistry, we performed our validation with the NP Atlas database. This database contains 32,552 manually curated natural product structures and their corresponding isomeric SMILES strings. PIKACHU failed to convert 22 SMILES from NP Atlas (~0.07%) to structure graphs all of which were erroneous SMILES describing nitrogen atoms with a valency of 5, which is impossible considering that nitrogen only has four electron orbitals available for bonding in its valence shell. This demonstrates how the detailed graph-based, object-oriented encoding of chemical structures down to the electron level in PIKACHU intrinsically ensures that all structures that are loaded are chemically valid. Of the remaining 32,530 SMILES-to-structure-graph-to-SMILES conversions, only one yielded a SMILES string that described different chemistry than the original: three carbon-13 atoms were interpreted as carbon-12 (online

Supplementary Table S1, third row). As PIKACHU does not yet support isotopic differentiation, this is not unexpected.

Additionally, we manually assessed the correctness of 22 SMILES-to-graph conversions by reading in and subsequently drawing the SMILES in PIKACHU. We chose the SMILES such that a variety of syntax representations and chemistries were represented, including rings, aromatic systems, charge, stereocentres and bond stereochemistry. Some SMILES describe the same structures but use a different syntax. PIKACHU handled all SMILES correctly, accurately detecting and visualising all aforementioned chemical properties (Figure 7.5).

PIKACHU is not suitable for reading in molecules with a high number of recursive cycles, such as buckminsterfullerene. As PIKACHU detects all possible cycles within a molecule to determine aromaticity of cyclic systems, this step takes so long to compute that the program appears to get 'stuck'. However, there exist only a handful of examples of such molecules, none of which have any real practical biological or chemical relevance.

7.3.2. Structure visualisation

Another key feature of PIKACHU is molecular visualisation from SMILES. PIKACHU's drawing software relies on similar logic to that of SmilesDrawer, a JavaScript SMILES drawing library. In our software, we added a few improvements: we fixed cis-trans stereochemistry detection (Figure 7.3C, D), included an extra overlap resolution step (Figure 7.4B), and implemented an improved version for finding the smallest subset of smallest rings, a key step in correctly depicting cycles (Figure 7.3A, B). There is always a bit of debate regarding the visualisation of molecular macrocycles. Many organic chemists opt for a 'honeycomb' architecture, as employed by ChemDraw and CDK, to better represent the 3D architecture of a molecule, hinting at long-distance interactions that may take place within the compound. However, this representation does not instantly draw the eye to sites of cyclisation, a drawback for natural product biologists and bioinformaticians who are often interested in the biosynthetic steps involved in a compound's assembly. As PIKACHU was created with natural product chemistry in mind, we chose to use a polygon representation for macrocycles, which clearly shows cyclisation sites (online Supplementary Figure S4). While PIKACHU always detects and interprets aromaticity internally, it currently only supports drawing structures in a kekulised format.

The most important aspect of automated molecular visualisation is accuracy: users need to be able to rely on the correctness of drawing software, especially when processing a large number of structures at once making it impossible to inspect each image independently. To this purpose, we visualised a chemically diverse set of structures from the ChEMBL and NP Atlas databases, and tested if RDKit could interpret correct chemistry from PIKACHU-generated atom coordinates. Out of the 32,552 structures in the NP Atlas database, only 40 (~0.12%) were drawn incorrectly. Of these, 33 were drawn wrongly due to incorrect depiction of cis-trans chemistry of double bonds adjacent to nested rings (online Supplementary Table S1). Additionally, PIKACHU failed to convert 29 structures to drawings (~0.09%). Of these, 22 were the same incorrectly defined SMILES describing nitrogens with a valency of 5 that were found previously. The remaining failures largely resulted from ChiralityErrors: errors raised by PIKACHU when it cannot correctly depict cis/trans chemistry of a double bond. With 32,483 correctly drawn structures, PIKACHU achieves a drawing accuracy of 99.79%. PIKACHU performed comparably on the 100,000 smallest molecules from the ChEMBL database (99.21% accuracy, 0.16% incorrect drawings, 0.63% unsuccessful conversions; Figure 7.6A).

Comprehensive lists and example depictions of SMILES leading to incorrect interpretations can be found in online Supplementary Table S1 and online Supplementary File S1.

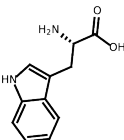
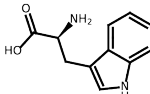
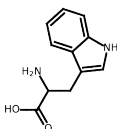
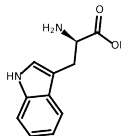
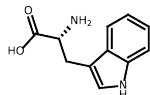
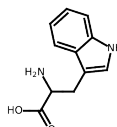
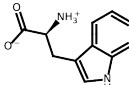















L-Tryptophan	 <chem>C1=CC=C2C(=C1)C(=CN2)C[C@H](C(=O)O)N</chem>	 <chem>c1[nH]c2ccccc2c1C[C@H](N)C(=O)O</chem>	 <chem>NC(Cc1c[nH]c2ccccc12)C(=O)O</chem>		
D-Tryptophan	 <chem>C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(=O)O)N</chem>	 <chem>c1[nH]c2ccccc2c1C[C@@H](N)C(=O)O</chem>	 <chem>NC(Cc1c[nH]c2ccccc12)C(=O)O</chem>		
L-Tryptophan zwitterion	 <chem>c1[nH]c2ccccc2c1C[C@H](N)C(=O)[O-]</chem>				
cis-difluoromethylethylene	 <chem>F/C=C(F)/C</chem>	 <chem>F/C=C(F)C</chem>	 <chem>F/C=C(F)C</chem>	 <chem>C/C(F)=C/F</chem>	 <chem>F/C(C)=C/F</chem>
trans-difluoromethylethylene	 <chem>F/C=C(F)/C</chem>	 <chem>F/C=C(F)C</chem>	 <chem>F/C=C(F)C</chem>	 <chem>C/C(F)=C/F</chem>	 <chem>F/C(C)=C/F</chem>
benzene	 <chem>c1ccccc1</chem>	 <chem>C1C=CC=CC=1</chem>	 <chem>C1=CC=CC=C1</chem>	 <chem>C=1C=CC=CC1</chem>	 <chem>C=1C=CC=CC=1</chem>

Figure 7.5. Assessment of PIKACHU's SMILES reader. Structures were drawn from the SMILES written beneath the molecule depictions. All 22 structures were correctly drawn.

We additionally assessed the readability of these PIKACHU-rendered drawings by automatically detecting steric clashes from PIKACHU-generated atom coordinate sets. Only 4.95% of NP Atlas SMILES renderings contained steric clashes, with ~ 1.62 clashes per clashing structure. PIKACHU was better at processing the ChEMBL database, with only $\sim 0.30\%$ of drawings containing clashing atoms. This makes sense, as NP Atlas contains a higher proportion of highly cyclised systems and large molecules, properties which make it more difficult to readably depict a molecule in a plane.

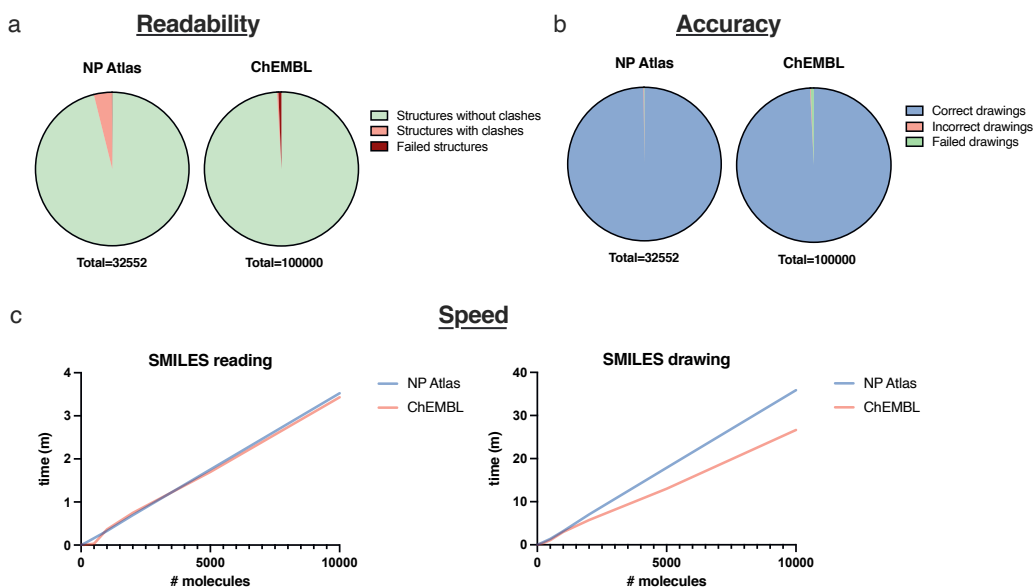


Figure 7.6. PIKACHU's performance tested on the NP Atlas and ChEMBL databases. For drawing readability and accuracy assessment, the entire NPAtlas database and the 100,000 smallest molecules of the ChEMBL database were tested. For speed assessment, 10,000 random molecules from each database were used. A. Structure readability expressed as the percentage of molecules with steric clashes. B. Drawing accuracy expressed as the percentage of drawings correctly interpreted by RDKit upon writing PIKACHU-calculated atom coordinates to a .mol file. C. PIKACHU's SMILES reading speed and drawing speed in molecules per minute.

In Figure 7.7, we show nine examples of structures rendered by PIKACHU. Due to the drawing algorithm that PIKACHU employs for complex ring systems, 5-membered and 6-membered rings often appear distorted, as observed for vancomycin and aplasmomycin B. Additionally, PIKACHU's overlap resolution step, while resolving a lot of steric clashes, sometimes results in carbon-carbon bonds being placed at an 180° angle which makes the structure less interpretable, as seen for PIKACHU's depiction of the molecule nanokid.

7.3.3. Speed assessment

We assessed PIKACHU's SMILES reading and structure drawing speed by drawing 10,000 random molecules from the NP Atlas database and the ChEMBL database (Figure 7.6C). With an average reading speed of $\sim 2,874$ SMILES per minute and an average drawing speed of ~ 279 molecules per

minute for the NP Atlas database and ~375 molecules per minute for the ChEMBL database on a single laptop core, it is clear that PIKACHU spends the bulk of its time rendering an image on atom positioning, not on SMILES reading. The discrepancy between the two databases can be explained by the nature of the molecules contained within them: typically, natural products are larger and more cyclised than the average small molecule. This makes PIKACHU's drawing speed one order of magnitude slower than RDKit's (online Supplementary Table S2), which is expected considering that PIKACHU is a pure Python package while RDKit generates drawings with pre-compiled C++ code. Also, PIKACHU's finetuning step is computationally expensive, likely leading to an increase in computational time. Still, PIKACHU is fast enough for integration into pure-Python bioinformatics and cheminformatics pipelines.

7.3.4. Substructure detection

A dedicated set of functions ensures that performing substructure searches using PIKACHU is straightforward. With a single line of code, users can visualise a single occurrence of a substructure, all occurrences of a substructure, or all occurrences of a range of substructures in a chemical compound (Figure 7.8A). Substructure searches are fast due to several pre-processing steps, ensuring that the expensive graph matching algorithm is only executed when a match is likely. Stereochemistry matching, activated by default, can be toggled on and off.

With PIKACHU's substructure matching algorithm, we visualised the amino acid composition of the cyclic peptides daptomycin and vancomycin, using only a single line of code for each (Figure 7.8B). Colours are fully and easily customisable and can be provided as hex codes or as colour names.

7.3.5. ECFP fingerprinting

To quickly determine the approximate similarity between two molecules, PIKACHU employs ECFP fingerprinting²⁵⁵. PIKACHU hashes each molecule into a set of unique identifiers, each of which represents a substructure. Collectively, these identifiers make up a molecule's fingerprint. Then, PIKACHU calculates the Jaccard/Tanimoto similarity between two molecules by comparing their fingerprints, giving a measure of molecular similarity and/or distance.

Here, we showcase PIKACHU's ECFP fingerprinting by calculating and subsequently constructing a tSNE plot of the molecular distances between 36 calcium-dependent lipopeptides. Lipopeptides of the same family grouped together (online Supplementary Figure S5), confirming that PIKACHU's ECFP fingerprinting yields reliable measures of chemical similarity.

Additionally, PIKACHU's ECFP fingerprinting makes it possible to generate bit vectors from molecule sets, where each element in the vector represents the presence/absence of a specific substructure. These can subsequently be used as interpretable molecular featurisations for machine learning.

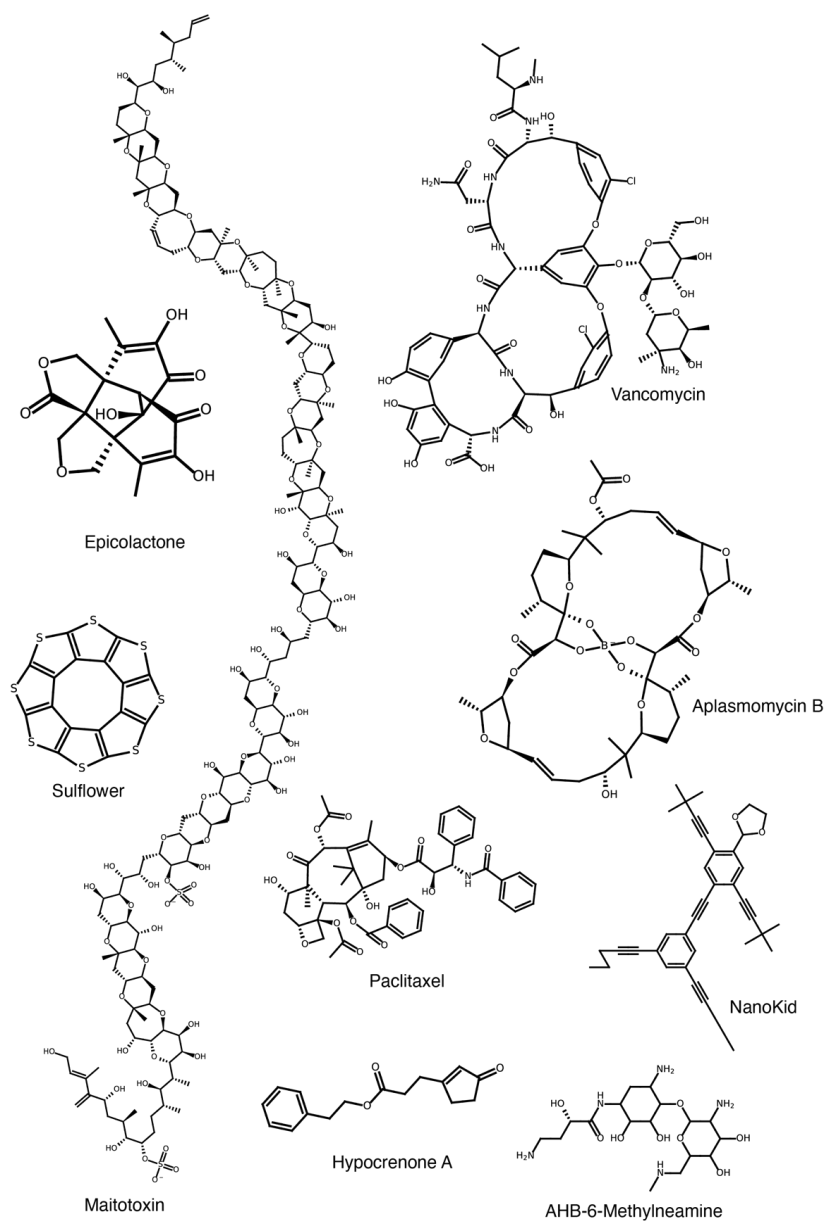


Figure 7.7. Various molecules rendered by PIKACHU.

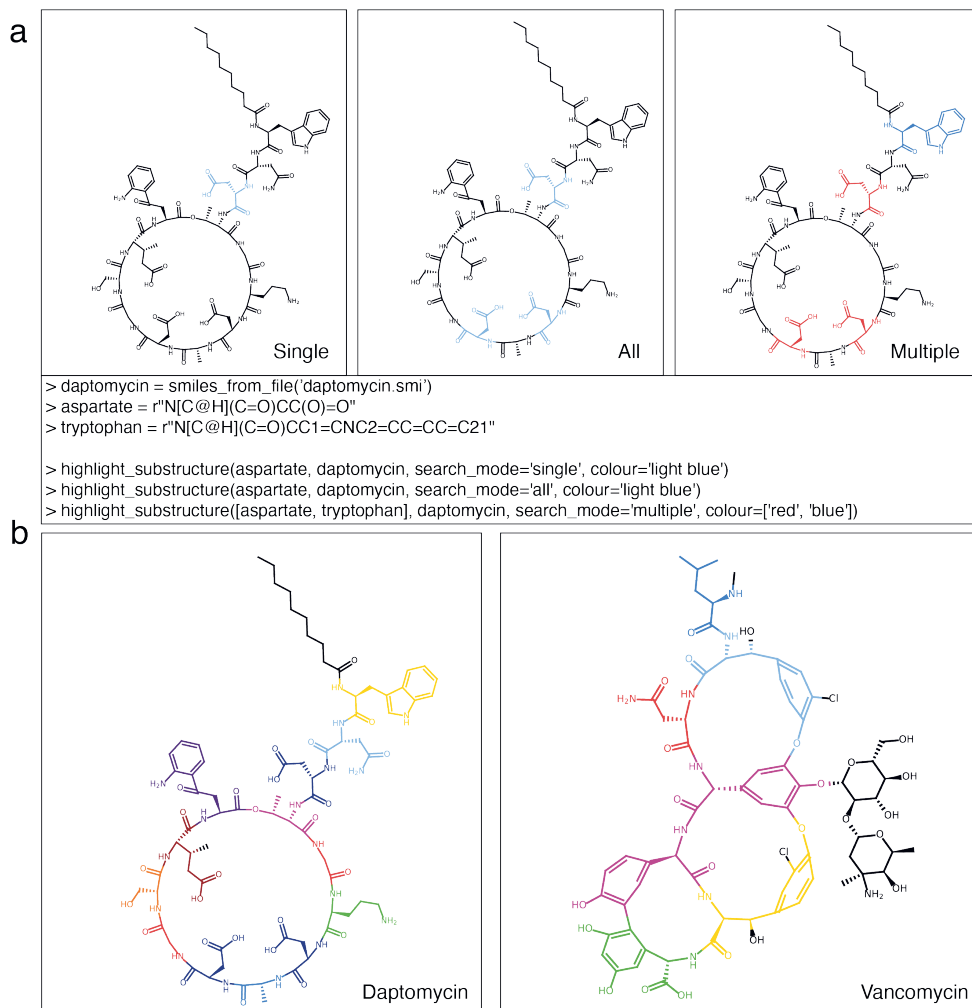


Figure 7.8. Substructure matching with PIKACHU. A. From left to right: examples of highlighting a single instance of a substructure, all instances of a substructure, or all instances of multiple substructures. In the example, occurrences of aspartic acid and tryptophan were searched in the superstructure daptomycin. The code used to generate the images is displayed underneath the panels. B. PIKACHU's substructure matching algorithm using to visualise all amino acid components of the antibiotics daptomycin (left) and vancomycin (right).

7.3.6. Building *in silico* reactions using PIKACHU

PIKACHU provides a platform for the creation and visualisation of reaction mechanisms by providing a range of reaction functions that can be used to make or break molecular bonds, add, or remove atoms and alter the chirality of stereocentres. In addition to these built-in reaction building blocks, PIKACHU allows users to easily define more complex reactions through the manipulation of atom- and bond object attributes. Additionally, PIKACHU supports fully customisable structure annotation,

which is useful for keeping track of reaction steps, reaction targets, or atom origin. As a proof of principle, we used PIKACHU to define and visualise a polyketide ketoreduction reaction, catalysed by a ketoreductase polyketide synthase domain during polyketide synthesis, employing both built-in and custom reaction functions (online Supplementary Figure S6). This example, as well as a comprehensive guide containing instructions on how to build reaction mechanisms using PIKACHU, can be found in the online documentation.

While creating reaction pathways with PIKACHU enforces chemically correct conversions by checking at each step if a structure is chemically correct, it is more laborious than similar functionalities in other cheminformatics kits such as RDKit, which use reaction SMILES and atom mapping to perform chemical reactions. We intend to implement reaction SMILES and atom mapping into PIKACHU in the future.

7.3.7. *PIKACHU compared to state-of-the-art chemical drawing software*

Finally, we assessed how PIKACHU performs compared to existing chemical drawing software. To this purpose, we visualised various structures in PIKACHU (v1.0.4), RDKit (v2020.09.1.0), ChemDraw (v20.1.0.112) and SmilesDrawer (v1.2.0), and manually assessed drawing quality and correctness (Figure 7.9). Only SmilesDrawer occasionally produced an incorrect structure, confusing cis-trans stereochemistry when stereochemistry is defined in or after a branch (Figure 7.9A). For heavily cyclised molecules (Figure 7.9B-E), we clearly see the difference between the ‘honeycomb’ (RDKit and ChemDraw) and the ‘polygon’ (PIKACHU and SmilesDrawer) approaches of cycle positioning. The honeycomb approach ensures minimal distortion of microcycles, even when they are part of larger systems; as such, RDKit and ChemDraw render molecules such as vancomycin significantly better than SmilesDrawer and PIKACHU (Figure 7.9B). However, when the honeycomb approach does not work because of steric constraints, forcing microcycles into regular polygons can distort the macrocyclic structure to the extent that the drawing becomes unreadable. This is the case for aplasmomycin B and epicolactone (Figure 7.9C, D), which are drawn considerably better by SmilesDrawer and PIKACHU. In structures where microcycles and macrocycles are separate, there is little difference in structure rendering between the two approaches (Figure 7.9E).

PIKACHU has a slight advantage over RDKit in drawing molecules of varying sizes, automatically adjusting the canvas size based on the size of the molecule to be drawn. This also means that PIKACHU’s font size, bond length and bond thickness maintain a constant ratio across different drawings, which is not the case for RDKit (Figure 7.9D, E). While it is possible to manually adjust canvas size in RDKit, some extra coding steps are required to achieve this.

PIKACHU’s visual output is far more customizable than that of SmilesDrawer, allowing for molecule rotation, drawing multiple molecules on a single canvas, and custom colouring of each individual bond and atom, supporting hex-codes as well as a range of descriptive strings.

While ChemDraw accommodates high-quality and highly customisable visualisation, it is not open source. This makes PIKACHU more suitable for integration into automated open-source pipelines required by many projects.

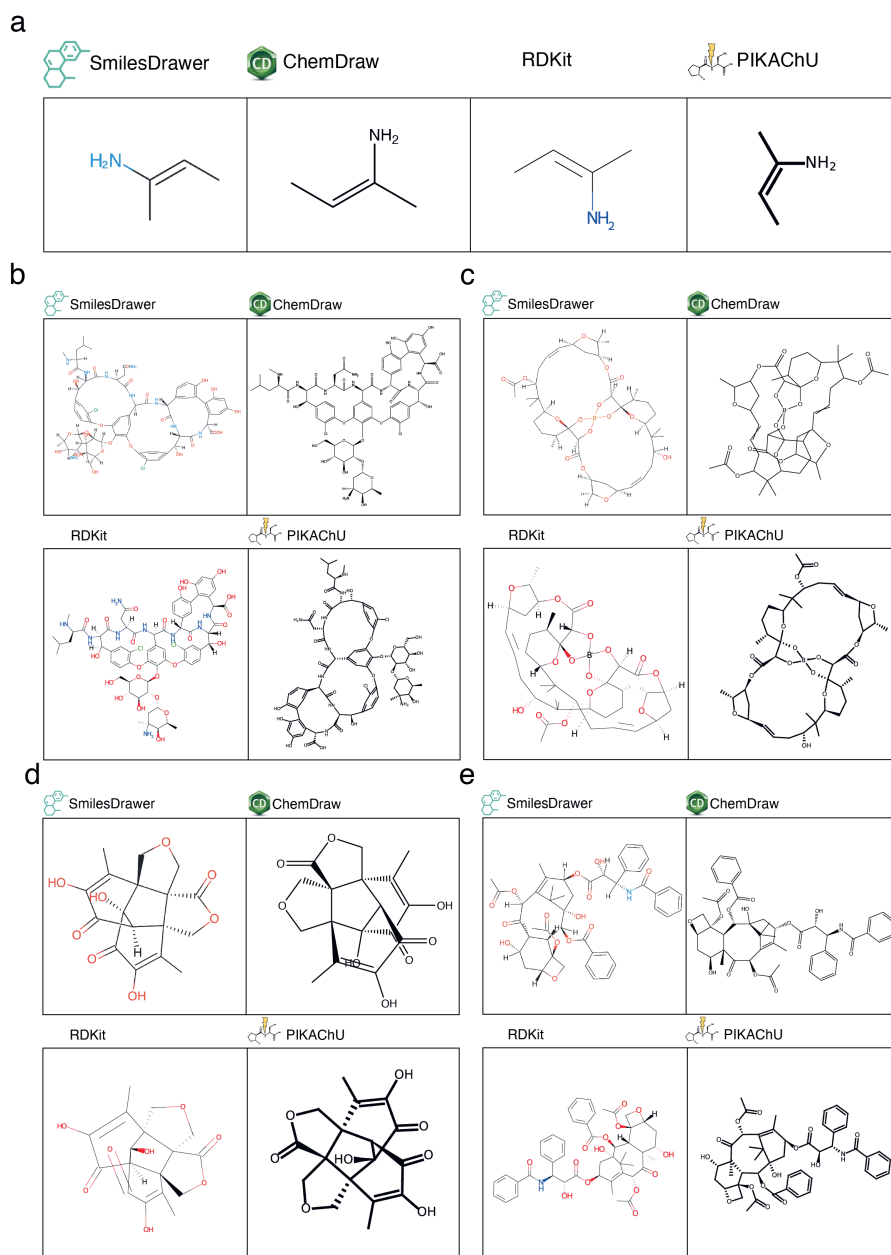


Figure 7.9. Comparison of PIKACHU to other chemical drawing software. A. SmilesDrawer, ChemDraw, RDKit and PIKACHU drawings given the SMILES string 'C/C=C(\N)/C'. While ChemDraw, RDKit and PIKACHU all draw the cis-trans stereochemistry of the double bond correctly, with the amino group cis of the methyl group, SmilesDrawer draws the stereobond in the wrong orientation. B. SmilesDrawer, ChemDraw, RDKit and PIKACHU drawings of the heavily cyclised molecule vancomycin. C. SmilesDrawer, ChemDraw, RDKit and PIKACHU drawings of the molecule aplasmomycin B. D. epicolactone, and E. paclitaxel.

