RESEARCH ARTICLE

Knowledge-guided data mining on the standardized architecture of NRPS: Subtypes, novel motifs, and sequence entanglements

Ruolin He[1], **Jinyu Zhang**[2,3], **Yuanzhe Shao**[4], **Shaohua Gu**[1,4], **Chen Song**[1,4], **Long Qian**[1], **Wen-Bing Yin**[2,3]*, **Zhiyuan Li**[1,4]*

1 Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China, **2** State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, PR China, **3** Savaid Medical School, University of Chinese Academy of Sciences, Beijing, PR China, **4** Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

* yinwb@im.ac.cn (WBY); zhiyuanli@pku.edu.cn (ZL)

OPEN ACCESS

Citation: He R, Zhang J, Shao Y, Gu S, Song C, Qian L, et al. (2023) Knowledge-guided data mining on the standardized architecture of NRPS: Subtypes, novel motifs, and sequence entanglements. PLoS Comput Biol 19(5): e1011100. https://doi.org/10.1371/journal.pcbi.1011100

Editor: Turkan Haliloglu, Bogazici University, TURKEY

Received: August 17, 2022

Accepted: April 12, 2023

Published: May 15, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1011100

Data Availability Statement: NRPS Motif Finder is an online platform for construction of NRPS motif-and-intermotif architecture (http://www.

Abstract

Non-ribosomal peptide synthetase (NRPS) is a diverse family of biosynthetic enzymes for the assembly of bioactive peptides. Despite advances in microbial sequencing, the lack of a consistent standard for annotating NRPS domains and modules has made data-driven discoveries challenging. To address this, we introduced a standardized architecture for NRPS, by using known conserved motifs to partition typical domains. This motif-and-intermotif standardization allowed for systematic evaluations of sequence properties from a large number of NRPS pathways, resulting in the most comprehensive cross-kingdom C domain subtype classifications to date, as well as the discovery and experimental validation of novel conserved motifs with functional significance. Furthermore, our coevolution analysis revealed important barriers associated with re-engineering NRPSs and uncovered the entanglement between phylogeny and substrate specificity in NRPS sequences. Our findings provide a comprehensive and statistically insightful analysis of NRPS sequences, opening avenues for future data-driven discoveries.

Author summary

NRPS, a gigantic enzyme that produces diverse microbial secondary metabolites, provides a rich source for important medical products such as antibiotics and antitumor agents. Despite the extensive knowledge gained about its structure and a large amount of sequencing data available, the frequent failure of re-engineering NRPS in synthetic biology highlights that much still needs to be discovered. In this work, we applied existing knowledge to data mining of NRPS sequences, using well-known conserved motifs to partition NRPS sequences into motif-intermotif architectures. This standardization allows for integrating large amounts of sequences from different sources, providing a comprehensive overview of NRPSs across different kingdoms. Our findings included new C domain subtypes, novel conserved motifs with implications in structural flexibility, and

bdainformatics.org/page?type=NRPSMotifFinder). The source codes data can be found in S4 and S5 Files.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2021YFF1200500 to ZL), National Natural Science Foundation of China (No. 32071255 to ZL, No. 42107140 to SG), Clinical Medicine Plus X - Young Scholars Project, Peking University, the Fundamental Research Funds for the Central Universities (No. PKU2022LCXQ009 to ZL), The Biological Resources Program, Chinese Academy of Sciences (KFJ-BRP-009-005 to WBY), Key Research Program of Frontier Sciences, Chinese Academy of Sciences (ZDBS-LY-SM016 to WBY) and National Postdoctoral Program for Innovative Talents (No. BX2021012 to SG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

insights into why NRPSs are so difficult to re-engineer. To facilitate researchers in related fields, we constructed an online platform, “NRPS Motif Finder”, for parsing the motif-and-intermotif architecture and C domain subtype classification. We believe that this knowledge-guided approach not only advances our understanding of NRPSs but also provides a useful methodology for data mining in large-scale biological sequences.

Introduction

Non-ribosomal peptide (NRP) synthesized by non-ribosomal peptide synthetase (NRPS) are a diverse family of natural products widespread in fungi and bacteria [1–5]. According to the BiG-FAM database (as of 2022/8), NRPS is the most prevalent class (29.30%) out of the 1,060,938 biosynthetic gene clusters (BGCs) detected from 170,585 bacterial genomes, and 46.81% out of the 123,939 BGCs from 5,588 fungal genomes [6]. Microbes utilize NRP for various niche-construction activities, such as communication [7,8], defense [9–13], and resource-scavenging [14]. Diverse NRPs serve as reservoirs for pharmaceutical innovations, with anti-cancer, anti-bacterial, anti-fungal, anti-viral, cytotoxic, and immunosuppressive activities [15–18]. To date, nearly thirty NRP structural medicines have been approved for commercialization [15], including actinomycin [19], penicillin [20,21], and vancomycin [22].

Many re-engineering efforts have been inspired by the pharmacological potentials of NRPS and its modular architecture re-engineering [23,24]. NRPS synthesizes peptides in an assembly-line fashion using its repeating module units. Each module is generally comprised of three key domains, including the adenylation (A) domain, which recognizes the substrate and activates it as its aminoacyl adenylate; the condensation (C) domain, which catalyzes amide bond formation between two substrates; and the thiolation (T) domain, which shuttles the substrates and peptide intermediates between catalytic domains. Other optional domains are also important, such as the thioesterase (TE) domain or terminal condensation-like (C_T) domain [25], which cyclize or release peptides from the NRPS and generally occurs at the end of an NRPS; epimerization (E) domain, which catalyzes the conversion of L-amino acids substrate to D-amino acids and usually follows the C domain in case it occurs.

In order to create novel products, many synthetic efforts have been made to re-arrange these building blocks of NRPSs. For example, due to the central roles of A domains in substrate recognition and activation, early works focused on the A domain, including A domain substitution, reprogramming A domain substrate coding residues, and A subdomain substitution [23]. Later, considering the possible roles of C domains in substrate selectivity, researchers attempted to re-engineer catalytic domains together with the A domain, such as substituting a complete module, or the C+A domain and T+C+A domain [23]. Recently, Bozhüyük et al. advanced the exchange unit concept and recombined NRPS at a fusion point between the C and A domains which they named as “C-A linker” [26]. They also found another cutting point inside the C domain [27]. Calcott et al. also attempted to recombine at another cutting point between the C and A domains [28], almost at the start of the C-A linker defined by Bozhüyük et al. However, in most cases, re-engineering often substantially reduces the product yield. Furthermore, conversions of these units were typically performed between two substrates with similar chemical properties, based on well-studied NRPS systems with a small number of sequences [29,30]. A systematic understanding of the rational design of NRPS has yet to be developed.

The massive expansion of microbial sequencing data has created opportunities for a comprehensive understanding of NRPSs [31]. As of October 2021, the Integrated Microbial

Genomes Atlas of Biosynthetic gene Clusters (IMG-ABC) v5 database contains a total of 411,027 BGCs based on sequence annotation [32], and Minimum Information about a Biosynthetic Gene Cluster (MiBiG) v2 database contains 1926 BGCs confirmed to be biologically active [33]. Together with the accelerated expansion of databases, bioinformatics platforms that specifically predict and annotate the secondary metabolites are also in rapid development [34], such as antiSMASH [35], PRISM [36], and SeMPI [37]. Pfam, a widely utilized database of protein families and domains since 1997 [38,39], has also been commonly used for NRPS domain detection by profile hidden Markov models (HMM). They have been used to aid in the discovery of new NRPS antibiotics [16,40]. However, different bioinformatics tools and platforms employ different annotation standards and algorithms, and their definitions of domains and inter-domains in NRPS need to be clarified.

Aside from annotating the module-domain architecture, predicting the substrate specificity of A domains in NRPS has been a focus of research for many years, with various algorithms being developed to address the challenge [37,41–44]. Some use supervised learning and rely on a training set of full-length A domain sequences and their known substrates [42]. Other well-known approaches incorporate structural information of the A domain [45]. With advances in protein structures, the substrate-specifying residues of modular mega-enzymes, such as NRPS [29,46] and polyketide synthase (PKS) [47], were identified. Based on the first crystal structure of the A domain (PheA, PDB 1AMU) [29], Stachelhaus et al. identified ten substrate-specifying residues in the A domain's binding pocket. These ten residues, known as the "specificity-conferring code", strongly associate with the substrate specificity, and have been widely exploited by researchers to improve prediction efficiency [45]. However, the diversity of the specificity-conferring code for the same substrate remains challenging, making the performance of current machine learning algorithms heavily dependent on the training set. Actually, diverse forms of the specificity-conferring code could exist for the same substrate [45]. Such diversity of the specificity-conferring code causes barriers in extending the prediction algorithm: without a one-to-one mapping between the specificity-conferring code and substrate specificity, the performance of current machine learning algorithms heavily relies on the training set. Indeed, the performances of these algorithms in fungal NRPSs are unsatisfactory, as fungi have a significantly smaller number of A domains with known substrates. Furthermore, re-engineering based on modifying the specificity-conferring code does not operate as expected [46], highlighting the complexities in the A domain's function.

Coevolution analysis has been used as a general strategy to uncover protein interactions and functional couplings from a large number of sequences [48]. Existing methods in coevolution analysis include normalized mutual information (MI) of amino acid frequencies [49], direct coupling analysis (DCA) [50], protein sparse inverse covariance (PSICOV) [51], and statistical coupling analysis (SCA) [52,53]. Among them, SCA has been utilized to identify appropriate boundaries for protein re-engineering [54]. The basic idea of SCA is to identify groups of residues sharing the same modes of covariations, referred to as "sectors" [55]. Residues projecting to a single sector can be viewed as "evolving together", and are likely to cooperate in function [56]. Therefore, sectors could serve as the "evolutionary units" for re-engineering, as the borders of sectors provide rational cutting-points for preserving functional connections between residues [55].

Applying SCA to NRPS sequences to gain insights on re-engineering is a straightforward idea. However, one obstacle to this simple notion is the need for manually curated sequences: the majority of sequences in databases were automatically annotated by machine learning algorithms, and have yet to be reviewed by human experts. Moreover, different algorithms with varying standards for annotation were adopted by different researchers, which muddled the exact definitions of domains and interdomains in NRPSs. For example, A and C domains annotated by the default option of antiSMASH are substantially shorter than Pfam-annotated domains [38], and are usually different from sequences from the Protein Data Bank (PDB) [57–59]. Even the domain annotations from antiSMASH, one of the most popular annotation platforms for NPRSs, slight changes happened between versions.

One solution for this "confusion by the annotation" is to focus on the invariant sequence pieces in NRPS architecture, known as "core motifs" [60]. Despite the enormous diversity in NRPs, highly conserved sequence elements have been discovered across domains, and are usually essential to the domain function. These well-established core motifs include: 7 motifs in the C domain, 10 motifs in the A domain (of note, these are not the specificity-conferring code), 7 motifs in the E domain, 1 motif in the T domain, and 1 motif in TE domain [60]. However, these motifs have been proposed for more than two decades. Since then, these motifs have yet to be systematically updated or confirmed. These conserved motifs could be used to standardize NRPS domains, allowing for statistical analysis of NRPS sequences from large databases without manual curation.

In our work, we constructed a "coordinate system" for NRPSs based on well-known conserved core motifs. Such motif-and-intermotif architecture allowed us to integrate sequence information of NRPSs from multiple databases and extract global statistics on the length, variation, and functional properties of each NRPS sequence. Following that, we were able to systematically evaluate a large number of sequences for their statistical properties and function-sequence couplings. For example, we identified new C domain subtypes and obtained the most comprehensive cross-kingdom classification of C domains to date. In addition, we uncovered several previously unrecognized conserved motifs, which could have functional implications and serve as candidates for novel core motifs. One of these novel motifs is located close to the A domain's specificity pocket with implications in structural flexibility, and mutation experiments supported its importance. Then, using coevolution analysis, we dissected two significant challenges in re-engineering NRPSs: first, there is no cutting point to separate multiple overlapping coevolving sectors, limiting domain/subdomain swapping; second, the specificity of the A domain intertwines with not only the specificity-conferring code but also the length of five loops in the A3-G intermotif, complicating the rational modification of A domain. Our findings provide a comprehensive overview and statistical insights into the sequence features of NRPSs, paving the way for future data-driven discoveries. In the end, an online platform entitled "NRPS Motif Finder" is offered for parsing the motif-and-intermotif architecture and C domain subtype classification (http://www.bdainformatics.org/page?type=NRPSMotifFinder).

Result

1. Overview of the motif-and-intermotif architecture of NRPS

1.1. Sequence properties of motifs and intermotifs in NRPS domains. To establish the foundation of our study, we first partitioned the NRPS architecture based on well-known core motifs. The sequence representations of 26 well-studied motifs were curated from the literature [60], then located to 7,329 NRPS domains from MiBiG database by alignment (see Method for details). As expected, most known core motifs are highly conserved, exhibiting clear sequence logos (first panel, Figs 1A–1D and S1–S5). Motifs from the A, E, and T domains have sequence identities higher than 60%. However, C domain motifs exhibit a lower sequence identity than A, T, and E domains. Among the C motifs, motif C3 is the most conserved, with 55% sequence identity across all C subtypes (Fig 1A, first panel). High conservation of C3 is reasonable due to its critical role in catalysis [61,62]. Except for the Heterocyclization subtype, where the motif C3 is DxxxxD, the motif C3 in other subtypes is almost always HHxxxD.

