`synthaser`: a CD-Search enabled Python toolkit for analysing domain architecture of fungal secondary metabolite megasynth(et)ases

Cameron L.M. Gilchrist*[1], Yit-Heng Chooi*[1]

[1]School of Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, 6009

E-mail: cameron.gilchrist@research.uwa.edu.au, yitheng.chooi@uwa.edu.au

Abstract

Background: Fungi are prolific producers of secondary metabolites (SMs), which are bioactive small molecules with important applications in medicine, agriculture and other industries. The backbones of a large proportion of fungal SMs are generated through the action of large, multi-domain megasynth(et)ases such as polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs). The structure of these backbones is determined by the domain architecture of the corresponding megasynth(et)ase, and thus accurate annotation and classification of these architectures is an important step in linking SMs to their biosynthetic origins in the genome.

Results: Here we report `synthaser`, a Python package leveraging the NCBI's conserved domain search tool for remote prediction and classification of fungal megasynth(et)ase domain architectures. `synthaser` is capable of batch sequence analysis, and produces rich textual output and interactive visualisations which allow for quick assessment of the megasynth(et)ase diversity of a fungal genome. `synthaser` uses a hierarchical rule-based classification system, which can be extensively customised by the user through a web application (`http://gamcil.github.io/synthaser`). We show that `synthaser` provides more accurate domain architecture predictions than comparable tools which rely on curated profile hidden Markov model (pHMM)-based approaches; the utilisation of the NCBI conserved domain database also allows for significantly greater flexibility compared to pHMM approaches. In addition, we demonstrate how `synthaser` can be applied to large scale genome mining pipelines through the construction of an *Aspergillus* PKS similarity network.

Conclusions: `synthaser` is an easy to use tool that represents a significant upgrade to previous domain architecture analysis tools. It is freely available under a MIT license from PyPI (`https://pypi.org/project/synthaser`) and GitHub (`https://github.com/gamcil/synthaser`).

Keywords: secondary metabolism, domain architecture, polyketide synthase, nonribosomal peptide synthetase, bioinformatics, software

1 Introduction

Domains are distinct functional and structural units that serve as the evolutionary building blocks of proteins. The majority of proteins found across all kingdoms of life consist of multiple functional domains, with the growth in number of multi-domain protein families far outpacing that of single domain protein families (Vogel et al., 2004; Levitt, 2009). Multi-domain proteins predominantly arise from the incorporation of new domains at the N or C terminus by genetic events such as gene fusion, fission, duplication and exon shuffling (Marsh and Teichmann, 2010; Pasek et al., 2006; Moore et al., 2008). Extensive domain rearrangement over time has led to the diversification of existing proteins, as well as the emergence of novel protein families (Bornberg-Bauer and Albà, 2013). Through this process, domains are placed into new molecular contexts where, via their interactions with different combinations of domains, novel functionality can be birthed (Bashton and Chothia, 2007; Jin et al., 2009). In eukaryotes, many multi-domain proteins have evolved from separate single domain proteins catalysing successive steps of biological pathways in prokaryotes, resulting in improved flux and stability of the pathway (Ostermeier and Benkovic, 2001). Thus, a study of functional domains, as well as a broader analysis of the domain architectures of proteins in which they are found, can be a fruitful approach for identifying novel functionality.

A good case study for the evolution of multi-domain proteins with domain architectures and functions can be found in the biosynthesis of secondary metabolites, which are bioactive small molecules with important applications in medicine, agriculture and other industries (Keller et al., 2005; Newman and Cragg, 2016). Secondary metabolites are produced by many microorganisms and plants, but are particularly abundant in filamentous fungi (Keller, 2015). Indeed, recent genomic work has made obvious the extent of the biochemical arsenal encoded by microbial, and particularly fungal, genomes (Vesth et al., 2018; de Vries et al., 2017). The biosynthesis of these compounds is orchestrated primarily through the action of large, multi-domain megasynthases; polyketides are synthesized by polyketide synthases (PKSs) and nonribosomal peptides by nonribosomal peptide synthetases (NRPSs). These multi-domain megasynthases generate the chemical backbones of the compound, which are then modified by 'tailoring' enzymes, typically encoded by genes neighbouring the megasynthases in the genome, in what are referred to as biosynthetic gene clusters (BGCs). Much of the work done by natural product researchers in past decades has been focused on the hunt for, and characterisation of, novel BGCs, in hopes of finding the next great drug lead.

Megasynthases can be easily identified by the presence of key functional domains. For example, PKSs typically contain a β-ketoacyl synthase (KS) domain, which is responsible for building the carbon

backphone of polyketides through repetitive condensation of short-chain carboxylic acids (Chooi and Tang, 2012). There are also deeper levels of classification based on the presence of other functional domains. Iterative PKSs from fungi, for instance, can be classified as highly-, partially- or non-reducing given the absence or presence of domains that catalyse reduction reactions of the polyketide chain. A highly-reducing PKS will synthesize a reduced polyketide chain, whereas a non-reducing PKS would synthesize an unreduced chain. Other domains present within the megasynthase also affect the synthesized product. For example, the PKSs involved in lovastatin biosynthesis, LovF and LovB, both contain methyltransferase domains which add methyl groups during synthesis of the polyketide product (Hutchinson et al., 2000).

This link between domain architecture and compound has several useful applications. Firstly, given some isolated metabolite, one can narrow down to the synthases likely responsible for its production by looking for a domain architecture that matches the structure of that metabolite. This is one of the first steps when taking a 'retro-biosynthetic' approach to identifying a BGC (Cacho et al., 2015). Indeed, we have used this approach to identify the megasynthases encoding numerous compounds isolated from Australian fungi (Lacey et al., 2019; Li et al., 2019, 2020a, 2021). Inversely, we can predict that synthases with unique domain architectures could potentially produce unique compounds. Previously we outlined genome mining strategies for the discovery of novel secondary metabolites (Gilchrist et al., 2018). One strategy is to prioritise BGCs which have partial similarity to known BGCs, in hopes of finding find new analogues of known bioactive compounds; another is to prioritise completely unique BGCs in order to find novel compounds. In either case, analysis of the domain architectures of secondary metabolite megasynthases plays a key role.

There are currently many databases dedicated to the analysis and functional classification of domains. Pfam (Mistry et al., 2021), SMART (Letunic et al., 2021) and PROSITE (Sigrist et al., 2013) are three such databases, each storing information about domain family structure and function. There are also larger resources such as the Conserved Domain Database (CDD; Lu et al., 2020) from the National Center for Biotechnology Information (NCBI), or the InterPro database (Blum et al., 2021), which integrate many of the smaller domain databases. The CDD contains over 50,000 curated entries taken from seven different sources, and the InterPro database stores over 30,000 entries from thirteen different sources, thus making them the most comprehensive tools for domain analysis available today.

However, there are comparatively few resources dedicated to the analysis of domain architecture. The NCBI offers several tools built on the CDD, most notably CD-Search (Marchler-Bauer and Bryant, 2004), which searches protein sequences against the CDD to identify the functional domains

they contain. CD-Search generates graphical outputs which make it easy to visually discern the domain architectures of query sequences. Likewise, the InterPro database can be searched using the InterProScan tool (Blum et al., 2021), generating similar output. While adequate for analysing individual sequences, these tools quickly become cumbersome when dealing with larger collections of sequences. Additionally, the output generated by these tools, particularly for larger enzymes, can contain hundreds of conserved domain hits, making it difficult to parse. The NCBI also offers other tools linked to CD-Search: the conserved domain architecture retrieval tool (CDART; Geer et al., 2002), which can be used to find proteins with similar domain architectures; and the subfamily protein architecture labeling engine (SPARCLE; Marchler-Bauer et al., 2017), which groups protein sequences with similar domain architectures and links them to curated functional classifications. Sequences are automatically placed into classification groups from SPARCLE after being analysed by CD-Search. More recently, TREND was developed (Gumerov and Zhulin, 2020), which allows analysis of domain architecture in an evolutionary context. TREND predicts domains by searching either the CDD or Pfam databases, whilst also generating a phylogeny of the input sequences. However, these tools do precisely annotate all domains within query sequences, with smaller domains often being obscured by hits to larger fused multidomain profiles.

Several tools have been developed specifically for the analysis of secondary metabolite megasynthases. One of the original tools built for this purpose was SEARCHPKS (Yadav et al., 2003), which was subsequently rolled into NRPS-PKS (Ansari et al., 2004) and is now available as a part of the structure based sequence analysis of PKS and NRPS (SBSPKS) webserver (Khater et al., 2017). It offers prediction of domain architecture for up to 10 sequences at a time via alignment to curated hidden Markov model (HMM) profiles, as well as predictions of substrate specificity and chemistry and comparison to sequences in a database of characterised PKS and NRPS gene clusters. However, it is not available for local installation, nor is it accessable programatically, and at the time of writing, several pieces of functionality are unavailable. The antibiotics and Secondary Metabolite Analysis Shell (antiSMASH) performs rule-based prediction of biosynthetic gene clusters in genomes based on the presence of key seed domains (Blin et al., 2019). The domain architectures of megasynthases in predicted BGCs are determined by searching a local database of curated profile HMMs. Occasionally domains are missed in the predicted architecture, particularly smaller domains which typically achieve lower scores during searches (e.g. acyl-carrier protein domains). Additionally, as antiSMASH takes genome sequence as input, it may be unsuitable for analysis of single proteins. The use of internally curated HMM profiles in both SBSPKS and antiSMASH, while greatly improving speed and specificity

of predictions, also makes them inflexible to prediction of new domain types.

Here we describe `synthaser`, a Python based software package leveraging the NCBI's CD-Search API which can automatically annotate and classify the domain architectures of multi-domain proteins based on a flexible, user-definable ruleset system. `synthaser` produces interactive visualisations of proteins grouped by their classification, making proteins with interesting architectures immediately apparent. Below, we extensively detail the `synthaser` search workflow and other functionality in the package, including modules for downloading search databases and extracting domain sequences, as well as a web application for easily building rule sets. As a proof of concept, we detail the process of building a `synthaser` ruleset using the web application for the classification of fungal secondary metabolite megasynthases, specifically polyketide synthases and nonribosomal peptide synthases. To evaluate this rule set, we analyse all available PKS and NRPS sequences deposited in the MIBiG repository (Kautsar et al., 2019) and compare the domain architectures predicted by `synthaser` to the corresponding antiSMASH-generated predictions stored in each MIBiG entry. Finally, we build a similarity network of polyketide synthases in publicly available *Aspergillus* genomes, and link it to `synthaser` domain architecture predictions to demonstrate how `synthaser` can be used to quickly identify interesting sequence groups for further investigation. We show `synthaser` to be a useful addition to the genome mining toolbox, particularly within the context of natural products research; however, given the programmable nature of `synthaser`, we can foresee much broader applications of the software.

2 Materials and methods

2.1 Software implementation and availability

`synthaser` is implemented in Python 3, and only requires the requests library to perform remote searches. `synthaser` is open source and is made freely available on GitHub (`https://github.com/gamcil/synthaser`) and PyPI (`https://pypi.org/project/synthaser`) under a MIT license. To perform local searches, `synthaser` requires that both Reverse Position Specific BLAST (RPS-BLAST) as well as `rpsbproc`, the command line utility that formats local RPS-BLAST results to resemble those returned by the CD-Search web service, are installed and accessable on the system `$PATH`. (Marchler-Bauer et al., 2002).

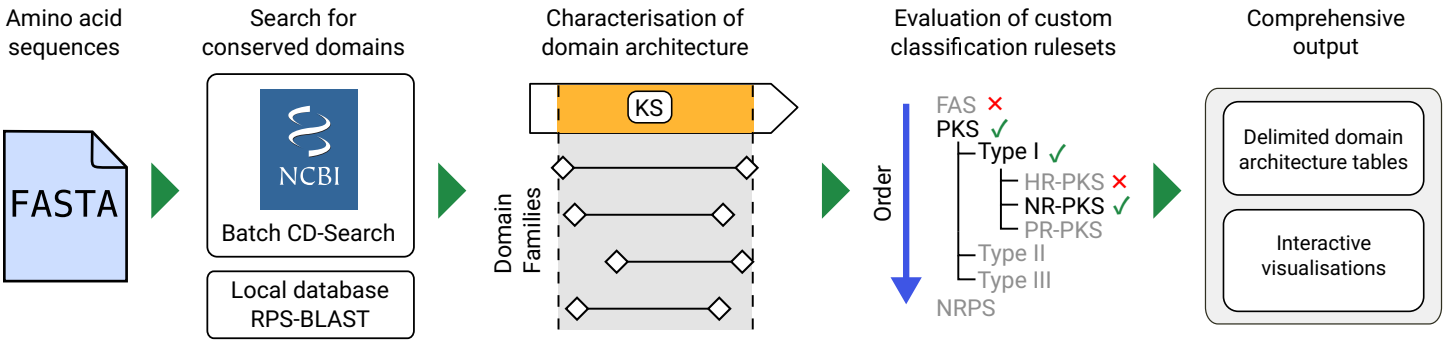

Figure 1: Outline of the `synthaser` workflow.

2.2 The synthaser search workflow

The `synthaser` search workflow is detailed in Figure 1. Briefly, query sequences are read from FASTA files and sent to the NCBI's CD-Search API to search for functional domains (or RPS-BLAST in local searches). Domain architectures of query sequences are annotated based on an analysis of overlapping domains. Sequences are classified according to a hierarchy of rules encoded in a programmable rule file. Finally, synthaser produces comprehensive text and visual outputs. These steps are described in more detail below.