7.4. Conclusions

We developed PIKACHU, a dependency-light cheminformatics library implemented entirely in Python. Having extensively validated our software, we conclude that, while RDKit heavily outperforms PIKACHU in terms of speed, PIKACHU performs sufficiently fast and reliably to be suitable for cheminformatics and bioinformatics pipelines. Backed by extensive online documentation, easy and straightforward installation, and state-of-the-art automated visualisation software, we hope that PIKACHU can provide a convenient alternative for chem- and bioinformaticians programming in Python.

Chapter 8

RAIChU: automating the visualisation of natural product biosynthesis

Barbara R. Terlouw*, Friederike Biermann*, Sophie P.J.M. Vromans*, Eric J. N. Helfrich, Marnix H. Medema

* These authors contributed equally to this work

Abstract

Natural products are molecules that fulfil a range of important pharmaceutical and agricultural functions. In contrast to many other natural products, the products of modular non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) systems can often (partially) be predicted from the sequence of the biosynthetic gene cluster (BGC) that encodes their production, as their biosynthetic pathways follow consistent rulesets. This allows automated visualisation of these pathways, which would accelerate pathway drawing and reduce human error. Still, there currently exists no software that leverages these rules to automate visualisation of (predicted) NRP and polyketide biosynthesis.

To enable high-quality automated visualisations of natural product biosynthetic pathways, we developed RAICHU (Reaction Analysis through Illustrating Chemical Units), which produces depictions of detailed biosynthetic mechanisms of PKS, NRPS, and hybrid PKS/NRPS systems from predicted or experimentally verified module architectures. RAICHU can be integrated into Python pipelines, or can be installed as a stand-alone interactive application, allowing users to upload and edit results from antiSMASH, a widely used BGC detection- and annotation tool, or to build PKS/NRPS systems from scratch. In addition, RAICHU boasts a library of functions to perform and visualise reactions and pathways that are still difficult to predict, including 32 tailoring reactions that are prevalent in the biosynthesis of terpenoids, RiPPs, alkaloids, NRPs and polyketides. RAICHU's cluster drawing correctness (100%) and drawing consistency (100%) were validated on 1500 randomly generated PKS/NRPS systems, and on the MIBiG database.

RAICHU is available at <https://github.com/SophieVromans/RAICHU>. Online supplementary material is available at <https://zenodo.org/record/8009605>.

8.1. Introduction

Natural products are abundant in nature, produced by a range of microbes across different domains of life^{38,371}. They are structurally highly diverse, fulfilling various important pharmaceutical, agricultural, and ecological functions^{46,372,373}. Famous examples include the fungal NRP antibiotic penicillin¹³, the bacterial polyketide insecticide Spinosad^{32,374}, the polyketide rapamycin³⁷⁵, an immunosuppressant, and the antibacterials daptomycin²⁷ and erythromycin³⁷⁵, a NRP and a polyketide respectively.

Typically, natural products are encoded by biosynthetic gene clusters (BGCs): groups of genes that are physically collocated in a genomic region and collectively encode a biosynthetic pathway. BGC architecture is highly varied and depends on natural product class. Broadly, we can subdivide BGC architectures into modular and non-modular systems.

Modular BGCs typically contain one or more core genes that encode enzymes producing the main scaffold of a natural product. For NRPs and polyketides, two natural product classes whose production is encoded by modular BGCs, these enzymes are called non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), respectively. The core enzymes of NRPSs and PKSs consist of multi-domain modules, each of which incorporates a peptide or polyketide monomer into the natural product scaffold in an assembly-line-like

fashion^{46,373}. The modularity of NRPS and PKS enzymes generates a rich combinatorial diversity, with evolution driving the inclusion, removal, exchange, and recombination of modules and domains resulting in new NRPS, PKS and even NRPS-PKS hybrid enzymes^{46,373,376}. Structural diversity is further enhanced by accessory genes in NRPS and PKS BGCs, which encode enzymes that catalyse post-assembly tailoring reactions such as cyclisation, halogenation, formylation, and glycosylation^{377,378}.

While the chemistry is different for NRPS and PKS modules, their architectures are remarkably similar. For both, a minimal module comprises three domains: a non-catalytic carrier domain that functions as a sulphur-based anchor for the to-be-incorporated building block; a recognition domain that selects which building block is incorporated and covalently attaches it to the carrier domain; and a synthesis domain, that removes the building block from the carrier domain and covalently fuses it to the growing natural product scaffold. For NRPS and PKS systems respectively, carrier domains are called peptidyl carrier protein (PCP) domains and acyl carrier protein (ACP) domains, recognition domains are called adenylation (A) and acyltransferase (AT) domains, and synthesis domains are called condensation (C) and ketosynthase (KS) domains^{45,379,380}.

In addition to these core domains, NRPS and PKS modules can also contain in-line tailoring domains that modify the incorporated building blocks. For NRPS modules, the most prevalent of these are epimerization (E) domains, which change the chirality of the α -carbon of amino acid building blocks, and N-methylation (nMT) domains, which methylate the amino group of amino acid subunits. Sometimes, cyclisation (Cyc) and oxidation (Ox) domains can also occur. These domains typically work sequentially, and catalyse cyclisation between a threonine, serine or cysteine residue with the amino acid backbone and subsequent oxidative bond formation on the resulting ring, respectively. In PKS enzymes, tailoring domains act sequentially as well: the ketoreductase (KR) domain reduces the β -ketoacyl-S-ACP formed by the previous module to a β -hydroxyacyl-S-ACP, the dehydratase (DH) domain forms a double bond by removing the β -hydroxy group and forming a α,β -enoyl-S-ACP intermediate, and the enoylreductase (ER) domain reduces this double bond, creating a saturated acyl-S-ACP. Terminal NRPS and PKS modules may additionally contain a thioesterase (TE) or terminal reductase (TD) domain, which catalyses the release of the natural product scaffold from the enzyme, either as a linear free acid or a macrocyclic lactone/lactam^{379–381}. The most commonly occurring PKS/NRPS domains and the reactions they catalyse are summarised in Figure 8.1.

The structure of a natural product scaffold is not only determined by domain and module composition, but also by the substrate specificity of the recognition domains, which select a building block from a wide variety of substrates each resulting in a scaffold with different properties. This is especially true for NRPS A domains, which recognize as many as five hundred different substrates, including L- and D-amino acids, a wide variety of non-proteinogenic amino acids, and even fatty acids, aryl acids and hydroxy acids, greatly exceeding the level of peptide sequence diversity achievable through ribosomal peptide synthesis^{61,382}. In contrast, the AT domains of PKS modules select a modest variety of acyl-CoA thioester substrates. In most cases, malonyl-CoA and its methylated variant methylmalonyl-CoA are used as building blocks²⁹⁴. Software tools exist that predict the specificity of modular NRPS and PKS recognition domains from their protein sequence, such

as the AT domain specificity predictor published by Minowa *et al.*²⁵⁶, the A domain substrate predictors SANDPUMA⁸⁰, NRPSPredictor2⁷⁷ and AdenylPred⁸², several of which have been incorporated into larger BGC detection and analysis tools such as antiSMASH⁶⁷ and PRISM⁷⁸.

In *trans*-AT PKS systems, modules lack the AT domain. Instead, an *in-trans* acting AT domain outside of the PKS enzyme, often encoded within the same BGC by a stand-alone gene, activates malonyl-CoA and loads it onto the assembly line. Often, these substrates are modified prior to incorporation into the polyketide scaffold. Recently, the phylogenetic tool transATor³⁸³ was developed, which computationally infers these modifications from the substrate specificity of the KS domain of the module.

As the catalytic domains of modular NRPS and PKS systems are so well-characterised and building blocks selected by recognition domains can be predicted, it is possible to predict the biosynthetic pathway and thus the core scaffold of the natural products they produce with reasonable accuracy. While BGC detection platforms like antiSMASH and PRISM provide predicted scaffolds, they do not supply detailed visualisations of their biosynthesis process^{67,78}. Such visual insight into the reaction mechanism as well as the final product is desirable for a number of reasons: it gives an instant overview to researchers not familiar with natural product biosynthesis of the underlying biosynthetic chemistry; it facilitates the identification of mistakes in structure predictions and links them to specific mispredicted biosynthetic reactions. However, manually drawing (predicted) reaction mechanisms is a laborious and error-prone process, which, if automated, could save researchers a lot of time and could remove human error, especially when processing large (meta)genome datasets.

To enable automated, high-quality visualisation of NRPS and PKS reaction mechanisms, we developed RAICHU: Reaction Analysis through Illustrating Chemical Units. Given a set of input domains and predicted or experimentally verified A/AT-domain specificities, RAICHU automatically produces ‘spaghetti diagrams’ (Figure 8.2) depicting detailed biosynthetic mechanisms of PKS, NRPS, and hybrid PKS/NRPS BGCs at module-resolution. RAICHU can also be run as an interactive application, where antiSMASH results can be uploaded and/or edited, NRPS/PKS BGCs can be built from scratch, and intermediate results can be visualised. RAICHU thus provides a solution for biologists and bioinformaticians looking to automatically visualise biosynthetic pathways to facilitate the study of natural products.

8.2. Results and Discussion

We created RAICHU: a software package which automatically visualises biosynthetic pathways encoded by *cis*-AT PKS, *trans*-AT PKS and NRPS BGCs as ‘spaghetti diagrams’ (Figure 8.2). Additionally, RAICHU can visualise pathways of RiPP, amin-acid-derived alkaloid, and terpenoid biosynthesis and boasts a collection of tailoring enzyme reactions to decorate natural product scaffolds. RAICHU comprises two main libraries: the first is a reaction library which performs the *in silico* chemistry required for NRPS, PKS, alkaloid, RiPP, and terpenoid assembly and tailoring; and the second renders cluster visualisations as saveable images. Below, we discuss each of these libraries and assess RAICHU’s speed.

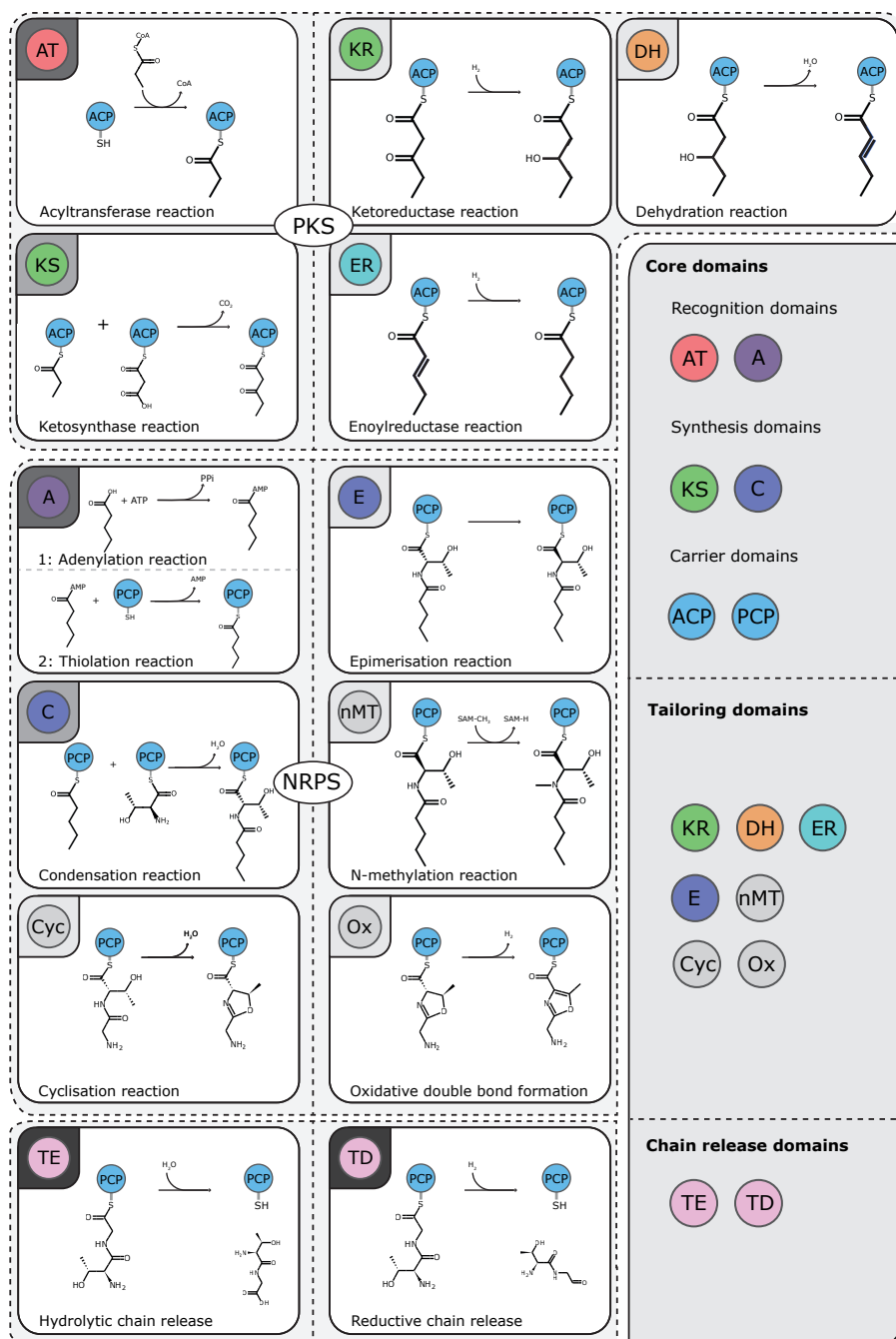


Figure 8.1. Reactions catalysed by core domains and common tailoring domains in PKS and NRPS biosynthetic gene clusters, which have been implemented into RAICHU. Carrier domains (light blue) are not catalytic, but function as tethers to which NRP and polyketide intermediates are covalently linked.

8.2.1. *In silico reactions*

Biosynthetic assembly lines catalyse many biochemical conversions. To ensure the correctness of these conversions, we implemented the most commonly occurring elongation and tailoring reactions using PIKACHU, a tool which enforces chemical correctness of reaction intermediates by monitoring electron availability for the formation of bonds and lone pairs²⁵⁴. This is different from the approach taken by antiSMASH, which relies on SMILES⁸⁵ concatenation to compute putative products, at times leading to the presence of extra atoms at both termini of the linear molecule⁶⁷. As in PRISM, we opted for a graph-based approach, which better represents the underlying chemistry and enables the implementation of reactions that are difficult to achieve through linear SMILES concatenation, like cyclisations⁷⁸.

8.2.1.1. *Modular systems*

For assembly line-like pathways like NRPS/PKS systems, reactions are determined and computationally executed per module based on domain composition and involve electron redistribution to break and form bonds. For both NRPS and PKS systems, we defined three module types: starter modules, elongation modules, and termination modules. In RAICHU, every cluster should be defined with exactly one starter module, followed by any number of elongation modules, and optionally completed with one termination module. Cis-activating NRPS and PKS starter modules minimally require a recognition (A/AT) domain and a carrier (PCP/ACP) domain, elongation modules additionally require a synthesis (C/KS) domain, and termination modules are elongation modules with a terminal TE or TD domain.

Trans-activating PKS modules do not require a recognition (AT) domain, but minimally require a KS and ACP domain for starter, elongation, and termination modules. Additionally, each module can contain tailoring domains, each corresponding to a tailoring reaction executed by RAICHU. NRPS modules can contain E, nMT, Cys and Ox tailoring domains, executing epimerisation, N-methylation, cyclisation, and oxidative double bond formation respectively; and PKS modules can contain KR, ER, and DH tailoring domains, catalysing ketoreduction, enoylreduction and dehydration (Figure 8.1). Any domain can be set to 'inactive', which will ignore the domain in scaffold assembly. Within a module, only the first active domain of a domain type will be considered as active; any subsequent domains of the same type are automatically set to inactive. In trans-AT systems, tailoring reactions predicted from the KS domain are automatically performed by assigning one of the 44 KS domain subtypes supported by antiSMASH. Tailoring reactions are only executed if the reaction is chemically possible; otherwise, the domain executing the reaction is automatically set to 'inactive'.

In addition to module architecture, RAICHU also requires the substrate selectivity of each recognition domain. To ensure compatibility between RAICHU and antiSMASH, a widely used BGC prediction and annotation resource, we selected those substrates that antiSMASH can predict, in addition to substrates that occurred in the training sets of the A domain specificity predictors AdenylPred⁸², NRPSPredictor²⁷⁷, SANDPUMA⁸⁰, and PARAS/PARASECT (see Chapter 5). For A domains in NRPS elongation and termination modules, RAICHU supports 126 substrates: the 20 proteinogenic amino acids, their 19 D-forms (as glycine is not chiral), 43 derivatives of proteinogenic amino acids, 43 non-proteinogenic L- and D-amino acids, and 1 wildcard. There are an additional 17 non-amino acid substrates and a customisable set of fatty acids up to C20 that can only be incorporated in NRPS starter modules. For AT domains in PKS

elongation and termination modules, there is a choice of 5 malonyl-CoA derivatives and 1 wildcard. PKS starter modules support a wider variety of 15 chemically diverse substrates (Supplementary File 1).

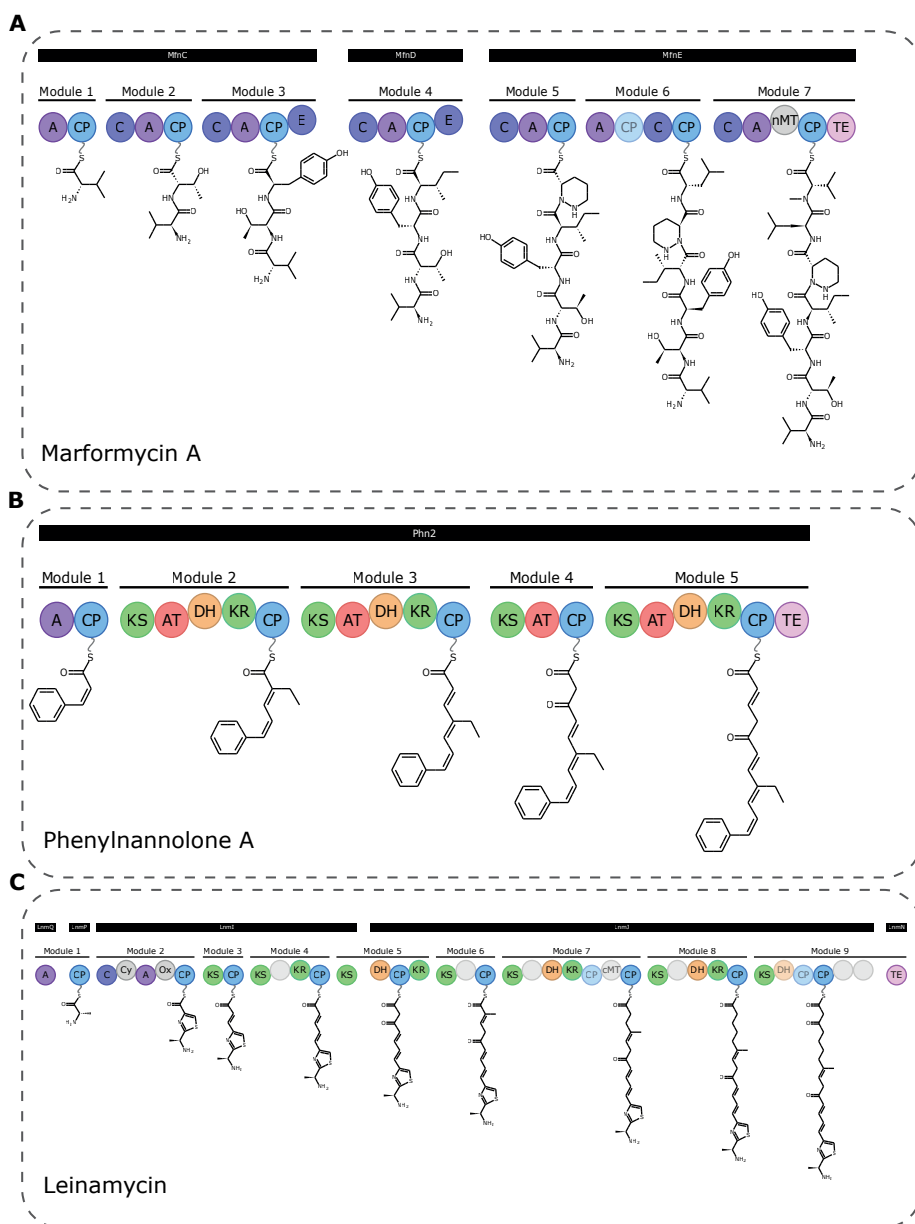
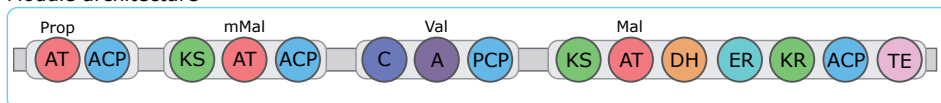


Figure 8.2. Examples of ‘spaghetti diagrams’ commonly used to depict biosynthetic pathways encoded by NRPS⁶⁶ (A), *cis*-PKS³⁸⁴ (B), and NRPS-*trans*-AT PKS hybrid³⁸⁵ (C) biosynthetic gene clusters. The above spaghetti diagrams were automatically generated by RAICuU.

Module architecture



Python input

```
from raichu.run_raichu import ClusterRepresentation, ModuleRepresentation, DomainRepresentation,
draw_cluster

cluster = ClusterRepresentation([ModuleRepresentation('PKS', 'PKS_CIS', "PROPIONYL_COA",
[DomainRepresentation('gene 1', 'AT'),
DomainRepresentation('gene 1', 'ACP')]),
ModuleRepresentation('PKS', 'PKS_CIS', "METHYLMALONYL_COA",
[DomainRepresentation('gene 1', 'KS'),
DomainRepresentation('gene 1', 'AT'),
DomainRepresentation('gene 1', 'ACP')]),
ModuleRepresentation("NRPS", None, "valine",
[DomainRepresentation("gene 1", 'C'),
DomainRepresentation("gene 1", 'A'),
DomainRepresentation("gene 1", 'PCP')]),
ModuleRepresentation('PKS', 'PKS_CIS', "MALONYL_COA",
[DomainRepresentation('gene 1', 'KS'),
DomainRepresentation('gene 1', 'AT'),
DomainRepresentation('gene 1', 'DH'),
DomainRepresentation('gene 1', 'ER'),
DomainRepresentation('gene 1', 'KR'),
DomainRepresentation('gene 1', 'ACP'),
DomainRepresentation('gene 1', 'TE')])
])

draw_cluster(cluster, 'visualisation.svg')
```

Executed reactions

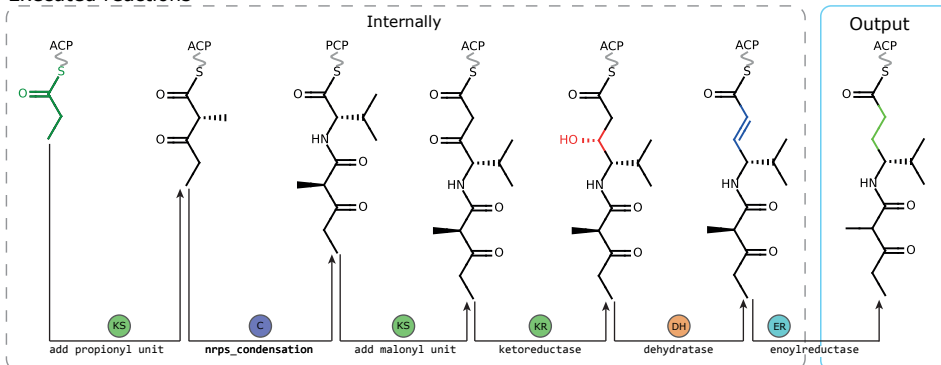


Figure 8.3. A biosynthetic pathway executed *in silico* by RAICHU. The hybrid NRPS/PKS gene depicted at the top is processed one module at the time. C and KS domains catalyse elongation reactions, extending the scaffold with the subunit recognised by A and AT domains respectively. In-line tailoring reactions are processed after elongation reactions. Internally, RAICHU reads in cluster architectures as cluster representation objects (grey box).

One module at a time, RAICHU executes all elongation and in-line tailoring reactions catalysed by the domains in the module on the selected substrate and any previously synthesised scaffold intermediate. First, the elongation reaction catalysed by the C or KS domain is

executed, and the resulting product is covalently tethered to the module's carrier domain. Then, all tailoring reactions are performed (Figure 8.3). When a TE or TD domain is encountered in a terminal module, the reaction product is detached from the carrier domain through a hydrolysis reaction. Then, potential sites of macrocyclization are determined, and candidate macrolactone and macrolactam products as well as the linear product are computed.