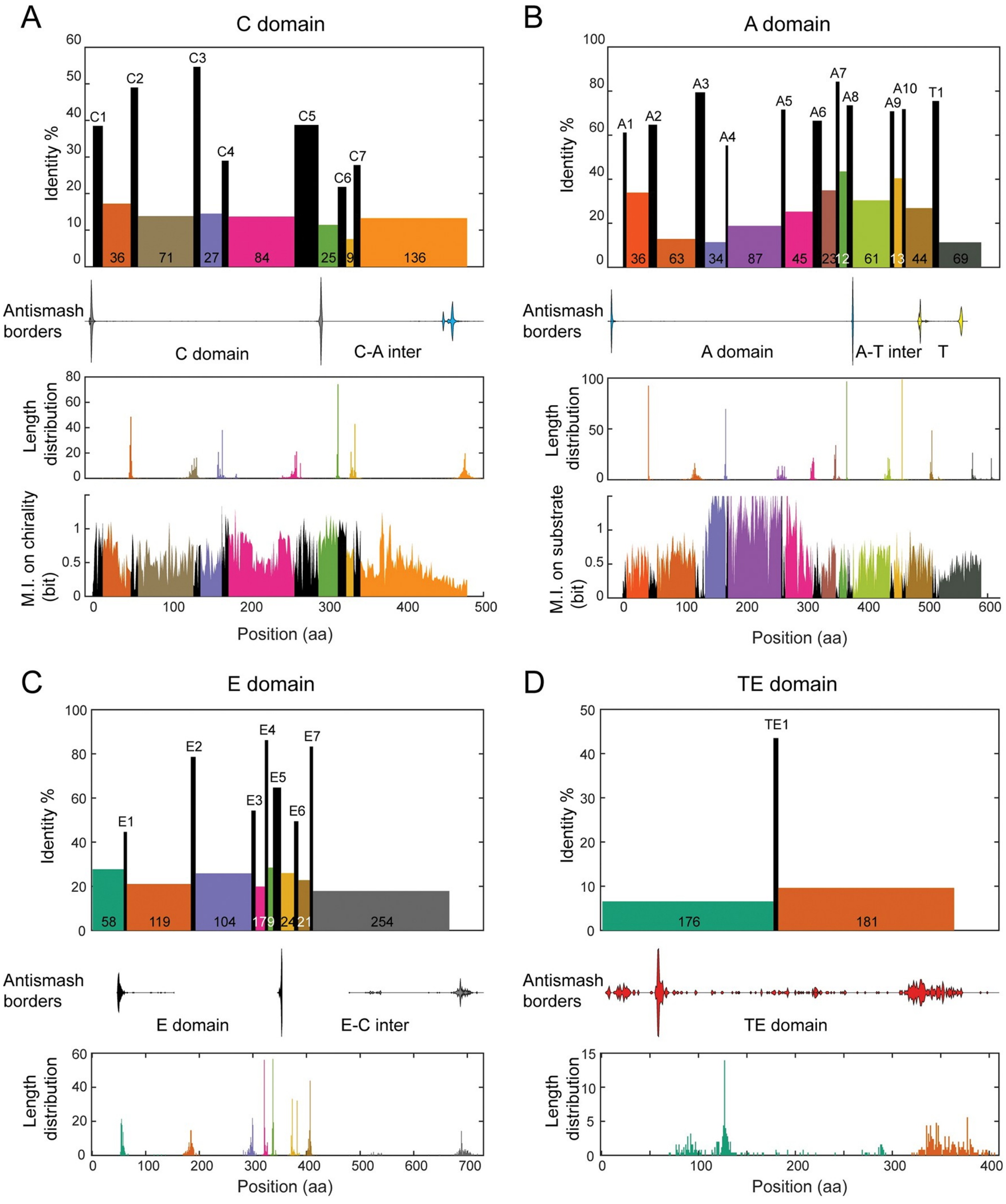


Fig 1. Overview of the motif-and-intermotif architecture of canonical NRPS domains. A. Organization of C domain. Rectangles in the first panel show the seven C motifs (black) and eight inter-motif regions (colored), including the last one as the inter-motif region between the C7 and A1 motifs. The heights of the rectangles indicate the average sequence identities. The widths of the rectangles are their average lengths in amino acid sequences, with the average lengths of inter-motif regions printed on their rectangles. Violin plots in the second panel mark the border positions of domains annotated by antiSMASH 5.1 by the default option (the start and end borders of the C domain: gray; the start of the A domain: blue). The third panel shows the distributions of the ending positions of the eight inter-motif regions, counting from their preceding conserved motifs. The fourth panel shows the mutual information (M.I.) between residues in the C domain and chirality subtypes. **B.** Same as that in **A**, but for the A and T domains. It starts with the A1 motif and ends before the C1 motif. In the violin plot, the start and end borders of the A domain are colored blue; the start and end borders of the T domain are colored yellow. The fourth panel shows the mutual information between residues and A domain substrate specificity (see Method for details). **C.** Same as that in **A**, but for the E domain. It starts after the T1 motif, and ends before the C1 motif. In the violin plot, the start and end borders of the E domain are colored black; the start of the C domain is colored gray. **D.** Same as that in **A**, but for the TE domain. It starts after the T1 motifs, and ends with regions after the annotated TE domain. In the violin plot, the start and end borders of the TE domain are colored red.

https://doi.org/10.1371/journal.pcbi.1011100.g001

Motifs C6 and C7 are less conserved, 22% and 28% identities, respectively. This lower conservation in C motifs is likely due to various functional subtypes within the C domain [61], which we will discuss in further detail in the following sections 1.3–1.5.

The sequence region between the conserved motifs, referred to as the "intermotif region", has a comparatively lower level of sequence identity, with values ranging from 10%-25%. Most intermotif regions display narrow length distributions, with a standard deviation smaller than 10 aa (third panel, Fig 1A–1C), demonstrating the architecture invariance of NRPSs. However, the TE motif at the end of the cluster exhibits a much wider distribution (320-380aa, Fig 1D). An additional noteworthy exception is observed in the intermotif region between the motif T1 and its next motif C1, which exhibits a bi-modal distribution (Fig 1B). Upon closer examination of the domain annotations, it was found that the length of the T1-C1 region depends on the MiBiG annotation of the following C domain (S3B Fig): Specifically, regions followed by the LCL subtype tend to be 65–70 amino acids in length, whereas regions followed by the Dual subtype are typically 95–105 amino acids in length. These observations suggest that the architecture of intermotif regions may have functional implications.

It's worth mentioning that by default, antiSMASH annotations for the C domain do not include the motif C6 and C7; these for the A domain do not include the motif A9 and A10; and these for the E domain do not include the motif E6 and E7 (second panel, Fig 1A–1C). We noted the motif E6 and E7 are located in the TIGR01720 domain annotated by antiSMASH. This explains the difference in NRPS domain length between MiBiG and Pfam (S6 Fig).

1.2. Exploring the coupling between sequence and function in the C and A domains by mutual information. Another focus of this study is how sequences couple with functions. Conserved motifs serve as "anchors" in multiple sequence alignment, facilitating a detailed evaluation of information content. Given that the sequences of the C domain have been previously demonstrated to be indicative of chirality subtypes [61], we first quantified the Shannon mutual information (MI) between the aligned sequences and functional subtypes in the C domain, residue by residue (see Method for details). The region between the C1 and C7 motifs, represented in the last panel of Fig 1A, contains a substantial amount of information about chirality subtypes. After the C7 motif, the information about chirality gradually decreases until it reaches the baseline level as the sequence progresses toward the A1 motif (Fig 1A, last panel).

Generally, conserved motifs contain low levels of mutual information (Fig 1). Nevertheless, it's noted that the C5-C6 region, including the C5 and C6 motif, has a high MI value with C domain subtypes (Fig 1A, last panel). We checked this region in the structure of LgrA, the only structure that contains multi-module NRPS in multiple conformations so far characterized by Reimer et al.[63]. We noticed that this region is located at the interface between the donor T domain and the recipient C domain of the next module (T_n:C_{n+1}), which has been suggested to play roles in chirality [64]. Multiple coevolution positions between the T_n domain and C_{n+1} domain were suggested by DCA in Reimer et al.'s work, with multiple key residues located in the C5 and C6 motif. Therefore, T_n and C_{n+1} are functionally linked by the C5-C6 motif in the T_n:C_{n+1} interface, which explains the high MI value between C5-C6 intermotif region and C domain subtypes.

In light of the critical role played by the A domain in substrate selectivity, we also employed the same mutual information measurement to evaluate the relationship between each position in the A domain and substrate specificity (see Method for details). Results in the last panel of Fig 1B show that positions with high information content start from the A3 motif and end before the A6 motif. This region overlaps the substrate's binding pocket [57]. It is worth noting that the A4-A5 region has been well-known for containing the specificity-conferring code [29].

1.3. Cross-kingdom validation and comparison of motif properties in a larger dataset. The MiBiG database provides a solid foundation for discovering the fundamental properties of NRPSs, as all BGCs in the database have been confirmed to be active. Then, to further validate the conservation of the NRPS motifs, we expanded our analysis to a larger dataset. With that aim in mind, we downloaded all complete bacterial genomes (30,984) and all assembly-level (including "contig", "scaffold", "chromosome", and "complete" levels) fungal genomes (3,672) from NCBI. After removing genomes with 99.6% similarity, we had 16,820 bacterial and 2,505 fungal genomes (as of 2022/10/23). BGCs in these genomes were annotated using antiSMASH v6. Unlike the MiBiG database, where the majority of pathways have been experimentally verified to be active, an unknown number of the domains in this dataset might be inactive. To account for this, we employed a preliminary dead domain filter by eliminating domains with non-standard motif lengths (see Method for details). In this large dataset, we analyzed 83,489 C domains, 95,582 A domains, 86,688 T domains, 14,502 E domains, and 23,590 TE domains from bacteria; and 34,269 C domains, 40,458 A domains, 26,651 T domains, 3,982 E domains, and 4,008 TE domains from fungi.

The prevalence of NRPS pathways extends across multiple kingdoms, with fungi being the only known eukaryotic producers. Examining the similarities and differences in the NRPS sequences of bacteria and fungi can help elucidate the conservation and divergence of NRPS functions. In contrast to the MiBiG database, where most pathways are of bacterial origin, the larger dataset contains a comparable number of C, A, and T domains from both fungi and bacteria, providing a valuable platform for validating and comparing motif properties.

Overall, the motif logo and intermotif length of the A, T, E, and TE domains were found to be similar between bacteria and fungi, albeit with some subtle differences (S7–S14 Figs). For example, the sequence logo of bacteria E6 was identified as RX(V/I/L)PXXGXG(Y/F)G, while fungi E6 was RX(V/I/L)PXXGXXYF (S11 Fig); the fungal intermotif A10-T1 was found to be 13 amino acids longer than that of bacteria by the median (S13 Fig); the fungal intermotif E2-E3 is 15 aa shorter than bacteria by the median (S14 Fig). The disparity in the E6 motif highlights a potential bias in our current understanding of NRPS, which is predominantly informed by sequences collected from bacteria: Previous research reported that the E6 consensus sequence was PXXGXGYG, only consistent with the E6 sequence logo we observed in bacteria [60]. A more comprehensive understanding of NRPS motif and intermotif across kingdoms is required to expand our knowledge.

1.4. Identification of new C subtypes in fungi guided by conserved motifs. The C domain of NRPS exhibits notable differences between bacteria and fungi. Previous research has established that the C domain can exist in multiple subtypes, each displaying distinct conserved motifs [61,65,66]. In order to conduct a comprehensive comparison of NRPS between fungi and bacteria, it is imperative to first clarify these subtypes. Currently, antiSMASH v6 can distinguish LCL, DCL, Cglyc, Dual, Starter, Cyc, and X subtypes [35], and NaPDoS v2 can distinguish LCL, DCL, Dual, Starter, Cyc, modAA, and Hybrid subtypes [67]. However, these tools are primarily based on bacterial subtypes and do not consider fungal subtypes such as CT and Iterative [25,68]. This lack of annotation for fungal C domain subtypes hinders genome mining of fungal NRPS.

In this study, we curated sequences representing 18 subtypes of the C and E domains, 14 of which were sourced from the conserved domain database (CDD) [69], and the remaining 4 were obtained from the literature [33,66,70,71] (S1 Table). Phylogenetic analysis of these subtypes revealed their relatedness (Fig 2A). These subtypes form two major clades: L-clade and D-clade, consistent with previous reports [65]. Furthermore, our results revealed the evolutionary landscape of all known C-domain subtypes for the first time, offering deeper insights into several sub-clades. It can be observed that, the CT subtype may evolve from the FUM14

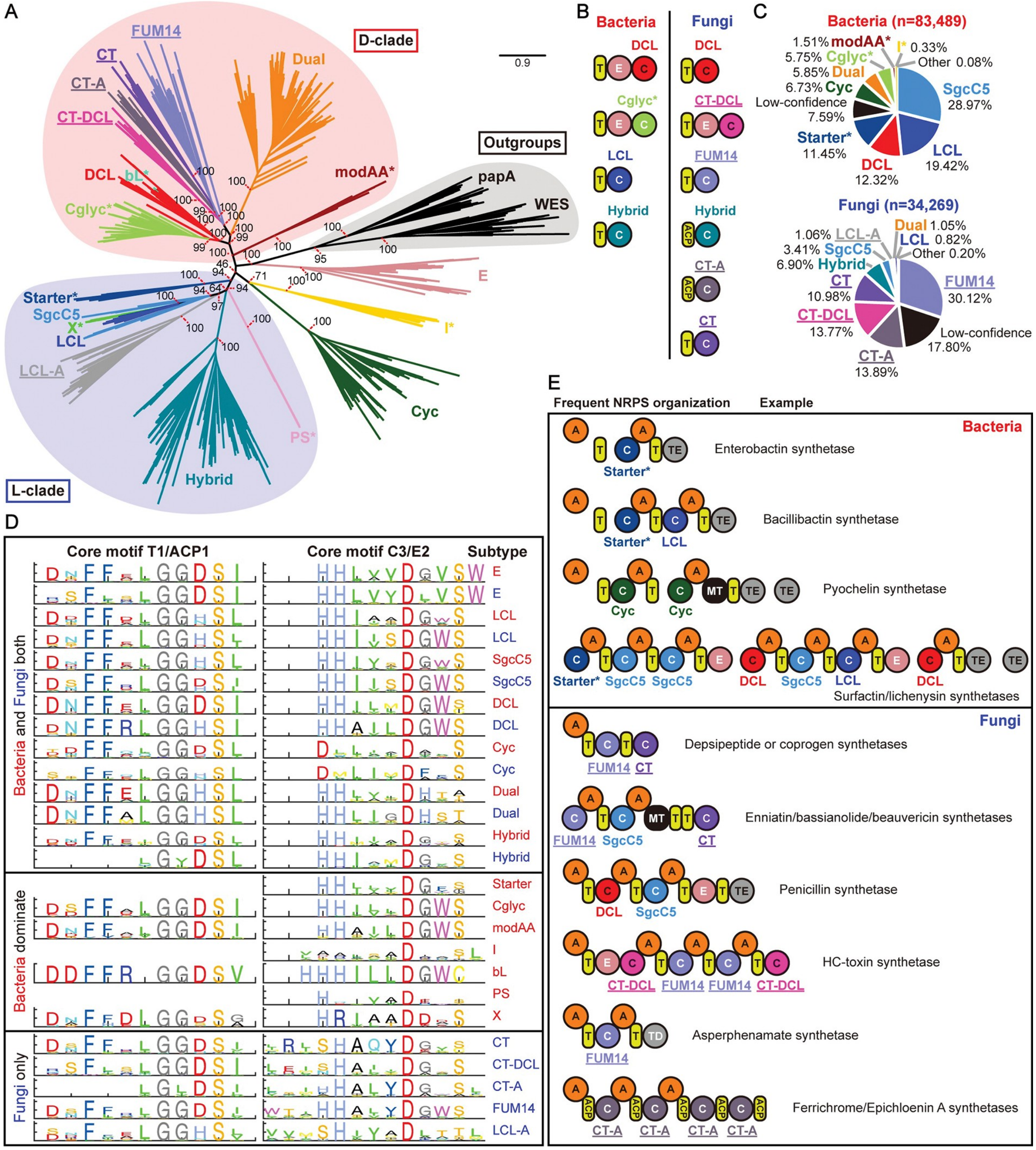


Fig 2. C domain subtype analysis and representative NRPS organizations in bacteria and fungi. A. Maximum-likelihood phylogenetic tree of the condensation domain superfamily. Subtype classification and sequences are described in the main text and the Method. Different subtypes are indicated by colors, with subtypes exclusive to fungi marked by underlines, and subtypes found predominantly in bacteria marked by asterisks. This tree is rooted, taking papA and WES as outgroups [65] (black shading). L-clade and D-clade are indicated by blue and red shading, respectively. **B.** Domains adjacent to different C domain subtypes in bacteria and fungi. **C.** The statistics of subtype distribution in 83,489 bacterial C domains and 34,269 fungal C domains. C domains with HMM scores above the empirical threshold of 200 were annotated by their predictions, otherwise marked as "Low-confidence". **D.** The sequence logo for the C3 or E2 motif from different C domain subtypes and the T1 or ACP1 motif adjacent to each subtype. Sequences from bacteria were marked by red, while sequences from fungi were marked by blue. **E.** Frequent NRPS organizations with known representative examples in bacteria and fungi.

https://doi.org/10.1371/journal.pcbi.1011100.g002

subtype; the LCL subtype may evolve from SgcC5, and Cglyc may evolve from DCL. The FUM14 subtype was named from *Fusarium verticillioides* FUM14 (also known as NRPS8), a bi-domain protein with an ester-bond forming NRPS C-domain [72]. SgcC5 is a free-standing NRPS condensation enzyme (rather than a modular NRPS), which catalyzes the formation of both ester and amide bonds [73]. Cglyc subtype is the glycopeptide condensation domain [61].