2.2.1 Accepted input

`synthaser` accepts files in FASTA format, as well as collections of valid NCBI sequence identifiers specified either in newline-separated text files or directly passed to the command line interface. Query sequences are parsed directly from FASTA files using BioPython (Cock et al., 2009), whereas sequences corresponding to NCBI identifiers are retrieved using the Entrez API (NCBI Resource Coordinators, 2014). Additionally, `synthaser` provides a module, `genbank`, which allows users to extract all PKS and NRPS sequences from GenBank format files generated by antiSMASH (version 5.0 and above) to a FASTA file ready for `synthaser`analysis.

2.2.2 Remote searches via NCBI Batch CD-Search API

In remote searches, query sequences are uploaded to the BATCH CD-Search API (Marchler-Bauer et al., 2011). Every search is assigned a unique CD-Search identifier (CDSID) that is saved and reported in the output. Each CDSID remains valid for 36 hours, and can be used to directly re-start a `synthaser` run at any point during this period. The CDSID is polled against the API continuously until the search has completed and results can be retrieved.

2.2.3 Local searches using RPS-BLAST and rpsbproc

The underlying search for any remote CD-Search run is performed using Reverse Position Specific BLAST (RPS-BLAST), a variant of Position-Specific Iterated BLAST (PSI-BLAST), which searches protein sequences against a database of domain profiles (Marchler-Bauer and Bryant, 2004). By default, RPS-BLAST output resembles the output of other BLAST variants. The NCBI offers another tool, rpsbproc, which processes RPS-BLAST results to resemble those returned by the CD-Search web server (Marchler-Bauer et al., 2017). `synthaser` provides a local search mode which wraps RPS-BLAST and rpsbproc, enabling searches against local profile databases. Here, input sequences are searched against a local profile database using RPS-BLAST, and search results are post-processed by rpsbproc such that they can be analysed like remote CD-Search results. The domain family profile databases used in CD-Search are available as pre-formatted RPS-BLAST databases from the NCBI FTP server, which can be retrieved using the `getseq` module.

2.2.4 The central `synthaser` rule file

Once CD-Search results have been retrieved, `synthaser` undergoes two phases: identification of domains in query sequences, and the classification of query sequences. Underlying these phases is a central rule file in JSON format which specifies i) conserved domains that `synthaser` should attempt to identify in query sequences, ii) rules for assigning classifications to sequences based on identified domains, and iii) a hierarchy that determines the order of rule evaluation. The full schema of the rule file is detailed in Figure S1.

2.2.5 Identification of functional domain 'islands'

Domain hits in CD-Search results naturally segregate into distinct 'islands' of overlapping related domain families (Figure 2). `synthaser` attempts to characterise the domain architecture of query sequences by programmatically identifying these islands. This is done by defining sets of conserved domain families that correspond to broader functional classes (as specified in the rule file). For example, the KS island in Figure 2 consists of a variety of individual conserved domain families (e.g. PKS, PKS_KS, KAS_I_II). These values are used at several stages in the `synthaser` workflow.

During the domain identification phase, `synthaser` discards domain hits not specified by any broader domain class and filters remaining hits for quality. Every domain family in the CDD has an underlying position-specific scoring matrix (PSSM) describing the amino acid makeup of the family, as well as a threshold bit-score value used to determine if a given hit is a specific (i.e. high confidence)

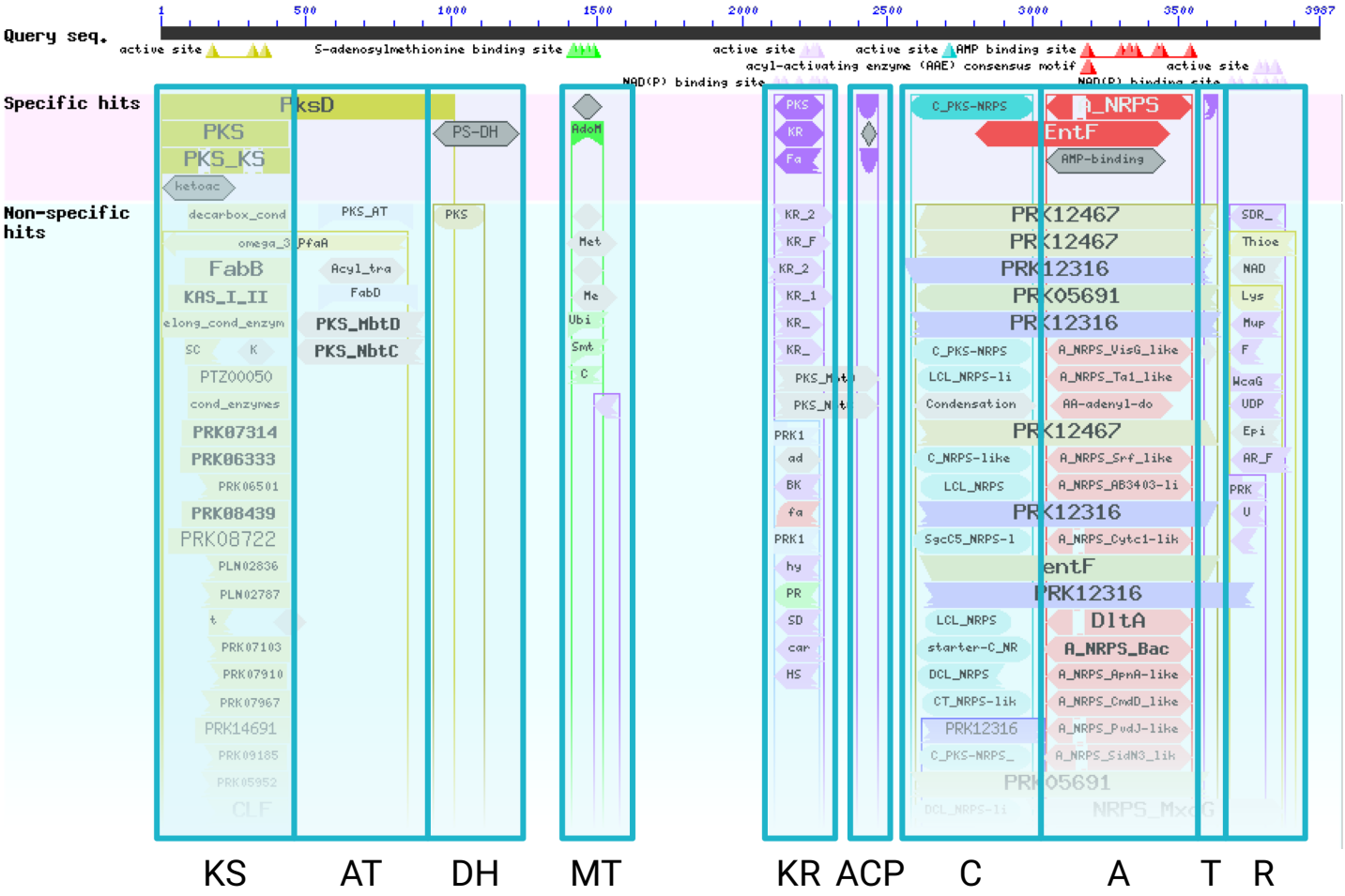

Figure 2: Naturally segragating 'islands' of functional domains found in CD-Search analysis of BuaA, the hybrid PKS-NRPS involved in the biosynthesis of the burnettramic acids (Li et al., 2019). The islands correspond to the domain architecture KS-AT-DH-MT-KR-ACP-C-A-T-R.

hit. By default, `synthaser` will discard hits that do not meet the thresholds for PSSM length (30% of the family PSSM), or bitscore (30% of specific-hit bitscore); users can freely adjust these parameters. After filtering, `synthaser` identifies groups of overlapping domain hits, choosing a representative hit based on maximum length, maximum bitscore or minimum e-value.

Occasionally, a single-domain can be reported as short, discontiguous, low-scoring hits. To resolve such cases, `synthaser` explicitly checks for adjacent, truncated domain hits of the same or equivalent types. `synthaser` uses two threshold parameters to determine if merging should occur: i) the length of each hit as a proportion of their corresponding PSSM lengths, termed coverage, and ii) the bitscore of each hit as a proportion of the specific-hit PSSM. Two hits are merged if both occur within the space of a single PSSM length (±10%), their combined bitscore is above the threshold bitscore, and the combined lengths are above a given query coverage threshold.

Finally, `synthaser` reports the domain architecture of each query sequence.

2.2.6 Functional classification based on domain architecture

Once the domain architectures of query sequences have been characterised, `synthaser` has the option to evaluate the classification rules defined in the rule file. This allows for the division of multidomain proteins into subgroups based on the absence or presence of specific domains, which can provide additional insights into the functional differences between them. Each rule must contain i) a name, ii) a list of domain types, and iii) a logical expression used to evaluate the rule, hereafter termed an evaluator. The rule name is transferred to the sequence upon successful evaluation; each sequence has a classification array which can contain any number of rule names (i.e. multiple rules satisfied in hierarchy). The list of domain types contains domains which are referenced by, though not necessarily required for satisfaction of, the rule. The evaluator is a logical expression which determines if a rule is satisfied by a collection of domains. It is comprised of a series of numbers referring to the indices of each domain in the list of domain types and logical operators that connect them. When a rule is evaluated on a collection of domains, `synthaser` checks that domains referred to in the evaluator are found, substituting the corresponding numerical index in the evaluator with the result (True if the domain type is found, otherwise False). The final expression is then evaluated to determine if the rule has been satisfied or not.

Continuing the example shown in Figure 2, we may wish to create a PKS-NRPS rule which requires domain types KS and A. In this rule, the domains list may resemble 1:

$$[KS,\ A] \tag{1}$$

As both domain types are required for the rule, the evaluator would then resemble 2:

$$0\ and\ 1 \tag{2}$$

Here, 0 refers to the KS domain and 1 refers to the A domain. If analysing a PKS containing a KS domain but not an A domain, the evaluator after substitution would resemble 3:

$$True\ and\ False \tag{3}$$

As 3 evaluates to False, the rule is not satisfied. However, classifying the sequence in Figure 2 would yield the expression 4:

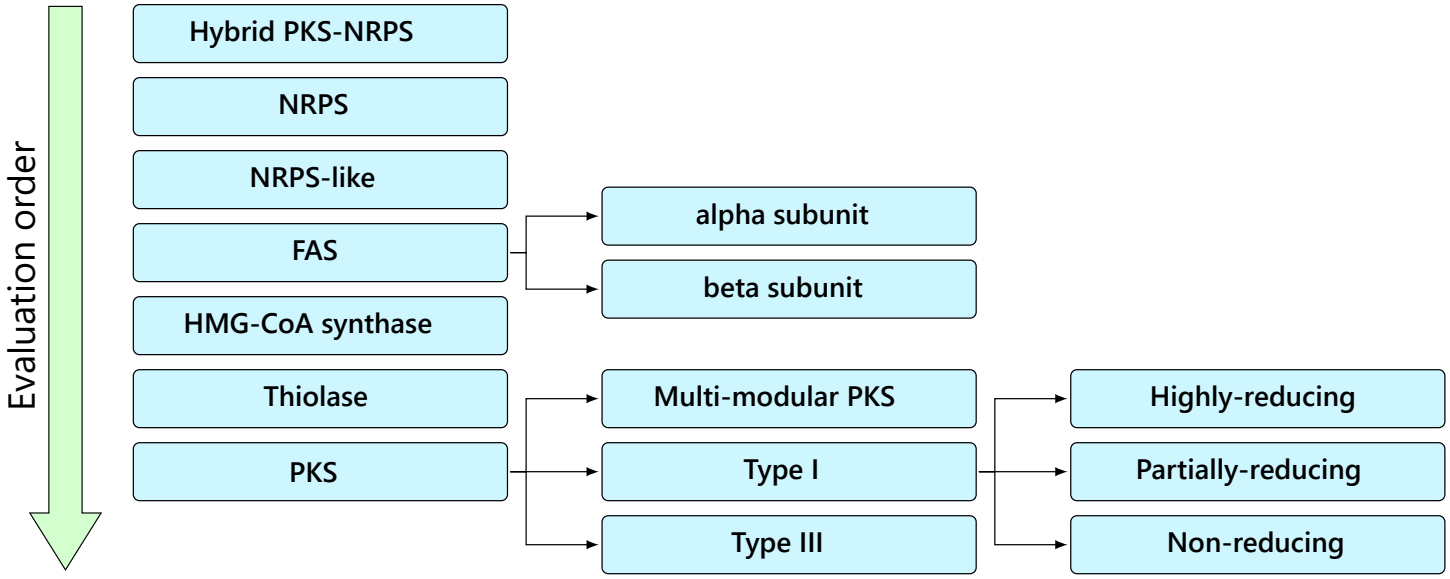

Figure 3: Evaluation hierarchy of sequence classification rules in fungal secondary metabolite megasynthases. Rules in this hierarchy are evaluated top-to-bottom; if a rule evaluates unsuccessfully, the next rule in the hierarchy is evaluated. If a rule with children (relationship depicted by arrows) is successfully evaluated, the child rules will be evaluated. Upon successful evaluation, the full classification array, including the parent and all child rules, is saved on the sequence.

$$True\ and\ True \tag{4}$$

Thus successfully classifying the sequence as a PKS-NRPS.