We sought to improve upon existing drawing software used by antiSMASH and PRISM to visualise predicted NRP and polyketide products by incorporating stereochemistry of chiral centres. This is especially relevant for the α -carbons of NRP amino acid building blocks, often dictated by the presence/absence of E domains, and the chiral polyketide products of ketoreductase reactions, which depends on the subtype of KR tailoring domains. When an E domain is present in a module, RAICHU performs epimerization on the α -carbon of L-amino acid substrates. We also implemented the KR domain subtypes A1, A2, B1, and B2, which each yield different polyketide stereoisomers; subtype C1, which is inactive; and subtype C2, which has no ketoreductase activity but does catalyse epimerisation (Figure 8.4).

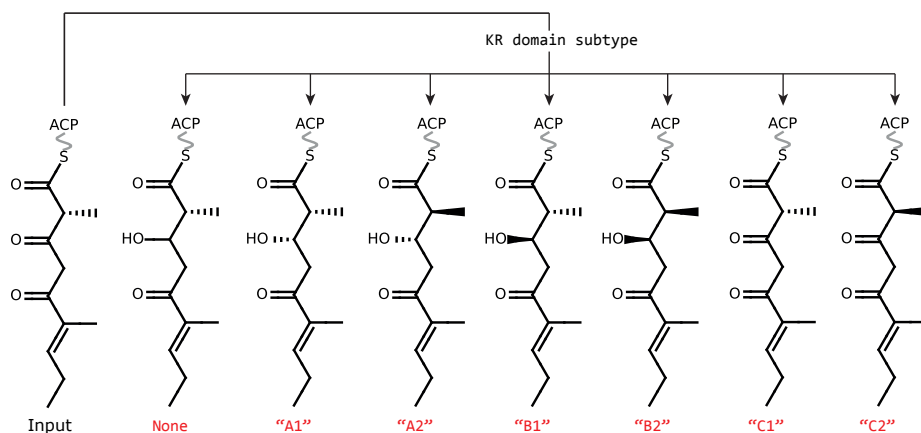


Figure 8.4. RAICHU's implementation of six different KR subtypes, each leading to different stereochemical conformations.

8.2.1.2. Non-modular systems

Another class of natural products for which the scaffold structure can be predicted by current tools are RiPPs. RAICHU computes RiPP structures in three steps: first, it constructs an initial scaffold from the amino acid sequence of the precursor peptide; second, RAICHU tailors the scaffold with user-specified reactions representing class-specific post-translational modifications; and third, it cleaves off any follower or leader peptides by hydrolysing peptide bonds.

In preparation for the development of better predictive tools, RAICHU also supports a library for automating the chemistry of non-modular systems, including RiPPs, terpenoids and alkaloids. RAICHU can build the backbone of terpenoids by allowing the user to pick one of 6 isoprene precursors. The user then defines cyclisation sites and the terpene cyclase class (I or II) to cyclise the precursor. When a class I terpene cyclase is specified, the precursor is

dephosphorylated in the process. The backbone of alkaloids is built from a starting amino acid that can then be further tailored.

8.2.1.3. Tailoring reactions

Most natural products are modified by tailoring reactions after initial scaffold assembly. To implement this important step of natural product chemistry, we implemented 31 tailoring reactions (Table 8.1) that can be called within the context of class-specific cluster frameworks for both modular and non-modular systems (Figure 8.5, Figure 8.6). These reactions include the class-specific posttranslational modifications of 25 RiPP classes, making advanced RiPP chemistry possible. As it is currently not possible to predict the site of action of most of these tailoring enzymes, it is necessary to specify which atom(s) the reaction should be performed on. To this end, RAICHU provides a function to automatically detect and visually highlight all atoms that tailoring reactions can be performed on (see RAICHU wiki). In modular systems, tailoring enzymes perform reactions after the growing NRP/PK chain is released from the assembly line by a TE or TD domain. In RiPP and terpene systems, tailoring, cyclisation and/or proteolytic cleavage tailoring are performed in any desired order, as long as the reactions within each category are performed sequentially. If a different reaction order is desired, for instance if cyclisations need to be performed both before and after tailoring, the PROTEASE or MACROLACTAM_FORMATION tailoring enzymes can be used instead of the default built-in proteolytic cleavage or cyclisation functions. Given any set of reactions and reaction targets, RAICHU can automatically visualise the resulting biosynthetic pathway (Figure 8.5, Figure 8.6).

As many reactions, such as cationic rearrangements in terpenes, are difficult to program in a standardised manner, RAICHU takes various programmatic shortcuts to perform reactions. For this reason, it is possible that some reaction intermediates in RAICHU may not represent a true reaction intermediate in nature. For this reason, RAICHU also provides the option to display a summarised reaction pathway, which only displays the first and last reaction intermediate (Figure 8.5B).

8.2.2. Automatic cluster visualisation of modular systems

After performing chemical reactions according to a modular BGC's module architecture, RAICHU produces two visual outputs: a spaghetti diagram depicting the complete biosynthetic pathway of a natural product at module resolution (Figure 8.7A), and images of all possible linear and cyclic products (Figure 8.7B). Biosynthetic modules can be sorted into genes, which show up as labels above the spaghetti diagram (Figure 8.7A). RAICHU also supports the visualisation of modules that are split across multiple genes (Figure 8.2C).

To produce a large number of cluster images at once, it is possible to script RAICHU in Python. There are three options for this: either the user can load antiSMASH output in .gbk format and produce the desired output (Figure 8.8A); the user can directly script module architectures into Python as cluster representations, which can be interpreted by RAICHU (Figure 8.3, grey box); or the user can provide a tab-separated text file containing the gene and module architecture of custom clusters, the template for which is available on the RAICHU wiki (Figure 8.8B; Table 8.2). With a single function call (Figure 8.8C), the user can then compute the cluster's product(s) (Figure 8.8D). Independent of the data format chosen, the

user can toggle each of RAChU's two visualisation modes to render spaghetti diagrams and/or reaction products. SMILES strings of reaction products can also be stored.

Table 8.1. Tailoring reactions that can be performed on the scaffold using RAChU, their target atoms and required substrate(s). Atoms must always be given in the order given below.

Tailoring Enzyme	Performed reaction	Required atoms	Substrate
Group transfer			
METHYLTRANSFERASE	Methylation	Atom to add methyl group to	None
C_METHYLTRANSFERASE	Methylation	Carbon atom to add methyl group to	None
N_METHYLTRANSFERASE	Methylation	Nitrogen atom to add methyl group to	None
O_METHYLTRANSFERASE	Methylation	Oxygen atom to add methyl group to	None
HYDROXYLATION	Hydroxylation	Atom to add hydroxy group to	None
EPOXIDATION	Epoxidation of a double bond	Both atoms of the double bond to be epoxidised	None
PRENYLTRANSFERASE	Prenylation	Atom to add prenyl group to	Name of prenyl group ("DIMETHYLALLYL", "3_METHYL_1_BUTENYL", "GERANYL", "FARNESYL", "GERANYLGERANYL", "SQUALENE" OR "PHYTOENE")
ACETYLTRANSFERASE	Acetylation	Atom to add acetyl group to	None
ACYLTRANSFERASE	Acylation	Atom to add acyl group to	SMILES string of the substrate to add, with the first atom being the one to be directly bound to the natural product (e.g. <chem>C(=O)CCCCCCC/C=C/CC</chem>)
AMINOTRANSFERASE	Transamination of a keto group	Oxygen of the keto group	None
HALOGENASE	Halogenation	Atom to add halogen to	Halogen atom
Oxidoreduction			
DOUBLE_BOND_REDUCTION	Reduction of double bond	Both atoms of the double bond that is reduced	None

DOUBLE_BOND_SHIFT	Isomerization of double bond	Both atoms of the double bond to be shifted Both atoms of the new double bond	None
DOUBLE_BOND_FORMATION	Formation of double bond	Both atoms of the bond that is oxidised	None
KETO_REDUCTION	Reduction of keto group	Oxygen of the keto group	None
ALCOHOL_DEHYDROGENASE	Oxidation of hydroxy group to keto group	Oxygen of the hydroxy group	None
Elimination			
PEPTIDASE	Peptide bond cleavage	Carbon and nitrogen atom of the peptide bond to be cleaved	None
PROTEASE	Protein cleavage	Carbon and nitrogen atom of the peptide bond to be cleaved	None
MONOAMINE_OXIDASE	Deamination	Nitrogen of the amine group to be removed	None
DEHYDRATASE	Dehydration	Carbon with hydroxy group Carbon to remove hydrogen from. Must be adjacent to first carbon	None
THREONINE_SERINE_DEHYDRATASE	Dehydration of threonine or serine	Oxygen atom of the threonine/serine	None
DECARBOXYLASE	Decarboxylation	Carbon atom of the carboxyl group to be removed	None
SPLICEASE	Removal of a part of the structure between two atoms (as seen in spliceotides)	Atoms flanking the substructure to be removed (input atoms are not removed)	None
ARGINASE	Arginine cleavage to L-ornithine	Secondary nitrogen of arginine	None
Cyclisation			
OXIDATIVE_BOND_FORMATION	Oxidative bond formation	Atoms to form bond between (each needs to neighbour at least one hydrogen)	None

MACROLACTAM_SYNTHETASE	Macrolactam formation between carboxy group and N-terminus	Hydroxy oxygen atom of the carboxy group	None
CYCLODEHYDRATION	Cyclodehydration of serine, threonine, or cysteine	Sulphur or oxygen atom of the sidechain in serine, threonine, or cysteine	None
LANTHIPEPTIDE_CYCLASE	Lanthipeptide cyclization	Sulphur of cysteine or β -carbon of dehydrated serine/threonine (for carbon bridge) β -carbon of dehydrated serine/threonine	None
LANTHIONINE_SYNTHETASE	Lanthionine formation	Sulphur of cysteine or β -carbon of serine/threonine (for carbon bridge) β -carbon of serine/threonine	None
THIOPEPTIDE_CYCLASE	Formation of nitrogen-containing six-membered ring in thiopeptides	β -carbon of dehydrated serine/threonine for which the upstream amino acid is not used in the 6 membered ring β -carbon of dehydrated serine/threonine for which the upstream amino acid is used in the 6 membered ring	None
Epimerisation			
AMINO_ACID_EPIMERASE	Amino acid epimerization	α -carbon of the amino acid to be epimerised	None

To validate RAICHU's visualisation accuracy and readability, we generated 500 NRPS BGCs, 500 PKS BGCs, and 500 hybrid NRPS/PKS BGCs, randomly selecting number of modules, module type (NRPS or PKS), domains within the modules, substrate specificities of A and AT domains, and KR domain subtype such that the rules we laid out for NRPS and PKS cluster and module architectures were obeyed. We manually assessed each spaghetti diagram on three characteristics: the correctness of the depicted structures; the angle of bonds attached to the backbone, which should be 0° or 180° with respect to the drawing plane; and overlap of functional groups with each other and with other drawing elements such as gene, module, and domain depictions. We achieved 100% readability and accuracy on all randomly generated sets of clusters. Spaghetti diagrams are available as separate .png files and as .gif files (online Supplementary File 8.2), showing an animation of all generated clusters per validation set.

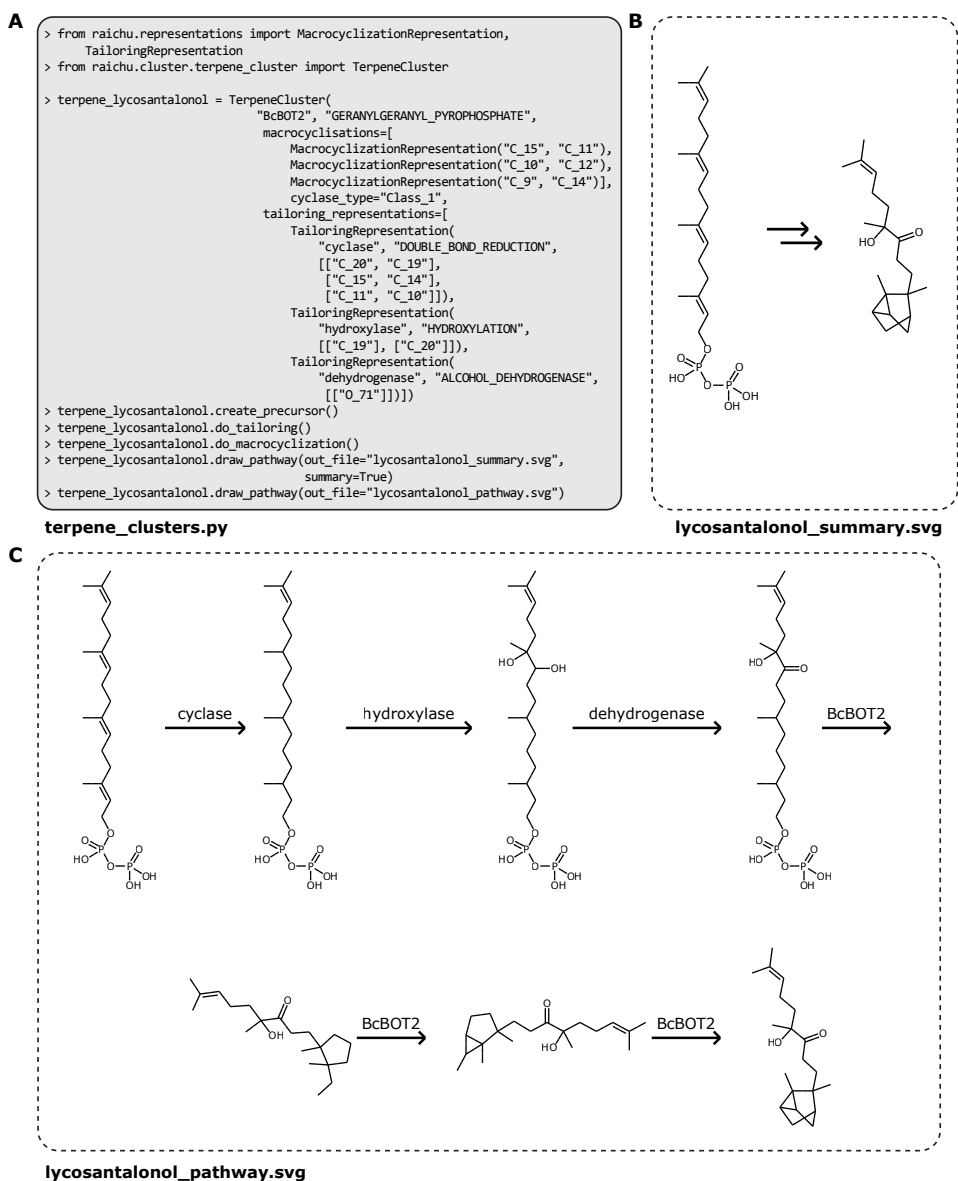


Figure 8.5. Automatic visualisation of the reaction pathway of the terpene lycosantalanol as implemented in RAICHU. Implemented code is shown in A. As some reactions, such as cationic rearrangements in terpenes, are difficult to program in a standardised manner, RAICHU provides the option to visualise the pathway as it was implemented (C) as well as a summarised pathway which only shows the first and last reaction intermediate (B). Note that RAICHU places all structures in the pathway next to one another from left to right, rather than visualising them across multiple rows as shown in C. However, individual elements of the SVG image are easily movable in image editors for publication purposes.

A

```

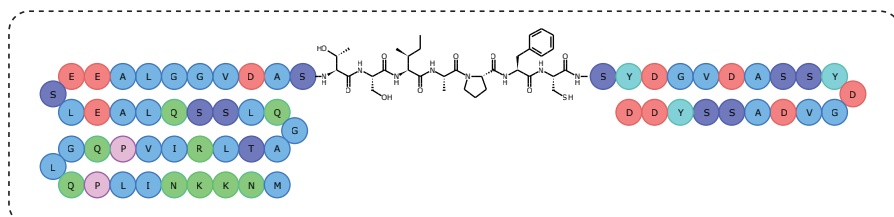
> from raichu.cluster.ripp_cluster import RiPPCluster
> from raichu.representations import MacrocyclizationRepresentation, TailoringRepresentation, CleavageSiteRepresentation
> from raichu.run_raichu import draw_ripp_structure

> cyanobactin_cluster_trunkamide = RiPPCluster(
    "truE",
    "MNKKNILPQLGQPVIRLTAGQLSSQLAELEEALGGVDASTSIAPFCSDYGVGDASSYDGVASSYDD",
    "TSIAPFC",
    macrocyclisations=[
        MacrocyclizationRepresentation(
            "N_0", "O_59", "condensative"),
    ],
    tailoring_representations=[
        TailoringRepresentation(
            "truD", "CYCLODEHYDRATION", [[["S_56"]]]),
        TailoringRepresentation(
            "truF", "PRENYLTRANSFERASE", [[['O_13'], ['O_5']]],
            "3_METHYL_1_BUTENYL"))
    ]
> draw_ripp_structure(cyanobactin_cluster_trunkamide, "cyanobactin")

```

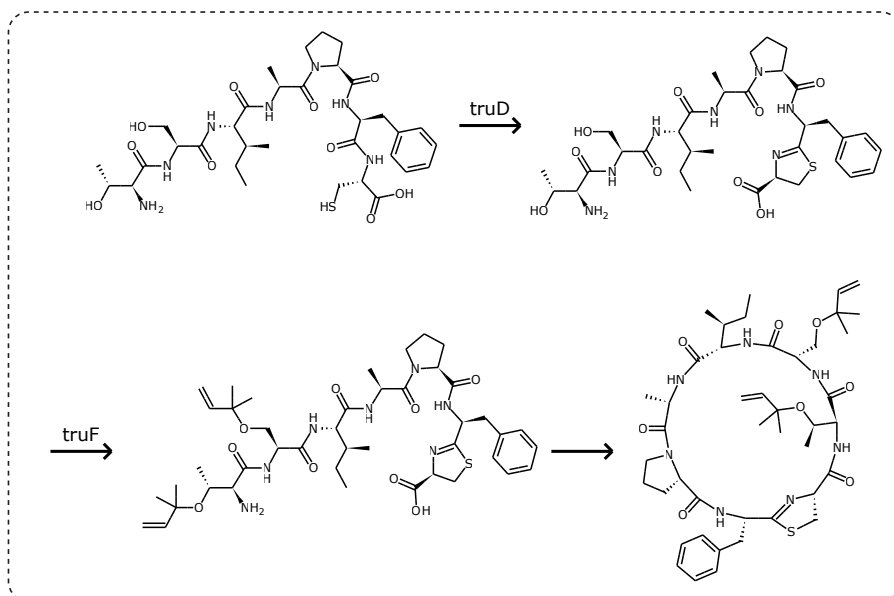
ripp_clusters.py

B



cyanobactin/ripp_inline.svg

C



cyanobactin/ripp_pathway.svg

Figure 8.6. Automatic visualisation of the reaction pathway of the RiPP trunkamide as implemented in RAICHU. Code is shown in A, core peptide visualisation in B, and tailoring modifications and cyclisations in C.

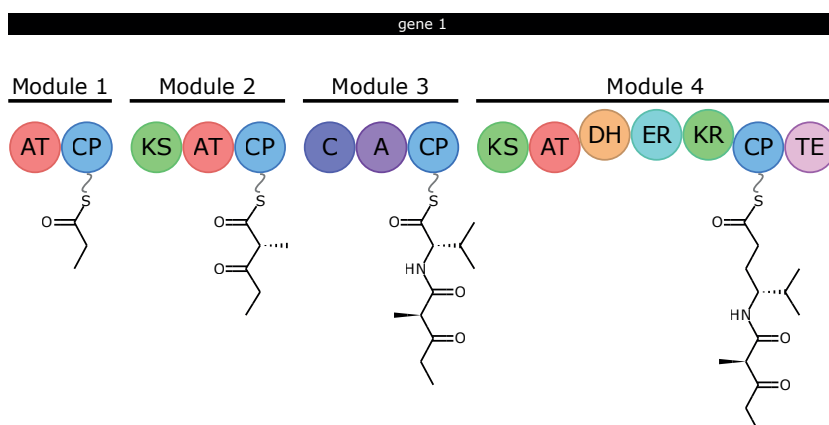
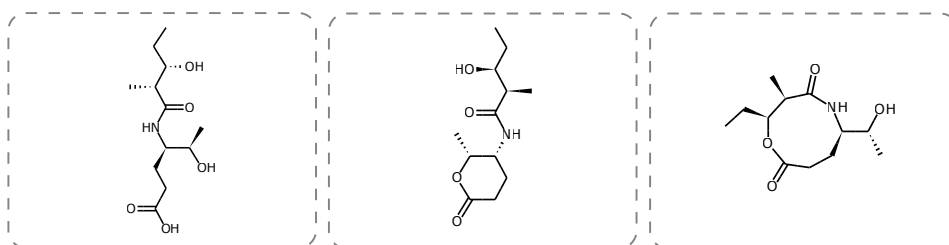
A**B**

Figure 8.7. RAICHU-rendered spaghetti diagram (A) and predicted products (B) of a 4-module hybrid NRPS/PKS cluster.

We additionally tested accuracy on 10 type I PKS clusters, 10 NRPS clusters, and 10 NRPS/PKS hybrid clusters from the MIBiG database: a database of experimentally verified biosynthetic gene clusters for which both DNA sequence and produced compound are known^{70–72} (Table 8.3). We found that RAICHU's spaghetti diagrams matched with the core scaffold biosynthesis pathways published in papers, and among the computed products there was often at least one that was very similar to or identical to the real product (Figure 8.9). Discrepancies usually arose for one of two reasons: post-assembly tailoring modifications, such as the glycosylations in erythromycin³⁷⁵ (Figure 8.9B) which RAICHU can execute but cannot predict; and absence of a required substrate in RAICHU, such as the first (tailored) subunit of hormaomycin³⁸⁶ (Figure 8.9D). RAICHU's product prediction worked especially well for NRPS BGCs (Figure 8.9A, D), which produce relatively unreactive scaffolds compared to polyketide products which often contain many accessible hydroxy- and keto-groups. Despite these shortcomings, scaffold assembly is highly accurate both stereochemically and structurally, with epimerisation catalysed by KR domains and epimerisation domains always inferred correctly given correct annotation. RAICHU even pointed us to erroneous structures in databases, such as the entry for daptomycin in the PubChem database²⁴⁷, which contains an L-asparagine residue instead of a D-asparagine residue at position 2 which should be epimerised due to the presence of an epimerisation domain in module 2²⁷ (BGC0000336). Similarly, RAICHU identified that the colistin A BGC in the MIBiG database points to the wrong PubChem entry, as this cluster likely produces an analogue of colistin A with a D-2,4-diaminobutyric acid residue at position 3 instead of an L-2,4-diaminobutyric acid, as also postulated by the authors

who originally published the BGC³⁸⁷ (Figure 8.9A). This demonstrates the need for automatic visualisation to guide researchers when they make chemical diagrams to remove as much human error as possible.

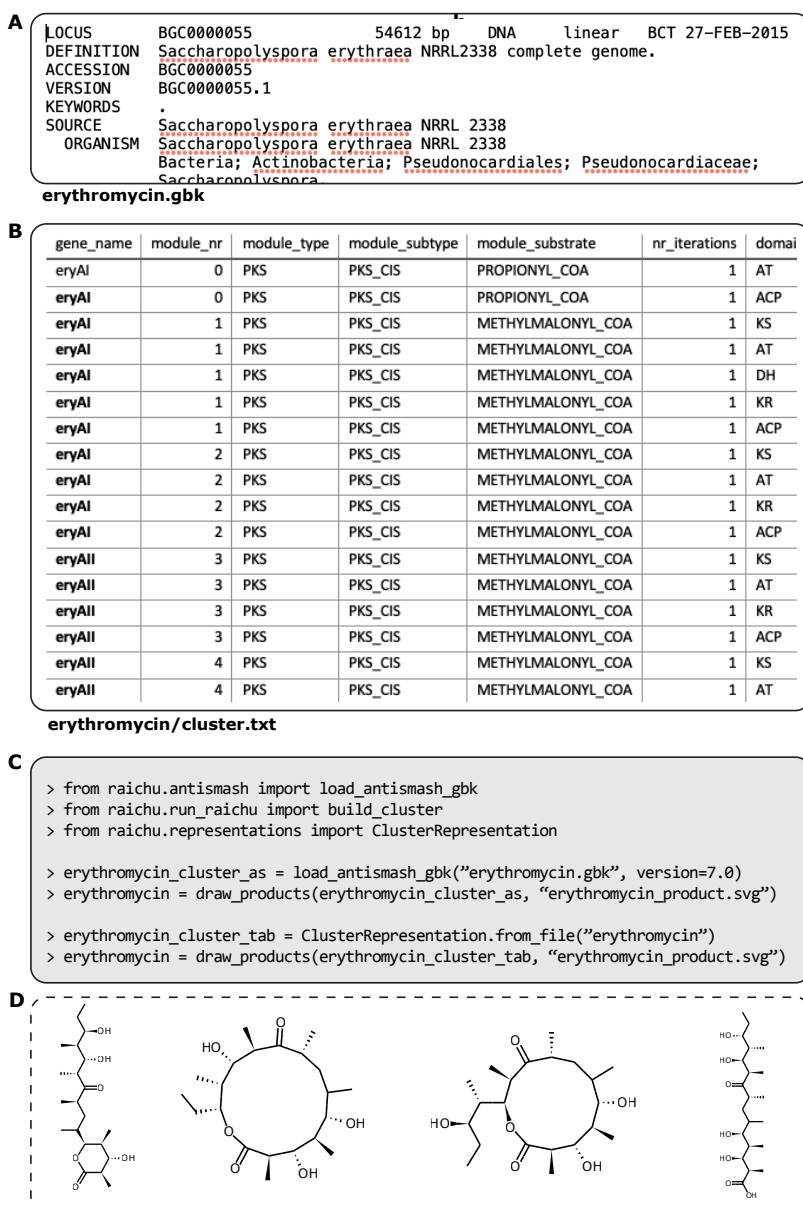


Figure 8.8. Options for scripting RAICHU that facilitate visualisation of multiple BGCs at once. From two different input formats (A, B), RAICHU can generate cluster and product visualisations with a single line of code (C). Resulting output products for the erythromycin BGC are shown in D.