New insights on subclasses in C domains can be obtained by precisely characterizing this C domain phylogeny, particularly in fungal-related domains. For each C domain subtype, we constructed MSA from curated sequences by Muscle v5 [5] and built reference HMM models by hmmer v3 [74]. We then utilized these HMM models to classify C domain subtypes, and extracted motifs for each subtype by the previously mentioned method.

During this process, we noticed that the sequence motifs of the curated CT subtype are remarkably variable, suggesting that this "CT" may not be a uniform subtype. Guided by the motif sequence and phylogenetic analysis, our analysis suggests that the original "CT" domain can be further classified into three subtypes by sequence similarity, which is also highly related to their locations in the NRPS pathway (Fig 2B): One of the subtypes is always located at the end of the NRPS, and we termed it "CT" because it fits the original definition of the CT subtype. Another subtype was termed "CT-DCL" because it is always behind an E domain and may function similarly to the DCL in bacteria. Of note, we also observed that the annotated fungal DCL subtype is always after the T domain. The third CT subtype is atypical because it is always behind an ACP (acyl carrier protein) domain rather than a T domain. Both the ACP and T domains are phosphopantetheinyl carrier domains, but the ACP domain has a conserved XGXDSL motif rather than the T domain's GG(D/H)S(I/L) motif [60,75]. Therefore, we termed this subtype "CT-A" (meaning atypical CT). These three CT subtypes form three different clades in the phylogenetic tree.

After clarifying subtypes, the overall statistics regarding the distribution of C domain subtypes in bacteria and fungi can be analyzed. Our subtype prediction covered 92.41% C domains in bacteria and 82.20% C domains in fungi, by an empirical score threshold of 200 in HMM models (Fig 2C). Our predictions for the Starter and Dual subtypes are consistent with antiSMASH. However, antiSMASH frequently assigns incorrect annotations of "LCL" and "DCL" to subtypes that are uncommon in bacteria, which were corrected in our annotation (S2 Table).

By the updated subtype classification, we redrew sequence logo of C domain in MiBiG. The conservations of motifs increased significantly compared to the prior classification relying on MiBiG (S1 and S15 Figs). Regarding subtype distribution, bacterial C domains mainly consist of SgcC5, LCL, DCL, Starter, Cyc, Dual, and Cglyc, while fungal C domains are mainly composed of FUM14, CT-A, CT-DCL, CT, and Hybrid. Among them, I (interface) only existed in bacterial C domains. Starter, Cglyc and modAA are almost exclusively found in bacteria, with a handful of occurrences in fungi (S3 and S4 Tables). FUM14, CT-A, CT-DCL, CT, and an atypical LCL (designated as LCL-A, meaning atypical LCL which usually has a non-canonical conserved SHXXXDXX(T/S), rather than HHXXXDGXS) only exist in fungi.

These distinctions between fungal and bacterial C domains and differences between subtypes are readily apparent in the comparison of typical motif logos (S16 and S17 Figs), such as the C3 and T1 (Fig 2D). The C3 motif is essential for catalysis, and the T1 motif has been suggested to exhibit covariation with its preceding C domain [65]. Our findings highlight the variation in the location and motif of C domains with the T domain between subtypes in fungi and bacteria (Fig 2D). Firstly, some C domain subtypes do not directly precede a T domain: the fungus Hybrid and CT-A subtypes are adjacent to an ACP domain rather than a T domain; the bacteria DCL and fungus CT-DCL subtypes are located after an E domain, consistent with previous reports [65]; For the Cyc subtype, only 29.34% of them are adjacent to a T domain, while the rest are located at the start of the pathway. The Starter, I, and PS subtypes are also exclusively found at the start of the pathway. Secondly, we observed that the coupling between the C domain and its adjacent T domain cannot be solely explained by the L- and D-clades. Previous research mainly based on bacterial data suggested that the LCL subtype (located in the L-clade of the phylogenetic tree in our analysis) is adjacent to a T domain with the LGGHSL motif, and the DCL subtype (located in the D-clade of the phylogenetic tree in our analysis) is adjacent to a T domain with the LGGDSI motif [64,65]. Our analysis not only confirmed these relationships but also showed that not all subtypes in the L- and D-clades follow this pattern: The X subtype in the L-clade is adjacent to a T domain with the LGGDSG motif; in the D-clade, and the Dual and fungal DCL subtypes in the D-clade are adjacent to T domains with the LGGHSI and LGGHSL motifs, respectively. These observations demonstrate that the T1 motifs are primarily linked to the specific subtypes of their adjacent C domains, rather than their clades.

In addition to the conserved motifs, the highly variable intermotif regions also differentiate fungi and bacteria based on some of their lengths. In the standardized motif-and-intermotif architecture, we compared the intermotif length distribution in different C domain subtypes between bacteria and fungi (S18 Fig). The most notable difference is in the T1/ACP1-C1 region. C domain subtypes with more complex functions, such as Dual (92aa), modAA (74aa), bL (82aa), CT (86aa), CT-A(82aa), and FUM14 (85aa), have a substantially longer T1-C1 intermotif compared to C domains with simpler functions, such as LCL/ScgC5 (60aa in bacteria, 59aa in fungi). This finding implies that longer interdomain length may provide the necessary space for coordination between different domains in the megasynthetase. This distinction in C domain subtype compositions and intermotif lengths may offer potential applications in the future, allowing us to distinguish between bacterial and fungal NRPS in fragmented metagenomic sequencing data.

1.5. The prevalent NRPS organizations in bacteria and fungi. This detailed annotation of C domain subtypes allows us to evaluate how domains are organized within NRPS pathways in bacteria and fungi. In this work, “organization” was defined as the composition and arrangement of domains within an NRPS pathway. For the accuracy of statistics, we focused on 82.47% (12,364/14,992) bacterial and 61.00% (6,438/10,555) fungal NRPS pathways that only contain the high-confidence C domains. Our findings show that in bacteria, 24 out of the 30 most frequent organizations are involved in the production of siderophores, accounting for 28.21% of the 12,364 pathways. Among these 24 siderophore pathway organizations, enterobactin [76] (BGC0002476), the siderophore with the highest binding affinity to iron, has the highest occurrence (21.35% of 12,364). The second most frequent siderophore NRPS is bacillibactin synthetase [77] (BGC0000309), making up 2.55% of the bacterial pathways, with an organization similar to enterobactin synthetase but with an additional module containing LCL subtype C domain. The third most frequent siderophore NRPS is pyochelin synthetase (1.16%, BGC0000412), containing two Cyc subtype C domains [78]. In addition to siderophores, biosurfactant surfactin/lichenysin synthetases are also frequent [79,80] (BGC0000433 and BGC0000381), representing 2 out of the 30 most frequent organizations and accounting for 1.24% of all pathways. This biosurfactant is comprised of three NRPS genes, with 1 Starter, 2 DCL, and 4 LCL/SgcC5 C domains. Interestingly, both pyochelin synthetase and surfactin/lichenysin synthetases contained 2 TE domains, with the function of the second TE domain suggested for proofreading for the NRPS product [81].

In fungi, the most frequent NRPS organization is also related to siderophore NRPS. Of the 20 most frequent NRPS organizations, 8 are depsipeptide or coprogen synthetases [68] (BGC0001249), accounting for 11.90% of the 6,438 pathways. These synthetases typically consist of one NRPS gene with one FUM14 and one CT subtyped C domains, with some having an additional NRPS gene with an A domain (Fig 2E). These depsipeptide or coprogen synthetases are known for their iterative features, and their pathway organization suggests that the FUM14 subtype could also participate in the iterative process. The second most frequent siderophore NRPS is Ferrichrome/Epichloenin A synthetase [82,83] (0.79% of the pathways, BGC0000901/BGC0001250, Fig 2E). It's noted that all subtypes of the C domains in all Ferrichromes (type I-VI) are CT-A, except for the first C domain in several cases [82]. Therefore, detecting Ferrichrome-type siderophore in the genome is possible simply by C domain subtype. The other two frequent NPRS organizations ranked within the top 20 are HC-toxin and Asperphenamate synthetase (1.29% and 0.87% of the pathways, BGC0001166 and BGC0001517, respectively, Fig 2E). HC-toxin is a virulence factor produced by plant pathogenic fungi [84], and Asperphenamate exhibits antitumor activity [85]. It's noted that Asperphenamate synthetase ends with a TD domain, one of the systematic differences between bacteria and fungi [86]. The representation of the third pathway is unknown, and the most similar in MiBiG is cyclo-(D-Phe-L-Phe-D-Val-L-Val) annotated by antiSMASH v6 (BGC0000357). Other well-known fungal NRPS include enniatin/bassianolide/beauvericin synthetases [87–89] (29th most frequent organization, 0.36% of the pathways, BGC0000342, BGC0000312 and BGC0000313), which are also known for their iterative feature and have been the focus of re-engineering for new products [90,91]. These synthetases consist of one NRPS gene with one FUM14, one SgcC5, and one CT C domain, again suggesting FUM14 subtype C domain may have an iterative function. Another known pathway (34th, 0.33%) is the penicillin/isopenicillin N/benzylpenicillin/phenoxymethylpenicillin synthetase pathway [92] (BGC0000404, BGC0000405), which has a unique organization with its DCL located behind a T domain rather than an E domain, and its E domain positioned before a TE domain instead of a C domain. Almost all DCL in fungi have a similar organization to penicillin synthetase.

Interestingly, in fungi, the function of DCL in bacteria might be replaced by CT-DCL. For example, the HC-toxin synthetase features an E+CT-DCL arrangement (Fig 2E). In the entire database, such E+CT-DCL arrangements are common, with 66.14% (3121/4719) CT-DCL located behind an E domain, and 78.38% of E domain proceeding a CT-DCL (3121/3982). And in some cases, such as HC-toxin synthetase (Fig 2E), fungal NRPS ends with a CT-DCL domain. It suggests that the CT-DCL domain may act as a CT domain to terminate the synthesis of NRP. The results highlight the limitations of current annotations in large dataset, despite that the large dataset provided insights for confirming previously discovered information. Therefore, after analyzing this large dataset, we shifted our focus back to the well-studied and annotated MiBiG database, where the well-established annotation offers a more suitable platform for in-depth discovery.

2. Discovery and functional identification of new conserved motifs

The presence of conserved motifs acts as anchor points in multiple sequence alignments, enabling an in-depth examination of the level of conservation across the entire sequence. From the MiBiG database, we calculated the amino acid frequency and gap frequency in a sequence alignment of 1,161 C+A+T NRPS modules (1053 from bacteria, 81 from fungi, and 27 from "others"). The results are presented in Fig 3A, with a lower panel showing the conservation of the alignment. As expected, most highly conserved positions are found within or close to known core motifs, and thus can be seen as the extension of known motifs (S19 Fig). However, three highly conserved regions are relatively distant from well-known motifs (Fig 3A, upper panel), therefore, cannot be considered extensions of these well-known motifs, suggesting potential new motifs.

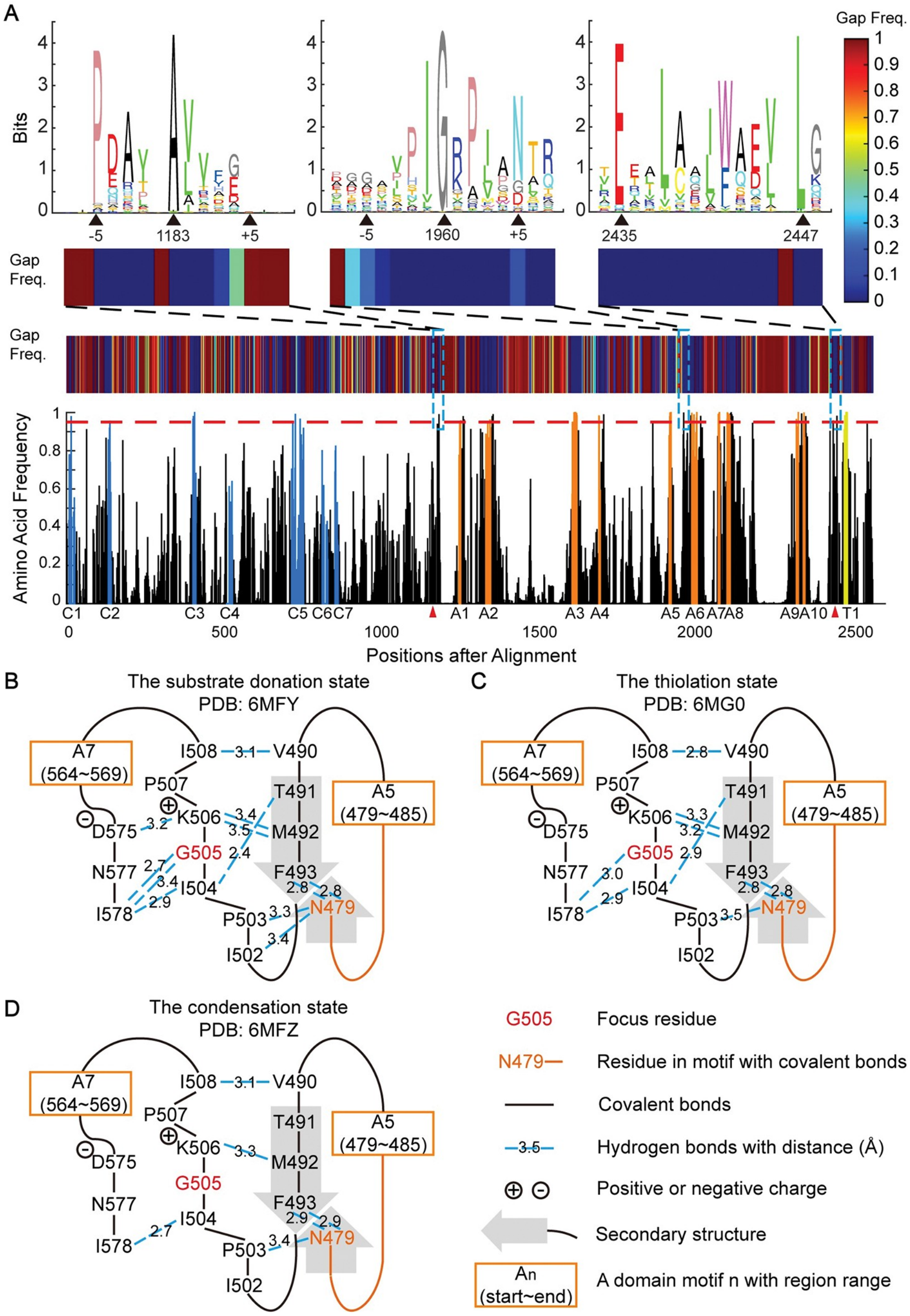


Fig 3. Analysis of amino acid frequency reveals potential new motifs with implications for structural flexibility. A. Amino acid frequency and gap frequency along the multiple sequence alignment of the NRPS C+A+T modules. In the bottom panel, bar heights indicate the frequency of the most frequent amino acid. Bars in the known core motifs from the C, A, and T domains were colored blue, orange, and yellow, respectively. The horizontal red dashed line represents the 0.95 frequency level. Domain boundaries annotated by Pfam are divided by red triangles. The colored patch above the amino acid frequency indicates gap frequency. Three potential new motifs (position 1183, 1960, and 2435/2447 in MSA) are marked by the blue dashed box. The upper panel shows the sequence logo and the gap frequency near the three potential new motifs. **B.** Chemical interactions and secondary structures surrounding the second potential new motif in (A) at the substrate donation state (PDB: 6MFY). Hydrogen bonds near the most conserved Gly were shown in blue dashed lines. Covalent bonds were shown as black lines. Secondary structures, such as beta-sheets, were demonstrated as bold gray arrows. Known A domain motifs adjacent to related residues were shown in the orange box. **C.** Same as that in **B**, but in the thiolation state (PDB: 6MG0). **D.** Same as that in **B**, but in the condensation state (PDB: 6MFZ).

https://doi.org/10.1371/journal.pcbi.1011100.g003

2.1. The new "G-motif" in the substrate binding pocket relates to conformation flexibility. A conserved glycine located in the middle of the A5-A6 intermotif region has caught our attention. We also observed this conserved glycine in the A domain MSA in the large dataset (S7 and S10 Figs). It is 21(±2) aa from the last residue of the A5 motif, and 23(±0.2) aa from the first residue of the A6 motif. This glycine is highly conserved in over 99% of the aligned A domains, with surrounding residues being moderately conserved (as shown in the second logo in the upper-most panel of Fig 3A). Notably, this glycine divides the A5-A6 intermotif region into two halves regarding substrate information, where the sequence preceding this glycine contains substantially more mutual information about substrate (S20 Fig). For convenience, we will collectively refer to this conserved glycine and its surrounding residues as the "G-motif".