Rules can additionally include domain type filters and renaming rules. Domain type filters indicate that a rule only accepts a domain in a sequence if the representative hit is of a specific domain family. This allows for differentiation between specific families that fall under the same broader classes (e.g. KS domains from FAS and PKS). Renaming rules allow domain types to be renamed in `synthaser` output. This is useful in cases where functionally equivalent domains have different nomenclature based on context. For example, acyl carrier proteins (ACP) of PKSs and peptidyl carrier proteins (PCP) of NRPSs are closely related and typically hit the same domain families in a CD-Search, but convention dictates they are denoted by ACP in PKSs and T (thiolation) or PCP in NRPSs. Renaming rules can optionally include *before* or *after* domains, which specify that the renaming target should only be renamed if it is found before or after certain domains. For instance, an ACP in a hybrid PKS-NRPS should only be renamed to T within the NRPS module, which can be accounted for in the rule by adding key NRPS domains (e.g. A or C) as *after* domains.

The final element of the classification rule system is the hierarchy. This takes the form of a tree structure, with each node in the tree containing the name of a classification rule as well as a list of any child rules (Figure 3). `synthaser` uses this tree to determine the order of evaluation during

sequence classification. If a rule is not satisfied, `synthaser` will proceed to the next rule of the same depth within the tree; if it is satisfied, `synthaser` will recurse into the children of that rule, and so on. During this stage of the `synthaser` workflow, the rule hierarchy is evaluated on each sequence, which is then assigned a classification array containing the names of all rules which were succesfully evaluated. For instance, the default rule file would assign a highly-reducing PKS the array: [*PKS, Type I PKS, Highly-reducing PKS.*] After sequences have been classified, domain architectures of each query sequence is reported, and an interactive visualisation is generated.

The generic nature of this rule system means that although `synthaser` was written primarily to analyse secondary metabolite megasynthases, it can be readily repurposed for the analysis of any multi-domain protein family.

2.3 Building rule files using the synthaser rule generator web application

Given how cumbersome it is to manually assemble the `synthaser` rule file, we provide a web application which can generate rule files in three easy steps (Figure 4).

In the leftmost pane, the collection of domain types are built by specifying names (e.g. *KS*) and domain families (*PKS* and *PKS_KS*). The *Families* selection box is linked to a file containing information about every family in the CDD, so families can be found simply by searching their names or accessions in the box.

Once domain types have been created, classification rules can be built in the middle pane. Each rule requires a name, a list of domain types it requires and the evaluator. The *Domains* box allows for selection of the domain types created in the *Domain types* pane. Rule names and evaluators can be added simply by writing in the relevant field. Domain type filters can be added inside the *Domain filters* section of each rule. Within each filter, the domain type can be specified in the *Domain name* selection field, and the domain families in the *Domain types* selection field. Renaming rules can be added in the *Rename domains* section. Within each renaming rule, the renaming target domain can be selected in the *From* field, any after domains in the *After* selection field, and the new name in the *To* input field.

Finally, the classification rule hierarchy can be established in the rightmost pane. When a rule is added or updated in the *Classification rules* pane, it is automatically added to, or updated in, the rule hierarchy. Each rule can be dragged and dropped anywhere within the hierarchy, and can be easily

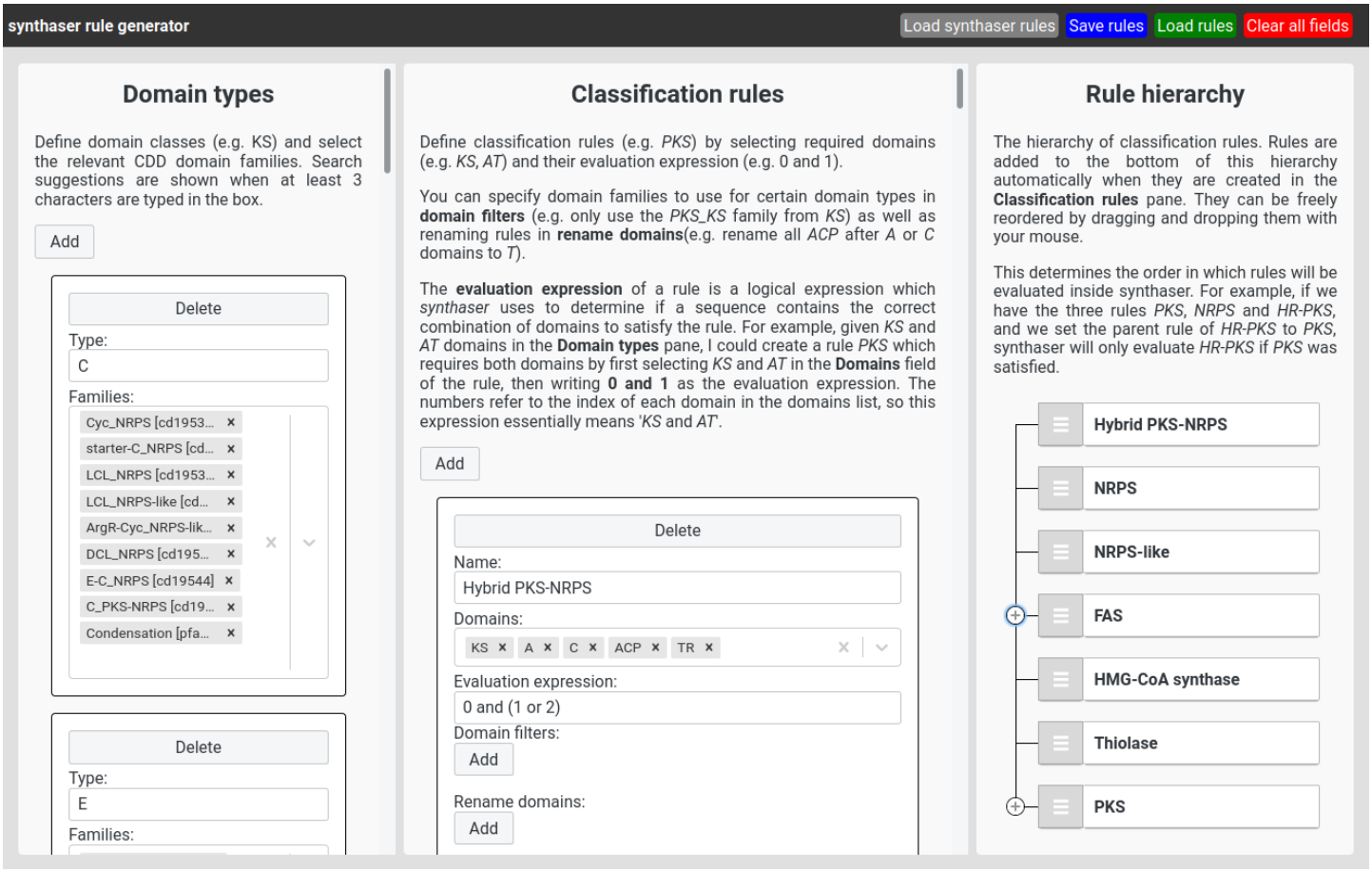

Figure 4: Web application for designing and editing `synthaser` rule files. Rule files are built in three stages: domain classes are defined in the 'Domain types' pane; sequence classification rules are built in the 'Classification rules' pane; and the hierarchy of rule evaluation is determined by re-arranging the rules in the 'Rule hierarchy' pane. The web application is hosted at `https://gamcil.github.io/synthaser`.

nested to form parent-child rule relationships.

The default ruleset bundled with `synthaser` can be loaded by clicking the *Load default rules* button (Figure 4d), so users can quickly understand how the rule generator system works. After building the ruleset in the web application, the JSON rule file can be generated by clicking the *Save rules* button; this can be easily loaded back into the application using the *Load rules* button, enabling easy updates to pre-existing rule files. This file can be passed to the `synthaser` search module using the `-uf / -rule_file` argument, where it will be loaded in place of the default ruleset.

The web application is implemented using the React framework and is hosted on GitHub pages (`https://gamcil.github.io/synthaser`).

2.4 Reporting and visualisation

`synthaser` provides comprehensive textual and visual outputs. By default, textual results are generated and printed to the command line. In this output, query sequences are listed alongside their predicted

domain architecture, grouped by their classification groups and in descending order of sequence length. Tabular, long form versions of this output can also be easily generated, allowing further analysis in spreadsheet software.

`synthaser` can also generate an interactive visualisation by using the `-plot` argument (Figure 5). Each query sequence is drawn to scale, grouped by classification and shown in descending sequence length order. Box annotations are drawn for each classification group; nested classifications are shown in reverse order (deepest classifications in the hierarchy shown first), making broader classification groups immediately obvious. Each sequence contains a collection of domains; all unique domain types within the visualisation are assigned a colour, which is used to colour both the domain element within the sequence, as well as the corresponding entry within the legend. Hovering over a domain will produce a tooltip which contains a summary of information about the domain hit, including the specific CDD family, its superfamily (if applicable), the domain class as determined by the rule file, its position within the query sequence, and the E-value and bitscore values from the CD-Search results. The sequence of the domain hit within the query can be copied to the system clipboard directly by clicking a button within the tooltip; the entire query sequence can also be copied in this way. FASTA files containing all sequences of a certain domain type can easily be generated within the *Download domain sequences* section of the settings panel, by first selecting the desired domain type and then clicking the download button.

The `synthaser` visualisation has various settings available to tweak its appearance. The shape, size and positioning of each synthase, as well as the vertical spacing between synthases, can be manipulated; maximum sequence length in pixels can be adjusted to control the width of the plot; and the font size of the various text elements within the plot, as well as text and box elements within the legend, can be changed. Once the user is satisfied with the appearance of the plot, a scalable vector graphics (SVG) image file can be generated by clicking the *Save SVG* button at the top of the settings panel. This file can be directly imported into vector image software for further manual editing.

`synthaser` can also generate a static HTML document containing all data and code required to display the visualisation when a file name is provided to the `-plot` argument. This enables `synthaser` results to easily be shared between individual computers.

The visualisation is implemented using the D3 JavaScript library (Bostock et al., 2011), and is available as a standalone reusable chart library under the MIT license (`https://www.github.com/gamcil/synthaser.js`).

2.5 Analysis of characterised synth(et)ases

Characterised PKS clusters of fungal origin were obtained from MIBiG 2.0 (Medema et al., 2015). The corresponding GenBank and JSON files (obtained from the full database dumps for each format) were then parsed for PKS and NRPS sequences, as well as their corresponding domain architectures as annotated on MIBiG, using a Python script (available from `https://github.com/gamcil/synthaser_scripts`) and then added to the dataset. Sequences were then analysed using `synthaser` with default settings (maximum E-value 1.0, domain family PSSM length threshold percentage 40%, domain family bitscore threshold percentage 40%, domain family coverage 60%, tolerance 10%), and compared to the antiSMASH domain architecture predictions shown by MIBiG. Sequences were compared for overall domain architecture matches between `synthaser` and antiSMASH predictions stored in MIBiG. Discrepencies were sorted into groups based on the specific type of mismatch between each prediction: domains identified in `synthaser` but not in MIBiG; equivalent domains found but mis-named in MIBiG; and mis-named in `synthaser` predictions. Count data was analysed and visualised using R.

2.6 Building a network of *Aspergillus* polyketide synthases

Sequences containing ketosynthase (KS) domains were identified by querying the NCBI Protein database for entries linked to the `cond_enzymes` (CDD UID: 238201) superfamily using Entrez Direct (Kans, 2019). The results were filtered to only include GenBank sequences from *Aspergillus* species. All remaining sequences were retrieved, and were analysed for PKS domains with the built-in ruleset using default settings with the command `synthaser search -qf sequences.fa`. Regions corresponding to KS domains in the identified sequences were extracted using the `extract` module in `synthaser`. The extracted sequences were formatted as a DIAMOND database and aligned against themselves using DIAMOND 0.9.17 with the `-more-sensitive` flag (Buchfink et al., 2021-04). An edge table was generated by summing the bitscores of individual high-scoring segment pairs (HSPs) of each unique query and target sequence pair using a Python script (available from `https://github.com/gamcil/synthaser_scripts`). This table was imported into Cytoscape 3.7.2, and clusters were predicted using the Markov Clustering (MCL) algorithm with an inflation parameter of 2.5 via the implementation in the clusterMaker2 1.3.1 plugin (Shannon, 2003; Morris et al., 2011). Representative domain architectures of each cluster were identified by mapping `synthaser` results to the extracted domains in the Cytoscape network. A full explanation of the creation of the network is provided in the supplementary information.

3 Results

3.1 A classification framework for fungal megasynthases

Two of the major classes of natural products are polyketides and nonribosomal peptides, synthesized by polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS), respectively (Keller et al., 2005). There is significant interest in the genome mining of new polyketide and nonribosomal pathways for their potential in making new drugs (Newman and Cragg, 2016). These megasynthases are large enzymes consisting of multiple functional domains, each responsible for a different step in the biosynthesis of the products backbone.

PKSs, similar to fatty acid synthases (FAS), build the carbon backbone of polyketides through repetitive condensation of short-chain carboxylic acids, catalyzed by a β-ketoacyl synthase (KS) domain (Chooi and Tang, 2012). KS domains belong to a broader family of condensing enzymes, which includes enzymes catalysing decarboxylating and non-decarboxylating reactions (Heath and Rock, 2002). The decarboxylating enzymes are further broken into the 'initiation' enzymes, which include chalcone synthases (CHS) of Type III PKSs and hydroxymethylglutaryl (HMG)-CoA synthases, and 'elongation' enzymes, which include β-ketoacyl-ACP synthases (type I and II) of FAS and KS domains of PKS. The non-decarboxylating enzyme group is comprised of biosynthetic and degradative thiolases.