Table 8.2. Example input table for the erythromycin BGC depicted in Figure 8.2B.

gene_name	module_nr	module_type	module_subtype	module_substrate	nr_iterations	domain_type	domain_subtype	domain_name	domain_used	domain_active
eryAI	0	PKS	PKS_CIS	PROPIONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	0	PKS	PKS_CIS	PROPIONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	1	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	1	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	1	PKS	PKS_CIS	METHYLMALONYL_COA	1	DH	None	DH	FALSE	FALSE
eryAI	1	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	B2	KR	TRUE	TRUE
eryAI	1	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	2	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	2	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	2	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	A1	KR	TRUE	TRUE
eryAI	2	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	3	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	3	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	3	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	C2	KR	TRUE	TRUE
eryAI	3	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	DH	None	DH	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	ER	None	ER	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	None	KR	TRUE	TRUE
eryAI	4	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	5	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	5	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	5	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	A1	KR	TRUE	TRUE
eryAI	5	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	6	PKS	PKS_CIS	METHYLMALONYL_COA	1	KS	None	KS	TRUE	TRUE
eryAI	6	PKS	PKS_CIS	METHYLMALONYL_COA	1	AT	None	AT	TRUE	TRUE
eryAI	6	PKS	PKS_CIS	METHYLMALONYL_COA	1	KR	A1	KR	TRUE	TRUE
eryAI	6	PKS	PKS_CIS	METHYLMALONYL_COA	1	ACP	None	ACP	TRUE	TRUE
eryAI	6	PKS	PKS_CIS	METHYLMALONYL_COA	1	TE	None	TE	TRUE	TRUE

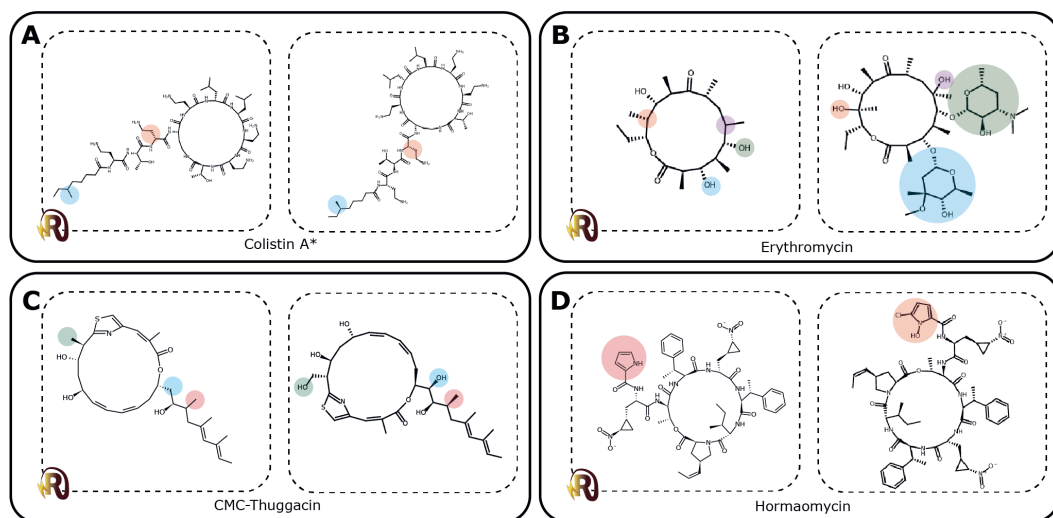


Figure 8.9. Examples where RAICHU-generated products (left) deviated from actual products (right). Inconsistencies are indicated with coloured circles.

* The colistin A cluster reported in MIBiG likely encodes the production of an analogue of colistin A, causing the stereochemical inconsistency observed in module 3.

8.2.3. Speed assessment

We measured RAICHU's speed by randomly generating spaghetti diagrams for 500 PKS, 500 NRPS, and 500 NRPS/PKS hybrid clusters, containing anywhere between 2 and 10 modules. RAICHU performed fastest for PKS clusters, with an average drawing time of 2.73 seconds per cluster. Drawing NRPS and NRPS/PKS hybrid clusters takes slightly longer: 9.62 seconds and 5.21 seconds per cluster respectively. This is likely due to the larger size of NRPS side chains, which require more overlap resolution steps to avoid clashes. Computing time increased with module number (Figure 8.10).

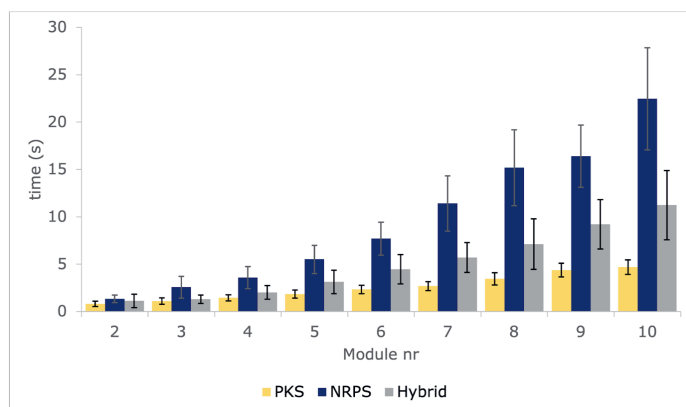


Figure 8.10. Speed assessment of RAICHU tested on 500 randomly generated PKS (yellow), NRPS (dark blue), and NRPS/PKS hybrid (grey) clusters.

Table 8.3. BGCs from the MIBiG database used for validation.

MIBiG BGC identifier	BGC type	Product	DOI
BGC0000028	PKS type I	Bafilomycin B1	10.1002/cbic.201200743
BGC0000038	PKS type I	Coelimycin P1	10.1039/C2SC20410J
BGC0000055	PKS type I	Erythromycin	10.1021/cr9600316
BGC0000072	PKS type I	Gulmirecin A	10.1002/chem.201404291
BGC0000122	PKS type I	Phenylannolone A	10.1002/cbic.201300676
BGC0000144	PKS type I	Salinomycin	10.1128/AEM.06701-11
BGC0000159	PKS type I	Tautomycin	10.1074/jbc.M804279200
BGC0000165	PKS type I	Tiacumicin B	10.1021/ja109445q
BGC0001119	PKS type I	Divergolide A	10.1016/j.gene.2014.04.052
BGC0001856	PKS type I	Caniferolide A	10.1039/C8OB03115K
BGC0000336	NRPS	Daptomycin	10.1099/mic.0.27757-0
BGC0000374	NRPS	Hormaomycin	10.1016/j.chembiol.2010.12.018
BGC0000381	NRPS	Lichenysin D	10.1128/JB.181.1.133-140.1999
BGC0000389	NRPS	Massetolide A	10.1128/JB.01563-07
BGC0000415	NRPS	Quinomycin	10.1371/journal.pone.0056772
BGC0000439	NRPS	Teicoplanin	10.1099/mic.0.26507-0
BGC0000447	NRPS	Tolaasin I	10.1002/cbic.201300553

BGC0000449	NRPS	Tridecaptin A1	10.1002/cbic.201300595
BGC0001192	NRPS	Colistin A	10.1007/s00203-015-1084-5
BGC0001214	NRPS	Marformycin A	10.1021/acs.orglett.5b00389
BGC0000346	NRPS/PKS hybrid	Epoxomicin	10.1021/cb400699p
BGC0000955	NRPS/PKS hybrid	Althiomycin	10.1002/cbic.201100154
BGC0000963	NRPS/PKS hybrid	Bleomycin	10.1016/S1074-5521(00)00011-9
BGC0001047	NRPS/PKS hybrid	Syringolin A	10.1094/MPMI.2004.17.1.90
BGC0001052	NRPS/PKS hybrid	Tirandamycin	10.1002/cbic.200900658
BGC0001165	NRPS/PKS hybrid	Curacin A	10.1021/np0499261
BGC0001212	NRPS/PKS hybrid	Nannocystin A	10.1002/anie.201505069
BGC0001230	NRPS/PKS hybrid	Salinamide A	10.1002/anie.201508576
BGC0001331	NRPS/PKS hybrid	BE-43547	10.1038/nchem.2657
BGC0001342	NRPS/PKS hybrid	CMC-Thuggacin	10.1016/j.chembiol.2010.02.013

8.2.4. Accessibility and future challenges

RAICHU can be installed as a Python pip package for scripting purposes. This makes RAICHU easily available for programmers looking to automate modular NRPS/PKS visualisation as a part of larger BGC prediction pipelines, researchers who want a quick look at the biosynthetic pathway of a cluster they have found, database maintainers who want to cross-reference compounds with BGCs to record the substrate specificities of recognition domains, and bioengineers who want to explore how existing BGCs could be edited to produce novel compounds. With an average cluster rendering time of under 6 seconds, RAICHU is also suitable for large-scale automatic rendering of biosynthetic pathways and NRP/polyketide product scaffolds.

Currently, one of RAICHU's main shortcomings is the requirement for providing reaction targets for post-assembly tailoring reactions, which makes it challenging to automatically script full reaction pathways using RAICHU. One option would be to take a combinatorial approach as seen in PRISM,

which performs all possible tailoring reactions on all possible reaction targets yielding often large libraries of putative structures⁷⁸. As RAICHU is a visualisation tool rather than a prediction tool, we envision that the incorporation of automatic reaction site detection will be added to RAICHU as the community develops better tools to predict the specificity and reaction targets of tailoring enzymes, a process that RAICHU can hopefully play a part in by facilitating the study of novel natural product BGCs. Ultimately, we hope to integrate RAICHU into antiSMASH to directly couple BGC prediction to visualisation.

8.3. Conclusion

We developed RAICHU: a novel tool for rapid automatic visualisation of natural product biosynthesis. With options to render spaghetti diagrams and putative BGC products, RAICHU can produce valuable visual output that can guide researchers within and outside the natural product field to better understand novel BGCs they encounter. By also implementing biosynthetic reactions that cannot be predicted as-of-yet, RAICHU is ready to be deployed once prediction methods for these classes of natural products are developed. Additionally, RAICHU lets researchers encode and visualise biosynthetic pathways that were not predicted from the associated BGCs but instead were elucidated by experimental research. We hope that RAICHU can provide a solution for those looking to get a quick overview of natural product biosynthetic pathways and are currently implementing RAICHU into antiSMASH.

8.4. Methods

8.4.1. Software description

RAICHU is implemented in Python (v3.10.0). It utilises three external Python dependencies: matplotlib³⁶² (v3.5.0), PIKACHU²⁵⁴ (v1.0.14), and BioPython³²³ (v1.79).

8.4.2. Reaction implementation

Scaffold assembly of modular systems

Each NRPS and PKS reaction was implemented as a single function in Python. First, we used the PIKACHU cheminformatics library to convert SMILES strings of NRPS and PKS building blocks to graph-based structure objects. Valid SMILES are those describing any carboxylic acid for NRPS starter modules; any amino acid or β -amino acid for NRPS elongation and termination modules; any thioester for PKS starter modules; and in-built SMILES representing malonyl-CoA, methylmalonyl-CoA, ethylmalonyl-CoA, methoxymalonyl-ACP or a wildcard extender unit for PKS elongation and termination modules. Acceptable module architectures start with a starter module and end with a termination module, with any number of elongation modules in between.

First, the starter unit is covalently attached to a sulphur atom that in turn is tethered to an ACP or PCP Domain instance: a RAICHU object that largely behaves as a PIKACHU Atom object. For NRPS starters, this is done through PIKACHU's in-built condensation function. For PKS starters, the subunit is transferred from the CoA sulphur atom to the sulphur atom attached to the carrier domain (Figure 8.1, thiolation reaction and acyltransferase reaction).

To perform NRPS elongation reactions, two reaction targets are identified with PIKACHU's substructure search function: a thioester within the upstream chain intermediate, as this structure is

already tethered to the carrier domain through a sulphur atom; and the nitrogen atom acid of an amino acid or β -amino acid backbone in the new building block. Next, using PIKACHU's in-built hydrolysis and condensation functions, the thioester is hydrolysed to release the previous chain intermediate from the carrier domain, and the released scaffold is covalently linked to the nitrogen atom of the new building block through a condensation reaction. Finally, the new chain intermediate is tethered to a new sulphur atom attached to a PCP Domain instance through another condensation reaction (Figure 8.1, condensation reaction).

For PKS elongation reactions, we worked with SMILES that represent truncated malonyl-CoA derivatives that lack their terminal carboxyl group, as the ketosynthase reaction removes this CO_2 group anyway. One hydrogen each is removed from the carbon (carbon 1) adjacent to the remaining keto-group and from a second carbon (carbon 2) adjacent to the first. Next, the thioester bond linking the existing chain intermediate to its carrier domain is identified through a substructure search, and the bond between the sulphur and the carbon (carbon 3) is broken to release the chain intermediate. Carbon 1 is then covalently linked to a sulphur attached to a new ACP Domain instance, and a bond is made between carbon 2 and carbon 3. This extends the chain intermediate and attaches it to a carrier domain (Figure 8.1, ketosynthase reaction).

8.4.2.1. Tailoring reactions

The tailoring reactions implemented into RAICHU fall into five subgroups: group transfers, eliminations, oxidoreductions, epimerisations, and cyclisations. They can be deployed as either in-line tailoring reactions that are typically catalysed by tailoring domains within the core enzymes of modular systems; and post-assembly tailoring reactions that are catalysed by enzymes encoded by stand-alone genes. The reaction order of in-line tailoring reactions is hard-coded according to known biosynthetic rules, while the reaction order of post-assembly reactions is always defined by the user. Tailoring reactions will only be performed if the reaction is possible on the target structure. A comprehensive list of RAICHU's tailoring reactions can be found on the RAICHU wiki.

In group transfer reactions such as N-methylations in NRPS systems (Figure 8.1: N-methylation), RAICHU first breaks the bond between the target atom to which the group is to be appended and a randomly chosen adjacent hydrogen atom. It also removes a hydrogen atom from the transferred group. Subsequently, a new bond is formed between the group and the target atom. Elimination reactions are encoded the reverse: first, RAICHU breaks the bond adjacent to the to-be-removed group and then adds hydrogen atoms to both the removed group and the scaffold.

Oxidoreductions deal with the formation, reduction and shifts of double and/or aromatic bonds. This typically involves the addition and removal of hydrogen atoms and changing the hybridisation state of atoms adjacent to the modified bonds. Examples are the four stereo-specific ketoreductions performed by in-line KR domains (Figure 8.1: ketoreduction; Figure 8.4) and double bond shifts.

Epimerisation reactions involve the identification of an α -carbon of an amino acid or β -amino acid in RiPPs and NRPs through a substructure search. Then, the chirality flag of this atom is reversed to yield the correct stereochemistry (Figure 8.1, epimerisation reaction).

For cyclisation reactions, RAICHU forms a new bond between two non-adjacent target atoms. Examples are two NRPS chain release mechanisms: oxidative cyclisation catalysed by the TD domain and macrolactam formation by the TE domain.

8.4.2.2. Scaffold release of modular systems

For modular systems, the chain intermediate is released from the carrier domain and returned as a product once all elongation and in-line tailoring reactions have been performed in the first-encountered termination module. RAICHU supports two options for chain release: linear and circular. For NRPS and PKS systems, RAICHU provides the option to automatically cyclise the released scaffold. It does this by assessing the presence of intramolecular hydroxyl and amino groups that could be used by the TE/TD domain to release the product by means of cyclization, generating lactone or lactam compounds, respectively. Hydroxyl groups that lead to the formation of β -lactone compounds are identified via substructure searches and excluded, as β -lactone formation through TE domain activity is not known to occur in nature¹⁵. All resulting circular products are returned as PIKACHU Structure objects.

8.4.3. Central chain determination

For generating readable spaghetti diagrams, it is key that the central polyketide/NRP chain is drawn vertically. The same problem applies to horizontally drawn RiPP and their abbreviated leader and precursor peptides. To this purpose, RAICHU employs six different strategies for finding the central chain in polyketide/NRP building blocks and RiPPs: one for polyketide starter units, one for polyketide extender units, one for amino acids, one for β -amino acids, one for the polypeptide backbone and one for non-amino NRP starter units. In all cases, atoms belonging to the central chain are flagged with PIKACHU annotations, which can later be used by RAICHU's drawing algorithm to correctly draw the central chain.

Identification of the central chain for amino acids, polypeptide backbones, β -amino acids, and polyketide extender units is straight-forward: substructure searches are performed using PIKACHU to identify all atoms of the central chain. Cyclic β -amino acids where three backbone atoms partake in the cycle are treated as non-amino NRP starter units. The central chains of starter amino acids and β -amino acids that contain a chain of six or more carbons in the side chain are also determined using the non-amino NRP starter unit function. Central chain identification for polyketide starter units was hardcoded for each starter unit supported by RAICHU.

We defined the central chain in non-amino NRP starter units as the longest chain of carbons that are attached to exactly two atoms that are either hydrogens or terminal atoms: atoms that are attached to exactly one non-hydrogen atom. When a ring is encountered in the central chain, as is the case for substrates like hydrocinnamic acid and 2-carboxyquinoxaline, the central chain extends two atoms into the ring and then terminates.

8.4.4. Visualisation

RAICHU draws two types of images for assembly line-like pathways: spaghetti diagrams and end products. For discrete multi-enzymatic assemblies, the precursor and the cyclized, tailored, and final products are visualized. RAICHU largely relies on PIKACHU's drawing engine for all diagram types, but required an important addition for spaghetti diagrams: the central chains of the molecules need to be positioned vertically (or horizontally for RiPPs), such that all bonds in the central chain sit at alternating angles of 60° and -60° with respect to the drawing plane, with the exception of cyclic bonds, which sit at a 90° angle such that the polygon of the side chain can protrude symmetrically

from the central chain (Figure 8.2A, module 5). To this purpose, we created a *RaichuDrawer* class which inherits from *PIKACHU*'s *Drawer* class. It first positions atoms with *PIKACHU*'s drawing algorithm, then adjusts the angles of the molecule's central chain, placing the carrier domain at the top, the attached sulphur atom directly beneath it, and the carbon of the thioester at an angle of 60° below it. Next, the rest of the atoms of the central chain are placed at alternating angles of -60° and 60°, rotating side chains to sit at 0° or 180° with respect to the drawing plane. For cyclic bonds in the central chain, which are placed at 90°, the entire cycle is mirrored if it sits on the wrong side of the bond, positioning it left of the bond if the previous angle was 60°, and right otherwise. Finally, *RAICHU* resolves steric clashes using *PIKACHU*'s secondary overlap resolution function and a custom finetuning function which prevents the rotation of bonds in or attached to the central chain whilst finding the optimal angle for other bonds in the structure to minimise clashes.

Spaghetti diagrams are encoded in scalable vector graphic (SVG) format. In spaghetti diagrams, domains are represented as circles and genes as rectangles. Structures are lined up with the circles representing carrier domains by moving each atom of the structure by the same translational operation such that the coordinates of the Domain instance in the structure match those of the circle representing the module's carrier domain. The module architecture and all structures are then rendered to the canvas and saved in SVG format. Biosynthetic end products are drawn using *PIKACHU*'s drawing library and saved in SVG format.

8.4.5. Validation and speed assessment

Drawing readability and speed were assessed by randomly generating 500 NRPS, 500 PKS, and 500 hybrid NRPS/PKS clusters. Clusters were generated such that every cluster starts with an NRPS or PKS starter module, ends with an NRPS or PKS terminator module, and contains anywhere between zero and eight NRPS or PKS elongation modules. Substrate specificities were randomly selected from substrate subsets depending on the module type: the full set of 143 possible amino acid, β -amino acid and carboxylic acid substrates for NRPS starter modules; 126 amino acid and β -amino acid substrates for NRPS elongation and terminator modules; 15 CoA-tethered starter substrates for PKS starter modules; and 5 malonyl-CoA derivatives for PKS elongation and terminator modules. Custom fatty acids were not tested. Speed was measured with the 'time' module, an in-built module in Python on a laptop running MacOS (Apple M1 chip; one core). Drawing readability and accuracy were manually assessed for each of the 1500 randomly generated clusters, assessing three characteristics: overlaps (none allowed), central chain (60° or -60°) and side chain (0° or 180°) angles, and correctness of stereocentres.

For validation against real biosynthetic end products, we selected 10 NRPS, 10 PKS, and 10 NRPS/PKS hybrid clusters from the MIBiG database. We collected SMILES strings directly from the MIBiG database where possible, and otherwise retrieved them from PubChem²⁴⁷ or NP Atlas³⁸⁸ in that order of preference. We also downloaded antiSMASH-generated GenBank files from the MIBiG database, loaded them into *RAICHU*, and added and changed substrate specificities of recognition domains and subtypes of KR domains to match experimentally verified KR domain subtypes and substrate specificities. We manually compared the resulting spaghetti diagrams with biosynthetic diagrams published in research articles for any inconsistencies and compared *RAICHU*'s rendered end products with the real end products visualised by *PIKACHU* from SMILES strings. For our figures, we chose the predicted product that best matched the cyclisation site found in the real product.

Chapter 9

General discussion

9.1. Summary

In this thesis, we discussed the development of various computational methods that can aid in the discovery of novel non-ribosomal peptides (NRPs) and other natural products. These tools closely mirror the biological steps from the DNA of the producing organism to the molecule whose production is encoded. In chapter 2, we explored which actinobacterial genera might be most suited to producing natural products of novel function, and how current computational tools may help identify these natural products and the biosynthetic gene clusters (BGCs) encoding their production. Next, we described the development of the third iteration of the MIBiG database, which is the largest database for experimentally characterised BGCs with 2502 unique entries and stores many types of metadata that can help gain insight into natural product biosynthesis. In chapter 4, we studied the transition of DNA to mRNA, and demonstrated that not only does DNA encoding have an immense impact on protein production, but it is also possible to predict protein production levels based on the DNA sequence of just the first eight codons of the reporter gene we studied. In chapter 5, we outlined the development of a pair of tools, PARAS and PARASECT, which predict the substrate selectivity of A-domains, a key step in predicting the structure of NRPs from BGC sequence. We compared the performance of models trained on protein sequence features and (predicted) protein structure features and concluded that sequence features consistently perform better. We investigated better ways of organising and sharing multi-faceted natural product datasets in chapter 6, where we described the development of Turterra, an enzyme analysis portal that can automatically build an interactive web-portal, which users can leverage to study the phylogeny, protein sequence, protein structure and substrate/product of natural product enzyme families. Finally, in chapters 7 and 8, we presented two dependency-light cheminformatics suites, PIKACHU and RAICHU, which can be used to automatically visualise natural product biosynthetic pathways given a set of (antiSMASH) predictions.

In summary, we developed a pipeline that starts with sequence acquisition, annotation, and storage and ends with the automatic visualisation of molecules and pathways that are predicted to be encoded by those sequences, particularly for NRP-producing systems. Below, we will discuss how this pipeline can contribute to the goal of predicting natural product structure and bioactivity from sequence data. We will explore how and where things can be improved and assess the feasibility of such pipelines to eventually provide accurate bioactivity predictions given a DNA sequence. We will structure this discussion as we did the chapters: starting with the producing microorganism and ending with the bioactivity of the produced compound.

9.2. Genomes and metagenomes

9.2.1. Familiar or obscure: which (meta)genomes should we sequence?

The question of where to look for natural product producers is an ongoing debate. In chapter 2, we explored the well-studied phylum of Actinobacteria, the members of which produce a wealth of natural products. Even within that phylum, it is apparent that there is no single best strategy for finding novel compounds. It can be fruitful to sequence underexplored genera such as *Verrucosispora*, which led to the discovery of structurally novel abyssomicins³⁸⁹, or species that thrive in unusual environments, such as a strain of *Streptomyces* that inhabits insect microbiomes and produces cyphomycin, a novel antifungal molecule⁴. At the same time, unknown pathways are

continually discovered in strains that have been studied for decades, such as *Streptomyces coelicolor*, which despite being studied for over 50 years still harbours biosynthetic pathways of which we do not know the function¹²⁷. Then, there is the choice to focus sequencing efforts on talented phyla, which we know are likely to harbour a large number of biosynthetic pathways, some of which might produce novel compounds; or to sequence underexplored phyla instead, which are more likely to contain pathways that are completely chemically novel, but for which the risk that the selected strain does not produce any molecules with desired properties is also greater. Another option that has gained traction is to sequence metagenomes instead, which has multiple benefits: it can capture the biosynthetic potential of unculturable strains; and it makes it possible to leverage the ecological context of the microbe and its biosynthetic pathways when predicting the function of any produced natural products. It has drawbacks as well, the largest being challenges surrounding a lack of access to the DNA of individual organisms, the assembly of metagenomes, and the subsequent separation of sequence data into genomic bins. This is especially problematic for BGCs that are split across different loci or different genomic scaffolds, as seen for frankobactin³⁹⁰, a NRP metallophore from *Frankia sp.* CH37, and the tridecaptins, a family of NRPs that target Gram-negative bacteria (unpublished work). As with genome sequencing, it is unclear if it is more lucrative to sequence metagenomes from extreme environments or to sequence the soil from your backyard^{218,220}.