We then located this G-motif to the G505 position in a recently solved NRPS structure, the linear gramicidin synthetase subunit A (LgrA) [63]. Interestingly, the G-motif resides on a loop structure, which is typically considered non-conserved. The LgrA structure has been solved with three function-related states: 6MFY, the substrate donation state; 6MG0, the thiolation state; and 6MFZ, the condensation state [63]. By calculating hydrogen bonds in these structures, we noticed that the G-motif (comprising residues I502, P503, I504, G505, K506, P507, and I508 in LgrA) interacts with N479 in the A5 motif (where the A5 spans N479-E485, with sequence NGYGPTE) and its backside residues (V490, T491, and M492). It also interacts with the backside residues of the A7 motif (D575 and I578; where the A7 spans Y564-R569 with sequence YRTGDR) (Fig 3B–3D). Of note, the number of hydrogen bonds for these interactions changes remarkably in different function-related states (Figs 3B–3D and S21): In the substrate donation state 6MFY, the G-motif has 12 hydrogen bonds (average bond length 3.13±0.27Å); the thiolation state 6MG0 has 9 bonds (average length 3.09±0.24Å); and the condensation state 6MFZ has only 6 bonds (average length 3.13±0.27Å).

These changes observed in hydrogen bonds during confirmation changes prompted us to examine whether these interactions are functionally relevant. One prediction is that, besides the highly conserved glycine, the chemical properties of other residues involved in these interactions should also be conserved. Therefore, we turned to the 1,161 MiBiG C+A+T alignment to check the chemical conservation of the residues interacting with the G-motif (S5 Table). The cross-species conservation of chemical properties supported our prediction (Table 1). Of note, among these residues, only T491 and N479 use the hydroxyl or amide group in their side chains to form hydrogen bonds. In contrast, the other hydrogen bonds are formed by the commonly occurring α-carboxyl group and α-amino group in the main chains. The high chemical conservation among these related residues suggests a strong selective pressure to maintain desired functions.

The small size and lack of a side chain of glycine make it unique among proteinogenic amino acids. We observed that two amino acids, N577 and F493 in LgrA, locate near the glycine in the G-motif (G505). One possible explanation for the high conservation of glycine in the G-motif is that its small size provides greater structural flexibility, reducing the likelihood of collisions with neighboring residues. To check this hypothesis in known structures of A domains, we analyzed all available structures of AMP-binding-domain-containing proteins from the PDB database (39 sequences existing in a total of 95 structures, including different conformations or ligands, S6 Table). Of these 39 sequences, 30 are from NRPS pathways (selecting 20 different substrates), 6 from NRPS-like pathways (selecting 5 different substrates, with the substrates of the carboxylate reductase all being considered carboxylate), and 3 are from D-Ala-ligases (DltA) pathways (selecting substrate D-Ala). Notably, NRPS-like carboxylate reductases (CARs) do not contain the G-motif, showing a distinct evolution path despite also containing an AMP-binding domain (S6 Table, "overview" sheet). Besides CARs, the G-

Table 1. Representative chemically conserved residues in or interacting with the G-motif.

The position of conserved residues		Amino acid composition	Properties of amino acid residues (side chains)
In the LgrA protein	In MiBiG C+A+T module MSA*	In MiBiG C+A+T module MSA	
T491	T1927	T:61.67% S:15.16%	Polar, -OH; the donor and acceptor of hydrogen bonds
I504	I1959	I:88.29% V:5.25% L:3.45%	Hydrophobic
G505	G1960	G:99.66%	
K506	R1961	R:46.25% K:23.84%	Positively charged; the donor of hydrogen bonds
I578	L2094	L:60.9% I:29.03% V:7.67%	Hydrophobic

Conserved residues in or interacting with the G-motif have significant chemical properties. * This first letter is the prevalent amino acid in this position of MSA.

https://doi.org/10.1371/journal.pcbi.1011100.t001

motif is conserved in all 30 NRPS A domains, and 3 NRPS-like A domains (S22 Fig). However, except for the NRPS-like protein poly-ε-Lys synthetase (Pls-A, PDB: 7WEW), all AMP-binding-domain-containing proteins have equivalent N577 and F493 in LgrA, even for those without a G-motif (S23 Fig). The equivalent N577 and F493 have side chain sizes (S5 Table). In these known structures, the G-motif is near the adenylate part of the ligand (S23 Fig), also suggesting a potential gatekeeper role of the G-motif.

By simulated mutagenesis using PyMOL, we found that the position of glycine in the G-motif only permits small amino acids such as glycine and alanine. In simulations, mutating this residue into amino acids with larger volume or inflexible loops resulted in a disturbance of the residue N397 and S491 in Fig 4A and 4B (equivalent F493 and N577 in LgrA, Fig 3B–3D).

To further verify the importance of this highly conserved glycine, we performed mutation experiments on the "G-motif" in a monomodular NRPS FmqC involved in the biosynthesis of fumiquinazoline C (FQC) in *Aspergillus fumigatus* (Fig 4C). In this system, a tripeptide precursor fumiquinazoline F (FQF) was first synthesized by FmqA, and then the indole side chain of FQF oxidized by the FmqB. Then, the monomodular NRPS FmqC, activated L-Ala to create the product of fumiquinazoline A (FQA), thereby constructing a new C_2–N bond and finally forming FQC [93] (Fig 4D). The *fmqC* gene deletion resulted in the disappearance of FQC as the pathway diverged towards a putative compound (termed compound **1**), as indicated in the first two rows of S4 Fig. This is consistent with a previous study in *A. fumtigatus* AF293 [94]. We reintroduced *fmqC* into the original locus in the Δ*fmqC* background and generated one control strain with a full-length copy of *fmqC* in *A. fumtigatus*. The control strain could produce FQC just as the wild type (S24 Fig). Then, the glycine in the G-motif of FmqC (G409) was mutated to six other representative amino acids (A, alanine, R, arginine, D, aspartic acid, P, proline, W, tryptophan and Y, tyrosine), to generate six FmqC (G409) mutants in *A. fumtigatus*, respectively. These mutations cover all combinations of structural effects estimated by Missense3D [95] (S7 Table).

Following LC-MS analysis, the FQC yield of all mutants is lower than the control. While the G-motif mutants G409A or G409R didn't strongly reduce the yield of FQC, the other four mutants substantially reduced the yield of FQC (Fig 4E and 4F). For example, when the glycine was mutated into proline (G409P), the yield of FQC was only around 3% of the control (Fig 4F and Table A in S1 Text). Although we observed no accumulation of the precursor FQF in strains with substantially decreased FQC, we detected compound **1** as the divergent product in all these strains instead (Fig 4D–4F). Our mutation experiment is consistent with the potentially important roles of the conserved glycine in the G-motif.

2.2. Conserved motif at the start of the A domain. A conserved alanine located at the start of the Pfam annotated A domain was also identified through standardized multiple sequence alignment (Fig 3A, the first sequence logo in the first panel). This residue is

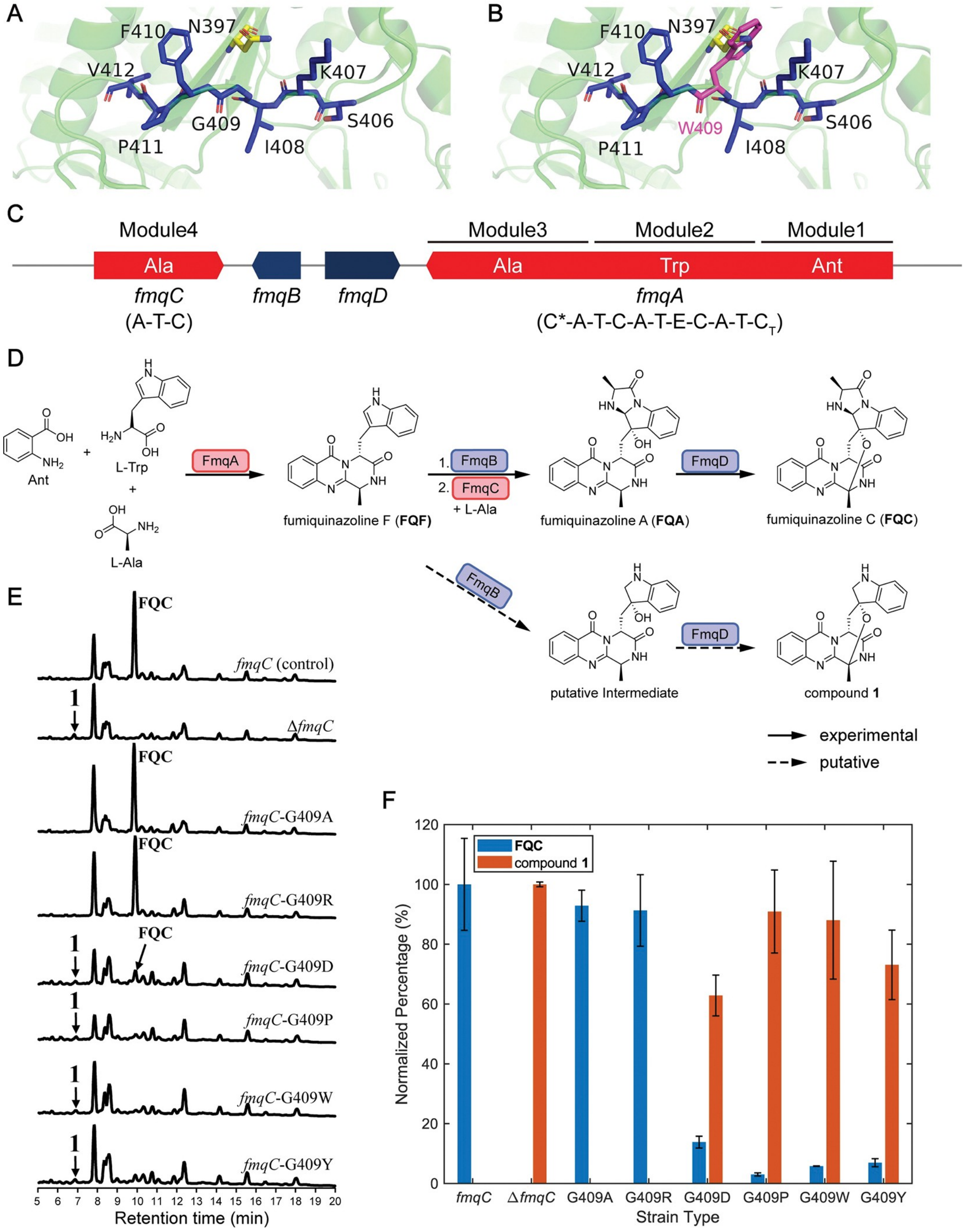


Fig 4. Mutations of G-motif G409 in FmqC support the importance of the conserved domain in the biosynthesis of fumiquinazoline C. A. Residues in and near the G-motif in the predicted structure of FmqC by Phyre2 [96]. G409 is the conserved glycine in the G-motif. Residues in G-motif are marked by blue. The residues N397 and S491 (equivalent to F493 and N577 in LgrA, Fig 3B–3D), which may collide with G409 are marked by yellow and cyan. **B.** Same as that in A, but with simulated mutation of G409W. The mutated tryptophan is marked in magenta. **C.** The *fmq* gene cluster responsible for the production of FQC. Two NRPSs, *fmqA* and *fmqC*, are filled in red with their substrate selectivity marked. Ant: non-proteinogenic amino acid anthranilate. C* represents a truncated and presumably inactive C domain. C$_T$ represents a terminal condensation-like domain that catalyzes macrocyclization reaction. **D.** The biosynthetic pathway for FQC is depicted, along with how it diverges into the production of compound **1** in the absence of functional FmqC. **E.** LC-MS analysis of the control *fmqC* (first row), Δ*fmqC* (second row), and six point mutation strains (3rd to 8th row). **F.** Normalized yield of FQC and compound **1** in different strains. For FQC, the yield is normalized by its production in the wild-type strain. For compound **1**, the yield is normalized by its production in the *fmqC* gene deletion strain. Error bars show standard deviations.

https://doi.org/10.1371/journal.pcbi.1011100.g004

positioned 11(±0.4) aa upstream of the A1 motif, and 124(±4) aa downstream of the C7 motif. This residue is found at A221 in the LgrA sequence. By analyzing hydrogen bonds based on the LgrA structures, we noticed that this conserved A221, V222, and another relatively conserved proline (P217 in LgrA), interact with residues in the A1 motif (L229, T230, Y231, and K232, where the A1 motif spans 229–234 in LgrA with sequence LTYKQL), as well as the back side of the A4 motif (L412 and I414; where the A4 motif spans 395–398 in LgrA with sequence FDGS). Direct interactions were also observed between A410, L406, and Y231 (S25 and S26 Figs). The chemical characteristics of residues anticipated to interact with the conserved alanine are also conserved in the alignment of 1,161 sequences, similar to the G-motif (S5 Table). We termed this motif "Aα1", marking the start of the A domain. Previous studies have reported Aα1 as a natural recombination site in Myxobacteria [97] and applied it in synthetic biology for diverse NRPS products [98].

2.3. Conserved motif at the first helix of the T domain. The other two conserved residues distant from known motifs are glutamic acid and leucine, both found at the start of the Pfam-anropnotated T domain (Fig 3A, the third sequence logo in the first panel). They were found at E700 and L711 in the LgrA sequence. The conserved glutamic acid is 10(±0.04) aa from the conserved leucine, which is 8(±0.6) aa before the first residue of the T motif. The T domain is known to be a distorted four-helix bundle with an extended loop between the first two helices [99]. According to the LgrA structure, the E700 and L711 are located at the N and C terminus of the first helix of the T domain, respectively. Distinct from the G-motif and the conserved A221, we found no hydrogen bonds outside the interior of this alpha helix. Instead, the E700, L711, and other conserved aliphatic and aromatic amino acids between E700 and L711, take part in the formation of the hydrophobic core (S27 Fig), agreeable with previous reports that the four helices were bound together by the hydrophobic interaction [100]. We termed this motif "Tα1", marking the start of the T domain.