PKSs are generally classified as types I, II or III (Table 1), though only types I and III are found in fungi. A minimal PKS consists of the KS, as well as acyltransferase (AT) and acyl-carrier (ACP) domains required for chain extension. While bacterial type I PKS are typically modular, with each chain extension step encoded by a distinct module, fungal PKS are typically iterative, with a single module being used repeatedly; though examples of modular PKS have been identified in fungi (Keller et al., 2005; Thynne et al., 2019). Iterative type I PKS are further classified as highly-reducing (HR), partially-reducing (PR) or nonreducing (NR) based on the presence of reductive β-keto processing domains. HR-PKSs typically produce aliphatic or alicyclic compounds, and will contain enoylreductase (ER), ketoreductase (KR) and dehydratase (DH) domains, which catalyze reduction reactions on the β-keto group during each chain extension step. Notable HR-PKSs include LovB, responsible for the production of lovastatin in *Aspergillus terreus* (Ma et al., 2009) and the prosolanapyrone synthase (PSS) involved in biosynthesis of solanapyrones in *Alternaria solani* (Kasahara et al., 2010). PR-PKS contain at least one, but not all, of these reductive domains (Kroken et al., 2003). For instance, the well known 6-methylsalicylic acid synthase (6-MSAS) from *Penicillium patulum* (Beck et al., 1990) and mellein synthase responsible for production of (R)-mellein in *Parastaganospora nodorum* (Chooi

et al., 2015) possess DH and KR domains but no ER domains. NR-PKSs have no reductive domains, and typically contain a starter unit:ACP transacylase (SAT), a product template (PT) and a releasing domain (thioesterase (TE) or thioreductase (R) domain) (Chooi and Tang, 2012). NR-PKSs almost always produce aromatic compounds, where cyclisation is mediated by the PT domain. For example, *pksA*, involved in the biosynthesis of aflatoxin in *Aspergillus parasiticus* is an NR-PKS (Chang et al., 1995). Such classifications are useful as it gives an indication as to the type of compound that may be produced by the PKS.

Type III PKS are distinguished by their lack of ACP domain and are related to the chalcone and stilbene synthases found in plants (Hertweck, 2009). They are observed mostly in bacteria, though several have been characterised in fungi and have been shown to produce α-pyrones, resorcylic acids and resorcinols (Hashimoto et al., 2014; Navarro-Muñoz and Collemare, 2020).

NRPSs typically consist of multiple modules, each possessing a binding specificity to a specific amino acid, which can be proteinogenic or non-proteinogenic. A nonribosomal peptide is synthesized through the formation of peptide bonds between amino acids attached to adjacent modules (Keller et al., 2005). A minimal NRPS consists of adenylation (A), peptidyl carrier protein (PCP)/thiolation (T), condensation (C) and thioesterase (TE) domains, and can be modular or iterative.

Finally, it is possible to have hybrid enzymes that contain both PKS and NRPS modules (denoted PKS-NRPS, or NRPS-PKS depending on module order) which in turn produce polyketide-peptide metabolites, or pathways of a mixture of PKS types (Hertweck, 2009). Notable examples include the PKS-NRPSs involved in the biosynthesis of the burnettramic acids in *Aspergillus burnettii* (Li et al., 2019), the cytochalasins in *Aspergillus clavatus* (Qiao et al., 2011), and phomacins in *Parastaganospora nodorum* (Li et al., 2020b).

3.2 Building a `synthaser` ruleset

Multidomain protein families can often be divided into subgroups based on the absence or presence of certain domains, which can facilitate further functional predictions. Likewise, fungal type I PKSs have been subdivided into HR, PR and NR-PKS based on the absence or presence of reductive β-keto processing domains. This information provides insights into the nature of the polyketide products encoded by the PKS genes; for instance, NR-PKS are most likely to make aromatic compounds, while HR-PKS can make alicyclic or aliphatic compounds. As a proof of concept for the `synthaser` workflow, we designed a rule file for the classification of fungal megasynthases, namely PKS-NRPS, FAS, PKS and NRPS. Using the rule generator web application, we built the rule file according to

Table 1: Classification scheme of polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS).

Classification			
Level 1	**Level 2**	**Level 3**	**Key domains**
Hybrid PKS-NRPS			KS, A or C
NRPS			A, T, C
NRPS-like			A
Fatty acid synthase (FAS)			β-ketoacyl-ACP synthase
	alpha subunit		ACP, KR, KS
	beta subunit		SAT, ER, DH
HMG-CoA synthase			HMG-CoA synthase
Thiolase			Thiolase
Polyketide synthase (PKS)			KS
	Multi-modular		Multiple KS
	Type I		KS, AT
		Highly-reducing	ER, KR, DH
		Non-reducing	SAT, PT
		Partially-reducing	ER, KR or DH
	Type III		CHS

the classification scheme shown in Table 1. A `synthaser` ruleset is comprised of three elements: the domain classes that we wish to identify, rules to classify sequences based on the domain classes that are identified, and a hierarchy which determines the order in which rules are evaluated. The domain classes, as well as the domain families that comprise them and scoring information is shown in Table 2. In sum, 58 domain families were placed into 16 different domain classes, covering domains frequently observed in fungal secondary metabolite megasynthases. This included classes for adenylation (A), acyl-carrier protein (ACP), ACP synthase (ACPS), acyltransferase (AT), condensation (C), dehydrogenase (DH), epimerization (E), enoylreductase (ER), ketoreductase (KR), beta-ketoacyl synthase (KS), methyltransferase (MT), product template (PT), starter unit:acyl carrier protein transacylase (SAT), thioesterase (TE), thioester reductase (TR) and carnityl acyltransferase (cAT). Domain families for each class were manually chosen by performing online CD-Search searches with characterised megasynthase sequences and analysing which domain families appear in each domain 'island' observed in the visual output (see Figure 2).

Once domain classes had been established, functional classification rules could be created. Following the framework outlined in Table 1, we generated a collection of rules covering each unique megasynthase classification (Table 3). In total, 15 rules were created, covering the spectrum of fungal PKS, fatty acid synthase (FAS) and NRPS sequences. These consist of 7 top-level rules, including those for Hybrid PKS-NRPS, Thiolases, HMG-CoA synthases, FAS, NRPS, NRPS-like and PKS sequences.

Within these top-level rules, there are further child rules. For example, FAS sequences are further classified as alpha or beta subunit. Similarly, PKS sequences can be classified as multi-modular

Table 2: Domain classes and domain families defined in the default `synthaser` rule file.

	Domain family				
Domain class	**Accession**	**Name**	**PSSM ID**	**PSSM Length**	**Threshold bitscore**
A	cd05930	A_NRPS	341253	444	356.838
	pfam00501	AMP-binding	366135	361	184.727
ACP	smart00823	PKS_PP	214834	86	33.3777
	CHL00124	acpP	177047	82	85.8428
	pfam14573	PP-binding_2	373139	96	112.899
	pfam00550	PP-binding	376348	67	29.814
ACPS	COG0736	AcpS	223807	127	88.0777
	PRK00070	acpS	234610	126	87.108
AT	smart00827	PKS_AT	214838	298	201.477
C	cd19535	Cyc_NRPS	380458	423	348.324
	cd19533	starter-C_NRPS	380456	419	482.254
	cd19538	LCL_NRPS	380461	432	638.54
	cd19531	LCL_NRPS-like	380454	427	287.329
	cd20480	ArgR-Cyc_NRPS-like	380470	406	714.278
	cd19543	DCL_NRPS	380465	423	420.456
	cd19544	E-C_NRPS	380466	413	476.159
	cd19532	C_PKS-NRPS	380455	421	402.605
	pfam00668	Condensation	334202	455	345.862
DH	pfam14765	PS-DH	379688	291	132.479
	smart00826	PKS_DH	214837	167	76.8814
E	cd19534	E_NRPS	380457	428	363.495
ER	COG4981	COG4981	227314	717	1062.53
	smart00829	PKS_ER	214840	287	250.768
	cd05195	enoyl_red	176179	293	129.997
	cd08270	MDR4	176231	305	268.471
	cd05282	ETR_like	176645	323	224.079
KR	smart00822	PKS_KR	214833	180	83.3005
	cd08950	KR_fFAS_SDR_c_like	187653	259	368.441
KS	smart00825	PKS_KS	214836	298	241.079
	cd00833	PKS	238429	421	167.35
	cd00829	SCP-x_thiolase	238425	375	147.795
	TIGR01833	HMG-CoA-S_euk	273826	457	790.507
	cd00751	thiolase	238383	386	222.354
	PLN02287	PLN02287	215161	452	632.955
	PRK07314	PRK07314	235987	411	601.393
	TIGR03150	fabF	274452	407	525.512
	COG0304	FabB	223381	412	152.8
	cd00832	CLF	238428	399	503.431
	cd00834	KAS_I_II	238430	406	285.201
	cd00831	CHS_like	238427	361	242.515
	cd00830	KAS_III	238426	320	212.4
MT	pfam08241	Methyltransf_11	369777	93	53.4302
	pfam08242	Methyltransf_12	369778	96	37.7331
	pfam13489	Methyltransf_23	372616	162	59.7422
	pfam13649	Methyltransf_25	379312	96	36.008
	pfam13847	Methyltransf_31	316372	150	67.449
	cd02440	AdoMet_MTases	100107	107	29.3203
	smart00828	PKS_MT	214839	224	133.695
PT	TIGR04532	PT_fungal_PKS	275325	324	202.466
SAT	pfam16073	SAT	374347	239	110.757
TE	smart00824	PKS_TE	214835	212	155.846
	pfam00975	Thioesterase	366397	223	155.591
	COG0657	Aes	223730	312	59.5636
	pfam00561	Abhydrolase_1	366166	245	93.3378
TR	TIGR01746	Thioester-redct	273787	367	305.493
	cd05235	SDR_e1	187546	290	229.845
cAT	pfam00755	Carn_acyltransf	376382	577	255.551

Table 3: Overview of rules for classification of fungal secondary metabolite megasynthases used in `synthaser`. Each rule is comprised of a name, a set of domain classes, an evaluator, domain filters which specify valid domain families for a given domain class, and rename rules, which specify domain classes which should be renamed in certain contexts.

			Domain rules		Rename rules			
Rule name	**Domains**	**Evaluator**	**Class**	**Families**	**From**	**To**	**Before**	**After**
Hybrid PKS-NRPS	KS, A, C, ACP	0 and (1 or 2)			ACP	T		A, C
					T	ACP	A, C	KS
Thiolase	KS	0	KS	cd00829, cd00751, PLN02287				
HMG-CoA synthase	KS	0	KS	TIGR01833				
beta subunit	AT, ER, DH	0 and 1 and 2	ER	COG4981				
alpha subunit	AT, KR, ACPS	0 and 1 and 2	KR	cd08950				
FAS	KS	0	KS	COG0304, cd00834, cd00830, TIGR03150				
NRPS-like	A, C, ACP	0 or 1			ACP	T		
NRPS	A, ACP, C	0 and 2			ACP	T		
Type III	KS	0	KS	cd00831				
Non-reducing PKS	SAT, PT	0 and 1						
Partially-reducing PKS	DH, ER, KR	0 or 1 or 2	KR	smart00822, cl00100				
			ER	smart00829, cd05195				
			DH	smart00826				
Highly-reducing PKS	DH, ER, KR	0 and 1 and 2	KR	smart00822, cl00100				
			ER	smart00829, cd05195				
Type I PKS	AT	0	AT	smart00827, cl08282				
Multi-modular PKS	KS, KS	0 and 1						
PKS	KS	0	KS	smart00825, cd00833, cd00831				

(containing multiple KS domains), Type I or Type III; Type I sequences can be further classified into highly-, partially- or non-reducing PKS.

Finally, a rule evaluation hierarchy was created (Figure 3). `synthaser` evaluates from the first listed rule to the last, recursing into child rules if successful. This makes it simple to define hierarchies with any number of levels where rules incrementally build on other rules to assign more specific classifications.

The final rule file is distributed alongside the source code and is freely available from the GitHub repository.

3.3 Analysis of megasynthases in characterised biosynthetic gene clusters

In order to verify the accuracy of our fungal secondary metabolite megasynthase rule set, we decided to test it against previously characterised megasynthases deposited in the MIBiG database (Kautsar et al., 2019). BGCs of fungal origin were retrieved from the MIBiG database, and 284 sequences covering the spectrum of fungal megasynthase classifications were extracted (Table S1). Domain architectures of a subset of these sequences is shown in Figure 5. This collection consisted of 137 PKS, 61 NRPS, 31

NRPS-like, 24 FAS and 31 hybrid PKS-NRPS sequences (as determined by `synthaser` classification). Of the 137 PKS sequences, 48 were further classified as highly-reducing, 70 as non-reducing and 16 as partially-reducing. Similarly, of the 24 FAS sequences, 23 could be further classified into separate alpha (12 sequences) and beta-subunit (11 sequences) encoding genes, typical of the Ascomycetes, with the remaining sequence, *fas2* from the ustilagic acid BGC in *Ustilago maydis*, being a complete single-chain FAS (Teichmann et al., 2010), common to the Basidiomycetes and mycobacteria (Maier et al., 2010).