For the computational discovery of novel NRPs, an advantage of targeting genomes from extensively studied genera such as *Streptomyces* is that our predictive algorithms for adenylation (A) domain selectivity perform better on them. As we discussed in chapter 5, the evolution of A domain selectivity appears to be convergent: A domains recognising the same substrates often do not clade together, even when aligning only the A domain active sites. We also saw that A domains from different phyla, such as the tryptophan-recognising A domains from the tyrocidine BGC (*Bacillus*) and the tryptopeptin BGC (*Streptomyces*) explored in chapter 5, may sometimes recognise an identical substrate in a completely different manner. In underexplored genera, it is more likely that we find A domains that recognise substrates with active site architectures that are not represented in the training data of our predictive algorithms and are therefore challenging to make a reasonably accurate prediction for. As many NRP systems contain multiple A domains, this is a compounding issue: for every A domain with a low prediction accuracy, the confidence of the final NRP scaffold diminishes. At the same time, A domains of underexplored genera do not necessarily accept rarer substrates and thus may not point to increased structural diversity. For example, the A domains of the cyanobacterial hassallidin BGCs, which at the time of their discovery were so distinct from data points in training datasets of A domain predictors that no meaningful substrate predictions could be made, only select proteinogenic amino acids³⁹¹.

9.2.2. DNA, mRNA, and amplicons: how should we be sequencing?

There exist various sequencing strategies that can facilitate natural product discovery, each with their unique use-cases, benefits, and drawbacks. DNA sequencing to obtain (meta)genomes is the most common, and with good reason: sequenced genomes can be easily mined for natural product pathways, including silent BGCs that may not be expressed under laboratory conditions. Also, as all genes involved in a natural product pathway should be located on the genome (barring some rare exceptions), sequencing an entire genome is a good way to ensure that the entirety of the pathway(s) of interest can be found. However, it can still be difficult to obtain genomes of sufficient quality and coverage to detect full BGCs, which is especially problematic when the goal is to detect BGCs of novel function for which no reference sequences are available. Long-read sequencing such as PacBio or

Nanopore sequencing, especially in combination with high-quality Next Generation Sequencing (NGS) reads, can provide a solution, but this is often costly, especially when many strains are analysed in tandem. For this reason, researchers sometimes opt to only obtain high-quality sequence data for genomic regions where they expect a BGC of interest to reside.

Alternatively, one can first prioritise strains or metagenomes using other sequencing methods such as amplicon sequencing, where degenerate primers are used to PCR amplify DNA fragments of interest and subsequently sequence them. NRP discovery lends itself particularly well to this strategy, as A domains are of a suitable size (~400 AA) and contain sufficient conserved motifs for degenerate PCR amplification, and their sequence can yield direct insights into the structure of natural products that are produced by their BGCs. This strategy has led to the discovery of the malacidin lipopeptides by prioritising soils that contained A domains selecting aspartic acid and glycine, which are required building blocks for the conserved Asp-X-Asp-Gly motif of calcium-dependent lipopeptides³⁹². Amplicon sequencing of A domains has also been employed to link amino acid building blocks to phenotypes such as the disease-suppressive properties of some soils³⁹³, to map the microbial diversity and biosynthetic potential of various global soils³⁹², and to create co-occurrence networks of biosynthetic domains that pointed to the presence of currently undiscovered NRPS BGC architectures³⁹³. However, it is not possible to deduce complete NRP scaffold structures from amplicon sequencing alone, as the biosynthetic order of the domains is lost in this type of analysis. Also, if multiple NRPS BGCs exist within a (meta)genome, it is not always obvious which A domains belong together. As such, NRP scaffold composition cannot be reliably determined unless the researcher is looking for specific motifs, like the Asp-X-Gly-Asp motif in calcium-dependent lipopeptides. Therefore, amplicon sequencing and subsequent analysis of those amplicons with tools such as PARAS is a good method for prioritising strains or samples but is insufficient for computational prediction of NRP structures.

Instead of looking at a producing strain's DNA, it can sometimes be more informative to study the sequence of the microbe's mRNA with high-throughput RNA sequencing. While the genome gives an overview of the organism's potential to produce novel molecules, the transcriptome indicates which biosynthetic enzymes are actually produced, in which quantities, and under which environmental conditions. This can yield valuable information: understanding when a biosynthetic pathway is activated can give insights into the ecological role of the produced molecule and point to its potential medical and societal applications. Furthermore, transcriptional networks may point to new elicitors of BGC expression that can then be applied in different strains to activate BGCs that are silent under laboratory conditions. Naturally, silent BGCs that are not transcribed under tested conditions will not show up in transcriptomics analyses, and as such it is advisable to also have a genome sequence available.

When paired with metabolomics, transcriptomics can also aid in de-orphaning natural products for which no BGC is known, as the researcher can directly link the produced metabolite to the expression of a BGC. Such a multi-omics analysis led to the discovery of the BGC producing the antimalarial molecule salinipostin¹³⁸. This kind of analysis is less useful for NRP-producing systems: due to the modular nature of NRP synthetases (NRPSs), it is often fairly straight-forward to link an NRP to its producing BGC, especially when A domain substrate predictions are provided by PARAS and PARASECT or other adenylation domain predictors^{394,395}. However, NRPS systems are not always linear: modules can be skipped³⁹⁶, iterated³⁹⁷, activated by trans-acting adenylation domains³⁹⁸, or assemble the NRP in an order not consistent with the observed gene or module architecture⁵¹. Also,

linking an NRP to its producing BGC can still be challenging for bacterial and fungal phyla that PARAS and PARASECT cannot provide confident substrate predictions for. In such cases, multi-omics approaches may provide insights that can lead to the successful de-orphaning of NRPs and their BGCs.

9.3. mRNA encoding and natural product discovery

Instead of measuring the transcription rates of mRNA with RNA-seq, it may be possible to infer transcription levels from the codon usage of biosynthetic genes and use this information to find novel BGCs or learn about the ecological role of a BGC and the function of its product. In chapter 4, we demonstrated that codon usage has a major impact on transcription levels in *E. coli*, especially surrounding the 5' UTR where mRNA secondary structures may inhibit ribosome binding rates. While the mRNA expression predictor MEW only works specifically for the mRFP gene that we expressed in *E. coli*, we showed that protein expression prediction is possible, and similar methods could be applied to biosynthetic genes once a wider range of expression data becomes available.

It is possible that codon usage could hint at the function of the BGC: if a BGC's codon usage indicates low expression levels, its encoded product may not be needed in large quantities or may only need to be produced under specific conditions. There is one prominent example of how codon usage affects the regulation of biosynthetic genes: in *Streptomyces coelicolor*, the UUA codon which encodes leucine is uniquely translated by the tRNA encoded by the *bldA* gene, which regulates morphological differentiation and antibiotics production through codon usage. This codon is only found in a small proportion of *S. coelicolor* genes, many of which are biosynthetic genes encoding the production of antibiotics^{399,400}. Similarly, it may be possible to identify antibiotic-encoding genes in other genera by monitoring the usage of rare codons or by linking predicted expression levels to the production of specific compound classes.

9.4. Annotation of biosynthetic sequences: dos and don'ts

When developing computational tools for natural product structure prediction, we rely heavily on the annotation quality of biosynthetic sequences. As such, BGC databases such as MIBiG, the third iteration of which we described in chapter 3, are pivotal for providing data that can be used for training machine learning algorithms. Any good scientific database should satisfy seven conditions: it needs to be standardised, centralised, searchable, open source, up to date, accurate, and provide references to the initial data source. During the development of MIBiG 3.0, we found that data annotation accuracy, standardisation, and keeping the database current provided the greatest challenge. Largely, this could be attributed to the fact that in an academic setting, the development and maintenance of most open-source, high-quality databases is often only possible through community annotation events, especially for groups that do not have access to funding for dedicated annotators. Despite efforts to teach annotators how to submit high-quality, standardised, and reliable annotations, each researcher still has their different style, and human error is unavoidable. Annotation mistakes can also stem from the source article from which the data was retrieved: various misannotations could be traced back to experimental errors or unclear reporting. This problem is not unique to datasets assembled through community efforts: when assembling a training dataset for PARAS and PARASECT in chapter 5, we identified and corrected various errors in the training datasets of previous adenylation (A) domain substrate predictors: 137 in the training dataset of SANDPUMA and 31 in the dataset used for NRPSPredictor2, each amounting to about 10% of the total labelled

training data. We observed similar error rates in the A domain dataset that was initially collated for MIBiG 3.0 prior to the extra validation steps we performed before training PARAS and PARASECT.

To limit annotation errors, various strategies can be employed. The first is the development of well-structured submission systems with built-in checks that automatically detect potential annotation errors. MIBiG would benefit from a system that automatically performs various checks. An easy example is to assert that the GenBank accession that describes the genomic region on which the BGC is located is valid and is not already present in the database. In the context of A domains, we recommend implementing an A domain substrate predictor into the submission system which checks if the substrate submitted by the researcher or annotator is identical to any high confidence predictions and raises a warning urging the user to double-check the substrate if it is not. Also, the system should examine if the substrate name input by the annotator is among those listed in MIBiG, and if not, should require the annotator to enter a substrate structure in SMILES format. SMILES strings themselves should also be checked for validity and redundancy through SMILES parsing and substructure searches with cheminformatics software suites such as PIKACHU.

While the development of such a submission system would initially be time-consuming, it would improve data quality, enforce data standardisation, and speed up the annotation process, thus targeting three current challenges in the annotation of biosynthetic sequences. Another strategy that could be employed in the short-term is to annotate entries in duplicate and check for any inconsistencies in annotations. This approach was used by the Charkoudian group, who curated a fungal natural product dataset⁴⁰¹, and the developers of PRISM⁷⁸ to obtain high-quality BGC datasets. The drawback is a lack of resources: even when only one annotator looked at each BGC entry, we were not able to process all novel BGCs reported in literature over the span of two years with a team of 87 annotators contributing around 550 manhours in total.

Finally, it should be made easy for users of the database to report errors. This way, even if mistakes are made in the initial annotation of a BGC, they can be traced and fixed down the line. The MIBiG database already has an automatic error reporting system in place, which with a single mouse-click redirects the user to a pre-structured template of a GitHub issue where they can specify the nature of the error. This process could be further improved by linking out to a visual error reporting system instead, where the user can click-and-correct an entry interactively. Afterwards, the changes would only demand a quick approval from the database maintainers, instead of requiring them to manually edit database files. This is a much more time-consuming process and is therefore often delayed until the next database release, unnecessarily perpetuating an erroneous entry for up to two years.

Some of the errors we discovered in our A domain data could be traced back to the original publications. Frequently, molecular structures were misreported, usually due to the incorrect depiction of a chiral atom. As any laborious manual process, structure drawing is prone to human error. These errors perpetuate throughout databases and are difficult to fix as there is no reason for researchers to mistrust the figures put out by their peers, especially annotators who are usually not experts on the chemistry of the molecule in question. We sought to prevent such human error by automating the drawing process of natural product pathways with PIKACHU and RAICHU. As detailed in chapter 8, RAICHU helped us find a mistake in the structure of daptomycin in the PubChem database.

Aside from data quality, data quantity is also an issue: there are many more experimentally verified BGCs than are accounted for in the MIBiG database. All these data would be invaluable for training machine learning algorithms that can help accelerate natural products research. While community annotation marathons help, it would be much better for researchers that worked on the BGC to directly submit an entry to MIBiG themselves. Only a small fraction (~10%) of BGCs added to MIBiG since version 1.0 were direct submissions; as such, it is clear that we need to incentivise natural product researchers to add their data to centralised repositories such as MIBiG. The best way to do this would be for journals to require a database submission prior to manuscript acceptance: this is already standard for DNA, mRNA, and protein sequences, for which many journals require a GenBank accession. In addition, the submission process needs to be straightforward and centralised. As such, the natural products field would truly benefit from a ‘natural products hub’, where the user can deposit all their sequence, structure, and bioactivity data at once, including mass spectrometry data and the results of biological assays, which could then be automatically submitted to various relevant repositories such as MIBiG, GNPS⁴⁰², and NP Atlas³⁸⁸.

9.5. From sequence to structure to bioactivity: promise or pipe dream?

In chapter 5, we saw that mislabelling of only 10% of the data hugely impacts the performance of predictive algorithms. We run into a similar challenge with compounding predictions: even if two algorithms are 90% accurate, when they are used in sequence, the result will only yield an 81% accurate result for independent predictions. Throughout this thesis, we have come across a few examples of compounding predictions: mRNA secondary structure prediction followed by protein production prediction in chapter 4; protein structure prediction and subsequent A domain substrate prediction in chapter 5; and the prediction of full NRP scaffold structures, which often requires up to a dozen separate predictions, each with their own uncertainty. The goal this thesis works towards, bioactivity prediction from predicted structure, suffers from the same, exponential decline of predictive power. This begs the question: is the prediction of NRP bioactivity from structure from sequence even possible?

In order to judge the feasibility of predicting bioactivity from a predicted structure, we need to assess each predictive step individually. For NRPs, these steps are manifold. We will address the predictions involved in structure prediction and bioactivity prediction separately.

9.5.1. NRP structure prediction

In this thesis, we focussed on the prediction of NRP scaffold structures by identifying adenylation domain substrates. However, there are many more steps involved in NRP biosynthesis, many of which we are not yet able to predict. Broadly, there are 5 different aspects to NRP structure assembly which all influence the structure of the final NRP: building block availability; substrate selectivity; assembly order; in-line, co-assembly, and post-assembly tailoring; and cyclisation site(s).

9.5.1.1. Building block availability

While the scaffold assembly and tailoring enzymes that build NRPs are restrictive in the substrates they recognise, they are often promiscuous to a varying degree. For example, the tryptopeptin A domain discussed in chapter 5 had a strong preference for loading tryptophan, but also recognised other hydrophobic substrates such as phenylalanine, leucine, and to a lesser extent histidine in the absence of tryptophan. As a result, the structure of the produced NRP strongly depends on which

building blocks are available, which in turn relies on the presence of precursor enzymes within the BGC, primary metabolic pathways of the producing organism, and the producing organism's environment. For example, the second A domain of LptC, an NRPS in the A54145 BGC, usually selects 3-methoxy-aspartic acid. However, it will also activate 3-hydroxy-aspartic acid or aspartic acid upon the knockout of the precursor genes LptK, a methyltransferase, or LptJ, a hydroxylase, respectively⁴⁰³. When close homologs exist, it is usually possible to predict if certain precursor biosynthesis genes are present in the BGC through Pfam domain searches, but for most precursor synthesis enzymes this is not the case. For example, antiSMASH cannot detect that DptJ, a precursor biosynthesis enzyme in the daptomycin BGC, catalyses the dioxygenation of tryptophan in kynurenine biosynthesis. Fortunately, if an A domain shows a strong substrate preference for a tailored precursor, the A domain active site will often (but not always) reflect this, and in these cases we can rely on A domain substrate predictions to learn which precursor is incorporated. For instance, the active sites of A domains that select the heavily modified ornithine residues that are often seen in siderophores tend to clade by modification (Figure 9.1). In contrast, A domains 2 and 3 of Mlcl, a NRPS in the malacidin BGC, recognize 3R-hydroxyaspartic acid and aspartic acid, respectively, but have identical active sites and presumably have access to the exact same set of precursors. For all these reasons, precursor biosynthesis adds a layer of complexity to NRP biosynthesis that impacts the confidence of NRP structure predictions.

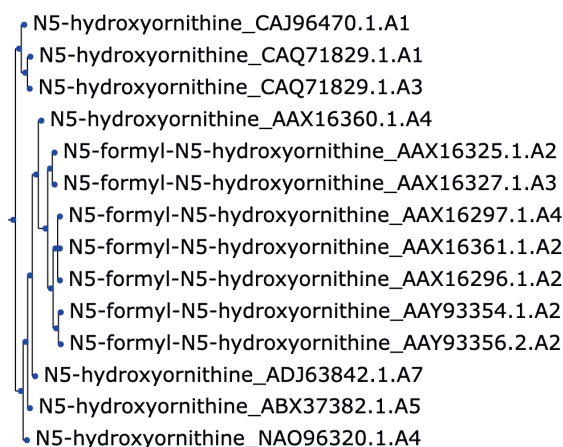


Figure 9.1. Phylogeny of the active site signatures of adenylation domains that select modified ornithine precursors in siderophore biosynthesis. The 34 residues of the adenylation domain active sites clade based on precursor.

Predicting precursors becomes even more challenging when evidence of their biosynthesis cannot be found inside the BGC. The effect of the producing organism's genome on NRP structure became especially evident in a study that heterologously expressed nine NRPS and NRPS-PKS hybrid BGCs in 25 different strains of proteobacteria. While the sequence of the BGC was not altered, the structures of their product varied depending on which organism the BGC was heterologously expressed in. For instance, a single BGC led to the production of eight related NRPs, including three GameXPeptides and five luminimides, with some strains producing all variants, others producing only a subset, and a few not producing any of the metabolites at all⁴⁰⁴. Notably, only two strains were able to produce GameXPeptide E, which contains a para-amino-phenylalanine (PAPA) as the fourth amino acid. Other strains incorporated proteinogenic amino acids at this position, including phenylalanine, leucine,

methionine, valine, and tyrosine. Likely, the two strains that produce GameXPeptide E possess genes encoding the production of PAPA elsewhere in the genome, while the other strains do not. Similar behaviour was observed for other BGCs that were tested, which demonstrates that looking at the BGC alone is not sufficient to understand which precursors an NRP assembly line has access to. It also poses another question: to what extent is heterologous expression a good way to find natural products in the form that they are produced in their native environment?

A related observation that further compounds matters is that not only the genome of the producing organism but also its environment dictates which precursors are present. A good example is a study done on lipopeptide production in *Bacillus subtilis* S499 and NT02, where bacteria were grown on media containing different amino acid compositions and the resulting lipopeptides were analysed. The lipid tails of the isolated fengycins, iturins, and surfactins varied based on the amino acid diet, showing how even genetically identical strains can produce different NRPs depending on the environment⁴⁰⁵. This becomes even more relevant when we consider that microbes have evolved as parts of complex communities, where not all building blocks for primary and secondary metabolism may be easily available and important biosynthetic precursors may even be synthesised and secreted by other microbes in the community. Therefore, lab-grown microbes may not produce the same metabolite that they would naturally produce in their ecological niche. This does not only make structure prediction difficult, but it also means that a lot of the data that our predictive algorithms rely on might not be entirely accurately annotated, as many natural product studies still rely on lab cultivation and heterologous expression. This stresses the need for metagenomics and meta-metabolomics data, which capture microbes in the context of their native community and environment, and therefore stand a much better chance of isolating a specialised metabolite in its natural state.

9.5.1.2. Substrate selectivity

In chapter 5, we presented PARAS and PARASECT: two structure-informed adenylation domain substrate predictors. With an accuracy of around 90% on average and 75% on phylogenetically distant sequences, a training dataset of over 3000 datapoints, and prediction time of under two seconds, it is the best adenylation domain predictor currently available. However, NRP scaffold prediction usually requires multiple substrate predictions, with many NRPS BGCs counting anywhere between 2 and 30 A domains. This means that the accuracy for the prediction of a full scaffold would actually be much lower: for an NRPS system with 10 A domains, a NRP scaffold prediction would only have an average accuracy of $0.9^{10} = 0.34$.

In practice, the kind of errors that PARAS and PARASECT make in scaffold predictions for well-represented phyla will generally not have a large impact on the NRP's bioactivity. Many of the mispredictions these tools make involve amino acids that have very similar properties and interchangeable functions. For instance, they often confuse the hydrophobic amino acids isoleucine, leucine, and valine, or the aromatic substrates tyrosine, phenylalanine, and tryptophan. As such, many scaffold predictions for these phyla will still give good insight into the nature of the produced NRP scaffold. Unfortunately, the same cannot be said for A domain predictions for taxa such as fungi, for which we have far fewer datapoints. Not only are the error probabilities for individual domains higher, but they are also 'more wrong', with predicted and actual substrates having little in common. It is possible that substrate selection in fungi is vastly different from bacterial substrate recognition, possibly by recognising the substrates in a different orientation or by involving other parts of the NRPS enzyme.

When engineering novel NRPS enzymes in bacteria, it was observed that switching out modules for new ones, exchanging adenylation domains to incorporate new substrates, and merging parts of different NRPS systems into hybrid NRPS enzymes is more difficult than initially thought. This prompted the idea that factors outside the module's A domain may influence substrate selectivity. Previously, the condensation (C) domains both upstream and downstream of the module's A domain have been implicated in proofreading activity^{53–55}; it is thought that while the A domain performs initial selection, the C domains ensure that only certain substrates can be built into the scaffold. We discussed briefly in the introduction how this might play a role in correct product assembly but did not address the ramifications for algorithms that predict NRP structure, which currently only consider the sequence of A domains. If the C domains indeed play a large role in substrate selection, they should be included as input to such algorithms. Also, if interpretable machine learning is used, the algorithm itself can help gain insight into which parts of the C domain play a role in substrate selectivity, which could in turn be used to guide NRPS design strategies. As such, we envision future versions of PARAS and PARASECT which also feature the sequence of other domains in the NRPS module, including the C domain, the peptidyl carrier protein (PCP) domain, and potentially the linker regions between these domains.

9.5.1.3. NRP assembly order

Even if substrates are predicted correctly at the domain or module level, this does not mean that the structure of the full NRP scaffold can be directly deduced. While assembly is linear in many systems, with NRPS enzymes adding modules in the order in which they are encoded on the genome, there are just as many examples of NRPS systems where this collinearity rule does not hold. In section 9.2.2., we mentioned briefly that there are many examples of non-linearity in NRPS systems. In the bleomycin BGC, modules within genes are activated in order, but the gene order does not match the assembly order⁵¹. In many siderophore BGCs, such as the BGC encoding the production of vicibactin⁴⁰⁶, (modified) ornithine substrates are joined backbone-to-sidechain instead of backbone-to-backbone, leading to a completely different scaffold than a linearly assembled peptide using the exact same substrates. In addition, the single module in the vicibactin NRPS works iteratively, incorporating the same substrate three times in a row to form the final natural product scaffold, a property which we can currently not predict. Like ornithine, glutamic acid and aspartic acid can be incorporated in a different orientation as well, with their sidechain acids used to form the scaffold backbone as observed in the microcystin BGC⁴⁰⁷. In our dataset for PARAS and PARASECT, we provided unique labels for domains which recognise substrates like glutamic acid and aspartic acid 'sidechain-first', such that our algorithms could predict their correct orientation in the final scaffold. However, there were not enough examples of these domains in nature to incorporate them into our training sets. In the WS9326A BGC, module 7 lacks an A domain and instead is trans-loaded by a trans-acting A domain elsewhere in the cluster. In the same BGC, module 8 is skipped entirely³⁹⁶. Finally, there are BGCs where modules and even the domains within a module are scattered across multiple genes on different strands of the DNA. Even the thioesterase domain, the terminal domain that is used by antiSMASH to predict module order, may be on a separate gene, as seen in the kedarcidin BGC⁴⁰⁸ (BGC0000081). In such cases, it is currently impossible to computationally determine in which order a NRP scaffold is assembled.

9.5.1.4. NRP tailoring and cyclisation

Often, even a correct NRP scaffold prediction is not sufficient for obtaining the structure of the final NRP. Many NRPs are heavily tailored during assembly or post-assembly by a plethora of reactions,

including cyclisation, methylation, halogenation, and glycosylation among many others. Even if the responsible tailoring enzyme is encoded in the cluster and we can predict its tailoring reaction, predicting the order and the site of action of the catalysed reactions remains a challenge. There are some exceptions to this, most of them catalysed by in-line tailoring domains that make up part of the NRPS itself. These include domains that catalyse epimerization, N-methylation, and threonine/serine/cysteine cyclisation and subsequent oxidation. All of these tailoring enzymes are automatically executed upon loading an NRPS system into RAChU, the cheminformatics suite we developed for automatically visualising biosynthetic pathways and computing the structures of encoded BGC products. For now, the prediction of other tailoring reactions remains out of reach, although we did include a reaction library in RAChU to encode 31 known tailoring reactions in NRPS and PKS chemistry, should the tools be developed that can accurately predict their site of action.

With so many sources of uncertainty, it may seem futile to attempt to predict the structure of an NRP at all. However, natural product researchers can often use their knowledge of experimentally validated NRPS systems to contextualise novel predictions and identify variants of known NRPs that may have slightly different properties. Also, high-confidence A domain predictions can help elucidate and/or correct the mechanism of NRP assembly: we used PARAS to identify various inaccuracies in previously published articles and suggested more likely mechanisms of scaffold assembly instead. Also, it is important to remember that structure prediction is still used to guide natural product discovery by prioritising BGCs of potential novelty: we have a long way to go yet before computational analysis can replace wet-lab validation. Finally, building block prediction can play an important role in de-orphaning structures for which no BGC is known, and BGCs for which no structure is known. For example, the Magarvey group combined building block prediction with retro-biosynthesis to successfully link BGCs to NRP products, even though not all scaffold predictions were correct³⁹⁵.