3. SCA reveals overlapped co-evolving sectors across the C+A+T module

Coevolution analysis relies on proper sequence alignment, standardized NRPS structures can facilitate. After examining the sequence properties of the standardized C+A+T module, we employed coevolution analysis to study the coupling between positions. Statistical Coupling Analysis (SCA), a commonly used technique for detecting "evolutionary units", was applied to the 1,161 aligned and standardized C+A+T sequences from MiBiG. In the SCA result, only a few top modes (representing the collective covariations of a set of residues known as "sectors") have eigenvalues (measures of the covariation magnitude) that are significantly larger than those obtained by random shuffling (26 out of 2560 total, S28 Fig). The dominant first mode of SCA displays global correlations, commonly regarded as the phylogenetic association and generally ignored in the subsequent analysis [55]. The remaining sectors, sorted in descending order based on eigenvalues, were referred to by their order and sign of the eigenvalues (see Method for details). Sectors II(+) and II(-) immediately after the phylogenetic mode contain 122 positions in the MSA with significant patterns of correlation (Fig 5A, two groups labeled green and magenta for II(+) and II(-), respectively). The II(+) and II(-) sectors are orthogonal because they are the opposite directions of the same eigenvector. The II(+) sector primarily contains residues in the C domain, while the II(-) sector spans the entire NRPS module but does not contain any residues located in the pocket region of the A domain (Fig 5A).

Then we searched for sectors with a high proportion of residues from the A domain pocket region. The sector densely covering the A domain pocket region appears in the sixth biggest mode (Fig 5A, labeled as IV(-) sector with red dots), which includes three residues in the specificity-conferring code [29]. While the IV(-) sector varies nearly independently from the II(-)

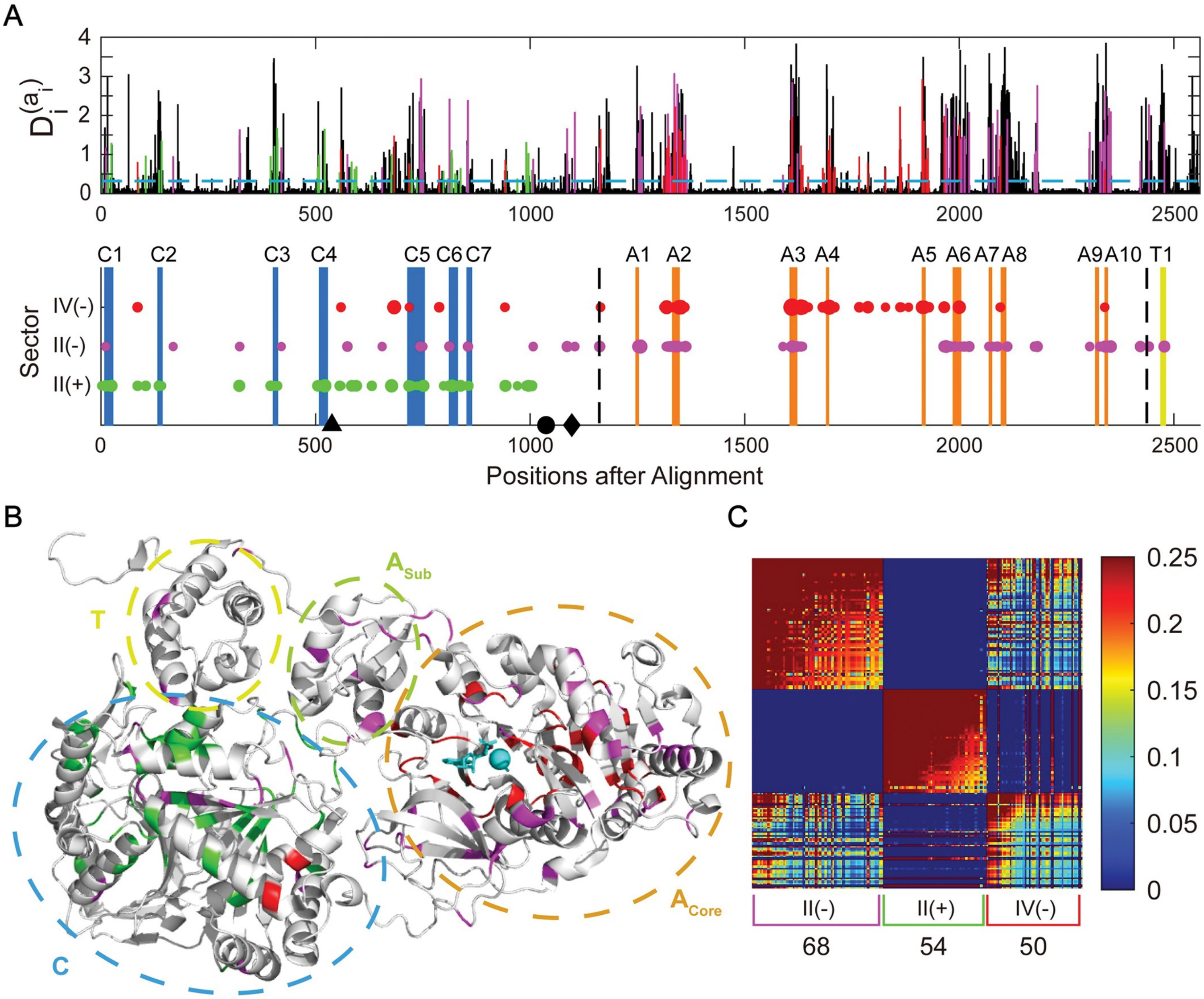


Fig 5. Statistical coupling analysis reveals overlapped sectors across the C+A+T module. A. The upper panel shows the conservation of residues in a multiple sequence alignment of 1,161 NRPS modules (containing the C, A, and T domains), quantified by the relative entropy $D_i^{(a_i)}$ in SCA method. The mean conservation level (0.32) is marked by the blue dashed line. In the lower panel, there are three groups of positions (II(+) with green, II(-) with magenta, and IV (-) with red, termed "sectors". Their corresponding conservations are marked in the same color in the upper panel. Blue bars mark C domain motifs from C1 to C7. Orange bars mark A domain motifs from A1 to A10. Yellow bar marks the T domain motif T1. Domain boundaries annotated by Pfam are divided by vertical black dashed lines. Black triangle marks the re-engineering point in the C domain reported by Bozhüyük et al. [27], black circle marks the re-engineering point in the C-A inter-domain reported by Calcott et al. [28] and black diamond marks the re-engineering point in the C-A inter-domain reported by Bozhüyük et al. [26]. **B.** Mapping three groups of correlated conservation positions into the three-dimensional structure of the NRPS module (PDB 4ZXI, containing the C, A, T, and TE domains. TE domain is hidden for clarity). Three sectors are marked in the same color as that in **(A).** C domain, A core domain, A sub domain, T domain are circled by blue, orange, yellow green, and yellow dotted line, respectively. Gly and AMP are substrates of this A domain. They and Mg^{2+} (for catalysis) are colored cyan. **C.** Heatmap of the SCA matrix after reduction of statistical noise and of global coherent correlations (see Method for details). Each sector is marked by the corresponding color bracket under the heatmap, with the number of contained residues listed. 68, 54, and 50 positions belong to the II(-), II(+), and IV(-) sectors, respectively. In each sector, residues are ordered by descending contributions, showing that sector positions comprise a hierarchy of correlation strengths.

https://doi.org/10.1371/journal.pcbi.1011100.g005

sector, residues contributing to the IV(-)and the II(-) sectors share a significant degree of covariation (Fig 5C), suggesting entanglements between substrate-specifying residues in the A domain and other regions of NRPSs. In detail, there are 5 positions (residue 85, 559, 718, 941, 942 in the MSA of Fig 5) shared between the II(+) and IV(-) sectors, and 6 positions (1164, 1351, 1358, 1605, 1631, 1635, 1965) shared between the II(-) and IV(-) sectors. In the II(+) sector, most positions are highly correlated. Therefore, overlapped positions with strong correlations were observed between the II(+) and IV(-) sectors (Fig 5C). We noted that some residues of the C4-C5 and C5-C6 intermotifs are located in the II(+) and IV(-) sectors. The N-terminal of the C5-C6 intermotif is adjacent to the G-A6 intermotif region in the crystal structure, which is part of the A domain substrate binding pocket (A3-A6, S29 Fig). Therefore, there is a possible coevolution in the overlapped region between the II(+) and IV(-) sectors.

Our coevolution analysis provides valuable insights for defining boundaries in NRPS re-engineering. According to previous works on re-engineering new proteins by recombination methods [23], it's recommended to avoid cutting into coevolution sectors when recombining sequences. However, our analysis indicated that there is no simple cutting point that can clearly separate all major sectors, as these sectors overlap with each other. Nevertheless, our findings are consistent with previously successful examples of re-engineering NRPSs in their cutting point for recombination. As shown in the second panel of Fig 5A by black marks, Bozhüyük et al. cut the NRPS at the beginning of the IV(-) sector in their 2019 work of the exchange unit condensation domain for NRPS re-engineering(XUC, black triangle); both Calcott et al. and Bozhüyük et al. in their 2018 work of defining the exchange unit (XU) as a functional unit to cut NRPS after the ending of II(+) sector (black dot and diamond, respectively).

In addition, we also analyzed 685 C+A+T+C sequences and 245 C+A+T+E sequences by SCA (S30A and S30B Fig). In both sets of sequences, we found consistent sectors with the C+A +T analysis (S31 Fig). In C+A+T+C, the first two sectors locate to the two C domains, respectively, and the third sector spans across the four domains. In C+A+T+E, one sector locates to the C domain, and another sector spans across the four domains. Intriguingly, the third sector spans across the T and E domains, suggesting a possible cut point located at the beginning of this sector, between the A and T domains (S30 Fig).

4. Factors influencing the mapping from the A domain sequence to its substrate specificity

4.1. Entanglement between substrate specificity and phylogenetic history. We then utilized a modified version of SCA to investigate the relationship between substrate specificity and residues in the A domain. Our analysis was based on 2,636 standardized A domain sequences with experimentally confirmed substrates, obtained from bacteria (2370 sequences), fungi (215), and other sources (51). These A domains were gathered from the supplemental material of SeMPI 2.0 research article [37] (S8 Table). These A domain sequences were aligned, and then their substrates were attached to the last column of the alignment (see Method for details). We primarily focused on modes that co-vary with the substrate column (Fig 6A). We found residues contributing to substrate-related sectors present throughout the A domain, with a higher concentration in the A3-A6 pocket region and a more scattered distribution in the A2-A3 region. Notably, the phylogenetic sector (Fig 6A, sector I) receives the second-largest contribution from the substrate column. It has long been recognized that phylogenetic relatedness results in consistent correlations across the whole sequence. As a result, substrate specificity, which is highly linked to the specificity-conferring code, is also linked to phylogenetic covariation throughout the A domain. This observation provides an explanation for the diversity of the specificity-conferring code even for the same substrate.

4.2. Architecture of the substrate binding pocket relates to the diversity of the specificity-conferring code. Generally, SCA method gives higher weight to conserved positions [55]. Interestingly, some substrate-related sectors tend to locate in highly variable regions (Figs 6A

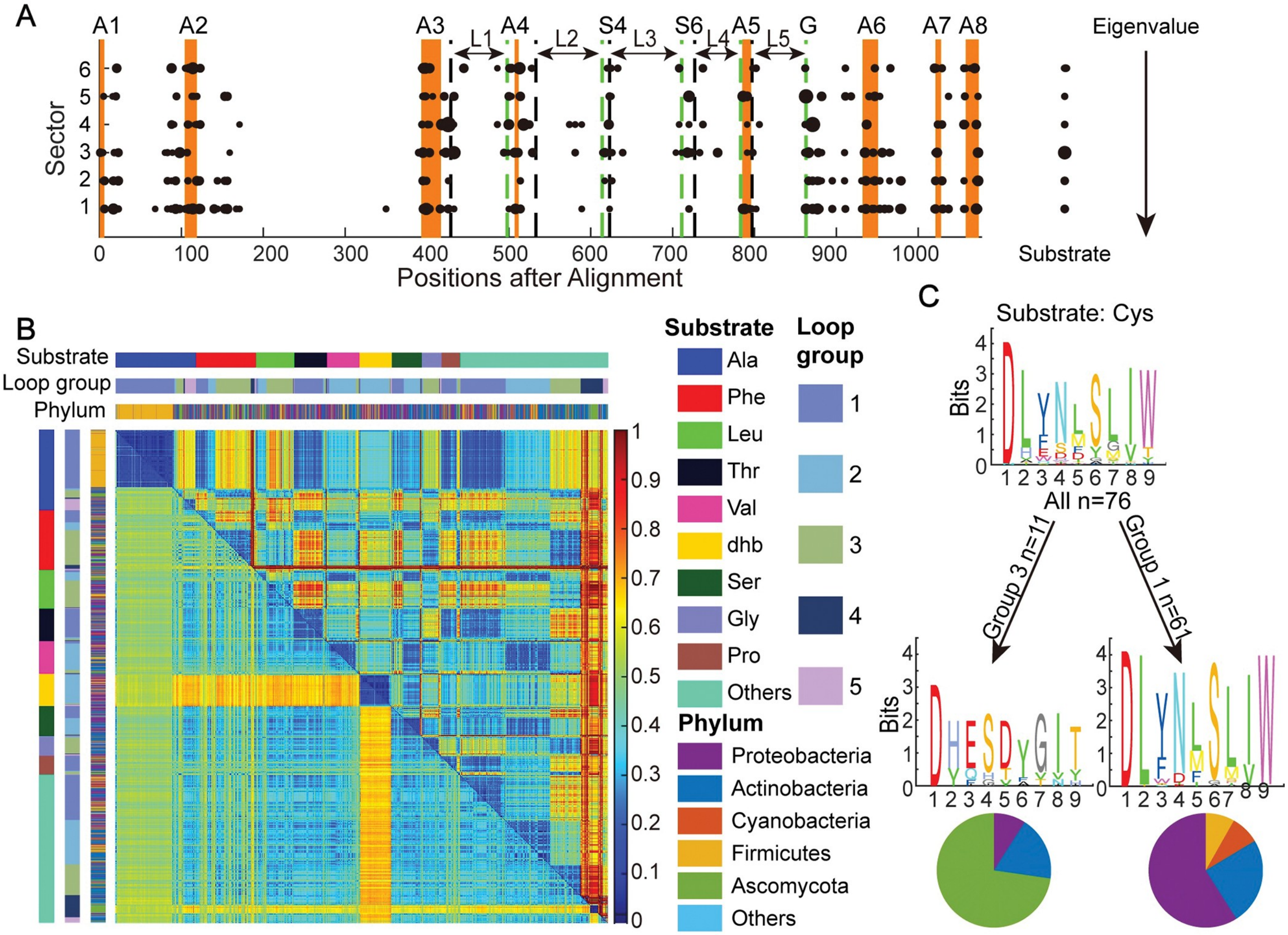


Fig 6. The specificity-conferring code of the A domain is correlated with loop length and phylogeny. A. SCA of 2,636 A domain sequences, together with their substrate specificities attached to the last column of the multiple sequence alignment. Six sectors with a high contribution from the substrate column (>0.05, the size of points on the left scales the substrate's contribution to the sector, see Method for details) are sorted by their eigenvalues. The size of points scales its contribution to the sector. Orange bars mark the A domain motifs from A1 to A8. The start and end of the five loop regions are marked by black and green dotted lines, respectively. S4 and S6 are the 4th and 6th of the specificity-conferring codes. G is the G-motif. **B.** Distance matrix of A domain. Upper right on the heatmap is the Euclidean distance of the loop length as a 5-element vector. Lower left on the heatmap is the sequence distance of the A domain. The matrix is sorted by the substrate specificity followed by the loop length group. Substrates, groups of loop length, and phylum of these A domains, are shown by colors in sidebars. **C.** Example showing that A domains conferring identical substrate exhibit distinct specificity-conferring codes, when they are categorized into different loop-length-groups. Phylum composition in each group is shown in the pie chart.

https://doi.org/10.1371/journal.pcbi.1011100.g006

and S31). In aligning to PDB structures 1AMU and 4ZXI, we observed that these variable regions represent five protein loops in the A domain binding pocket (Fig 6A, loops region marked by arrows between motif A3-A6): A3-A4, A4-S4, S4-S6, S6-A5, and A5-G. A_n is the n-th core motif of the A domain. S_n is the n-th residue in the specificity-conferring code. G is the highly conserved glycine in the G-motif aforementioned).