All sequences were correctly classified; domain architectures predicted by `synthaser` either matched completely (barring different naming schemes for like domains) or correctly identified more domains than the antiSMASH predictions reported on MIBiG. In total, 182 (35.92%) antiSMASH-generated domain architecture predictions exactly matched those from `synthaser`. Of the remaining 102 predictions (64.08%), 101 were mismatched due to domains being present in the `synthaser` predictions but not in the MIBiG records; 75 (73.53%) were missing a single domain, 38 (37.25%) were missing two domains, and 24 (23.53%) missing three domains. The most frequently missed were ACP/T domains, in PKS and NRPS, respectively, which were absent in 32 (31.37%) of the antiSMASH-generated predictions; other commonly missed domains were TE domains (16, 15.69%), KR domains (14, 13.73%) and SAT domains (14, 13.73%).

Notably, `synthaser` architecture predictions for the sordarin HR-PKS from *Sordaria araneosa* (Kudo et al., 2016), and the AF-toxin HR-PKS from *Alternaria alternata* (Ruswandi et al., 2005), both contain carnitine acyltransferase (cAT) domains, which are not present in the antiSMASH-generated domain architecture predictions. The cAT domain was recently shown to be capable of esterification of polyketide products in *Trichoderma virens* (Hang et al., 2017). While the Pfam database contains a profile HMM corresponding to the cAT domain (accession: PF00755), the NRPS/PKS analysis module in antiSMASH currently does not. This is one of the drawbacks of using manually curated profiles; though searches take significantly less time, new models must be built from scratch whenever new domains are to be analysed. The extensibility of the `synthaser` rule system allows for new domains to be easily added, provided an entry is available within the CDD.

Another notable case study is the starter unit:acyl carrier protein (ACP) transacylase (SAT) domain, a characteristic feature of NR-PKSs that is sometimes missed in both the `synthaser` and antiSMASH-generated domain architecture predictions. For instance, the PKS involved in the biosynthesis of the meroterpenoid paraherquonin in *Penicillium brasilianum* (Matsuda et al., 2016), *prhL*, though correctly classified as non-reducing, lacks an SAT domain in both the `synthaser` and antiSMASH predictions.

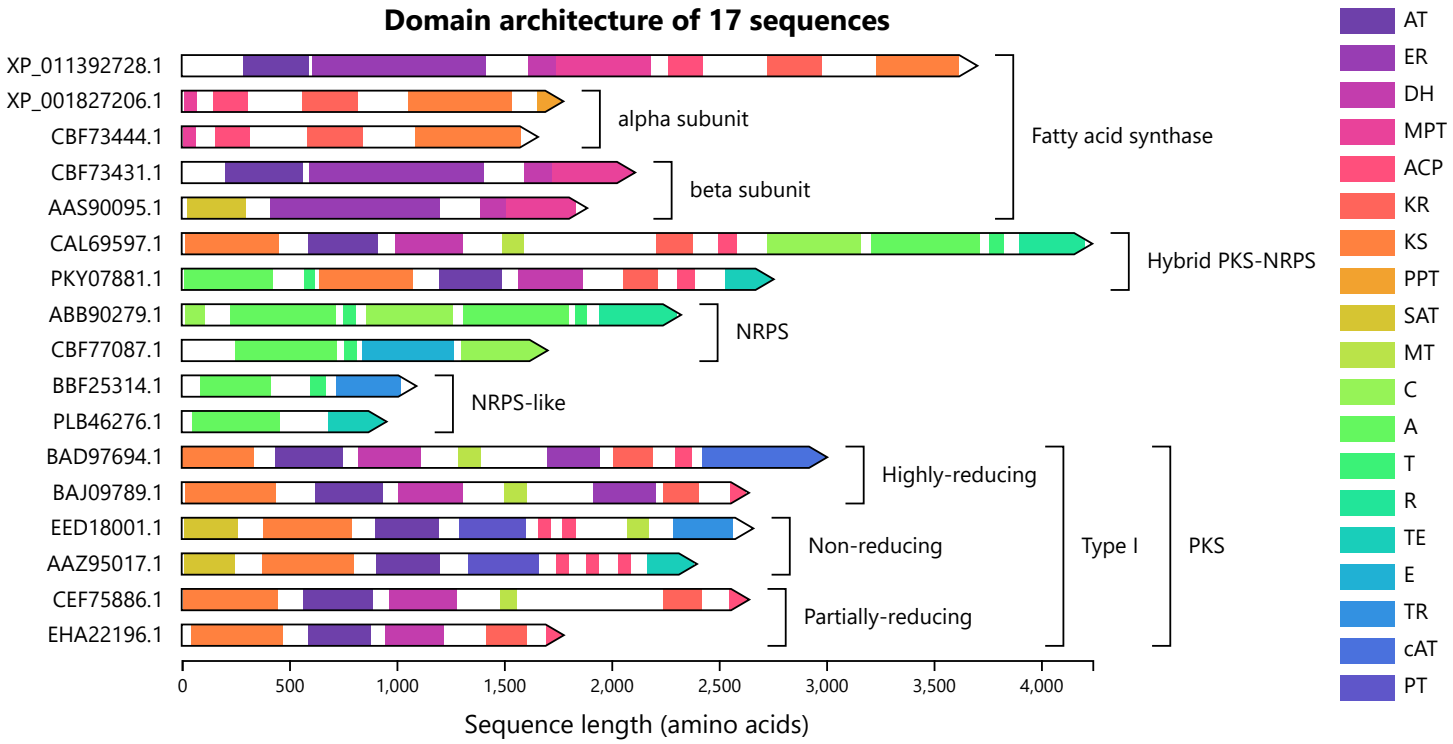

Figure 5: `synthaser` visualisation of a subset of the analysed synthase sequences sourced from the MIBiG database covering a spectrum of classification groups.

This is also observed in *trt4*, involved in the biosynthesis of another meroterpenoid compound, terretonin, in *Aspergillus terreus* (Guo et al., 2012). On the other hand, predictions for the NR-PKSs of the related andrastin A and novofumigatonin biosynthetic pathways in *Penicillium roqueforti* (Rojas-Aedo et al., 2017) and *Aspergillus novofumigatus* (Kjærbølling et al., 2018), respectively, do contain SAT domains. Closer inspection of the sequences with missing SAT domains showed annotation gaps in N-terminal regions, indicating that there were likely SAT domains that were missed (Figure S2). Sequence alignment of the N-terminal regions of NR-PKSs involved in fungal meroterpenoid biosynthesis, annotated both with and without SAT domains, revealed the presence of the characteristic SAT domain active site GXCXG motif (Crawford et al., 2006) in all sequences, confirming that the underlying CD-Search predictions did in fact miss the SAT annotations (Figure S3). In cases where a domain is missed due to low quality, this problem can be alleviated by simply raising the E-value cutoff used during a `synthaser` search; in other cases, missing domains may persist due to other reasons (e.g. structural variation, poor domain curation). That the quality of `synthaser` predictions and classifications is reliant upon the quality of the underlying search databases is a limitation of the tool. However, as the quality of domain profile HMMs increases, so to will the power of `synthaser` to predict and classify domain architectures.

3.4 Network analysis of PKS domain architectures in *Aspergilli* reveals interesting variation

`synthaser` can rapidly extract PKS and NRPS genes and generate domain annotations from genome files, making it extremely useful in providing an overview of the diversity of PKS/NRPS domain architectures encoded in an organism. For instance, we recently used `synthaser` to analyse synthases found in the genome of *A. burnettii*, which facilitated the linkage of expressed metabolites to their corresponding synthases (Gilchrist et al., 2020). However, we hypothesized that `synthaser` could also be incorporated into larger scale genome mining pipelines to guide the discovery of novel metabolites.

To test this hypothesis, we constructed a similarity network of ketoacyl synthase (KS) domains in PKS sequences from *Aspergillus* genomes (Figure 6). While the size and complexity of full PKS sequences complicates phylogenetic analyses, KS domains exhibit tight clustering patterns and are a useful proxy for exploring the evolutionary relationship of PKSs (Ziemert et al., 2012). To build the network, we first retrieved any sequences in the NCBI protein database from *Aspergillus* species containing hits to the `cond_enzymes` superfamily (accession: cd00327). This superfamily contains a variety of enzymes catalyzing decarboxylating and non-decarboxylating Claisen-like condensation reactions, covering the spectrum of FASs and PKSs. In total, 2923 sequences were retrieved. Using `synthaser`, we predicted and classified the domain architectures of each retrieved sequence, then extracted the sequence of each KS domain. This consisted of 95 FAS (25 alpha subunit), 292 hybrid PKS-NRPS, 1991 PKS (35 Type III, 221 PR-PKS, 583 NR-PKS and 960 HR-PKS, with 36 furthest annotation Type I, 156 furthest PKS) and 545 thiolases. This dataset was then extended with the PKS, PKS-NRPS and FAS sequences from the MIBiG database analysed above (137 PKS, 31 PKS-NRPS and 24 FAS).

All versus all sequence comparisons of the extracted KS domains were performed using DIAMOND (Buchfink et al., 2021-04), which were then used to construct a similarity network in CytoScape (Morris et al., 2011). Mapping of orthogonal data to sequence similarity networks has been shown to be a powerful approach for revealing themes within biological sequence data (Atkinson et al., 2009). Thus, we mapped domain architecture predictions of the parent PKS sequences generated using `synthaser` to the KS domain network (depicted in Figure 6 by colour scheme) to explore their relationships.

Four distinct subnetworks were formed within the KS domain network, corresponding to the four broad classification groups of KS domain-containing sequences. One subnetwork contained mostly highly-reducing and partially-reducing PKS (Figure 6 top-left), and was clearly separate from, but

related to, another subnetwork consisting of hybrid PKS-NRPS sequences (top-right). Non-reducing PKS formed another clear subnetwork (bottom-left), as did domains from fatty acid synthases (bottom-right).

Perhaps the most powerful aspect of the similarity network approach is its ability to reveal outliers; a characteristic we wished to exploit for the purpose of genome mining for biosynthetic novelty. We were immediately drawn to two specific sequence clusters, which were clearly demarcated from the other members of their respective subnetworks thanks to the architecture-mapped colour scheme (circled in Figure 6). The first cluster falls within the hybrid PKS/NRPS subnetwork and consists of sequences where the NRPS module precedes the PKS module, instead of the typical PKS-NRPS arrangement. Comparatively few NRPS-PKS have been characterised in the literature. The first reported fungal NRPS-PKS was the synthase involved in the biosynthesis of tenuazonic acid in *Magnaporthe oryzae*, TAS1, which has an NRPS module before a PKS module containing only a KS domain (Yun et al., 2015). Later, Cook et al. (2017) characterised the swainsonine BGC, containing the NRPS-PKS SwnK, in several fungal species. More recently, Hai et al. (2020) characterised a NRPS-PKS enzyme, AnATPKS, capable of producing the amino acid derived α-pyrone natural products pyrophen and campyrone B in *Aspergillus niger*. While the cluster contains sequences matching the domain architectures of TAS1, SwnK and AnATPKS, it also includes more variation that could be explored in further studies. Perhaps more interesting was the second cluster, which fell within the non-reducing PKS subnetwork and consisted of non-reducing PKSs with ketoreductase (KR) domains at the N-terminal. As previously outlined, a typical NR-PKS sequence starts with a SAT domain and contains a product template (PT) domain and no reductive (DH, ER, KR) domains (Keller et al., 2005). The sequences within this cluster match this template almost exactly, with SAT domains being substituted with KR domains, making them very abnormal. While outliers such as this could result from incorrect gene annotation (i.e. through fusion of separate coding regions), given the otherwise textbook NR-PKS domain architectures, proximity of the KR domain to the KS domain, and the number of homologues that were identified, we do not believe this to be the case. One biosynthetic hypothesis might be that the KR domain performs similar reductive processing steps as they do in HR-PKS and PR-PKS. Future work is required to further characterise these synthases; however, the discovery of such sequences highlights the value of `synthaser` to genome mining pipelines.