9.5.2. Bioactivity prediction

While NRP structure prediction is uncertain, we can pinpoint where the uncertainty comes from and pursue avenues that can improve our predictions. This is different for bioactivity prediction. The bioactivity of a molecule is tightly linked to the interaction with its target. This involves the 3D conformation and dynamics of both the natural product and the target; interaction with any cofactors or ions, such as Ca^{2+} in calcium-dependent lipopeptides; potential oligomerization of the compound prior to or during interaction with the target, as seen for daptomycin⁴⁰⁹; and the bioactive form of the molecule, which may be different from the product that is initially synthesised, as seen for the prexencoumacins, which are cleaved to form their active final form⁴¹⁰. Without a clear indication of a molecular target, these properties are difficult to study: while we can quite easily determine the broad activity of a compound, such as antifungal, antibacterial, or antiviral properties, the exact mechanism of action is often challenging to uncover. Without a known protein target, obtaining X-ray crystallography data for drug-target complexes is not trivial or cheap. Also, crystal structures are static, while true drug-target complexes are dynamic. NMR could provide a solution, but while 1D NMR methods that are currently used for high-throughput screening can detect binding events and can be used to complement X-ray crystallography data, they do not provide the resolution required to resolve the structures of the drug-target complexes by themselves^{411,412}. As a result, we do not know the exact mechanism of action for a large proportion of bioactive molecules. Even the exact mechanism of daptomycin, a NRP antibiotic used in the clinic, is not fully understood⁴⁰⁹.

This creates a data problem: while we can train our models to predict general bioactivities, our lack of knowledge of the mechanism of action of most compounds means that we can say very little about how a molecule is likely to behave in the setting of its desired application. High-throughput mechanism of action annotation pipelines as developed by Potts et al. and Hight et al. might help increase the proportion of annotated molecules, but these annotations are often general, indicating a target pathway or protein but not a mode of binding. Therefore, we cannot make an informed decision on how to featurise our data such that we capture the most relevant properties. This issue is perfectly demonstrated by the work done by the Grisoni group on activity cliffs: pairs of highly similar compounds with different bioactivities. They showed that machine learning methods struggle with datasets that include such cliff pairs, as they typically use structural similarity, a feature they cannot rely on in this case, to predict bioactivity⁴¹³. This highlights that featurisation methods of existing machine learning algorithms do not capture all the factors that contribute towards a bioactivity. Despite this, machine learning has still been successfully applied to discover novel compounds with desired properties, such as the antibacterial drug halicin²²⁵. For now, it seems that machine learning algorithms can be leveraged to predict if a structure belongs to a compound group that has the potential for a certain bioactivity but cannot conclusively determine if the structure itself possesses the bioactivity.

Even when both NRP structure prediction and bioactivity prediction methods still require a lot of development before we can use the sequence-structure-function paradigm to reliably predict bioactivity, we can still use the predictive tasks that current methods are good at to find novel molecules with desired mechanisms of action. We can attempt to link pharmacophores, substructures that are responsible for certain functional properties, to building blocks, such as the beta-lactam ring in penicillin-like antibiotics^{414,415} or the Asp-X-Asp-Gly motif in lipopeptide antibiotics^{25,392} mentioned before. While this does not ensure that every molecule with these substructures has the desired bioactivity, such motifs can help prioritise strains that have the capacity to make certain molecule pools. Also, we can look for BGCs that might be responsible for the production of congeners of known molecules: it is often easier to predict the differences between molecules than it is to predict a molecule from scratch, similar to how it is easier to assemble a genome when you have a good reference available.

For now, predicting the bioactivity of molecules whose function results from novel mechanisms remains a challenge. Likely, it will require a much deeper understanding of the involved biological processes and chemical interactions than we currently have. This is the key reason that we used three-dimensional approaches throughout this thesis: to better approximate reality by using a representation that mirrors the real world.

9.6. Sequence or 3D structure: why do structure-based methods underperform?

In this thesis, we used three-dimensional representations in chapter 4 and chapter 5 for two distinct predictive tasks. In chapter 4, we predicted the secondary structure of mRNA as a set of base pairing probabilities along the length of the molecule and used these probabilities as features. In chapter 5, we predicted structural models for A domains, converted those models to voxel clouds around the A domain active site, and summarised the variation within those voxel clouds with principal component analysis. Against expectations, we found that structure-based approaches underperformed compared to sequence-based approaches: one-hot encoding, a sequence-based method, performed ~30% better than featurisation of base-pair probabilities for protein production prediction from

mRNA sequence, and sequence-based featurisation of the A domain active-site outperformed our voxel-based featurisation approach by 6-9% for A domain selectivity prediction. This begs the question why: as three-dimensional structure should be more directly related to a molecule's function than sequence, we expected the reverse.

The most likely reason for this observation is the effect of compounding predictions that we already saw in the context of NRP structure and bioactivity prediction: with each slightly uncertain prediction, the overall prediction becomes exponentially worse. mRNA secondary structure predictions are notoriously poor, and this is reflected by the large drop in Pearson correlation between predicted and observed protein production for models trained on these predictions. With AlphaFold 2, protein structure predictions have become a lot better, but they are still not perfect. Most importantly, they suffer from the same shortcoming as X-ray crystal structures: they paint a static picture of a highly dynamic system, with many transient interactions that are critical for the behaviour of the protein in its natural environment. In fact, the function of most enzymes highly relies on their mobility and flexibility. This is also true for A domains: they have two subdomains that are highly mobile with respect to one another to allow the phosphopantetheine arm of the downstream PCP domain to interact with its active site. Molecular dynamics (MD), which can model the behaviour of a molecule or set of molecules through time, can likely help improve structural approaches. However, the computational cost associated with such analyses even for small enzyme regions means that for now it is not feasible to use MD in high-throughput fashion, especially for modelling NRP synthesis, which requires the coordination of some of the largest enzymes that exist in bacteria and fungi. A large contributor to computational cost for MD is the modelling of probabilities of certain molecular events, a task that our current binary computers are not well-suited for. In the next few decades, we might find a solution in quantum computing, which by design is extremely efficient at probabilistic modelling.

To explore another reason for the underperformance of structure-based methods, we need to understand the concept of embedding in machine learning. Prior to training models, it is important to represent data in a format that machine learning algorithms can work with. Most machine learning algorithms work with vectors: sequences of numbers or strings. Even machine learning frameworks that appear to work on different objects, such as graph neural networks or convolutional neural networks, embed these objects as vectors under the hood. Therefore, we are predicting structure from sequence, and then embed that structure as a sequence again. This seems counterproductive: while providing a different view of the data, chances are that we are losing information as we are hand-picking which (often static) structural features we deem important. As such, perhaps we can view biological sequences as highly accurate embeddings of DNA, mRNA, and protein structures that provide near-perfect one-dimensional representations of structural features that we can predict as well as those we cannot.

This does not mean that structure-based methods do not have a place in biological research. They are especially powerful in the alignment of divergent sequences: structure alignments can detect similar protein folds in homologues across great evolutionary distances⁴¹⁶, and therefore can aid in predicting the function of proteins for which no close homologues have been studied. While A domains show high sequence similarity across the majority of the protein, their active sites are highly divergent, and as such, sequence-based alignments struggle to detect the exact architecture of the active site. For this reason, we used structure-based alignment in chapter 5 to extract the A domain active site signatures and demonstrated that structure-based extraction led to more powerful

models. This way, we achieved the best of both worlds: we used structure to identify which residues we needed, and then used the biological sequence of those residues as an embedding of the active site.

Also, while structure-based approaches underperform, they still do have some predictive power. This means that they can be leveraged for feature inference, especially in combination with interpretable AI. This was the main purpose of the protein production predictors we trained in chapter 4: we meant to understand which regions of the mRNA molecule contribute towards beneficial protein production. From feature importance analysis on models trained on mRNA secondary structure alone, we could discern that low binding probability surrounding the ribosome binding site (RBS) and the transcription start site (TSS) was instrumental in achieving high protein production levels, with the exception of two bases in between the RBS and TSS. This led to the hypothesis that the involvement of these bases in secondary structures might actually be beneficial for ribosome binding, an observation that can henceforth be used in designing sequences with specific expression profiles.

9.7. Conclusion

Above, we discussed the potential of leveraging the sequence-structure-function paradigm for the prediction of NRP structure and bioactivity and explored the potential role for structural approaches in this process. We saw that NRPS systems are incredibly complex, with many biological factors at play within the BGC, the producing organism's genome, and even in the organism's environment. As it stands, there are still many limitations that prevent us from getting accurate predictions for NRP structure and bioactivity: data quantity and quality, biological conversions and chemical interactions that we cannot yet predict, and the exponential uncertainty that comes with stacking predictions on top of predictions. We showed that in general, structure-based approaches do not perform as well as sequence-based approaches on this problem but demonstrate their utility in alignment of divergent sequences and feature inference to gain mechanistic understanding into the processes underlying NRP biosynthesis.

If one day we hope to replace experimental NRP discovery with bioinformatics, a lot of work needs to be done and not just in the field of natural products. Natural product biosynthesis is just one part of a much larger, ecological system, and in order to accurately predict NRP structure and bioactivity from sequence, we need a much broader and deeper mechanistic understanding of that full system. We believe that the acquisition of metagenome and meta-metabolome data will be pivotal in this endeavour, as they give a more complete picture of the producing organism's environment, heighten the probability of the organism producing specialised metabolites in their native state, and provide an ecological context into natural product biosynthesis and the bioactivities associated with that ecological state. For now, even predictions that are only partially correct can help accelerate the discovery of novel natural products, especially for congeners of known molecules and for compounds produced by well-explored phyla.

References

1. Nemergut, D. R. *et al.* Structure and function of alpine and arctic soil microbial communities. *Res. Microbiol.* **156**, 775–784 (2005).
2. Tripathi, B. M. *et al.* Distinctive tropical forest variants have unique soil microbial communities, but not always low microbial diversity. *Front. Microbiol.* **7**, 1–11 (2016).
3. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
4. Chevrette, M. G. *et al.* The antimicrobial potential of *Streptomyces* from insect microbiomes. *Nat. Commun.* **10**, 1–11 (2019).
5. Turley, C. Bacteria in the cold deep-sea benthic boundary layer and sediment-water interface of the NE Atlantic. *FEMS Microbiol. Ecol.* **33**, 89–99 (2000).
6. Molina-Menor, E., Gimeno-Valero, H., Pascual, J., Peretó, J. & Porcar, M. High Culturable Bacterial Diversity From a European Desert: The Tabernas Desert. *Front. Microbiol.* **11**, 1–15 (2021).
7. Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nat. Rev. Microbiol.* **18**, 152–163 (2020).
8. Guerinot, M. L. Microbial iron transport. *Annu. Rev. Microbiol.* **48**, 743–772 (1994).
9. Liu, M. *et al.* Microbial production of ectoine and hydroxyectoine as high-value chemicals. *Microb. Cell Fact.* **20**, 1–11 (2021).
10. Czech, L. *et al.* Role of the extremolytes ectoine and hydroxyectoine as stress protectants and nutrients: Genetics, phylogenomics, biochemistry, and structural analysis. *Genes (Basel)*. **9**, 1–58 (2018).
11. Sadeghi, A. *et al.* Diversity of the ectoines biosynthesis genes in the salt tolerant *Streptomyces* and evidence for inductive effect of ectoines on their accumulation. *Microbiol. Res.* **169**, 699–708 (2014).
12. Gao, Q. & Garcia-Pichel, F. Microbial ultraviolet sunscreens. *Nat. Rev. Microbiol.* **9**, 791–802 (2011).
13. Fleming, A. Penicillin. *Br. Med. J.* **2**, 386 (1941).
14. Miller, E. L. The penicillins: A review and update. *J. Midwifery Women's Heal.* **47**, 426–434 (2002).
15. Gond, S. K., Bergen, M. S., Torres, M. S. & White, J. F. Endophytic *Bacillus* spp. produce antifungal lipopeptides and induce host defence gene expression in maize. *Microbiol. Res.* **172**, 79–87 (2015).

16. Santoyo, G. How plants recruit their microbiome? New insights into beneficial interactions. *J. Adv. Res.* **40**, 45–58 (2022).
17. Abrudan, M. I. *et al.* Socially mediated induction and suppression of antibiosis during bacterial coexistence. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11054–11059 (2015).
18. Keller, L. & Surette, M. G. Communication in bacteria: An ecological and evolutionary perspective. *Nat. Rev. Microbiol.* **4**, 249–258 (2006).
19. Davies, J. Are antibiotics naturally antibiotics? *J. Ind. Microbiol. Biotechnol.* **33**, 496–499 (2006).
20. Romero, D., Traxler, M. F., Daniel, L. & Kolter, R. Antibiotics as Signal Molecules - Chemical Reviews (ACS Publications). *Pubs.Acs.Org* 5492–5505 (2011).
21. Fleming, A. On the antibacterial action of cultures of a penicillium with special reference to their use in the isolation of *B. influenzae*. *Br. J. Exp. Pathol.* **10**, 226–236 (1929).
22. Ikuta, K. S. *et al.* Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **400**, 2221–2248 (2022).
23. Nothias, L. F., Knight, R. & Dorrestein, P. C. Antibiotic discovery is a walk in the park. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14477–14479 (2016).
24. Rubinstein, E. & Keynan, Y. Vancomycin revisited - 60 years later. *Front. Public Heal.* **2**, 1–7 (2014).
25. Heidary, M. *et al.* Daptomycin. *J. Antimicrob. Chemother.* **73**, 1–11 (2018).
26. Baltz, R. H., Miao, V., Wrigley, S. K. & Miao, V. Natural products to drugs: daptomycin and related lipopeptide antibiotics. *Nat. Prod. Rep.* **22**, 717–741 (2005).
27. Miao, V. *et al.* Daptomycin biosynthesis in *Streptomyces roseosporus*: Cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151**, 1507–1523 (2005).
28. Umezawa, H., Maeda, K., Takeuchi, T. & Okami, Y. New antibiotics, bleomycin A and B. *J. Antibiot. (Tokyo)*. **19**, 200–209 (1966).
29. Froudarakis, M. *et al.* Revisiting bleomycin from pathophysiology to safe clinical use. *Crit. Rev. Oncol. Hematol.* **87**, 90–100 (2013).
30. Pawar, S., Chaudhari, A., Prabha, R., Shukla, R. & Singh, D. P. Microbial pyrrolnitrin: Natural metabolite with immense practical utility. *Biomolecules* **9**, (2019).

31. Tani, K., Usuki, Y., Motoba, K., Fujita, K. & Taniguchi, M. UK-2A, B, C, and D, novel antifungal antibiotics from *Streptomyces* sp. 517-02 VII. Membrane injury induced by C9-UK-2A, a derivative of UK-2A, in *Rhodotorula mucilaginosa* IFO 0001. *J. Antibiot. (Tokyo)*. **55**, 315–321 (2002).
32. Sparks, T. C. *et al.* The spinosyns, spinosad, spinetoram, and synthetic spinosyn mimics - discovery, exploration, and evolution of a natural product chemistry and the impact of computational tools. *Pest Manag. Sci.* **77**, 3637–3649 (2021).
33. Achan, J. *et al.* Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar. J.* **10**, 144 (2011).
34. Takano, H. K. & Dayan, F. E. Glufosinate-ammonium: a review of the current state of knowledge. *Pest Manag. Sci.* (2020).
35. Watson, J. & Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
36. Crick, F. H. The genetic code--yesterday, today, and tomorrow. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 1–9 (1966).
37. Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13**, 509–523 (2015).
38. van Bergeijk, D. A., Terlouw, B. R., Medema, M. H. & van Wezel, G. P. Ecology and genomics of Actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* **18**, 546–558 (2020).
39. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* (2005) doi:10.1016/j.gene.2005.06.037.
40. Behloul, N. *et al.* Effects of mRNA secondary structure on the expression of HEV ORF2 proteins in *Escherichia coli*. *Microb. Cell Fact.* **16**, 1–14 (2017).
41. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **10**, 770–770 (2014).
42. Nieuwkoop, T. *et al.* Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* **51**, 2363–2376 (2023).
43. Höllerer, S. & Jeschek, M. Ultradeep characterisation of translational sequence determinants refutes rare-codon hypothesis and unveils quadruplet base pairing of initiator tRNA and transcript. *Nucleic Acids Res.* **51**, 2377–2396 (2023).
44. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505

(1999).

45. Walsh, C. T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* **33**, 127–135 (2016).
46. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angew. Chemie - Int. Ed.* **56**, 3770–3821 (2017).
47. Mootz, H. D. & Marahiel, M. A. The Tyrocidine Biosynthesis Operon of *Bacillus brevis*: Complete Nucleotide Sequence and Biochemical Characterization of Functional Internal Adenylation Domains. *Microbiology* **179**, 6843–6850 (1997).
48. Smith, H. G., Beech, M. J., Lewandowski, J. R., Challis, G. L. & Jenner, M. Docking domain-mediated subunit interactions in natural product megasynth(et)ases. *J. Ind. Microbiol. Biotechnol.* **48**, (2021).
49. Watzel, J., Hacker, C., Duchardt-Ferner, E., Bode, H. B. & Wöhnert, J. A New Docking Domain Type in the Peptide-Antimicrobial-Xenorhabdus Peptide Producing Nonribosomal Peptide Synthetase from *Xenorhabdus bovienii*. *ACS Chem. Biol.* **15**, 982–989 (2020).
50. Hahn, M. & Stachelhaus, T. Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15585–15590 (2004).
51. Du, L., Sánchez, C., Chen, M., Edwards, D. J. & Shen, B. The biosynthetic gene cluster for the antitumor drug bleomycin from *Streptomyces verticillus* ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chem. Biol.* **7**, 623–642 (2000).
52. Calcott, M. J., Owen, J. G. & Ackerley, D. F. Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. *Nat. Commun.* **11**, 4554 (2020).
53. Meyer, S. *et al.* Biochemical Dissection of the Natural Diversification of Microcystin Provides Lessons for Synthetic Biology of NRPS. *Cell Chem. Biol.* **23**, 462–471 (2016).
54. Bozhüyük, K. A. J. *et al.* Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nat. Chem.* **11**, 653–661 (2019).
55. Linne, U. & Marahiel, M. A. Control of directionality in nonribosomal peptide synthesis: role of the condensation domain in preventing misinitiation and timing of epimerization. *Biochemistry* **39**, 10439–10447 (2000).
56. Ames, B. D. & Walsh, C. T. Anthranilate-activating modules from fungal nonribosomal

- peptide assembly lines. *Biochemistry* **49**, 3351–3365 (2010).
57. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7**, 78 (2007).
 58. Keating, T. A. *et al.* Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-S-enzyme breakdown in antibiotic and siderophore biosynthesis. *Chembiochem* **2**, 99–107 (2001).
 59. Miller, B. R. & Gulick, A. M. Structural Biology of Nonribosomal Peptide Synthetases. in *Nonribosomal Peptide and Polyketide Biosynthesis: Methods and Protocols* (ed. Evans, B. S.) 3–29 (Springer New York, 2016). doi:10.1007/978-1-4939-3375-4_1.
 60. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
 61. Caboche, S., Leclère, V., Pupin, M., Kuchеров, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: Towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
 62. Blodgett, J. A. V, Zhang, J. K., Yu, X. & Metcalf, W. W. Conserved biosynthetic pathways for phosalacine, bialaphos and newly discovered phosphonic acid natural products. *J. Antibiot. (Tokyo)*. **69**, 15–25 (2016).
 63. Du, L., Sánchez, C. & Shen, B. Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab. Eng.* **3**, 78–95 (2001).
 64. Noike, M. *et al.* A peptide ligase and the ribosome cooperate to synthesize the peptide pheganomycin. *Nat. Chem. Biol.* **11**, 71–76 (2015).
 65. Ding, Y. *et al.* Genome-based characterization of two prenylation steps in the assembly of the stephacidin and notoamide anticancer agents in a marine-derived *Aspergillus* sp. *J. Am. Chem. Soc.* **132**, 12733–12740 (2010).
 66. Liu, J. *et al.* Biosynthesis of the anti-infective marformycins featuring pre-nrps assembly line N -formylation and O -methylation and post-assembly line C -hydroxylation chemistries. *Org. Lett.* **17**, 1509–1512 (2015).
 67. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* (2023) doi:10.1093/nar/gkad344.
 68. Blin, K. *et al.* AntiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).

69. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
70. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
71. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
72. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2022).
73. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
74. Walker, A. S. & Clardy, J. A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* **61**, 2560–2571 (2021).
75. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).
76. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799–5808 (2005).
77. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, 362–367 (2011).
78. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
79. Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* **8**, e62136 (2013).
80. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
81. Wood, T. M. & Martin, N. I. The calcium-dependent lipopeptide antibiotics: structure, mechanism, & medicinal chemistry. *Medchemcomm* **10**, 634–646 (2019).

82. Robinson, S. L. *et al.* Global analysis of adenylate-forming enzymes reveals b-lactone biosynthesis pathway in pathogenic nocardia. *J. Biol. Chem.* **295**, 14826–14839 (2020).
83. Blin, K. *et al.* AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
84. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
85. Weininger, D. SMILES, a Chemical Language and Information System. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
86. Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
87. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
88. Goodfellow, M. *et al.* *Bergey's Manual of Systematic Bacteriology: Volume 5: The Actinobacteria*. (Springer Science & BusinessMedia, 2012).
89. van der Meij, A., Worsley, S. F., Hutchings, M. I. & van Wezel, G. P. Chemical ecology of antibiotic production by actinomycetes. *FEMS Microbiol. Rev.* **41**, 392–416 (2017).
90. Barka, E. A. *et al.* Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2016).
91. Bérdy, J. Bioactive Microbial Metabolites. *J. Antibiot. (Tokyo)*. **58**, 1–26 (2005).
92. Hopwood, D. A. *Streptomyces in Nature and Medicine: the Antibiotic Makers*. (Oxford University Press, 2007).
93. Vrancken, K. & Anné, J. Secretory production of recombinant proteins by *Streptomyces*. *Future Microbiol.* **4**, 181–188 (2009).
94. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **70**, 461–477 (2007).
95. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
96. Cooper, M. A. & Shlaes, D. Fix the antibiotics pipeline. *Nature* **472**, 32 (2011).
97. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).

98. Baltz, R. Antimicrobials from actinomycetes: Back to the future. *Microbe* **2**, 125–131 (2007).
99. Baltz, R. H. Renaissance in antibacterial discovery from actinomycetes. *Curr. Opin. Pharmacol.* **8**, 557–563 (2008).
100. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
101. Ikeda, H. *et al.* Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* **21**, 526–531 (2003).
102. van den Berg, M. A. *et al.* Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nat. Biotechnol.* **26**, 1161–1168 (2008).
103. Bode, H. B., Bethe, B., Höfs, R. & Zeeck, A. Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem* **3**, 619–627 (2002).
104. Romano, S., Jackson, S. A., Patry, S. & Dobson, A. D. W. Extending the 'One Strain Many Compounds' (OSMAC) Principle to Marine Microorganisms. *Mar. Drugs* **16**, (2018).
105. Wu, C. *et al.* Lugdunomycin, an Angucycline-Derived Molecule with Unprecedented Chemical Architecture. *Angew. Chem. Int. Ed. Engl.* **58**, 2809–2814 (2019).
106. Wright, E. S. & Vetsigian, K. H. Inhibitory interactions promote frequent bistability among competing bacteria. *Nat. Commun.* **7**, 11274 (2016).
107. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8**, 15–25 (2010).
108. Abrudan, M. I. *et al.* Socially mediated induction and suppression of antibiosis during bacterial coexistence. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11054–11059 (2015).
109. Traxler, M. F. & Kolter, R. Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.* **32**, 956–970 (2015).
110. Traxler, M. F., Watrous, J. D., Alexandrov, T., Dorrestein, P. C. & Kolter, R. Interspecies interactions stimulate diversification of the *Streptomyces coelicolor* secreted metabolome. *MBio* **4**, (2013).
111. Zhu, H., Sandiford, S. K. & van Wezel, G. P. Triggers and cues that activate antibiotic production by actinomycetes. *J. Ind. Microbiol. Biotechnol.* **41**, 371–386 (2014).
112. Li, C., Ji, C. & Tang, B. Purification, characterisation and biological activity of melanin from *Streptomyces* sp. *FEMS Microbiol. Lett.* **365**, fny077 (2018).

113. Kroiss, J. *et al.* Symbiotic Streptomyces provide antibiotic combination prophylaxis for wasp offspring. *Nat. Chem. Biol.* **6**, 261–263 (2010).
114. Raaijmakers, J. M. & Mazzola, M. Diversity and natural functions of antibiotics produced by beneficial and plant pathogenic bacteria. *Annu. Rev. Phytopathol.* **50**, 403–424 (2012).
115. Doroghazi, J. R. & Metcalf, W. W. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **14**, 611 (2013).
116. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
117. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat. Commun.* **8**, 1162 (2017).
118. Shimkets, L. *Bacterial Genomes: Physical Structure and Analysis*. (Chapman & Hall, 1998).
119. Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397–403 (1995).
120. Salem, H. *et al.* Drastic Genome Reduction in an Herbivore's Pectinolytic Symbiont. *Cell* **171**, 1520–1531.e13 (2017).
121. Davies, J. Millennium bugs. *Trends Cell Biol.* **9**, M2–M5 (1999).
122. Claessen, D., Rozen, D. E., Kuipers, O. P., Søgaard-Andersen, L. & van Wezel, G. P. Bacterial solutions to multicellularity: a tale of biofilms, filaments and fruiting bodies. *Nat. Rev. Microbiol.* **12**, 115–124 (2014).
123. Flärdh, K. & Buttner, M. J. Streptomyces morphogenetics: dissecting differentiation in a filamentous bacterium. *Nat. Rev. Microbiol.* **7**, 36–49 (2009).
124. Chater, K. & Losick, R. *Bacteria as Multicellular Organisms*. (Oxford University Press, 1997).
125. Merrick, M. J. A morphological and genetic mapping study of bald colony mutants of Streptomyces coelicolor. *J. Gen. Microbiol.* **96**, 299–315 (1976).
126. Hopwood, D. A., Wildermuth, H. & Palmer, H. M. Mutants of Streptomyces coelicolor defective in sporulation. *J. Gen. Microbiol.* **61**, 397–408 (1970).
127. Gomez-Escribano, J. P. *et al.* Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the cpk gene cluster of Streptomyces coelicolor M145. *Chem. Sci.* **3**, 2716–2720 (2012).