Constructing an efficient MSA from highly varied regions is challenging. Therefore, to further understand the functions of these specificity-related loops, we compared their basic architectures, which are quantified by their length in amino acids. Loops' starts and ends are highly conserved motifs and residues, so loop lengths are relatively independent with MSA. These highly conserved sequence pieces help to anchor the alignment, so the length of the variable regions between motifs (i.e., the loop length) does not influence the alignment much. In the variable regions, length discrepancies correlate with larger sequence distances. Therefore, we estimated the Euclidean distances between pairs of A domains using the lengths of the five loops as a five-element vector. We observed that this loop-based distance is connected with, but not totally determined by, the similarity of sequence alignments and the phylogeny of A domains (Fig 6B).

This loop-based distance matrix was clustered hierarchically, generating five groups based on their loop-length vectors (S32 Fig). Except for substrates Ala, dhb, and aad, one substrate typically corresponds to multiple loop-length groups. Importantly, the diversity of the specificity-conferring code is significantly reduced within each loop-length group. For example, for the substrate cysteine, the specificity-conferring code in the first loop-length group is readily distinguishable from that in the third loop-length group, despite these two groups containing A domains from distant phylogenies. Statistically, the diversity of the specificity-conferring code is reduced by identifying its loop-length vector group (S33 Fig). To visualize the reduction, we decoupled the specificity-conferring code for specific substrates along the phylum and loop group (S34–S41 Figs). In addition, we compared the loop length and loop group distribution between bacteria and fungi in MiBiG and the large dataset (S42 Fig). We found a significant difference in loop group preference: the loop group 3 is dominant in fungi (85.86%) but it prefers the loop group 1 (37.89%), 2 (28.43%), and 3 (25.10%) in bacteria.

In summary, we highlighted the importance of loop length for A domain substrate, which should be considered in the future A domain substrate prediction.

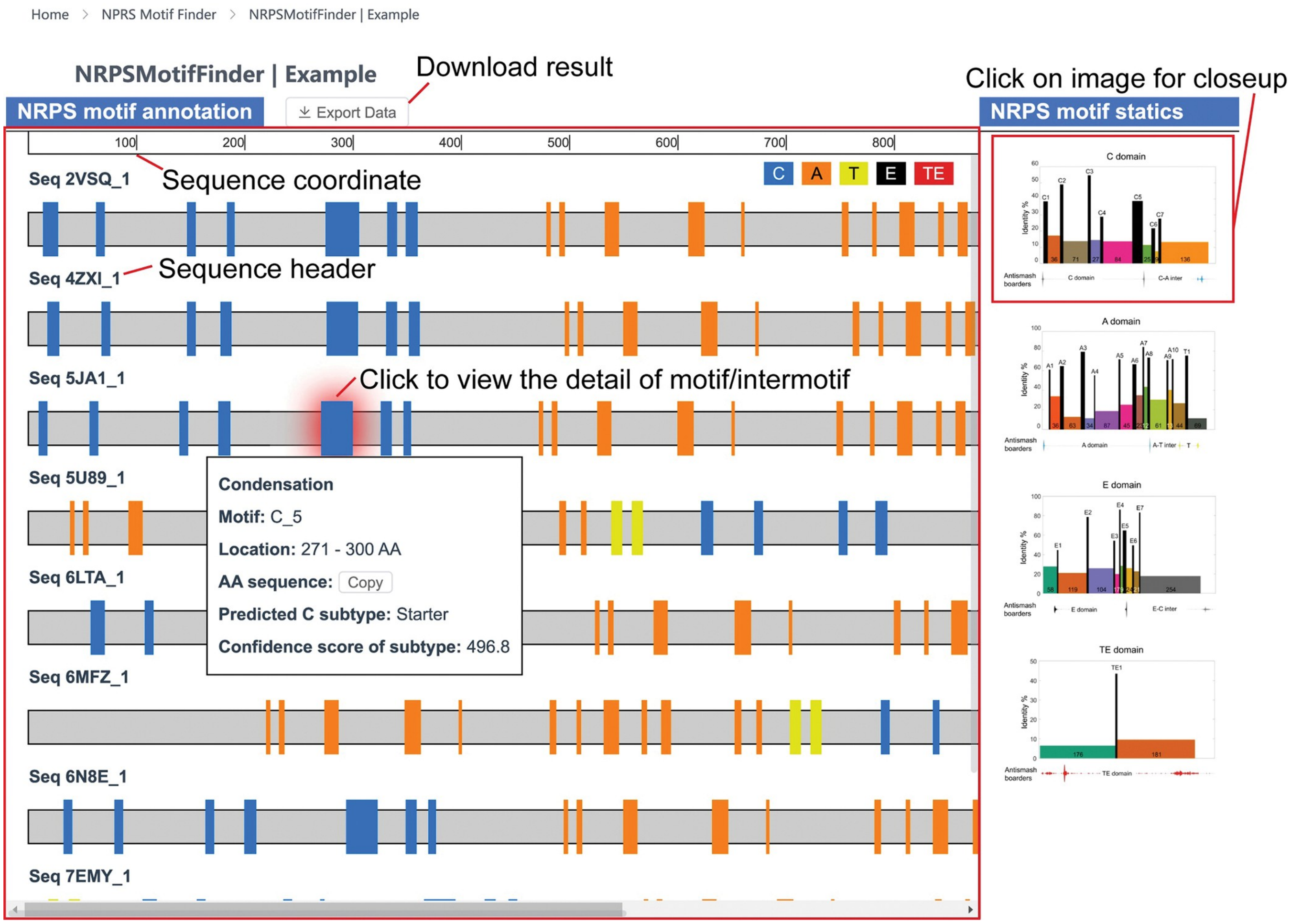
5. The NRPS Motif Finder online platform

Overall, such standardization enables statistical characterization of the sequence-function connections, supporting the use of known core motifs as "coordinates" of NPRS. To facilitate researchers in related fields, we constructed an online platform, "NRPS Motif Finder", for parsing the motif-and-intermotif standardized architecture of NRPS from its coding sequences (Fig 7, http://www.bdainformatics.org/page?type=NRPSMotifFinder). It takes any query amino acid sequences in FASTA format and can be interacted with by clicking on the website result page for more information. In addition to the 26 well-established motifs, the NRPS Motif Finder also supports the three new motifs proposed in this article. The NRPS Motif Finder online version is based on Python, and the result can be downloaded in Excel format for further analysis. For researchers with basic programming skills, we recommend the NRPS Motif Finder Matlab version for timely updates (the source code is available in S4 and S5 Files).

NRPS Motif Finder also supports the classification of 18 subtypes of the C domain, providing the first tool to annotate major fungal C domain subtypes (Fig 7). Furthermore, we provide all HMM files used in C domain subtype classification in S1 File for other researchers.

Discussion

About two decades ago, the core motifs of NRPSs were mapped out based on the early crystal structures and a limited number of sequences available at the time [60]. Since then, modules, domains, and motifs have been extensively utilized by annotation algorithms like antiSMASH [35] and Pfam [38], as well as experimental investigations [36,37], albeit with differing standards. Nowadays, with the rapid expansion of microbial sequencing data, systematic characterizations of NPRS are becoming possible. In this work, we presented a motif-and-intermotif architecture of NRPSs by partitioning the C, A, T, and TE domains by well-established core motifs. Guided by prior knowledge about these 19 motifs and domain characterizations, our standardized architecture presented a "common language" for processing NRPS sequences



Condensation
Motif: C_5
Location: 271 - 300 AA
AA sequence: Copy
Predicted C subtype: Starter
Confidence score of subtype: 496.8

Fig 7. Demonstration on the result panel of the NRPS Motif Finder. The NRPS Motif Finder result panel provides an interactive interface. The whole result could be navigated by scrolling the page, and details about motif and intermotif could be viewed by clicking the corresponding components. Especially, the predicted subtype and confidence score are displayed for C domains. The general statistics about the NRPS architecture are displayed on the right for comparison. The results could be downloaded in table format.

https://doi.org/10.1371/journal.pcbi.1011100.g007

from various sources, enabling us to gain statistical insights from a large number of sequences without the need for manual curation.

Evolution imprints protein functions into their sequences. With the help of such a standardized architecture, novel insights acquired from sequence statistics exhibit amazing concordance with our point mutation experiments, as well as earlier findings derived from structure or re-engineering efforts. For example, we identified novel conserved residues distant from known motifs, among which the G-motif seems the most intriguing. It centers on an exceptionally conserved glycine, where a single point mutation can abolish the enzyme's function, as described in Fig 4. Interestingly, by using visual cues from protein structure, the G-motif has been used as the cut point of A subdomain for successfully re-engineering A domains [101,102]. Crüsemann et al. first proposed the A domain re-engineering strategy by A subdomain swap guided by evolution in 2013 [101]. Subsequently, Kries et al. proposed another A subdomain swap strategy in 2015, as they discovered that the substrate binding region is a flavodoxin-like subdomain [102]. This subdomain starts from the middle of A3 and A4, and ends from the middle of A and A6 (the ending is exactly the G-motif, VPIGAPI in the

PheA). Recently, Thong et al. also used this A subdomain swapping strategy by CRISPR-Cas9 and successfully re-engineered A domain substrate specificity by subdomains from a range of NRPS enzymes of diverse bacterial origins [103].

NRPS complexes undergo substantial conformational changes as they exert multistep synthetase activity [63]. Identifying key residues associated with such structural flexibility is crucial for understanding NRPS functionality. In our analysis, three additional conserved motifs were identified. We hypothesized that the G-motif in the loop might be one of these key sites for structural flexibility, acting as a hinge when NRPS switches between distinct states. First, the number of hydrogen bonds around the conserved glycine changes substantially during the three function states. Also, glycine is the smallest of the 22 proteinogenic amino acids [104], allowing the greatest structural flexibility. Additionally, it has been reported that the G-X-Y or Y-X-G oligopeptides motif provides the flexibility necessary for enzyme conformation change for catalysis, where X and Y are polar and non-polar residues, respectively [105]. G-R/K-P in the G-motif fit this G-X-Y structure. Furthermore, several enzymes have demonstrated the important function of conserved glycine residues in structure flexibility. For example, conserved G76 contributes to active-site loop flexibility in the pepsin [106], and conserved G316 and G324 are the structural basis of efficient metal exchange in the cadmium carbonic anhydrase of marine diatoms [107]. In addition, the importance of this highly conserved glycine was supported in the fungal FQC biosynthesis system. We noted that the large amino acid in the 409 position mostly collides with N397 in simulated mutations (F493 in this position of MiBiG NRPS sequences, Fig 3B–3D). This may explain why mutating into large amino acids at the 409 position always reduces the FQC yield. Nevertheless, mutation into a small amino acid proline also strongly reduces the yield. It is possible that the proline's side chain locks the dihedral angle Φ of the protein backbone at approximately −65° and causes significant conformational restriction [108]. Our results implied that the flexibility of glycine in the G-motif might be important for conformational change during A domain functions, so conformational restrictions in the G-motif impede A domain function.

We have also identified other conserved residues that may play a role in the functionality of NRPS. The conserved Ala and Pro at the start of the A domain (Aα1) may bridge the A1 and A4 motifs. Similarly, the conserved Glu acid and Leu in the first helix of the T domain (Tα1) may contribute to the proper folding of the T domain, much like the three conserved Gly residues contribute to the folding and function of the green fluorescent protein [109]. From a functional perspective, these residues take part in the formation of the hydrophobic core [99,100]. In addition, 4'-phosphopantetheine cofactor required by the T domain function near residues Y748, L751, and F752 [110] (numbered in LgrA. They are L65, L68, and F69 in the original research for holo-TycC3-PCP). Besides, Y748 and F752 also are part of the hydrophobic interface, which interacts with the C domain in docking process (in LgrA. V2534 and F2538 in original research for PCP2-C3 didomain from fuscachelin) [111]. In summary, the conserved residues we identified may contribute to the stability and flexibility of NRPS proteins in large conformational changes for product synthesis. In the future, it would be meaningful to experimentally explore these new motif candidates more thoroughly, to gain deeper insights into the conformational dynamics during NRPS functioning.

Another point of consistency between our sequence characterization and earlier re-engineering studies is the SCA sectors that we identified. Among the three sectors we analyzed, the cutting point from Bozhüyük et al.'s 2019 work [27] locates right at the beginning of the red sector, which contains residues enriched in the binding pocket of the A domain. The cutting points of Calcott et al. and Bozhüyük et al.'s 2018 work [26,28] both locate to the end of the green sector, which primarily represents covariation mode in the C domain. Congruence between sector boundaries and successful re-engineering points substantiated SCA's potential to discern the "unit of evolution".

Nonetheless, our analysis also dissected and reemphasized the complexities in re-engineering NRPS. For example, sectors obtained from SCA intersperse with each other, even share covariation residues. Therefore, there is no universal cutting point in an NRPS module that does not disturb any of the major sectors. This observation is in line with difficulties in re-engineering NRPS [23]. The success of Bozhüyük et al. and Calcott et al. [26–28] may be partially attributed to their cutting points being right beside some of the major sectors, and they utilized sequences from NRPSs relatively close on phylogeny. It may be possible in the future to search phyla for NRPS systems with non-overlapping covariation sectors and use these systems as "building blocks" that can be recombined with greater freedom. As homologous recombination requires conserved sequence, the standardized architecture could aid in establishing the appropriate "cutting points" in such potential NRPS systems for re-engineering.

The phylogeny's entanglement with the substrate-specific binding pocket adds another layer of complexity in manipulating NPRSs. The contribution of the phylogeny sector to substrate specificity suggests non-degeneracy of the sequence space in determining substrate: the same substrate can be selected by distinctive binding-pocket sequences, which drift with phylogenies and may recombine with each other. Such non-degeneracy leads to the diversity of the specificity-conferring code, causing troubles in the targeted design of the A domain specificity. We demonstrated that knowing the length of the five loops in the pocket region reduces the specificity-conferring code diversity, and tried to provide an explanation by a causal diagram (S43 Fig, the causal diagram is a method for causal inference [112]). It suggests a connection between the specificity-conferring code and loop length group for a specific substrate (S34–S41 Figs). However, a mechanistic understanding of substrate selectivity is still needed to guide the rational design of the A domain. The architectural distinctions and loop length group preference between fungal and bacterial A domains discussed may shed light on the detailed mechanism of substrate selection, and help to establish the framework for developing substrate prediction algorithms that are less reliant on experimentally validated substrates.

The standardized motif-intermotif architecture not only enables efficient analysis of large datasets, but also provides a useful framework for in-depth examination of individual NRPS pathways. To facilitate researchers in related fields, we presented the NRPS Motif Finder online platform with a user-friendly tool to construct the motif-and-intermotif architecture of NRPS. The resulting architecture can be used in A domain substrate prediction based on phylogenies of A3-A6 or A4-A5. Also, this tool can be used in guiding re-engineering and new NRPS discovery, as G-motif and Aα1 both coincide with known cutting points [97,98,102]. In addition, NRPS Motif Finder supports the classification of C domain subtypes into 18 kinds, offering the first tool to cover most of the fungal C domain subtypes.

In terms of limitations, most of our results are based on computation. Despite the high degree of concordance between computational predictions and our point mutation experiments on the G-motif, more comprehensive experiments might be developed in the future to explore the roles of other conserved residues we identified. Additionally, given the diversity of NRPSs, despite our use of the largest A domain database with known substrates, A domains selecting rare substrates may be unrepresented. Moreover, the identification of the new C domain subtypes, while promoting the understanding of fungal NRPS, also highlighted the current limitation of the dataset, which is heavily biased towards bacterial BGCs. The catalytic mechanism and evolutionary history of new C domain subtypes are still unclear and require more investigation. Overall, our effort is a preliminary investigation into the possibilities of standardized architecture in modular enzymes. Much more discoveries could be achieved in the future with the rapid expansion of microbial sequencing data.