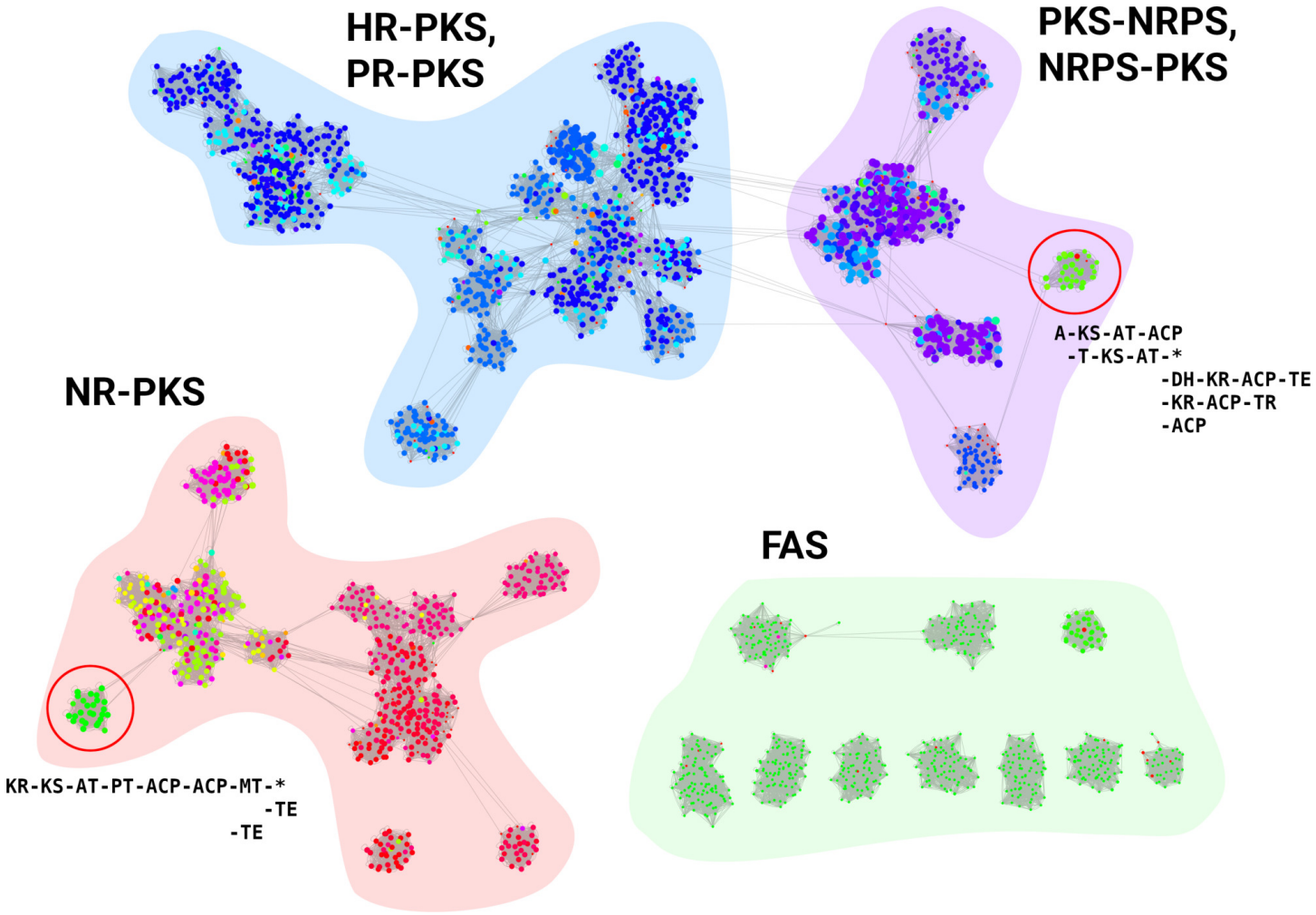

Figure 6: Ketosynthase (KS) domain similarity network of polyketide and fatty acid synthases in *Aspergillus* genomes. Colours of nodes correspond to the full domain architectures of each synthase.

4 Discussion

In this paper we describe `synthaser` a Python-based software package for automatic prediction, classification and visualisation of domain architectures of multi-domain proteins. `synthaser` is capable of fully remote searches using the NCBI's CD-Search tool, which searches query sequences against domain families stored in the conserved domain database (CDD). This is advantageous to other tools which rely on manually curated local profile HMMs for domain searches, as new domains can be added just by referencing the corresponding CDD identifier. Moreover, as the CDD and its sub-databases are continually curated, any improvements to domain profiles will automatically filter through to predictions generated by `synthaser`.

`synthaser` takes a unique approach in that it explicitly searches for specific domain 'islands' during domain architecture prediction (Figure 2). This differs from other tools that select purely for top scoring domains, which can include broader domain families encompassing multiple smaller domains. For instance, a CD-Search of any Type I PKS sequence will contain the domain family PksD, which consists of both KS and AT domains (visible in Figure 2). While this may be preferable when looking at the overall similarity of two proteins, the goal of `synthaser` is to instead predict exact domain architectures, reporting every distinct domain found within each sequence. Thus, `synthaser` can be superior when precise labelling of domains within a sequence is desired.

Another advantage of `synthaser` is its ability to analyse fungal megasynthases at (pan-)genomic scale. There is currently no tool available that can characterise, classify and display the domain architectures of all PKSs and NRPSs either within a genome or across multiple genomes. As we demonstrate above in our similarity network of KS domains, this can form the basis of a genome mining strategy for uncovering unique synthases encoding potentially novel metabolites.

Domain architecture prediction and classification in `synthaser` is controlled by an underlying rule file, which can be freely modified by the user. The rule file consists of three components: classes containing CDD domain families which correspond to domain islands, classification rules, and the rule evaluation hierarchy. This allows for a level of flexibility not possible in tools which rely on manually curated profile HMMs. In addition, we provide a web application (`https://gamcil.github.io/synthaser`) which allows users to easily add, delete or modify domains and classification rules. The default rule file can be loaded for editing by the click of a button, enabling users to tweak it as necessary for their purposes. Moreoever, we can foresee `synthaser` being completely repurposed, via its rule system, for the analysis of other multi-domain protein families, outside of the scope of secondary metabolites.

In the current paper, we extensively demonstrated the use of this system above in the context of fungal secondary metabolite megasynthases; the corresponding rule file is distributed with, and is the namesake of, the tool. Indeed, `synthaser` has already seen use in the analysis of fungal biosynthetic gene clusters in our own group (Li et al., 2019; Gilchrist et al., 2020). We previously outlined strategies for genome mining for BGCs encoding novel small molecules, or those encoding new or improved bioactivities (Gilchrist et al., 2018). Analysis of domain architecture is a key step in uncovering such molecules, as unusual domain architectures could potentially encode unusual chemistry. Such an approach has already been fruitful across several classes of synthase, including PKS, NRPS and terpene synthases (Hang et al., 2017; Baccile et al., 2016; Okada et al., 2016). `synthaser` makes this analysis significantly more convenient, automating both the prediction and classification stages for sequences in batch, without the need for curation of local domain profiles, or maintenance of local profile databases. In addition, `synthaser` provides the `genbank` module, which is capable of directly parsing antiSMASH-generated GenBank format files for megasynthase sequences. If local analysis is desired, `synthaser` does possess the ability to both download profile databases from the NCBI using its `getdb` module, as well as perform local searches using RPS-BLAST (provided it it is installed on the system).

`synthaser` generates comprehensive visual and text result outputs. The visualisations are fully interactive, allowing for changes to sequence size and shape, as well as other convenient functionalities such as the extraction of domain sequences to FASTA files. The text output reports the length and domain architectures of each query sequence, grouped by their classifications. This can also be generated in tabular formats, such that it can be easily imported into spreadsheet software or incorporated into larger bioinformatic pipelines.

The `synthaser` approach does have some caveats. While `synthaser`'s remote search capabilities are its biggest advantage, this also means that an internet connection is required to use the tool. Moreover, certain sequence features indicated by the web CD-Search tool, such as the active sites of certain domains, are not available in `synthaser` results. Perhaps the largest drawback is that the specificity of domain predictions is limited by the domain profiles within the CDD. This has a couple of consequences. Firstly, distinct but functionally related domains generally cannot be separated during a search. For example, acyl carrier protein (ACP) domains in FAS and PKS and peptidyl carrier protein (PCP)/thiolation (T) domains in NRPS, which are structurally and functionally related, hit the same CDD profiles in a CD-Search run. `synthaser` attempts to alleviate this issue by allowing domains to be renamed based on the classification of the protein; in the previous example, `synthaser` will keep

the ACP name within a PKS or FAS, but change it to a T (thiolation) if found in a NRPS. Secondly, certain domains may fail to be detected if the corresponding domain profiles are weakly defined. In these scenarios, `synthaser` will also fail to report the missing domains. However, this is made very clear in the `synthaser` visual output, as large gaps in sequence can be seen where missing domains should be (e.g. the NR-PKS N-terminal SAT domain, as shown in Figure S2), hopefully prompting further investigation. As curation of the CDD continues, and the quality of domain profiles improves, so to will the predictions given by `synthaser`.

In summary, `synthaser` is a powerful tool for the characterisation and classification of multi-domain protein architecture. `synthaser` offers both local and remote search capabilities, which utilise the curated domain profiles in the NCBI's conserved domain database. Its intuitive visualisations, as well as text summaries, allow interesting domain architectures to become immediately obvious. While `synthaser` is distributed with the fungal megasynthase rule set detailed in this paper, the flexibility of the rule system, as well as the easy to use rule generator web application, means `synthaser` could readily be repurposed for the study of any multi-domain protein family. Thus, `synthaser` is a valuable addition to not only the natural products genome mining toolbox, but potentially to any area where multidomain proteins are of interest.

5 Declarations

5.1 Ethics approval and consent to participate

Not applicable.

5.2 Consent for publication

Not applicable.

5.3 Availability of data and materials

`synthaser` is freely available from GitHub (`https://github.com/gamcil/synthaser`) and PyPI (`https://pypi.org/project/synthaser`) under a MIT license. The datasets and scripts generated and analysed during the current study are available in a GitHub repository (`https://github.com/gamcil/synthaser_scripts`).

5.4 Competing interests

Not applicable.

5.5 Funding

CLMG is supported by an Australian Government Research Training Project scholarship. YHC is supported by an Australian Research Council Future Fellowship (FT160100233). This work was funded in part by the Cooperative Research Centres Projects scheme (CRCPFIVE000119).

5.6 Author's contributions

CLMG developed the `synthaser` software, analysed the data and wrote the manuscript. YHC conceived the study and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

5.7 Acknowledgements

Not applicable.

References

Mohd Zeeshan Ansari, Gitanjali Yadav, Rajesh S. Gokhale, and Debasisa Mohanty. NRPS-PKS: A knowledge-based resource for analysis of NRPS-PKS megasynthases. *Nucleic Acids Research*, 32 (WEB SERVER ISS.):405–413, 2004. ISSN 03051048. doi: 10.1093/nar/gkh359.

Holly J. Atkinson, John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLOS ONE*, 4(2):e4345, February 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004345.

Joshua A Baccile, Joseph E Spraker, Henry H Le, Eileen Brandenburger, Christian Gomez, Jin Woo Bok, Juliane Macheleidt, Axel A Brakhage, Dirk Hoffmeister, Nancy P Keller, and Frank C Schroeder. Plant-like biosynthesis of isoquinoline alkaloids in Aspergillus fumigatus. *Nature Chemical Biology*, 12(6):419–424, June 2016. ISSN 1552-4450, 1552-4469. doi: 10.1038/nchembio.2061.

Matthew Bashton and Cyrus Chothia. The Generation of New Protein Functions by the Combination of Domains. *Structure*, 15(1):85–99, January 2007. ISSN 0969-2126. doi: 10.1016/j.str.2006.11.009.

Joachim Beck, Sabine Ripka, Axel Siegner, Emil Schiltz, and Eckhart Schweizer. The multifunctional 6-methylsalicylic acid synthase gene of Penicillium patulum. *European Journal of Biochemistry*, 192(2):487–498, 1990. ISSN 1432-1033. doi: 10.1111/j.1432-1033.1990.tb19252.x. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1990.tb19252.x.

Kai Blin, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, pages 1–7, apr 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz310. URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz310/5481154.

Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa977.

Erich Bornberg-Bauer and M Mar Albà. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23(3):459–466, June 2013. ISSN 0959-440X. doi: 10.1016/j.sbi.2013.02.012.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 Data-Driven Documents. *IEEE T Vis Comput Gr*, 17(12):2301–2309, December 2011. ISSN 1077-2626. doi: 10.1109/TVCG.2011.185.

Benjamin Buchfink, Klaus Reuter, and Hajk-Georg Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, 2021-04. ISSN 1548-7105. doi: 10.1038/s41592-021-01101-x. URL https://www.nature.com/articles/s41592-021-01101-x.

Ralph A. Cacho, Yi Tang, and Yit-Heng Heng Chooi. Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Frontiers in Microbiology*, 5(JAN):1–16, jan 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00774. URL http://www.frontiersin.org/Journal/Abstract.aspx?s=677{&}name=microbial{_}physiology{_}and{_}metabolism{&}ART{_}DOI=10.3389/fmicb.2014.00774http://journal.frontiersin.org/article/10.3389/fmicb.2014.00774/abstract.

Perng-Kuang Chang, Jeffrey W. Cary, Jiujiang Yu, Deepak Bhatnagar, and Thomas E. Cleveland. The Aspergillus parasiticus polyketide synthase gene pksA, a homolog of Aspergillus nidulans wA, is required for aflatoxin B1 biosynthesis. *Molecular and General Genetics MGG*, 248(3):270–277, August 1995. ISSN 1432-1874. doi: 10.1007/BF02191593.

Yit-Heng Chooi and Yi Tang. Navigating the Fungal Polyketide Chemical Space: From Genes to Molecules. *The Journal of Organic Chemistry*, 77(22):9933–9953, nov 2012. ISSN 0022-3263. doi: 10.1021/jo301592k. URL `http://pubs.acs.org/doi/abs/10.1021/jo301592k`.

Yit-Heng Heng Chooi, Christian Krill, Russell A. Barrow, Shasha Chen, Robert Trengove, Richard P. Oliver, and Peter S. Solomon. An In Planta-Expressed Polyketide Synthase Produces (R)-Mellein in the Wheat Pathogen Parastagonospora nodorum. *Applied and Environmental Microbiology*, 81(1): 177–186, 2015. ISSN 10985336. doi: 10.1128/aem.02745-14. URL `http://aem.asm.org/content/81/1/177.abstract`.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp163.

Daniel Cook, Bruno G G Donzelli, Rebecca Creamer, Deana L Baucom, Dale R Gardner, Juan Pan, Neil Moore, Stuart B Krasnoff, Jerzy W Jaromczyk, and Christopher L Schardl. Swainsonine Biosynthesis Genes in Diverse Symbiotic and Pathogenic Fungi. *G3 Genes|Genomes|Genetics*, 7(6): 1791–1797, June 2017. ISSN 2160-1836. doi: 10.1534/g3.117.041384.