128. Bibb, M. J. Regulation of secondary metabolism in streptomycetes. *Curr. Opin. Microbiol.* **8**, 208–215 (2005).
129. van der Heul, H. U., Bilyk, B. L., McDowall, K. J., Seipke, R. F. & van Wezel, G. P. Regulation of antibiotic production in Actinobacteria: new perspectives from the post-genomic era. *Nat. Prod. Rep.* **35**, 575–604 (2018).
130. Manteca, Á., Fernández, M. & Sánchez, J. A death round affecting a young compartmentalized mycelium precedes aerial mycelium dismantling in confluent surface cultures of *Streptomyces antibioticus*. *Microbiology* **151**, 3689–3697 (2005).
131. Manteca, A., Mäder, U., Connolly, B. & Sanchez, J. A proteomic analysis of *Streptomyces coelicolor* programmed cell death. *Proteomics* **6**, 6008–6022 (2006).
132. Tenconi, E., Traxler, M. F., Hoebreck, C., van Wezel, G. P. & Rigali, S. Production of Prodiginines Is Part of a Programmed Cell Death Process in *Streptomyces coelicolor*. *Front. Microbiol.* **9**, 1742 (2018).
133. Ohnishi, Y. *et al.* Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **190**, 4050–4060 (2008).
134. Wu, C. *et al.* Expanding the chemical space for natural products by *Aspergillus*-*Streptomyces* co-cultivation and biotransformation. *Sci. Rep.* **5**, 10868 (2015).
135. Challis, G. L. & Hopwood, D. A. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc. Natl. Acad. Sci. U. S. A.* **100 Suppl 2**, 14555–14561 (2003).
136. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
137. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).
138. Amos, G. C. A. *et al.* Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E11121–E11130 (2017).
139. Machado, H., Tuttle, R. N. & Jensen, P. R. Omics-based natural product discovery and the lexicon of genome mining. *Curr. Opin. Microbiol.* **39**, 136–142 (2017).
140. Martinet, L. *et al.* A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. *MBio* **10**, (2019).
141. Seipke, R. F., Kaltenpoth, M. & Hutchings, M. I. *Streptomyces* as symbionts: an emerging and widespread theme? *FEMS Microbiol. Rev.* **36**, 862–876 (2012).

142. Jensen, P. R., Williams, P. G., Oh, D.-C., Zeigler, L. & Fenical, W. Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl. Environ. Microbiol.* **73**, 1146–1152 (2007).
143. Yang, A. *et al.* Nitrosporeusines A and B, unprecedented thioester-bearing alkaloids from the Arctic *Streptomyces nitrosporeus*. *Org. Lett.* **15**, 5366–5369 (2013).
144. Sayed, A. M. *et al.* Extreme environments: microbiology leading to specialized metabolites. *J. Appl. Microbiol.* **128**, 630–657 (2020).
145. Zipperer, A. *et al.* Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535**, 511–516 (2016).
146. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1130–9 (2014).
147. Adamek, M. *et al.* Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. *BMC Genomics* **19**, 426 (2018).
148. Andam, C. P., Choudoir, M. J., Vinh Nguyen, A., Sol Park, H. & Buckley, D. H. Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *ISME J.* **10**, 1731–1741 (2016).
149. Tidjani, A.-R. *et al.* Massive Gene Flux Drives Genome Diversity between Sympatric *Streptomyces* Conspecifics. *MBio* **10**, (2019).
150. McDonald, B. R. & Currie, C. R. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus *Streptomyces*. *MBio* **8**, (2017).
151. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
152. Joynt, R. & Seipke, R. F. A phylogenetic and evolutionary analysis of antimycin biosynthesis. *Microbiology* **164**, 28–39 (2018).
153. Chevrette, M. G. *et al.* Taxonomic and Metabolic Incongruence in the Ancient Genus *Streptomyces*. *Front. Microbiol.* **10**, 2170 (2019).
154. Bruns, H. *et al.* Function-related replacement of bacterial siderophore pathways. *ISME J.* **12**, 320–329 (2018).
155. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends Microbiol.* **24**, 968–977 (2016).
156. Chater, K. F. & Chandra, G. The evolution of development in *Streptomyces* analysed

- by genome comparisons . *FEMS Microbiol. Rev.* **30**, 651–672 (2006).
157. Ventura, M. *et al.* Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev.* **71**, 495–548 (2007).
 158. Choulet, F. *et al.* Evolution of the terminal regions of the Streptomyces linear chromosome. *Mol. Biol. Evol.* **23**, 2361–2369 (2006).
 159. Bilyk, B., Horbal, L. & Luzhetskyy, A. Chromosomal position effect influences the heterologous expression of genes and biosynthetic gene clusters in Streptomyces albus J1074. *Microb. Cell Fact.* **16**, 5 (2017).
 160. Letzel, A.-C. *et al.* Genomic insights into specialized metabolism in the marine actinomycete Salinispora. *Environ. Microbiol.* **19**, 3660–3673 (2017).
 161. Ghinet, M. G. *et al.* Uncovering the prevalence and diversity of integrating conjugative elements in actinobacteria. *PLoS One* **6**, e27846 (2011).
 162. Kinashi, H., Shimaji, M. & Sakai, A. Giant linear plasmids in Streptomyces which code for antibiotic biosynthesis genes. *Nature* **328**, 454–456 (1987).
 163. Medema, M. H. *et al.* The sequence of a 1.8-mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. *Genome Biol. Evol.* **2**, 212–224 (2010).
 164. Mochizuki, S. *et al.* The large linear plasmid pSLA2-L of Streptomyces rochei has an unusually condensed gene organization for secondary metabolism. *Mol. Microbiol.* **48**, 1501–1510 (2003).
 165. Hoskisson, P. A. & Fernández-Martínez, L. T. Regulation of specialised metabolites in Actinobacteria - expanding the paradigms. *Environ. Microbiol. Rep.* **10**, 231–238 (2018).
 166. Huang, J. *et al.* Cross-regulation among disparate antibiotic biosynthetic pathways of Streptomyces coelicolor. *Mol. Microbiol.* **58**, 1276–1287 (2005).
 167. McLean, T. C., Hoskisson, P. A. & Seipke, R. F. Coordinate Regulation of Antimycin and Candicidin Biosynthesis. *mSphere* **1**, (2016).
 168. Liu, G., Chater, K. F., Chandra, G., Niu, G. & Tan, H. Molecular regulation of antibiotic biosynthesis in streptomyces. *Microbiol. Mol. Biol. Rev.* **77**, 112–143 (2013).
 169. Wietzorrek, A. & Bibb, M. A novel family of proteins that regulates antibiotic production in streptomycetes appears to contain an OmpR-like DNA-binding fold. *Molecular microbiology* vol. 25 1181–1184 (1997).

170. Autret, S., Nair, R. & Errington, J. Genetic analysis of the chromosome segregation protein Spo0J of *Bacillus subtilis*: evidence for separate domains involved in DNA binding and interactions with Soj protein. *Mol. Microbiol.* **41**, 743–755 (2001).
171. Gramajo, H. C., Takano, E. & Bibb, M. J. Stationary-phase production of the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2) is transcriptionally regulated. *Mol. Microbiol.* **7**, 837–845 (1993).
172. Tomono, A., Tsai, Y., Yamazaki, H., Ohnishi, Y. & Horinouchi, S. Transcriptional control by A-factor of *strR*, the pathway-specific transcriptional activator for streptomycin biosynthesis in *Streptomyces griseus*. *J. Bacteriol.* **187**, 5595–5604 (2005).
173. Lawlor, E. J., Baylis, H. A. & Chater, K. F. Pleiotropic morphological and antibiotic deficiencies result from mutations in a gene encoding a tRNA-like product in *Streptomyces coelicolor* A3(2). *Genes Dev.* **1**, 1305–1310 (1987).
174. Fernández-Moreno, M. A., Caballero, J. L., Hopwood, D. A. & Malpartida, F. The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bldA* tRNA gene of *Streptomyces*. *Cell* **66**, 769–780 (1991).
175. Takano, E. Gamma-butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation. *Curr. Opin. Microbiol.* **9**, 287–294 (2006).
176. Willey, J. M. & Gaskell, A. A. Morphogenetic signaling molecules of the streptomycetes. *Chem. Rev.* **111**, 174–187 (2011).
177. Tahlan, K. *et al.* Initiation of actinorhodin export in *Streptomyces coelicolor*. *Mol. Microbiol.* **63**, 951–961 (2007).
178. Wang, L. *et al.* Autoregulation of antibiotic biosynthesis by binding of the end product to an atypical response regulator. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 8617–8622 (2009).
179. Willems, A. R. *et al.* Crystal structures of the *Streptomyces coelicolor* TetR-like protein ActR alone and in complex with actinorhodin or the actinorhodin biosynthetic precursor (S)-DNPA. *J. Mol. Biol.* **376**, 1377–1387 (2008).
180. Francis, I. M., Jourdan, S., Fanara, S., Loria, R. & Rigali, S. The cellobiose sensor CebR is the gatekeeper of *Streptomyces scabies* pathogenicity. *MBio* **6**, e02018 (2015).
181. Rigali, S. *et al.* Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by *Streptomyces*. *EMBO Rep.* **9**, 670–675 (2008).
182. Nazari, B. *et al.* Chitin-induced gene expression in secondary metabolic pathways of *Streptomyces coelicolor* A3(2) grown in soil. *Appl. Environ. Microbiol.* **79**, 707–713 (2013).

183. Craig, M. *et al.* Unsuspected control of siderophore production by N-acetylglucosamine in streptomycetes. *Environ. Microbiol. Rep.* **4**, 512–521 (2012).
184. Świątek-Połatyńska, M. A. *et al.* Genome-wide analysis of in vivo binding of the master regulator DasR in *Streptomyces coelicolor* identifies novel non-canonical targets. *PLoS One* **10**, e0122479 (2015).
185. Zhu, H. *et al.* Eliciting antibiotics active against the ESKAPE pathogens in a collection of actinomycetes isolated from mountain soils. *Microbiology* **160**, 1714–1725 (2014).
186. Hosaka, T. *et al.* Antibacterial discovery in actinomycetes strains with mutations in RNA polymerase or ribosomal protein S12. *Nat. Biotechnol.* **27**, 462–464 (2009).
187. Tanaka, Y. *et al.* Antibiotic overproduction by rpsL and rsmG mutants of various actinomycetes. *Appl. Environ. Microbiol.* **75**, 4919–4922 (2009).
188. Bertrand, S. *et al.* Metabolite induction via microorganism co-culture: a potential way to enhance chemical diversity for drug discovery. *Biotechnol. Adv.* **32**, 1180–1204 (2014).
189. Hoshino, S., Wakimoto, T., Onaka, H. & Abe, I. Chojalactones A-C, cytotoxic butanolides isolated from *Streptomyces* sp. cultivated with mycolic acid containing bacterium. *Org. Lett.* **17**, 1501–1504 (2015).
190. Sugiyama, R. *et al.* 5-Alkyl-1,2,3,4-tetrahydroquinolines, new membrane-interacting lipophilic metabolites produced by combined culture of *Streptomyces nigrescens* and *Tsukamurella pulmonis*. *Org. Lett.* **17**, 1918–1921 (2015).
191. Hsiao, N.-H., Gottelt, M. & Takano, E. Chapter 6. Regulation of antibiotic production by bacterial hormones. *Methods Enzymol.* **458**, 143–157 (2009).
192. Albright, J. C. *et al.* Large-scale metabolomics reveals a complex response of *Aspergillus nidulans* to epigenetic perturbation. *ACS Chem. Biol.* **10**, 1535–1541 (2015).
193. Craney, A., Ozimok, C., Pimentel-Elardo, S. M., Capretta, A. & Nodwell, J. R. Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* **19**, 1020–1027 (2012).
194. Moon, K., Xu, F. & Seyedsayamdost, M. R. Cebulantin, a Cryptic Lanthipeptide Antibiotic Uncovered Using Bioactivity-Coupled HiTES. *Angew. Chem. Int. Ed. Engl.* **58**, 5973–5977 (2019).
195. Moon, K., Xu, F., Zhang, C. & Seyedsayamdost, M. R. Bioactivity-HiTES Unveils Cryptic Antibiotics Encoded in Actinomycete Bacteria. *ACS Chem. Biol.* **14**, 767–774 (2019).

196. Niu, G., Chater, K. F., Tian, Y., Zhang, J. & Tan, H. Specialised metabolites regulating antibiotic biosynthesis in *Streptomyces* spp. *FEMS Microbiol. Rev.* **40**, 554–573 (2016).
197. Onaka, H., Mori, Y., Igarashi, Y. & Furumai, T. Mycolic acid-containing bacteria induce natural-product biosynthesis in *Streptomyces* species. *Appl. Environ. Microbiol.* **77**, 400–406 (2011).
198. Schroeckh, V. *et al.* Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in *Aspergillus nidulans*. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14558–14563 (2009).
199. Sung, A. A., Gromek, S. M. & Balunas, M. J. Upregulation and Identification of Antibiotic Activity of a Marine-Derived *Streptomyces* sp. via Co-Cultures with Human Pathogens. *Mar. Drugs* **15**, (2017).
200. Pérez, J. *et al.* *Myxococcus xanthus* induces actinorhodin overproduction and aerial mycelium formation by *Streptomyces coelicolor*. *Microb. Biotechnol.* **4**, 175–183 (2011).
201. Patin, N. V., Floros, D. J., Hughes, C. C., Dorrestein, P. C. & Jensen, P. R. The role of inter-species interactions in *Salinispora* specialized metabolism. *Microbiology* **164**, 946–955 (2018).
202. Ezaki, M. *et al.* Biphenomycin A production by a mixed culture. *Appl. Environ. Microbiol.* **58**, 3879–3882 (1992).
203. Kurosawa, K. *et al.* Rhodostreptomycins, antibiotics biosynthesized following horizontal gene transfer from *Streptomyces padanus* to *Rhodococcus fascians*. *J. Am. Chem. Soc.* **130**, 1126–1127 (2008).
204. Traxler, M. F., Seyedsayamdost, M. R., Clardy, J. & Kolter, R. Interspecies modulation of bacterial development through iron competition and siderophore piracy. *Mol. Microbiol.* **86**, 628–644 (2012).
205. Onaka, H., Tabata, H., Igarashi, Y., Sato, Y. & Furumai, T. Goadsporin, a chemical substance which promotes secondary metabolism and morphogenesis in streptomycetes. I. Purification and characterization. *J. Antibiot. (Tokyo)*. **54**, 1036–1044 (2001).
206. Yang, Y.-L., Xu, Y., Straight, P. & Dorrestein, P. C. Translating metabolic exchange with imaging mass spectrometry. *Nat. Chem. Biol.* **5**, 885–887 (2009).
207. Mendes, R. *et al.* Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097–1100 (2011).
208. Currie, C. R., Scott, J., Summerbell, R. & Malloch, D. Erratum: Fungus-growing ants use

- antibiotic-producing bacteria to control garden parasites (Nature (1999) 398 (701-704)). *Nature* **423**, (2003).
209. Heine, D. *et al.* Chemical warfare between leafcutter ant symbionts and a co-evolved pathogen. *Nat. Commun.* **9**, 2208 (2018).
 210. Spaepen, S. *Principles of Plant-Microbe Interactions*. (Springer, 2015).
 211. van der Meij, A. *et al.* Inter- and intracellular colonization of Arabidopsis roots by endophytic actinobacteria and the impact of plant hormones on their antimicrobial activity. *Antonie Van Leeuwenhoek* **111**, 679–690 (2018).
 212. Doroghazi, J. R. *et al.* A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
 213. Badri, D. V & Vivanco, J. M. Regulation and function of root exudates. *Plant. Cell Environ.* **32**, 666–681 (2009).
 214. Badri, D. V, Chaparro, J. M., Zhang, R., Shen, Q. & Vivanco, J. M. Application of natural blends of phytochemicals derived from the root exudates of Arabidopsis to the soil reveal that phenolic-related compounds predominantly modulate the soil microbiome. *J. Biol. Chem.* **288**, 4502–4512 (2013).
 215. Bulgarelli, D. *et al.* Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
 216. Lebeis, S. L. *et al.* PLANT MICROBIOME. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* **349**, 860–864 (2015).
 217. Alanjary, M. *et al.* The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.* **45**, W42–W48 (2017).
 218. Schorn, M. A. *et al.* Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075–2086 (2016).
 219. Charlop-Powers, Z. *et al.* Global biogeographic sampling of bacterial secondary metabolism. *Elife* **4**, e05048 (2015).
 220. Charlop-Powers, Z. *et al.* Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14811–14816 (2016).
 221. van Heel, A. J. *et al.* BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.* **46**, W278–W281 (2018).

222. Du, C. & van Wezel, G. P. Mining for Microbial Gems: Integrating Proteomics in the Postgenomic Natural Product Discovery Pipeline. *Proteomics* **18**, e1700332 (2018).
223. Goering, A. W. *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.* **2**, 99–108 (2016).
224. Soldatou, S., Eldjarn, G. H., Huerta-Urbe, A., Rogers, S. & Duncan, K. R. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. *FEMS Microbiol. Lett.* **366**, (2019).
225. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702.e13 (2020).
226. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743 (2016).
227. Tiwari, K. & Gupta, R. K. Rare actinomycetes: a potential storehouse for novel antibiotics. *Crit. Rev. Biotechnol.* **32**, 108–132 (2012).
228. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
229. Culp, E. J. *et al.* Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat. Biotechnol.* **37**, 1149–1154 (2019).
230. Hiard, S. *et al.* PREDetector: a new tool to identify regulatory elements in bacterial genomes. *Biochem. Biophys. Res. Commun.* **357**, 861–864 (2007).
231. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. & Stormo, G. D. A comparative genomics approach to prediction of new members of regulons. *Genome Res.* **11**, 566–584 (2001).
232. Rigali, S., Anderssen, S., Naômé, A. & van Wezel, G. P. Cracking the regulatory code of biosynthetic gene clusters as a strategy for natural product discovery. *Biochem. Pharmacol.* **153**, 24–34 (2018).
233. Carrión, V. J. *et al.* Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. *Science* **366**, 606–612 (2019).
234. 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.* **112**, 12580–12585 (2015).
235. Mohimani, H. *et al.* Dereplication of peptidic natural products through database

- search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
236. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* **9**, 4035 (2018).
 237. Ernst, M. *et al.* MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites* **9**, (2019).
 238. Wilson, M. C. & Piel, J. Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. *Chem. Biol.* **20**, 636–647 (2013).
 239. Smanski, M. J. *et al.* Synthetic biology to access and expand nature’s chemical diversity. *Nat. Rev. Microbiol.* **14**, 135–149 (2016).
 240. Sugimoto, Y. *et al.* A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science* **366**, (2019).
 241. Smanski, M. J. *et al.* Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **32**, 1241–1249 (2014).
 242. Shomar, H. *et al.* Metabolic engineering of a carbapenem antibiotic synthesis pathway in *Escherichia coli*. *Nat. Chem. Biol.* **14**, 794–800 (2018).
 243. Carroll, L. M. *et al.* Accurate de novo identification of biosynthetic gene clusters with GECCO. (2021).
 244. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, E110 (2019).
 245. Agrawal, P., Khater, S., Gupta, M., Sain, N. & Mohanty, D. RiPPMiner: A bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res.* **45**, W80–W88 (2017).
 246. Van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
 247. Kim, S. *et al.* PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
 248. Paoli, L. *et al.* *Biosynthetic potential of the global ocean microbiome*. *Nature* vol. 607 (Springer US, 2022).
 249. Nayfach, S. *et al.* A genomic catalog of Earth’s microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
 250. Izoré, T. *et al.* Structures of a non-ribosomal peptide synthetase condensation domain

- suggest the basis of substrate selectivity. *Nat. Commun.* **12**, 1–14 (2021).
251. Gavriilidou, A. *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat. Microbiol.* **7**, 726–735 (2022).
252. Kelly, R. & Kidd, R. Editorial: ChemSpider-a tool for Natural Products research. *Nat. Prod. Rep.* **32**, 1163–1164 (2015).
253. Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D. & ... *The LOTUS initiative for open natural products research: knowledge management through Wikidata.* *BioRxiv* (2021).
254. Terlouw, B. R., Vromans, S. P. J. M. & Medema, M. H. PIKACHU: a Python-based informatics kit for analysing chemical units. *J. Cheminform.* **14**, 1–17 (2022).
255. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
256. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *J. Mol. Biol.* **368**, 1500–1517 (2007).
257. Miller, B. R. & M., G. A. Structural Biology of Non-Ribosomal Peptide Synthetases. 3–29 (2017) doi:10.1007/978-1-4939-3375-4.
258. Nieuwkoop, T., Finger-Bou, M., van der Oost, J. & Claassens, N. J. The Ongoing Quest to Crack the Genetic Code for Protein Production. *Mol. Cell* **80**, 193–209 (2020).
259. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science (80-.)*. **324**, 255–259 (2009).
260. Boël, G. *et al.* Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* **529**, 358–363 (2016).
261. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol.* **36**, 1005–1015 (2018).
262. Harigaya, Y. & Parker, R. The link between adjacent codon pairs and mRNA stability. *BMC Genomics* **18**, 364 (2017).
263. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science (80-.)*. **342**, 475–480 (2013).
264. Looman, A. C. *et al.* Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *Escherichia coli*. *EMBO J.* **6**, 2489–2492 (1987).

265. Stenström, C. M., Jin, H., Major, L. L., Tate, W. P. & Isaksson, L. A. Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* **263**, 273–284 (2001).
266. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* **43**, 13–28 (2015).
267. Quax, T. E. F. *et al.* Differential translation tunes uneven production of operon-encoded proteins. *Cell Rep.* **4**, 938–944 (2013).
268. Sharp, P. M. & Li, W.-H. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
269. Welch, M. *et al.* Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* **4**, e7002 (2009).
270. Isacchi, A., Sarmientos, P., Lorenzetti, R. & Soria, M. Mature apolipoprotein AI and its precursor proApoAI: influence of the sequence at the 5' end of the gene on the efficiency of expression in *Escherichia coli*. *Gene* **81**, 129–137 (1989).
271. Kelsic, E. D. *et al.* RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Syst.* **3**, 563–571.e6 (2016).
272. Frumkin, I. *et al.* Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell* **65**, 142–153 (2017).
273. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
274. Hanson, G. & Collier, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* (2017) doi:10.1038/nrm.2017.91.
275. Radhakrishnan, A. *et al.* The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167**, 122–132.e9 (2016).
276. Parret, A. H., Besir, H. & Meijers, R. Critical reflections on synthetic gene design for recombinant protein expression. *Curr. Opin. Struct. Biol.* **38**, 155–162 (2016).
277. Höllerer, S. *et al.* Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nat. Commun.* **11**, 3551 (2020).
278. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022).
279. Nikolados, E.-M., Wongprommoon, A., Aodha, O. Mac, Cambray, G. & Oyarzún, D. A.

- Accuracy and data efficiency in deep learning models of protein expression. *Nat. Commun.* **13**, 7755 (2022).
280. Potapov, V. *et al.* Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth. Biol.* **7**, 2665–2674 (2018).
 281. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
 282. Cetnar, D. P. & Salis, H. M. Systematic Quantification of Sequence and Structural Determinants Controlling mRNA stability in Bacterial Operons. *ACS Synth. Biol.* **10**, 318–332 (2021).
 283. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
 284. Nieuwkoop, T., Claassens, N. J. & van der Oost, J. Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. *Microb. Biotechnol.* **12**, 173–179 (2019).
 285. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
 286. Kimchi-Sarfaty, C. *et al.* A ‘silent’ polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525–528 (2007).
 287. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).
 288. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, (2011).
 289. Sabi, R., Volvovitch Daniel, R. & Tuller, T. stAlcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* **33**, 589–591 (2017).
 290. Mirzadeh, K. *et al.* Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector-Coding Sequence Junction. *ACS Synth. Biol.* **4**, 959–965 (2015).
 291. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).
 292. Dittmann, J., Wenger, R. M., Kleinkauf, H. & Lawen, A. Mechanism of cyclosporin A biosynthesis. Evidence for synthesis via a single linear undecapeptide precursor. *J.*

Biol. Chem. **269**, 2841–2846 (1994).

293. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
294. Ladner, C. C. & Williams, G. J. Harnessing natural product assembly lines: structure, promiscuity, and engineering. *J. Ind. Microbiol. Biotechnol.* **43**, 371–387 (2016).
295. Zhu, M., Wang, L. & He, J. Chemical Diversification Based on Substrate Promiscuity of a Standalone Adenylation Domain in a Reconstituted NRPS System. *ACS Chem. Biol.* **14**, 256–265 (2019).
296. Rothe, M. Exploring epoxyketone synthases and their biosynthetic potential. (Warwick, 2022).
297. Phelan, V. V., Du, Y., McLean, J. A. & Bachmann, B. O. Adenylation enzyme characterization using gamma -(18)O(4)-ATP pyrophosphate exchange. *Chem. Biol.* **16**, 473–478 (2009).
298. Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. *Science* **291**, 1790–1792 (2001).
299. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
300. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
301. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
302. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).
303. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
304. Tooming-Klunderud, A., Rohrlack, T., Shalchian-Tabrizi, K., Kristensen, T. & Jakobsen, K. S. Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon from *Microcystis*: implications for evolution of cyanopeptolins. *Microbiology* **153**, 1382–1393 (2007).
305. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

306. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
307. Schrödinger, L. L. C. & DeLano, W. PyMOL.
308. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
309. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
310. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).
311. Akdel, M., Durairaj, J., de Ridder, D. & van Dijk, A. D. J. Caretta - A multiple protein structure alignment and feature extraction suite. *Comput. Struct. Biotechnol. J.* **18**, 981–992 (2020).
312. Wold, S. *et al.* Principal property values for six non-natural amino acids and their application to a structure–activity relationship for oxytocin peptide analogues. *Can. J. Chem.* **65**, 1814–1820 (1987).
313. Neumaier, A., Huyer, W. & Bornberg-Bauer, E. Hydrophobicity Analysis of Amino Acids.
314. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253–266 (1999).
315. Fauchère, J. L., Charton, M., Kier, L. B., Verloop, A. & Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (1988).
316. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
317. Radzicka, A. & Wolfenden, R. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**, (1988).
318. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).
319. Chou, P. Y. & Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 45–148 (1978).
320. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202-5 (2008).

321. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
322. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a ‘Kneedle’ in a Haystack: Detecting Knee Points in System Behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* 166–171 (2011). doi:10.1109/ICDCSW.2011.20.
323. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
324. Sechidis, K., Tsoumakas, G. & Vlahavas, I. On the stratification of multi-label data. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **6913 LNAI**, 145–158 (2011).
325. Trent, B. Iterstrat: a package for the stratification of multi-label data. (2022).
326. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
327. Forli, S. Meeko: preparation of small molecules for AutoDock. (2022).
328. Jérôme, E., Santos-Martins, D., Tillack, A. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **XXXX**, (2021).
329. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
330. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
331. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
332. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
333. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385 (2003).
334. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
335. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview

- Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
336. Velankar, S. *et al.* PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **38**, D308–17 (2010).
 337. Timonina, D., Sharapova, Y., Švedas, V. & Suplatov, D. Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and conformational plasticity in protein superfamilies. *Comput. Struct. Biotechnol. J.* **19**, 1302–1311 (2021).
 338. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621–1637 (2009).
 339. Schwarzer, D., Finking, R. & Marahiel, M. Nonribosomal Peptides: From Genes to Products. *Nat. Prod. Rep.* **20**, 275–287 (2003).
 340. Python Software Foundation. Python Language Reference, version 3.6. (1995).
 341. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
 342. Hossain, S. Visualization of Bioinformatics Data with Dash Bio. *Proc. 18th Python Sci. Conf.* 126–133 (2019) doi:10.25080/majora-7ddc1dd1-012.
 343. Dash-Bio: Dash components for bioinformatics.
 344. Dash-Cytoscape: A component library for Dash aimed at facilitating network visualization in Python, wrapped around Cytoscape.js.
 345. Dash-extensions: Extensions for Plotly Dash.
 346. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **27**, 1575–1577 (2011).
 347. Dash-bio-utils: Simple parsing tools that supplement dash-bio.
 348. Durairaj, J., Akdel, M., Ridder, D. de & Dijk, A. D. J. van. Fast and adaptive protein structure representations for machine learning. *bioRxiv* 2021.04.07.438777 (2021) doi:10.1101/2021.04.07.438777.
 349. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.* **68**, 365–369 (2019).
 350. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases.

Phytochemistry **158**, 157–165 (2019).

351. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, e1008197 (2021).
352. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science* **277**, 1815–1820 (1997).
353. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
354. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *J. Cheminform.* **13**, 1–13 (2021).
355. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 1–9 (2020).
356. Volkamer, A., Kuhn, D., Rippmann, F. & Rarey, M. Dogsitescorer: A web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* **28**, 2074–2075 (2012).
357. Alvarsson, J. *et al.* Large-scale ligand-based predictive modelling using support vector machines. *J. Cheminform.* **8**, 1–9 (2016).
358. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **9**, 1–19 (2017).
359. Morris, J. & Jiao, D. ChemViz2: Cheminformatics App for Cytoscape. <http://www.rbvi.ucsf.edu/cytoscape/chemViz2/> (2016).
360. Beisken, S., Meinl, T., Wiswedel, B., Figueiredo, L. F. De & Berthold, M. KNIME-CDK : Workflow-driven cheminformatics. *BMC Bioinformatics* **14**, 2–5 (2013).
361. Cass, S. Top Programming Languages 2021, IEEE Spectrum. <https://spectrum.ieee.org/top-programming-language> (2021).
362. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
363. Miles, L. H. Cycle detection. <https://github.com/qpwo/python-simple-cycles> (2019).
364. Johnson, D. Finding all the elementary cycles of a digraph. *SIAM J. Comput.* **4**, 77–84

- (1975).
365. Hückel, E. Quantentheoretische Beiträge zum Benzolproblem - I. Die Elektronenkonfiguration des Benzols und verwandter Verbindungen. *Zeitschrift für Phys.* **70**, 204–286 (1931).
366. Yorkyer. Python implementation of Edmonds' Blossom Algorithm. <https://github.com/yorkyer/edmonds-blossom> (2020).
367. Edmonds, J. Paths, trees, and flowers. *Can. J. Math.* **17**, 449–467 (1965).
368. Probst, D. & Reymond, J. L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J. Chem. Inf. Model.* **58**, 1–7 (2018).
369. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).
370. Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. ACM* **23**, 31–42 (1976).
371. Hutchinson, R. C. Polyketide and non-ribosomal peptide synthases: Falling together by coming apart. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3010–3012 (2003).
372. Boettger, D. & Hertweck, C. Molecular Diversity Sculpted by Fungal PKS-NRPS Hybrids. *ChemBioChem* **14**, 28–42 (2013).
373. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem. Rev.* **119**, 12524–12547 (2019).
374. Waldron, C. *et al.* Cloning and analysis of the spinosad biosynthetic gene cluster of *Saccharopolyspora spinosa*. *Chem. Biol.* **8**, 487–499 (2001).
375. Staunton, J. & Wilkinson, B. Biosynthesis of erythromycin and rapamycin. *Chem. Rev.* **97**, 2611–2629 (1997).
376. Du, L., Sánchez, C. & Shen, B. Hybrid peptide-polyketide natural products: Biosynthesis and prospects toward engineering novel molecules. *Metab. Eng.* **3**, 78–95 (2001).
377. Walsh, C. T. *et al.* Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines. *Curr. Opin. Chem. Biol.* **5**, 525–534 (2001).
378. Olano, C., Méndez, C. & Salas, J. A. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. *Nat. Prod. Rep.* **27**, 571–616 (2010).

379. Hur, G. H., Vickery, C. R. & Burkart, M. D. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat. Prod. Rep.* **29**, 1074–1098 (2012).
380. Weissman, K. J. Introduction to Polyketide Biosynthesis. *Methods Enzymol.* **459**, 3–16 (2009).
381. Du, L. & Lou, L. PKS and NRPS release mechanisms. *Nat. Prod. Rep.* **27**, 255–278 (2010).
382. Walsh, C. T., O'Brien, R. V. & Khosla, C. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angew. Chemie - Int. Ed.* **52**, 7098–7124 (2013).
383. Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).
384. Bouhired, S. M. *et al.* Biosynthesis of phenylannolone A, a multidrug resistance reversal agent from the halotolerant myxobacterium *Nannocystis pusilla* B150. *ChemBioChem* **15**, 757–765 (2014).
385. Pan, G. *et al.* Discovery of the leinamycin family of natural products by mining actinobacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E11131–E11140 (2017).
386. Höfer, I. *et al.* Insights into the biosynthesis of hormaomycin, an exceptionally complex bacterial signaling metabolite. *Chem. Biol.* **18**, 381–391 (2011).
387. Tambadou, F. *et al.* Characterization of the colistin (polymyxin E1 and E2) biosynthetic gene cluster. *Arch. Microbiol.* **197**, 521–532 (2015).
388. Van Santen, J. A. *et al.* The Natural Products Atlas 2.0: A database of microbially-derived natural products. *Nucleic Acids Res.* **50**, D1317–D1323 (2022).
389. Wang, Q. *et al.* Abyssomicins from the South China Sea deep-sea sediment *Verrucosipora* sp.: natural thioether Michael addition adducts as antitubercular prodrugs. *Angew. Chem. Int. Ed. Engl.* **52**, 1231–1234 (2013).
390. Mohr, J. F. *et al.* Frankobactin Metallophores Produced by Nitrogen-Fixing *Frankia* Actinobacteria Function in Toxic Metal Sequestration. *J. Nat. Prod.* **84**, 1216–1225 (2021).
391. Vestola, J. *et al.* Hassallidins, antifungal glycolipopeptides, are widespread among cyanobacteria and are the end-product of a nonribosomal pathway. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1909–17 (2014).
392. Hover, B. M. *et al.* Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive

- pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
393. Vittorio, T. *et al.* Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics. *mSystems* **6**, 10.1128/msystems.01116-20 (2021).
394. Kersten, R. D. *et al.* A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).
395. Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12**, 1007–1014 (2016).
396. Kim, M.-S. *et al.* Unprecedented Noncanonical Features of the Nonlinear Nonribosomal Peptide Synthetase Assembly Line for WS9326A Biosynthesis. *Angew. Chem. Int. Ed. Engl.* **60**, 19766–19773 (2021).
397. Yu, D., Xu, F., Zhang, S. & Zhan, J. Decoding and reprogramming fungal iterative nonribosomal peptide synthetases. *Nat. Commun.* **8**, 15349 (2017).
398. Bernhardt, M., Berman, S., Zechel, D. & Bechthold, A. Role of Two Exceptional trans Adenylation Domains and MbtH-like Proteins in the Biosynthesis of the Nonribosomal Peptide WS9324A from *Streptomyces calvus* ATCC 13382. *Chembiochem* **21**, 2659–2666 (2020).
399. Hackl, S. & Bechthold, A. The Gene *bldA*, a regulator of morphological differentiation and antibiotic production in streptomyces. *Arch. Pharm. (Weinheim)*. **348**, 455–462 (2015).
400. Chater, K. F. & Chandra, G. The use of the rare UUA codon to define “Expression Space” for genes involved in secondary metabolism, development and environmental adaptation in *Streptomyces*. *J. Microbiol.* **46**, 1–11 (2008).
401. Li, Y. F. *et al.* Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genet. Biol.* **89**, 18–28 (2016).
402. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
403. Alexander, D. C. *et al.* Production of novel lipopeptide antibiotics related to A54145 by *Streptomyces fradiae* mutants blocked in biosynthesis of modified amino acids and assignment of *lptJ*, *lptK* and *lptL* gene functions. *J. Antibiot. (Tokyo)*. **64**, 79–87 (2011).
404. Wang, G. *et al.* CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nat. Microbiol.* **4**, 2498–2510 (2019).

405. Akpa, E. *et al.* Influence of culture conditions on lipopeptide production by *Bacillus subtilis*. *Appl. Biochem. Biotechnol.* **91**, 551–561 (2001).
406. Heemstra, J. R. J., Walsh, C. T. & Sattely, E. S. Enzymatic tailoring of ornithine in the biosynthesis of the *Rhizobium* cyclic trihydroxamate siderophore vicibactin. *J. Am. Chem. Soc.* **131**, 15317–15329 (2009).
407. Rouhiainen, L. *et al.* Genes coding for hepatotoxic heptapeptides (microcystins) in the cyanobacterium *Anabaena* strain 90. *Appl. Environ. Microbiol.* **70**, 686–692 (2004).
408. Lohman, J. R. *et al.* Cloning and sequencing of the kedarcidin biosynthetic gene cluster from *Streptoalloteichus* sp. ATCC 53650 revealing new insights into biosynthesis of the enediyne family of antitumor antibiotics. *Mol. Biosyst.* **9**, 478–491 (2013).
409. Taylor, S. D. & Palmer, M. The action mechanism of daptomycin. *Bioorg. Med. Chem.* **24**, 6253–6268 (2016).
410. Reimer, D., Pos, K. M., Thines, M., Grün, P. & Bode, H. B. A natural prodrug activation mechanism in nonribosomal peptide synthesis. *Nat. Chem. Biol.* **7**, 888–890 (2011).
411. Balazs, A. Y. S. *et al.* Free Ligand 1D NMR Conformational Signatures To Enhance Structure Based Drug Design of a Mcl-1 Inhibitor (AZD5991) and Other Synthetic Macrocycles. *J. Med. Chem.* **62**, 9418–9437 (2019).
412. Shortridge, M. D., Hage, D. S., Harbison, G. S. & Powers, R. Estimating protein-ligand binding affinity using high-throughput screening by NMR. *J. Comb. Chem.* **10**, 948–958 (2008).
413. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
414. Tahlan, K. & Jensen, S. E. Origins of the β -lactam rings in natural products. *J. Antibiot. (Tokyo)*. **66**, 401–410 (2013).
415. Smith, P. W. *et al.* Pharmacokinetics of β -Lactam Antibiotics: Clues from the Past To Help Discover Long-Acting Oral Drugs in the Future. *ACS Infect. Dis.* **4**, 1439–1447 (2018).
416. Xie, L. & Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proc. Natl. Acad. Sci.* **105**, 5441–5446 (2008).

Summary

Evolution has equipped microbes with molecular technologies that help ensure their survival in often hostile environments. These technologies include specialised metabolites, also called natural products, which fulfil a wide range of functions, including nutrient acquisition, protection against environmental stressors, communication, and microbial warfare. Examples include antibiotics, pesticides, antimalarials and anticancer drugs. This wealth of functional diversity provides a natural well of chemical potential which we can draw from for pharmaceutical and other societal applications.

Natural product assembly is complicated: it often requires the coordination of multiple molecular machines called proteins to put together different subunits in an assembly-line-like fashion. This thesis describes the development of various computational methods that help us find and understand these biological pipelines, with as goal to discover natural products with interesting properties. In **chapter 2**, we discuss which microbes produce natural products and why, and suggest how we can use this knowledge to prioritise which microbes to research.

Once we have identified a microbial community of interest, we continue by finding and analysing biosynthetic gene clusters: instruction manuals for the production of natural products written in the language of DNA. In **chapter 3**, we describe the development of a database to store experimentally characterised biosynthetic gene clusters from natural-product-producing microbes, such that the research community can easily compare and learn from them. It is not only interesting to learn what an instruction booklet encodes, but also how efficient the instructions are. We explore this in **chapter 4**, where we correlate DNA code to the production levels of the protein machinery that is subsequently produced.

Next, we can go one step further and try to computationally decipher specific clues within these instruction manuals and relate them to building blocks that are incorporated into natural products. In **chapter 5**, we use machine learning to do this in the context of non-ribosomal peptides: a potent class of natural products which boasts an immense chemical and functional diversity. This chapter demonstrates that high-quality training data are the most important aspect of any machine learning algorithm.

Many different data types are relevant for natural products research: think of the evolutionary origin of the producing organism, the DNA sequence of a biosynthetic gene cluster, the molecular structures of building blocks, and the protein sequence and structure of the molecular machines that synthesise natural products. It can be challenging to analyse all these data types in tandem, especially for non-bioinformaticians. To this purpose, we developed a web portal that automatically and interactively integrates all these data types in **chapter 6**.

Finally, visual aids are essential for understanding key concepts in biology, including natural products research. Therefore, we developed software that can automatically visualise full natural product pipelines given annotated biosynthetic gene cluster data. The chemical engine required for this software is described in **chapter 7**, and the natural product drawing suite built upon it is introduced in **chapter 8**.

Acknowledgements

While this thesis bears only one name on its cover, it is the culmination of the efforts of many. Some directly contributed to the research chapters of this thesis, while others shared invaluable insights, sparked creative lines of thought, or provided much-needed distraction and support during the challenging process of finishing a PhD. I would like to use this space to thank a few people in particular without whom this book would not exist.

First, there is Sandra Smit: the person who, as my MSc thesis supervisor, reignited my passion for science, and without whom I would never have contemplated pursuing a PhD in the first place. Thank you for your patience with me and your belief in me during a turbulent yet defining time in my life.

Then there is my promotor Marnix Medema, who has continually impressed and inspired me in his role as my daily supervisor. You have established a rare balance of scientific involvement, collaborative prowess, approachability, mutual trust, and leadership, your group standing firm as my safe haven in an oftentimes stormy scientific landscape. Thank you so much for giving me the chance to develop myself as a researcher in such a stimulating environment; for all the hard and soft skills you taught me; the support you have given me throughout the years; and for the incredible scientific input that made my manuscripts, presentations, and posters that much better. Also, many thanks to my other promotor, Dick de Ridder, for giving me the opportunity to pursue a PhD in the Bioinformatics Group, which despite the many different research lines pursued within feels like a unified family. Also, thank you for all your scientific insights over the years, particularly regarding machine learning. I greatly enjoyed working together on Chapter 4 of this thesis and hope we can collaborate on a Mew-two someday.

Next, there are a lot of students and colleagues in the Bioinformatics Group, which went through many iterations since I first arrived. Many of you provided valuable contributions during work discussions and brainstorming sessions, and you all made the work feel lighter and more fun with coffee and lunch breaks, dinners, and Dungeons and Dragons sessions. Many thanks in particular to MSc students Sophie Vromans and Leron Kok, and BSc student Nico Louwen, whose excellent thesis work led to high-quality manuscripts and thesis chapters; to my wonderful paranympths Lotte Pronk and Hannah Augustijn, for their lovely friendship and for helping me navigate the stress that comes with finishing a PhD and organising a wedding; and to Frida Biermann, Catarina Loureiro, David Meijer, Serina Robinson, Mehmet Akdel, and Jay Durairaj for being wonderful collaborators as well as life-long friends. And there are so many others whose company I enjoyed thoroughly; far too many to name here.

There are also many collaborators outside of the Bioinformatics Group that contributed to this work. Most notable are Marc Chevette and Greg Challis, who made key contributions to Chapter 5, the core chapter of this thesis, devoting hours of their time to online discussions and a research exchange at the University of Warwick. Also special thanks to Shanshan Zhou, Matthew Jenner, Lona Alkhalaf, and Chris Fage from the University of Warwick: without your able hands, kind help, and expertise, my wet-lab experiments would never have been

possible. Then I would like to thank Thijs Nieuwkoop and Nico Claassens for an exciting scientific trip outside of the world of natural products; and Doris van Bergeijk, Karol Al Ayed, Nataliia Machushynets, and Gilles van Wezel, my collaborators and co-authors from the University of Leiden, the latter two of which will become a direct colleague and my new PI, respectively, in the near future. I really enjoyed hunting for antibiotics together – here is to many more in the future!

Then, there are many outside of the academic bubble who helped me throughout the past five years in more ways than I could count. My family have supported my scientific pursuits from a very young age, starting with my mother and father who have always stimulated logical thinking through play; and my two grandfathers, who as engineer for the European Space Agency and Physics PhD, respectively, inspired me by exemplifying two of many exciting scientific walks of life.

Finally, and most importantly, there is my wonderful husband Anthony Terlouw-Jacquet, who met me during one of the most stressful periods in my life and decided to marry me anyway. With your support, love, and kindness, I can face anything.

Education Statement of the Graduate School Experimental Plant Sciences



Issued to: Barbara Terlouw
Date: 26 September 2023
Group: Bioinformatics
University: Wageningen University & Research

1) Start-Up Phase	<u>date</u>	<u>cp</u>
► First presentation of your project An algorithm-based approach for discovering novel lipopeptide antibiotics	17 Jan 2019	1.5
► Writing or rewriting a project proposal ► Writing a review or book chapter van Bergeijk, D.A.*, Terlouw, B.R.* <i>et al.</i> Ecology and genomics of Actinobacteria: new concepts for natural product discovery. Nat Rev Microbiol 18, 546–558 (2020). https://doi.org/10.1038/s41579-020-0379-y	20 Apr 2020	3.0
► MSc courses FTE-35306 Machine Learning (WUR)	Feb 2019	6.0
<i>Subtotal Start-Up Phase</i>		10.5

2) Scientific Exposure	<u>date</u>	<u>cp</u>
► EPS PhD days Annual EPS PhD days GET2GETHER, Soest, The Netherlands Annual EPS PhD days GET2GETHER, Soest, The Netherlands	10-11 Feb 2020 03-04 May 2022	0.6 0.6
► EPS theme symposia EPS Theme 3 Symposium 'Metabolism and Adaptation', Wageningen, The Netherlands EPS Theme 4 Symposium 'Genome Biology', Online	05 Nov 2021 17 Jan 2022	0.3 0.3
► Lunteren Days and other national platforms Annual Dutch Bioinformatics and Systems Biology (BioSB) Conference, including BioSB PhD retreat, Online Annual Meeting Experimental Plant Sciences (EPS), Lunteren, The Netherlands	14-16 Jun 2021 11-12 Apr 2022	0.8 0.6
► Seminars (series), workshops and symposia B-Wise Seminar - Gurnoor Singh & Janani Durairaj B-Wise Seminar - Christian Gilissen & Mohammad Alanjari B-Wise Seminar - Rachel Cavill & Mehmet Akdel B-Wise Seminar - Rik van Rosmalen & Sevgin Demirci B-Wise Seminar - Gerben Hermes & Pariya Behrouzi B-Wise Seminar - Martijn Huijnen & Mark Sterken B-Wise Seminar - Sven Warris & Vittorio Tracanna B-Wise Seminar - Simon van Heeringen & Chiara Bortoluzzi B-Wise Seminar - Veronika Laine & Raul Wijffes B-Wise Seminar - Eliana Papoutsoglou & Roeland Voorrips B-Wise Seminar - Jingyuan Fu & Catarina Sales e Santos Loureiro B-Wise Seminar - Simon Rogers & Barbara Terlouw (as speaker) B-Wise Seminar - Mario Calus & Eef Jonkheer B-Wise Seminar - Carlos de Lannoy & Chaozhi Zheng B-Wise Seminar - Age Smilde & Cristina Furlan B-Wise Seminar - Nicola Segata B-Wise Seminar - Saskia Oosterbroek & Diana Hendrickx B-Wise Seminar - Marjolijn Hooykaas & Barbara Terlouw (as speaker)	04 Sep 2018 02 Oct 2018 04 Dec 2018 08 Jan 2019 05 Feb 2019 05 Mar 2019 02 Apr 2019 07 May 2019 03 Sep 2019 01 Oct 2019 05 Nov 2019 03 Dec 2019 07 Jan 2020 04 Feb 2020 03 Mar 2020 13 Sep 2021 14 Jun 2022 04 Oct 2022	0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.1 0.2 0.2 0.2 0.1 0.2 0.1
► Seminar plus ► International symposia and congresses Copenhagen Bioscience Conference - Natural Products: Discovery, Biosynthesis and Application, Hillerød, Denmark Lorentz Workshop - AI and Natural Products, Leiden, Netherlands Symposium Natural Products (Meta) Genome Mining, Copenhagen, Denmark International Conference of Chemical Structures, Noordwijkerhout, The Netherlands	05-09 May 2019 27 Sep - 01 Oct 2021 30-31 May 2022 12-16 Jun 2022	1.3 1.5 0.6 1.3
► Presentations Poster: Computational method to predict lipopeptide structure from a gene cluster sequence, Copenhagen Bioscience Conference, Hillerød, Denmark Poster: PIKACHU and RAICHU: new tools for automating modular PKS and NRPS biosynthesis visualisation, International Conference of Chemical Structures, Noordwijkerhout, The Netherlands Talk: From 1D to 3D: using structural features in NRPS A-domain substrate prediction, B-Wise Presentation, Wageningen, The Netherlands Talk: Deciphering the nonribosomal code: the language of antibiotics and other natural products, Annual EPS Meeting, Lunteren, The Netherlands Talk: PIKACHU and RAICHU, EPS GET2GETHER, Soest, The Netherlands Talk: Deciphering the nonribosomal code: the language of antibiotics and other natural products, B-Wise Presentation, Wageningen, The Netherlands	07 May 2019 13 Jun 2022 03 Dec 2019 11 Apr 2022 04 May 2022 04 Oct 2022	1.0 1.0 1.0 1.0 1.0 1.0
► IAB interview ► Excursions		
<i>Subtotal Scientific Exposure</i>		17.2

3) In-Depth Studies	<u>date</u>	<u>cp</u>
► Advanced scientific courses & workshops EMBL-EBI course Structural Bioinformatics, Hinxton, United Kingdom Netherlands eScience Center workshop Introduction to Deep Learning, Online	16-20 Sep 2019 05-07 Jul 2021	1.4 0.6
► Journal club Bioinformatics group, every 2 weeks	2018-2021	3.0
► Individual research training Laboratory training at University of Warwick	Nov 2020-May 2021	3.0
<i>Subtotal In-Depth Studies</i>		8.0

4) Personal Development		<u>date</u>	<u>cp</u>
► General skill training courses			
WGS PhD Competence Assessment, Wageningen, The Netherlands		03-04 Apr 2019	0.3
EPS Introduction Course, Wageningen, The Netherlands		29 Oct 2019	0.3
WGS course Career Perspectives, Wageningen, The Netherlands		01-29 Nov 2021	1.6
WGS course Supervising BSc and MSc thesis students, Wageningen, The Netherlands		28-29 Mar 2022	0.6
WGS workshop Scientific Publishing, Wageningen, The Netherlands		12 May 2022	0.3
► Organisation of meetings, PhD courses or outreach activities			
PhD representative Bioinformatics Staff meetings		Aug 2020-Jun 2022	1.5
► Membership of EPS PhD Council			

Subtotal Personal Development

4.6

TOTAL NUMBER OF CREDIT POINTS*		40.3
Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.		
* A credit represents a normative study load of 28 hours of study.		

The research described in this thesis was financially supported by the Dutch Research Council NWO (TTW 16440) and the Graduate School Experimental Plant Sciences.

Cover design by Barbara Terlouw

Printed by proefschriftmaken.nl || <https://www.proefschriftmaken.nl>