Method

Data acquisition

We downloaded the MiBiG database version 2.0 [33], then extracted sequence information from this database. In this research, only C/A/T/E/TE domains from NRPS clusters were considered. Finally, 326 NPRS sequences from 264 species were obtained, with 1,864 A domains, 1,765 C domains, 1,803 T domains, 310 E domains, 280 TE domains. In total, there are 1,161 C+A+T modules, 685 C+A+T+C modules, and 245 C+A+T+E modules.

A domains sequences with known substrates were gathered from the supplemental material of SeMPI 2.0 research article [37]. Some of the A domains from SeMPI overlap with those in MiBiG (http://sempi.pharmazie.uni-freiburg.de/database/MiBiG).Sequences which had deleted in MiBiG database or UniProt [113] were removed from our dataset. Finally, we collected 2,636 A domain sequences. Sequence details see S8 Table. Pfam seed alignments were downloaded for comparison with antiSMASH domains (PF00668 for C domain, PF00501 for A domain, PF00550 for T domain).

Sequences with high identity were removed (S44 Fig). All analysis was applied to the remaining sequences.

For the large dataset to confirm our results, we downloaded all complete bacterial genomes (30,984) and all assembly-level (including "contig", "scaffold", "chromosome", and "complete" levels) fungal genomes (3,672) from NCBI (as of 2022/10/23). Genomes were deduplicated by Mash distance with a cutoff of 0.004 (> = 99.6% genome similarity) using Mash tool v2.3 [114]. Representative genomes were chosen by picking the genome with the longest genome size. After removing redundancy, we had 16,820 bacterial and 2,505 fungal genomes. The information of used genomes was summarized in S9 Table. Species information was obtained by TaxonKit [115].

Length threshold for multi-domain NRPS from MiBiG

Considering the NRPS sequence length distribution, we filtered sequences by the following thresholds. For C+A+T, the sequence length should be more than 1000 and less than 1300; for C+A+T+C, the sequence length should be more than 1500 and less than 1750; for C+A+T+E, the sequence length should be more than 1350 and less than 2350. In practice, 14 sequences in the C+A+T+C type are removed because their secondary C domain, rather than the first C domain, align with the first C domain of other sequences. And a few sequences from C+A+T +E were removed because they are incomplete at start.

Multiple sequence alignment

If not specifically stated, the MSA in our work were constructed by Clustal Omega 1.2.4 [116]. For the construction of C domain subtype reference HMM files, we prepared MSA by Muscle v5 referring to previous studies [5,61].

antiSMASH annotation

Data obtained from MiBiG database has already been annotated by antiSMASH 5.1 [33]. BGCs in the large dataset were annotated by antiSMASH 6 [35].

Conserved motif detection for NRPS domain

As shown by S45 Fig, first, manual curation of the conserved motifs from the literature was performed for all domains. That resulted in 7 motifs for C, 10 motifs for A, 1 motif for T, 7 motifs for E, and 1 motif for TE [60]. A table of these conserved motifs according to the previous literature was summarized in S10 Table.

Second, we obtained reference sequences for each type of NRPS domain. The reference sequences of the A, T, and TE domains were from surfactin A synthetase C which has resolved crystal structure (PDB: 2VSQ) [58]. The reference sequence of the E domain was retrieved from the initiation module of tyrocidine synthetase A (PDB: 2XHG) [59]. The reference sequences of C domain are described in the section of C domain subtypes classification. The motif locations on all reference sequences above were manually identified, according to the results from step 1.

Third, for every query sequence, global alignment was performed against possible reference sequences. Regions on the query sequence that align with the conserved motifs in the reference sequence were recorded as "motifs".

For clarification, only sequences with the prevalent motif length were used in the construction of motif sequence logos. The prevalent motif lengths of each domain were recorded in S10 Table.

C domain subtypes classification

First, we curated sequences from the E domain and 18 subtypes of the C domains as references. 14 of 18 subtypes of the C domains were sourced from the CDD [69], and the remaining 4 were obtained from literatures [33,66,70,71] (S1 Table). The original "CT" subtype sequences were classified into CT, CT-DCL and CT-A subtypes guided by our phylogenetic anlaysis.

We prepared MSA for each subtype by Muscle v5 [5] and constructed reference HMM files by hmmer v3.1b2 [74]. The MSA results of each C domain subtype were attached in the S2 File.

Finally, we aligned the query sequence to profile using hidden Markov model alignment. In the NRPS Motif Finder online version (python), it is achieved by the hmmscan function in hmmer3. In the NRPS Motif Finder Matlab version, it is achieved by the hmmprofalign function in Matlab.

Calculation of mutual information

In generating the fourth panel of Fig 1A, mutual information between the residues and the chirality subtypes of C domains was calculated. The C domain subtype was defined by the antiSMASH annotation, including "LCL", "DCL", "Dual", "Starter", and "Heterocyclization". In the multi-alignment of C domains, the reduced Shannon entropy on the amino acid composition at position *i* given the subtype, was calculated as the mutual information about chirality at position *i*.

The fourth panel of Fig 1B was obtained by calculating the mutual information between substrate specificity and amino acid composition at each position of the multi-alignment of A domains.

Phylogenetic analysis

Phylogenetic trees of the condensation domain superfamily were reconstructed in Fig 2A. Details of sequences see **C domain subtypes classification**. Multiple sequence alignments of protein sequences were prepared with Muscle v5 with default parameter (muscle -super5) and trimmed sequence at both ends manually. The maximum-likelihood phylogenetic tree was reconstructed in IQ-TREE 2.1.2 [117], using the best-fit model of protein evolution for each alignment (LG+F+R10 model) as chosen by ModelFinder (Bayesian Information Criterion) [118]. Branch support was assessed by bootstrapping (1000 bootstrap replicates). Following previous research, papA and WES are taken as outgroups to root the phylogenetic tree [65]. See details in S3 File. The phylogenetic tree was visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Protein structure modeling

The protein structure of FmqC was predicted by Phyre2 [96]. Phyre2 used high-score PDB templates as 5U89, 6MFZ, and 6N8E for predicting the FmqC structure, with alignment coverages of 93%, 94%, 87%, and identities of 25%, 23%, 21%, respectively. Mutation of FmqC G-motif from Gly to Ala was simulated by mutagenesis function in the PyMOL (http://www.pymol.org/pymol).

The position of motif boundaries for multi-sequence

The motif boundaries in Figs 3, 5, 6, S30, and S31 show the average motif boundary positions from each single sequence. Motif boundary positions are consistent in different sequence after MSA other than the C domain, where the conserved motifs vary by subtypes (S10 Table).

The calculation of conservation in NRPS protein

For a quantitative perspective, we use ConSurf [119] and AACon [120] to calculate the conservation of specific positions in the NRPS protein. The results of two methods are similar in very conserved positions, but a little different in middle conserved positions. The results are displayed in S5 Table.

Statistical coupling analysis

The statistical coupling analysis was performed with the binary approximation method [55]. The cleaned correlation matrix is obtained using the method described in the same article [55].

To investigate substrate-related positions, we attached substrates as the last column of the multiple sequence alignment. Then we applied SCA to this modified MSA. The least frequent substrate was taken as background of this position.

Sector identification in SCA

We use the same notation in Halabi et al. [55]. Except two sectors of the second mode with opposite eigenvalues, third sector was chosen empirically. The bra-ket notation is such that $|k\rangle$ denotes the k^{th} eigenvector and $\langle i|k\rangle$ the weight for position i along eigenvector k. The threshold ε to separate significant weights along an eigenvector from statistical noise is 0.05, which was used in Halabi et al. [55].

For C+A+T, green sector is defined as $\langle i|2\rangle>\varepsilon$; magenta sector is defined as $\langle i|2\rangle<\varepsilon$; red sector is defined as $\langle i|6\rangle<\varepsilon$.

For C+A+T+C, green sector is defined as $\langle i|2\rangle>\varepsilon$; magenta sector is defined as $\langle i|4\rangle>\varepsilon$; red sector is defined as $\langle i|2\rangle<\varepsilon$.

For C+A+T+E, green sector is defined as $\langle i|2\rangle<\varepsilon$; magenta sector is defined as $\langle i|3\rangle>\varepsilon$; red sector is defined as $\langle i|2\rangle>\varepsilon$.

For A domain, from bottom to top, sectors are defined as $\langle i|1\rangle>\varepsilon$, $\langle i|2\rangle>\varepsilon$, $\langle i|3\rangle>\varepsilon$, $\langle i|4\rangle>\varepsilon$, $\langle i|5\rangle>\varepsilon$ and $\langle i|6\rangle<\varepsilon$.

Loop length clustering

The sum of the Euclidean distance of five loops between different A domain sequences was calculated for distance matrix. To avoid the influence of local maximum value on the heatmap, the upper-bound in distance was set to 12. Then hierarchical clustering analysis is applied to this distance matrix (S32 Fig).

Loop length profile calculation method

$$\bar{L_{i,j}} = \frac{L_{i,j} - \min(L_i) + 1}{mean(L_i) - \min(L_i) + 1}$$

Related to S32 Fig, $\bar{L_{i,j}}$ is normalized length of loop i in jth row. $L_{i,j}$ is actual length of loop i in jth row. L_i is a column vector containing all length of loop i.

Experimental Materials and General Methods

The plasmids and strains utilized in this work are listed in Table B in S1 Text. The oligonucleotide primers synthesized by Shanghai Sango Biotech are given in Table C in S1 Text. PCR reactions were carried out with FastPfu high-fidelity DNA polymerase (Transgene Biotech). *Escherichia coli* strain DH5α was propagated in LB medium with appropriate antibiotics for plasmid DNA, and plasmid DNA was prepared using the Plasmid Mini Kit I (Omega). Automated DNA sequencing was performed by Shanghai Sango Biotech. PCR screening for transformants was carried out using 2×GS Taq pcr mix (Genesand).

A. *fumigatus* strains were cultivated in glucose minimal medium (GMM) [121] culture medium at 28˚C for 3 days in the dark. The media and mycelia were extracted with ethyl acetate, and then evaporated under reduced pressure. The extract was dissolved in 1 mL methanol, and 5 μL of the solution was directly injected for LC-MS analysis. LC-MS analysis was performed on an Agilent HPLC 1200 series system equipped with a single quadrupole mass selective detector, an electrospray ionization (ESI) and an Agilent 1100LC MSD model G1946D mass spectrometer by using a Venusil XBP C18 column (3.0 x 50 mm, 3 μm, Bonna-Agela Technologies, China). Data collected in positive mode. Water with 0.1% (v/v) formic acid and acetonitrile were used as mobile phase with a flow rate of 0.5 mL/min. For analysis of the extracts, a linear gradient of acetonitrile in water (10–45%, v/v) in 30 min was used and washed with 100% (v/v) acetonitrile for 5 min.

Gene cloning, plasmid construction, and genetic manipulations

(a) Creation of fmqC deletion strains in A. *fumigatus*. For site-directed mutagenesis of G-motif in *A*. *fumigatus*, *fmqC* was fully deleted in Cea17.2 to increase the rate of homologous recombination and reduce the difficulty of screening transformants. The FmqC (accession: EDP49773.1) deletion strain was created in Cea17.2 by replacing the *fmqC* with hygromycin phosphotransferase gene (*hph*) using modified double joint PCR described previously [122] consisting of the following: 1 kb DNA fragment upstream of *fmqC*, a 1.9 kb DNA fragment of *hph*, and a 1 kb DNA fragment downstream of *fmqC*. Third round PCR product were purified for protoplast transformation (fragment concentration greater than 300ng/μL). Polyethylene glycol (PEG) based transformation of *A*. *fumigatus* was done as previously described [123]. The fragment concentration was 3 μg per 100 μL of protoplasts. The mutants were confirmed by using diagnostic PCR and the correct transformants were used for subsequent analysis.

(b) Creation of fmqC mutation strain in A. *fumigatus*. Seven plasmids were generated, one which included a full-length copy of *fmqC* (pYJY25), and the others were mutated *fmqC*

(pYJY26-31). The pYJY25 plasmid was assembled by amplifying 1 kb DNA fragment upstream of *fmqC*, *fmqC*, *A*. *fumigatus pyrG* as the selectable marker, and a 1 kb DNA fragment downstream of *fmqC*. The fragments were assembled into a full plasmid by homologous recombination using MultiS One Step Cloning Kit (Vazyme). pYJY26-31 were assembled using the same fragments and method, except that the targeted mutations of 409Gly in G-motif were introduced by using primers containing the mutation and amplifying *fmqC* in two separate PCR reactions. After plasmid construction, the fused 8kb PCR fragments with the point mutation were purified and used for transformation to TYJY81. Transformants were confirmed by sequencings to obtain TYJY82-88 for the subsequent experiments.

Supporting information

S1 Fig. Sequence logo of the seven C motifs among the multialignment of 1758 C domains (first row), and among six subtypes of C domains (second row) in MiBiG. The ranges of y-axis in sequence logo figures all are 0~4.4 bits. The numbers of each C domain subtypes are 809 (LCL), 385 (DCL), 114 (Starter), 300 (Dual), 114 (Cgly), and 36 (Heterocyclization).
(PNG)

S2 Fig. Sequence logo of the ten A motifs among the multialignment of 1859 A domains in MiBiG. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S3 Fig. Sequence logo of T1 motif and the length distribution of the T1-C1 region in MiBiG. A. Sequence logo of the T1 motif. The y-axis range in sequence logo figure is 0~4.4 bits. **B.** Length distribution of the T1-C1 region for different subtypes of C.
(PNG)

S4 Fig. Sequence logo of the seven E motifs among the multialignment of 310 E domains in MiBiG. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S5 Fig. Sequence logo of the TE1 motifs among the multialignment of 280 TE domains in MiBiG. The y-axis range in sequence logo figure is 0~4.4 bits.
(PNG)

S6 Fig. Length distributions of C, A, and T domains, in MIBIG and in Pfam seeds.
(PNG)

S7 Fig. Sequence logo of the twelve A motifs and two T motifs among 95,582 A domains and 86,688 T domains in bacteria. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S8 Fig. Sequence logo of the seven E motifs among 14,502 E domains in bacteria. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S9 Fig. Sequence logo of the TE1 motifs among 23,590 TE domains in bacteria. The y-axis range in the sequence logo figure is 0~4.4 bits.
(PNG)

S10 Fig. Sequence logo of the twelve A motifs and two T motifs among 40,458 A domains and 26,651 T domains in fungi. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S11 Fig. Sequence logo of the seven E motifs among 3,982 E domains in fungi. The y-axis ranges in sequence logo figures all are 0~4.4 bits.
(PNG)

S12 Fig. Sequence logo of the TE1 motifs among 4,008 TE domains in fungi. The y-axis range in sequence logo figures is 0~4.4 bits.
(PNG)

S13 Fig. Comparison of NRPS A and T domain architecture between bacteria and fungi source. For comparison, only A domains which have the same motif length with reference A domain are used. Sequence numbers of intermotifs (A1-A2, A2-A3, A3-A4, A4-A5, A5-G motif, G motif-A6, A6-A7, A7-A8, A8-A9, A9-A10, Tα1-T1) are 75,407 in bacteria and 17,890 in fungi. Sequence numbers of A1-Tα1 intermotif (actually interdomain) in bacteria source are 69,440 while they in fungi source are 12,194 because only part of A domains are adjacent with the T domain. Sequence numbers of Tα1-T1 intermotif in bacteria are 85,755 while they in fungi are 25,069.
(PNG)