Jason M. Crawford, Blair C. R. Dancy, Eric A. Hill, Daniel W. Udwary, and Craig A. Townsend. Identification of a starter unit acyl-carrier protein transacylase domain in an iterative type I polyketide synthase. *Proceedings of the National Academy of Sciences*, 103(45):16728–16733, November 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0604112103.

Ronald P. de Vries, Robert Riley, Ad Wiebenga, Guillermo Aguilar-Osorio, Sotiris Amillis, Cristiane Akemi Uchima, Gregor Anderluh, Mojtaba Asadollahi, Marion Askin, Kerrie Barry, Evy Battaglia, Özgür Bayram, Tiziano Benocci, Susanna A. Braus-Stromeyer, Camila Caldana, David Cánovas, Gustavo C. Cerqueira, Fusheng Chen, Wanping Chen, Cindy Choi, Alicia Clum, Renato Augusto Corrêa dos Santos, André Ricardo de Lima Damásio, George Diallinas, Tamás Emri, Erzsébet Fekete, Michel Flipphi, Susanne Freyberg, Antonia Gallo, Christos Gournas,

Rob Habgood, Matthieu Hainaut, María Laura Harispe, Bernard Henrissat, Kristiina S. Hildén, Ryan Hope, Abeer Hossain, Eugenia Karabika, Levente Karaffa, Zsolt Karányi, Nada Kraševec, Alan Kuo, Harald Kusch, Kurt LaButti, Ellen L. Lagendijk, Alla Lapidus, Anthony Levasseur, Erika Lindquist, Anna Lipzen, Antonio F. Logrieco, Andrew MacCabe, Miia R. Mäkelä, Iran Malavazi, Petter Melin, Vera Meyer, Natalia Mielnichuk, Márton Miskei, Ákos P. Molnár, Giuseppina Mulé, Chew Yee Ngan, Margarita Orejas, Erzsébet Orosz, Jean Paul Ouedraogo, Karin M. Overkamp, Hee-Soo Park, Giancarlo Perrone, Francois Piumi, Peter J. Punt, Arthur F. J. Ram, Ana Ramón, Stefan Rauscher, Eric Record, Diego Mauricio Riaño-Pachón, Vincent Robert, Julian Röhrig, Roberto Ruller, Asaf Salamov, Nadhira S. Salih, Rob A. Samson, Erzsébet Sándor, Manuel Sanguinetti, Tabea Schütze, Kristina Sepčić, Ekaterina Shelest, Gavin Sherlock, Vicky Sophianopoulou, Fabio M. Squina, Hui Sun, Antonia Susca, Richard B. Todd, Adrian Tsang, Shiela E. Unkles, Nathalie van de Wiele, Diana van Rossen-Uffink, Juliana Velasco de Castro Oliveira, Tammi C. Vesth, Jaap Visser, Jae-Hyuk Yu, Miaomiao Zhou, Mikael R. Andersen, David B. Archer, Scott E. Baker, Isabelle Benoit, Axel A. Brakhage, Gerhard H. Braus, Reinhard Fischer, Jens C. Frisvad, Gustavo H. Goldman, Jos Houbraken, Berl Oakley, István Pócsi, Claudio Scazzocchio, Bernhard Seiboth, Patricia A. VanKuyk, Jennifer Wortman, Paul S. Dyer, and Igor V. Grigoriev. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. *Genome Biology*, 18(1):28, dec 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1151-0. URL `http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1151-0https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty856/5124275`.

Lewis Y. Geer, Michael Domrachev, David J. Lipman, and Stephen H. Bryant. CDART: Protein homology by domain architecture. *Genome Research*, 12(10):1619–1623, 2002. ISSN 10889051. doi: 10.1101/gr.278202.

Cameron L. M. Gilchrist, Hang Li, and Yit-Heng Chooi. Panning for gold in mould: Can we increase the odds for fungal genome mining? *Organic & Biomolecular Chemistry*, 16(10):1620–1626, 2018. ISSN 1477-0520, 1477-0539. doi: 10.1039/C7OB03127K.

Cameron L. M. Gilchrist, Heather J. Lacey, Daniel Vuong, John I. Pitt, Lene Lange, Ernest Lacey, Bo Pilgaard, Yit-Heng Chooi, and Andrew M. Piggott. Comprehensive chemotaxonomic and genomic

profiling of a biosynthetically talented Australian fungus, Aspergillus burnettii sp. nov. *Fungal Genetics and Biology*, 143:103435, October 2020. ISSN 1087-1845. doi: 10.1016/j.fgb.2020.103435.

Vadim M. Gumerov and Igor B. Zhulin. TREND: A platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Research*, 48(W1):W72–W76, July 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa243.

Chun-Jun Guo, Benjamin P. Knox, Yi-Ming Chiang, Hsien-Chun Lo, James F. Sanchez, Kuan-Han Lee, Berl R. Oakley, Kenneth S. Bruno, and Clay C. C. Wang. Molecular Genetic Characterization of a Cluster in A. terreus for Biosynthesis of the Meroterpenoid Terretonin. *Organic Letters*, 14(22): 5684–5687, November 2012. ISSN 1523-7060. doi: 10.1021/ol302682z.

Yang Hai, Arthur Huang, and Yi Tang. Biosynthesis of Amino Acid Derived α-Pyrones by an NRPS–NRPKS Hybrid Megasynthetase in Fungi. *Journal of Natural Products*, 83(3):593–600, March 2020. ISSN 0163-3864. doi: 10.1021/acs.jnatprod.9b00989.

Leibniz Hang, Man-Cheng Tang, Colin J. B. Harvey, Claire G. Page, Jian Li, Yiu-Sun Hung, Nicholas Liu, Maureen E. Hillenmeyer, and Yi Tang. Reversible Product Release and Recapture by a Fungal Polyketide Synthase Using a Carnitine Acyltransferase Domain. *Angewandte Chemie International Edition*, 56(32):9556–9560, 2017. ISSN 1521-3773. doi: 10.1002/anie.201705237.

Makoto Hashimoto, Takamasa Nonaka, and Isao Fujii. Fungal type III polyketide synthases. *Nat. Prod. Rep.*, 31(10):1306–1317, 2014. ISSN 0265-0568. doi: 10.1039/C4NP00096J. URL `http://xlink.rsc.org/?DOI=C4NP00096J`.

Richard J. Heath and Charles O. Rock. The Claisen condensation in biology. *Natural Product Reports*, 19(5):581–596, September 2002. ISSN 1460-4752. doi: 10.1039/B110221B.

Christian Hertweck. The Biosynthetic Logic of Polyketide Diversity. *Angewandte Chemie International Edition*, 48(26):4688–4716, jun 2009. ISSN 14337851. doi: 10.1002/anie.200806121. URL `http://doi.wiley.com/10.1002/anie.200806121`.

C. Richard Hutchinson, Jonathan Kennedy, Cheonseok Park, Steven Kendrew, Karine Auclair, and John Vederas. Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. *Antonie van Leeuwenhoek*, 78(3):287–295, December 2000. ISSN 1572-9699. doi: 10.1023/A:1010294330190.

J. Jin, X. Xie, C. Chen, J. G. Park, C. Stark, D. A. James, M. Olhovsky, R. Linding, Y. Mao, and T. Pawson. Eukaryotic Protein Domains as Functional Units of Cellular Evolution. *Science Signaling*, 2(98):ra76–ra76, November 2009. ISSN 1937-9145. doi: 10.1126/scisignal.2000546.

Jonathan Kans. *Entrez Direct: E-Utilities on the UNIX Command Line*. National Center for Biotechnology Information (US), November 2019.

Ken Kasahara, Takanori Miyamoto, Takashi Fujimoto, Hiroki Oguri, Tetsuo Tokiwano, Hideaki Oikawa, Yutaka Ebizuka, and Isao Fujii. Solanapyrone synthase, a possible Diels-Alderase and iterative type I polyketide synthase encoded in a biosynthetic gene cluster from Alternaria solani. *ChemBioChem*, 11(9):1245–1252, may 2010. ISSN 14397633. doi: 10.1002/cbic.201000173. URL `http://doi.wiley.com/10.1002/cbic.201000173`.

Satria A Kautsar, Kai Blin, Simon Shaw, Jorge C Navarro-Muñoz, Barbara R Terlouw, Justin J J van der Hooft, Jeffrey A van Santen, Vittorio Tracanna, Hernando G Suarez Duran, Victòria Pascal Andreu, Nelly Selem-Mojica, Mohammad Alanjary, Serina L Robinson, George Lund, Samuel C Epstein, Ashley C Sisto, Louise K Charkoudian, Jérôme Collemare, Roger G Linington, Tilmann Weber, and Marnix H Medema. MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*, page gkz882, October 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz882.

Nancy P Keller. Translating biosynthetic gene clusters into fungal armor and weaponry. *Nature Chemical Biology*, 11(9):671–677, September 2015. ISSN 1552-4450, 1552-4469. doi: 10.1038/nchembio.1897.

Nancy P Keller, Geoffrey Turner, and Joan W Bennett. Fungal secondary metabolism - from biochemistry to genomics. *Nat Rev Micro*, 3(12):937–947, 2005. URL `http://dx.doi.org/10.1038/nrmicro1286`.

Shradha Khater, Money Gupta, Priyesh Agrawal, Neetu Sain, Jyoti Prava, Priya Gupta, Mansi Grover, Narendra Kumar, and Debasisa Mohanty. SBSPKSv2: Structure-based sequence analysis of polyketide synthases and non-ribosomal peptide synthetases. *Nucleic Acids Research*, 45(W1): W72–W79, 2017. ISSN 13624962. doi: 10.1093/nar/gkx344.

Inge Kjærbølling, Tammi C. Vesth, Jens C. Frisvad, Jane L. Nybo, Sebastian Theobald, Alan Kuo, Paul Bowyer, Yudai Matsuda, Stephen Mondo, Ellen K. Lyhne, Martin E. Kogle, Alicia Clum, Anna Lipzen, Asaf Salamov, Chew Yee Ngan, Chris Daum, Jennifer Chiniquy, Kerrie Barry, Kurt LaButti, Sajeet Haridas, Blake A. Simmons, Jon K. Magnuson, Uffe H. Mortensen, Thomas O.

Larsen, Igor V. Grigoriev, Scott E. Baker, and Mikael R. Andersen. Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proceedings of the National Academy of Sciences*, 115(4):E753–E761, January 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1715954115.

Scott Kroken, N. Louise Glass, John W. Taylor, O. C. Yoder, and B. Gillian Turgeon. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proceedings of the National Academy of Sciences*, 100(26):15670–15675, December 2003. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2532165100.

Fumitaka Kudo, Yasunori Matsuura, Takaaki Hayashi, Masayuki Fukushima, and Tadashi Eguchi. Genome mining of the sordarin biosynthetic gene cluster from Sordaria araneosa Cain ATCC 36386: Characterization of cycloaraneosene synthase and GDP-6-deoxyaltrose transferase. *The Journal of Antibiotics*, 69(7):541–548, July 2016. ISSN 1881-1469. doi: 10.1038/ja.2016.40.

Heather J. Lacey, Cameron L. M. Gilchrist, Andrew Crombie, John A. Kalaitzis, Daniel Vuong, Peter J. Rutledge, Peter Turner, John I. Pitt, Ernest Lacey, Yit-Heng Chooi, and Andrew M. Piggott. Nanangenines: Drimane sesquiterpenoids as the dominant metabolite cohort of a novel Australian fungus, Aspergillus nanangensis. *Beilstein Journal of Organic Chemistry*, 15(1):2631–2643, November 2019. ISSN 1860-5397. doi: 10.3762/bjoc.15.256.

Ivica Letunic, Supriya Khedkar, and Peer Bork. SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Research*, 49(D1):D458–D460, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa937.

Michael Levitt. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106 (27):11079–11084, July 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0905029106.

Hang Li, Cameron L.M. Gilchrist, Heather J. Lacey, Andrew Crombie, Daniel Vuong, John I. Pitt, Ernest Lacey, Yit Heng Chooi, and Andrew M. Piggott. Discovery and Heterologous Biosynthesis of the Burnettramic Acids: Rare PKS-NRPS-Derived Bolaamphiphilic Pyrrolizidinediones from an Australian Fungus, Aspergillus burnettii. *Organic Letters*, 21(5):1287–1291, 2019. ISSN 15237052. doi: 10.1021/acs.orglett.8b04042.

Hang Li, Cameron L. M. Gilchrist, Chin-Soon Phan, Heather J. Lacey, Daniel Vuong, Stephen A. Moggach, Ernest Lacey, Andrew M. Piggott, and Yit-Heng Chooi. Biosynthesis of a New Ben-

zazepine Alkaloid Nanangelenin A from Aspergillus nanangensis Involves an Unusual l-Kynurenine-Incorporating NRPS Catalyzing Regioselective Lactamization. *Journal of the American Chemical Society*, 142(15):7145–7152, April 2020a. ISSN 0002-7863. doi: 10.1021/jacs.0c01605.

Hang Li, Haochen Wei, Jinyu Hu, Ernest Lacey, Alexandre N. Sobolev, Keith A. Stubbs, Peter S. Solomon, and Yit-Heng Chooi. Genomics-Driven Discovery of Phytotoxic Cytochalasans Involved in the Virulence of the Wheat Pathogen Parastagonospora nodorum. *ACS Chemical Biology*, 15(1): 226–233, 2020b. ISSN 1554-8929. doi: 10.1021/acschembio.9b00791. URL `https://doi.org/10.1021/acschembio.9b00791`.

Hang Li, Alastair E. Lacey, Si Shu, John A. Kalaitzis, Daniel Vuong, Andrew Crombie, Jinyu Hu, Cameron L. M. Gilchrist, Ernest Lacey, Andrew M. Piggott, and Yit-Heng Chooi. Hancockiamides: Phenylpropanoid piperazines from Aspergillus hancockii are biosynthesised by a versatile dual single-module NRPS pathway. *Organic & Biomolecular Chemistry*, 19(3):587–595, January 2021. ISSN 1477-0539. doi: 10.1039/D0OB02243H.

Shennan Lu, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Mingzhang Yang, Dachuan Zhang, Chanjuan Zheng, Christopher J. Lanczycki, and Aron Marchler-Bauer. CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Research*, 48(D1):D265–D268, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz991.

Suzanne M Ma, Jesse W H Li, Jin W Choi, Hui Zhou, K K Michael Lee, Vijayalakshmi A Moorthie, Xinkai Xie, James T Kealey, Nancy A Da Silva, John C Vederas, and Yi Tang. Complete Reconstitution of a Highly-Reducing Iterative Polyketide Synthase. *Science (New York, N.Y.)*, 326 (5952):589–592, 2009. doi: 10.1126/science.1175602. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875069/`.

Timm Maier, Marc Leibundgut, Daniel Boehringer, and Nenad Ban. Structure and function of eukaryotic fatty acid synthases. *Quarterly Reviews of Biophysics*, 43(3):373–422, August 2010. ISSN 0033-5835, 1469-8994. doi: 10.1017/S0033583510000156.

A. Marchler-Bauer and S. H. Bryant. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, 32(Web Server):W327–W331, July 2004. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkh454.

A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, 39(Database): D225–D229, January 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq1189.

Aron Marchler-Bauer and Stephen H. Bryant. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, 32(WEB SERVER ISS.):327–331, 2004. ISSN 03051048. doi: 10.1093/nar/gkh454.

Aron Marchler-Bauer, Anna R. Panchenko, Benjamin A. Shoemaker, Paul A. Thiessen, Lewis Y. Geer, and Stephen H. Bryant. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, 30(1):281–283, January 2002. ISSN 0305-1048. doi: 10.1093/nar/30.1.281.

Aron Marchler-Bauer, Yu Bo, Lianyi Han, Jane He, Christopher J. Lanczycki, Shennan Lu, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, Lewis Y. Geer, and Stephen H. Bryant. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, 45 (D1):D200–D203, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1129.

Aron Marchler-Bauer, Yu Bo, Lianyi Han, Jane He, Christopher J. Lanczycki, Shennan Lu, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, Lewis Y. Geer, and Stephen H. Bryant. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, 45 (D1):D200–D203, January 2017. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkw1129.

Joseph A. Marsh and Sarah A. Teichmann. How do proteins gain new domains? *Genome Biology*, 11 (7):126, July 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-7-126.

Yudai Matsuda, Taiki Iwabuchi, Takayuki Fujimoto, Takayoshi Awakawa, Yu Nakashima, Takahiro Mori, Huiping Zhang, Fumiaki Hayashi, and Ikuro Abe. Discovery of Key Dioxygenases that Diverged the Paraherquonin and Acetoxydehydroaustin Pathways in Penicillium brasilianum. *Journal of*

the American Chemical Society, 138(38):12671–12677, September 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b08424.

Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene de Bruijn, Yit-Heng Chooi, Jan Claesen, R Cameron Coates, Pablo Cruz-Morales, Srikanth Duddela, Stephanie Düsterhus, Daniel J Edwards, David P Fewer, Neha Garg, Christoph Geiger, Juan Pablo Gomez-Escribano, Anja Greule, Michalis Hadjithomas, Anthony S Haines, Eric J N Helfrich, Matthew L Hillwig, Keishi Ishida, Adam C Jones, Carla S Jones, Katrin Jungmann, Carsten Kegler, Hyun Uk Kim, Peter Kötter, Daniel Krug, Joleen Masschelein, Alexey V Melnik, Simone M Mantovani, Emily A Monroe, Marcus Moore, Nathan Moss, Hans-Wilhelm Nützmann, Guohui Pan, Amrita Pati, Daniel Petras, F Jerry Reen, Federico Rosconi, Zhe Rui, Zhenhua Tian, Nicholas J Tobias, Yuta Tsunematsu, Philipp Wiemann, Elizabeth Wyckoff, Xiaohui Yan, Grace Yim, Fengan Yu, Yunchang Xie, Bertrand Aigle, Alexander K Apel, Carl J Balibar, Emily P Balskus, Francisco Barona-Gómez, Andreas Bechthold, Helge B Bode, Rainer Borriss, Sean F Brady, Axel A Brakhage, Patrick Caffrey, Yi-Qiang Cheng, Jon Clardy, Russell J Cox, René De Mot, Stefano Donadio, Mohamed S Donia, Wilfred A van der Donk, Pieter C Dorrestein, Sean Doyle, Arnold J M Driessen, Monika Ehling-Schulz, Karl-Dieter Entian, Michael A Fischbach, Lena Gerwick, William H Gerwick, Harald Gross, Bertolt Gust, Christian Hertweck, Monica Höfte, Susan E Jensen, Jianhua Ju, Leonard Katz, Leonard Kaysser, Jonathan L Klassen, Nancy P Keller, Jan Kormanec, Oscar P Kuipers, Tomohisa Kuzuyama, Nikos C Kyrpides, Hyung-Jin Kwon, Sylvie Lautru, Rob Lavigne, Chia Y Lee, Bai Linquan, Xinyu Liu, Wen Liu, Andriy Luzhetskyy, Taifo Mahmud, Yvonne Mast, Carmen Méndez, Mikko Metsä-Ketelä, Jason Micklefield, Douglas A Mitchell, Bradley S Moore, Leonilde M Moreira, Rolf Müller, Brett A Neilan, Markus Nett, Jens Nielsen, Fergal O'Gara, Hideaki Oikawa, Anne Osbourn, Marcia S Osburne, Bohdan Ostash, Shelley M Payne, Jean-Luc Pernodet, Miroslav Petricek, Jörn Piel, Olivier Ploux, Jos M Raaijmakers, José A Salas, Esther K Schmitt, Barry Scott, Ryan F Seipke, Ben Shen, David H Sherman, Kaarina Sivonen, Michael J Smanski, Margherita Sosio, Evi Stegmann, Roderich D Süssmuth, Kapil Tahlan, Christopher M Thomas, Yi Tang, Andrew W Truman, Muriel Viaud, Jonathan D Walton, Christopher T Walsh, Tilmann Weber, Gilles P van Wezel, Barrie Wilkinson, Joanne M Willey, Wolfgang Wohlleben, Gerard D Wright, Nadine Ziemert, Changsheng Zhang, Sergey B Zotchev, Rainer Breitling, Eriko Takano, and Frank Oliver Glöckner. Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*, 11(9):625–631, sep 2015. ISSN 1552-4450. doi: 10.1038/nchembio.1890. URL `http://www.ncbi.nlm.nih.gov/pubmed/26284661http://www.nature.com/doifinder/10.`

1038/nchembio.1890http://www.nature.com/articles/nchembio.1890.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1): D412–D419, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa913.

Andrew D. Moore, Åsa K. Björklund, Diana Ekman, Erich Bornberg-Bauer, and Arne Elofsson. Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences*, 33(9):444–451, September 2008. ISSN 0968-0004. doi: 10.1016/j.tibs.2008.05.008.

John H. Morris, Leonard Apeltsin, Aaron M. Newman, Jan Baumbach, Tobias Wittkop, Gang Su, Gary D. Bader, and Thomas E. Ferrin. ClusterMaker: A multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12(1):436, 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-436. URL http://www.biomedcentral.com/1471-2105/12/436.

Jorge Carlos Navarro-Muñoz and Jérôme Collemare. Evolutionary Histories of Type III Polyketide Synthases in Fungi. *Frontiers in Microbiology*, 10, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2019.03018. URL https://www.frontiersin.org/articles/10.3389/fmicb.2019.03018/full.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 42(D1):D7–D17, January 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1146.

David J Newman and Gordon M Cragg. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661, 2016. doi: 10.1021/acs.jnatprod.5b01055. URL http://dx.doi.org/10.1021/acs.jnatprod.5b01055.

Masahiro Okada, Yudai Matsuda, Takaaki Mitsuhashi, Shotaro Hoshino, Takahiro Mori, Kazuya Nakagawa, Zhiyang Quan, Bin Qin, Huiping Zhang, Fumiaki Hayashi, Hiroshi Kawaide, and Ikuro Abe. Genome-Based Discovery of an Unprecedented Cyclization Mode in Fungal Sesterterpenoid Biosynthesis. *Journal of the American Chemical Society*, 138(31):10011–10018, August 2016. ISSN 0002-7863, 1520-5126. doi: 10.1021/jacs.6b05799.

Marc Ostermeier and Stephen J. Benkovic. Evolution of protein function by Domain swapping. In *Advances in Protein Chemistry*, volume 55 of *Evolutionary Protein Design*, pages 29–77. Academic Press, January 2001. doi: 10.1016/S0065-3233(01)55002-0.

Sophie Pasek, Jean-Loup Risler, and Pierre Brézellec. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, 22(12):1418–1423, June 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl135.

Kangjian Qiao, Yit-Heng Chooi, and Yi Tang. Identification and engineering of the cytochalasin gene cluster from Aspergillus clavatus NRRL 1. *Metabolic Engineering*, 13(6):723–732, 2011. ISSN 10967176. doi: 10.1016/j.ymben.2011.09.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S109671761100098X`.

Juan F. Rojas-Aedo, Carlos Gil-Durán, Abdiel Del-Cid, Natalia Valdés, Pamela Álamos, Inmaculada Vaca, Ramón O. García-Rico, Gloria Levicán, Mario Tello, and Renato Chávez. The Biosynthetic Gene Cluster for Andrastin A in Penicillium roqueforti. *Frontiers in Microbiology*, 8, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.00813.

Sheila Ruswandi, Keiji Kitani, Kazuya Akimitsu, Takashi Tsuge, Tomonori Shiraishi, and Mikihiro Yamamoto. Structural analysis of cosmid clone pcAFT-2 carrying AFT10-1 encoding an acyl-CoA dehydrogenase involved in AF-toxin production in the strawberry pathotype of Alternaria alternata. *Journal of General Plant Pathology*, 71(2):107–116, April 2005. ISSN 1345-2630, 1610-739X. doi: 10.1007/s10327-004-0170-3.

Paul Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, nov 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303. URL `http://ci.nii.ac.jp/naid/110001910481/http://www.genome.org/cgi/doi/10.1101/gr.1239303`.

Christian J. A. Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(D1):D344–D347, January 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1067.

Beate Teichmann, Lidan Liu, Kay Oliver Schink, and Michael Bölker. Activation of the Ustilagic Acid Biosynthesis Gene Cluster in Ustilago maydis by the C2H2 Zinc Finger Transcription Factor Rua1. *Applied and Environmental Microbiology*, 76(8):2633–2640, April 2010. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.02211-09.

Elisha Thynne, Oliver L Mead, Yit-Heng Chooi, Megan C McDonald, and Peter S Solomon. Acquisition and Loss of Secondary Metabolites Shaped the Evolutionary Path of Three Emerging Phytopathogens

of Wheat. *Genome Biology and Evolution*, 11(3):890–905, 2019. ISSN 1759-6653. doi: 10.1093/gbe/evz037. URL https://academic.oup.com/gbe/article/11/3/890/5355066.

T.C. Vesth, J.L. Nybo, S. Theobald, J.C. Frisvad, T.O. Larsen, K.F. Nielsen, J.B. Hoof, J. Brandl, A. Salamov, R. Riley, J.M. Gladden, P. Phatale, M.T. Nielsen, E.K. Lyhne, M.E. Kogle, K. Strasser, E. McDonnell, K. Barry, A. Clum, C. Chen, M. Nolan, L. Sandor, A. Kuo, A. Lipzen, M. Hainaut, E. Drula, A. Tsang, J.K. Magnuson, B. Henrissat, A. Wiebenga, B.A. Simmons, M.R. Mäkelä, R.P. de Vries, I.V. Grigoriev, U.H. Mortensen, S.E. Baker, and M.R. Andersen. Investigation of inter- and intra-species variation through genome sequencing of Aspergillus section Nigri. *Nature Genetics*, page in press, 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0246-1.

Christine Vogel, Matthew Bashton, Nicola D Kerrison, Cyrus Chothia, and Sarah A Teichmann. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, 14(2):208–216, April 2004. ISSN 0959-440X. doi: 10.1016/j.sbi.2004.03.011.

Gitanjali Yadav, Rajesh S. Gokhale, and Debasisa Mohanty. SEARCHPKS: A program for detection and analysis of polyketide synthase domains. *Nucleic Acids Research*, 31(13):3654–3658, 2003. ISSN 03051048. doi: 10.1093/nar/gkg607.

Choong-Soo Yun, Takayuki Motoyama, and Hiroyuki Osada. Biosynthesis of the mycotoxin tenuazonic acid by a fungal NRPS–PKS hybrid enzyme. *Nature Communications*, 6(1):8758, October 2015. ISSN 2041-1723. doi: 10.1038/ncomms9758.

Nadine Ziemert, Sheila Podell, Kevin Penn, Jonathan H. Badger, Eric Allen, and Paul R. Jensen. The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE*, 7(3):1–9, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0034064.