S14 Fig. Comparison of NRPS E domain architecture between bacteria and fungi source. For comparison, only A domains which have same motif length with reference A domain are used. Sequence numbers of intermotifs (E1-E2, E2-E3, E3-E4, E4-E5, E5-E6, E6-E7) are 12,875 in bacteria source and 2,852 in fungi source. Sequence numbers of intermotifs (actually interdomain) in bacteria are 12,618 for T1-E1 and 8,353 for E7-C1 while they in fungi are 2,088 for T1-E1 and 2,530 for E7-C1.
(PNG)

S15 Fig. Sequence logo of the seven C motifs among the multialignment of 2,572 C domains (first row), and among 15 subtypes of C domains (second row) in MiBiG. There are 2,572 C domains with subtype prediction scores more than the threshold (200) and a count of domain subtype sequences of more than 3. The y-axis ranges in sequence logo figures all are 0~4.4 bits. The numbers of each C domain subtype are labeled at the end. For clarification, only motifs that have prevalent length are used in plotting. The actual total C domain number used in the figure for the specific motif is shown in the title. There are some gaps at both ends of the sequence for alignment. In motif C5, there are some interior gaps labeled in red to align with motif C5 of other subtypes.
(PNG)

S16 Fig. Sequence logo of the seven C motifs among 77,152 C domains (first row), and among 13 subtypes of C domains (second row) in bacteria. There are 77,152 C domains with subtype prediction scores more than the threshold (200) and a count of domain subtype sequences of more than 3. The y-axis ranges in sequence logo figures all are 0~4.4 bits, except it's 0~1 for the bL subtype and 0~2.1 for the PS subtype because these subtypes are few in the sequence number. The numbers of each C domain subtype are labeled at the end. For clarification, only motifs that have prevalent length are used in plotting. The actual total C domain number used in the figure for the specific motif is shown in the title. There are some gaps at both ends of the sequence for alignment. In motif C5, there are some interior gaps labeled in red to align with motif C5 of other subtypes.
(PNG)

S17 Fig. Sequence logo of the seven C motifs among 34,269 C domains (first row), and among 11 subtypes of C domains (second row) in fungi. There are 34,269 C domains with subtype prediction scores more than the threshold (200) and a count of domain subtype

sequences of more than 3. The y-axis ranges in sequence logo figures all are 0~4.4 bits. The numbers of each C domain subtype are labeled at the end. For clarification, only motifs that have prevalent length are used in plotting. The actual total C domain number used in the figure for the specific motif is shown in the title. There are some gaps at both ends of the sequence for alignment. In motif C5, there are some interior gaps labeled in red to align with motif C5 of other subtypes.
(PNG)

S18 Fig. Comparison of NRPS C domain architecture between bacteria and fungi. For comparison of intermotif length between different C domain subtypes and E domain, we chose the conserved positions which exist in all C domain and E domain as start and end of intermotif. T1/ACP1-C1/E1 ends before the conserved "Q" in C1 and E1 (the conserved "E" in LCL-A subtype). C1-C2 starts with the conserved "Q" in C1 (the conserved "E" in LCL-A subtype), and ends before the second conserved "R" in C2. C2-C3 starts before the second conserved "R" in C2 and ends before the conserved "D" in C3. C3-C4 starts before the conserved "D" in C3 and ends before the second conserved "Y" in C4. C4-C5 starts before the second conserved "Y" in C4 and ends before the conserved "G" in C5. C5-C6 starts before the conserved "G" in C5 and ends before the conserved "P" in C6. C6-C7 starts before the conserved "P" in C6 and ends before the conserved "F" in C7 (the conserved "F" in LCL). C7-Aα1 starts before the conserved "F" in C7 (the conserved "F" in LCL) and ends before Aα1. Sequence numbers of intermotifs (C1-C2, C2-C3, C3-C4, C4-C5, C5-C6, C6-C7) are 77,152 in bacteria and 34,269 in fungi. Sequence numbers of intermotifs (actually interdomain) in bacteria are 28,185 for T1-C1 and 33,176 for C7-Aα1 in LCL subtype C domain, 6,967 for E7-C1 and 6,752 for C7-Aα1 in DCL subtype C domain, 3,860 for T1-C1 and 4,495 for C7-Aα1 in Dual subtype C domain, 6,967 for T1-C1 and 6,752 for C7-Aα1 in starter subtype C domain and while they in fungi are 434 for T1-C1 and 616 for C7-Aα1 in LCL subtype C domain, 808 for E7-C1 and 1,931 for C7-Aα1 in DCL subtype C domain and 24 for T1-C1 and 21 for C7-Aα1 in Dual subtype C domain.S19 Fig.
(PNG)

S19 Fig. Sequence logo of highly conserved positions near known motifs. The black box shows known core motifs in the A domain. The black triangle shows highly conserved positions in multialignment from the 1,161 C+A+T NRPS sequences from MiBiG database.
(PNG)

S20 Fig. The mutual information between residues in the A5-A6 and A domain substrate specificity. Same as that in the fourth panel of Fig 1B, but this plot focuses on regions between motif A5 and motif A6. The position of conserved Gly is indicated by the black arrow.
(PNG)

S21 Fig. The G-motif in different function states of LgrA structure. Related to Fig 3B–3D. The view zoomed in toward the region near the G-motif emphasized by red sticks. Hydrogen bonds were shown in the blue dashed-line. Different domains marked by different colors (F: formylation domain, colored by magenta; T: thiolation domain, colored by cyan; A: adenylation domain, Acore (orange) covers A1-A8 of A domain, Asub (yellow) covers A9-A10, Apocket (yellow green) covers A3-A6 of A domain).
(PNG)

S22 Fig. The sequence logo of G-motif in A domains with known structures and FmqC. Non-NRPS means A domains from these proteins which aren't NRPS. Non-CAR means A domains from these proteins which aren't CAR. Non-NRPS&Non-CAR means A domains

from these proteins which aren't NRPS or CAR.
(PNG)

S23 Fig. The equivalents of N397, G409 (G-motif), and S491 in FmqC mapped to known structures. From the first to the 8-th figure, the known structures are CAR protein (PDB: 6OZ1), DltA (PDB: 7VHV), and NRPS (PDB: 1AMU, 2VSQ, 6IYK, 5WMM, 6OZV, and 7LY7). Residues in G-motif were marked by blue. The equivalent N397 and S491 were marked by yellow and cyan. The ligand molecule was marked in green. The PDB IDs of proteins are shown below. Only the CAR protein doesn't contain the G-motif, but it still has equivalents of N397 and S491. G-motif is in close proximity to the adenylate part of the ligand, suggesting a potential gatekeeper role.
(PNG)

S24 Fig. LC-MS analysis in the strain construction. Wild type (Cea17.2, first row), Δ*fmqC* (second row) and the control *fmqC* (third row).
(PNG)

S25 Fig. Aα1 motif in different states of LgrA structure. Similar to S21, but for the Aα1 motif. The view zoomed in towards the region near the Aα1 motif emphasized by red sticks.
(PNG)

S26 Fig. The interaction near Aα1 motif in different states of LgrA structure. Related to S25. Similar to Fig 3B, but shows chemical interactions and secondary structures surrounding the Aα1 motif at the substrate donation state (A), the thiolation state (B) and the condensation state (C). Of note, in these residues, only T230 and Y231 use the hydroxyl group in the side chain to form hydrogen bonds. Other hydrogen bonds, on the other hand, are formed by the common α-carboxyl group and α-amino group in the main chains.
(PNG)

S27 Fig. Conserved residues of T domain in the condensation state. The structure is obtained from LgrA in the condensation state (PDB: 6MFZ). Cyan color shows the T domain defined by Pfam. A small white region isn't covered by Pfam, although it is visually one part of the first helix of T domain. N-terminal and C-terminal were marked by green texts. Conserved resides were marked by red sticks with their one letter labels shown. Reside labels in the top view were hidden for visual clearness. **A.** Side view of T domain. **B.** Top view of T domain.
(PNG)

S28 Fig. Eigenvalue spectra for the SCA matrix of C+A+T modules MSA and random MSA. Eigenvalue spectra for the SCA matrix corresponding to the 1,161 C+A+T modules (top panel) and for 100 trials of randomizing sequences alignment (bottom panel). The randomization process scrambles the order of amino acids in each alignment column independently, which did not change amino acid frequencies at positions. The black dashed-line marked the maximum of eigenvalues from randomized alignments. This analysis shows that only a small part of the spectrum (26 out of 2560 total eigenvalues) is significant given sample size.
(PNG)

S29 Fig. The interaction of C5-C6 with other domains in LgrA structure. The formylation domain in the first module (F1) of LgrA is hidden for visual clearness. The colors of each domain are noted at the top of the figure. The C domain is split into the N-terminal subdomain (N-subdomain, covering C1-C4) and the C-terminal subdomain (C-subdomain, covering C5-C7), referring to previous research (PMID: 23756159). The active site histidine (the second histidine in C3 motif HHxxxD), the residues in C5-C6 intermotif interacting with T

domain or A domain, and the residues in T domain and A domain interacted by these residues are shown as stick format. Hydrogen bonds were shown in yellow dashed-line with distances by black.
(PNG)

S30 Fig. SCA analysis for C+A+T+C and C+A+T+E four domains NRPS sequences. Similar to Fig 3A, we also analyzed 685 C+A+T+C (**A**) and 245 C+A+T+E (**B**) module NRPS sequences by SCA. Although sequence numbers are less than C+A+T composition NRPS (1,161), they also could provide insights for NRPS reengineering. Except known cutting points, the SCA result of C+A+T+E NRPS sequences indicated the junction between A and T domains may be a potential cutting point. Of note, cutting point proposed by Mootz et al.[1]. was for C+A+T+C in original research.
(DOCX)

S31 Fig. The gap frequency in the MSA of 2,636 A domains. Related to Fig 6A, upper panel shows the gap frequency in the MSA of 2,636 A domains. Some residues in the substrate-related sectors were found in the highly variable loop regions (L1-L5). These regions contain high numbers of gaps in MSA, and are usually loops in the structure.
(PNG)

S32 Fig. Clustering and groups of five loops. A. Hierarchical clustering of the A domains based on the Euclidean distances of their lengths in five loops. A domains were categorized into five groups based on their loop-length vectors. For visual clearness, in calculation, the Euclidean distances which are more than 12 are set as 12 before normalizing. **B.** Loop length profiles of five groups shown in **A**. More details in Method.
(PNG)

S33 Fig. The entropy and conditional entropy of the specificity-conferring code given different constraints. The first column shows the entropy of the specificity-conferring code for different substrates. The proteinogenic amino acids are named according to the standard amino acid one-letter code. Abbreviations of non-proteinogenic amino acid substrate: aad = 2-amino-adipic-acid, bht = beta-hydroxy-tyrosine, dab = diaminobutyric acid, dhb = 2,3-dihydroxy-benzoic acid, dhbu = 2,3-dehydroaminobutyric acid, dhpg = 3,5-dihydroxy-phenyl-glycin, horn = hydroxy-L-ornithine, hpg = 4-hydoxy-phenyl-glycine, orn = ornithine and pip = pipecolic acid. Only A domains from 5 main phylum are used to calculate entropy (2564/2623 = 97.8% sequences). The sequence number of each substrate was marked in the bracket after substrate names on the y labels (the first is the number used in the calculation of entropy, and the second is the total number of this substrate in our datasets). Second to fourth columns show the conditional entropy of the specificity-conferring code given information about the phylum, the loop group, and the phylum with the loop group, respectively. Information from the phylum and the loop group could both reduce the uncertainty of the specificity-conferring code, and they together could further reduce the uncertainty.
(PNG)

S34 Fig. The sequence logo of the specificity-conferring code for substrate alanine in the dimension of phylum and loop group. 9 of 10 the specificity-conferring code are displayed. The last one is conserved lysine (K) in the A10 motif. It wasn't shown because our A domain sequences only cover A1-A8. Sequence logo will not be plotted, if the number of sequences is less than 3. Substrate abbreviation: A = alanine.
(PNG)

S35 Fig. The sequence logo of the specificity-conferring code for substrate phenylalanine in the dimension of phylum and loop group. Similar to S34, but for substrate phenylalanine (F).
(PNG)

S36 Fig. The sequence logo of the specificity-conferring code for substrate leucine in the dimension of phylum and loop group. Similar to S34, but for substrate leucine (L).
(PNG)

S37 Fig. The sequence logo of the specificity-conferring code for substrate valine in the dimension of phylum and loop group. Similar to S34, but for substrate valine (V).
(PNG)

S38 Fig. The sequence logo of the specificity-conferring code for substrate tyrosine in the dimension of phylum and loop group. Similar to S34, but for substrate tyrosine (Y).
(PNG)

S39 Fig. The sequence logo of the specificity-conferring code for substrate 2-amino-adipic-acid in the dimension of phylum and loop group. Similar to S34, but for substrate 2-amino-adipic-acid (aad).
(PNG)

S40 Fig. The sequence logo of the specificity-conferring code for substrate glutamine in the dimension of phylum and loop group. Similar to S34, but for substrate glutamine (Q).
(PNG)

S41 Fig. The sequence logo of the specificity-conferring code for substrate diaminobutyric acid in the dimension of phylum and loop group. Similar to S34, but for substrate diamino-butyric acid (dab).
(PNG)

S42 Fig. Loop length and group distributions in bacteria and fungi. A. Comparison loop length between bacteria and fungi in the MiBiG database. The numbers of A domains are 2,370 and 215 for bacteria and fungi, respectively. **B.** Comparison loop length between bacteria and fungi in the larger dataset. The numbers of A domains are 61,494 and 4,484 for bacteria and fungi, respectively. **C.** Loop group distribution in bacteria and fungi in the larger datasets. The numbers of A domains are the same as B. Their loop groups are predicted as the closest one loop group in the MiBiG database by calculating Euclidean distance. A small amount of data (<5%) is not counted because they are the same distance from multiple loop groups.
(PNG)

S43 Fig. Causal analysis of A domain substrate specificity. The specificity-conferring code distance used is alignment-score distance. A domain sequence distance used is p-distance. The loop length distance used is Euclidean distance. r is Pearson correlation coefficient. **A.** Causal diagram of A domain substrate specificity. **B.** Relationship between the specificity-conferring code distance and Loop length Euclidean distance. **C.** Relationship between Loop length Euclidean distance and A domain sequence distance. **D.** Relationship between the specificity-conferring code distance and A domain sequence distance. **E.** Relationship between the specificity-conferring code distance and Loop length Euclidean distance for A domains with substrate Ala. Similar to B, but for 430 A domains activating Ala as substrate.
(PNG)

S44 Fig. Protein sequence pairwise distance distribution. A. Pairwise distance distribution of 1,161 C+A+T NRPS sequences. The calculation was based on the p-distance, representing the fraction of amino acid being different after global alignment. **B.** Same as that in **A**, but for 685 C+A+T+C NRPS sequences. **C.** Same as that in **A**, but for 245 C+A+T+E NRPS sequences. **D.** Same as that in **A**, but for 2,636 A domain sequences.
(PNG)

S45 Fig. Workflow of detecting conserved motifs in NRPS domain. Illustration of the steps in locating known core motifs to query NRPS sequences. First, we curated known core motifs and reference sequences from the literature. Then known motifs on reference sequences were mapped according to previous research, with their locations recorded. Finally, multiple sequence alignment was performed between reference sequences and query sequences. Locations of core motifs in query sequences were inferred by the aligned reference sequences.
(PNG)

S1 Table. Source of C domain subtype reference sequences.
(XLSX)

S2 Table. Subtype prediction by NRPS Motif Finder and annotation by antiSMASH v5 for C domains in MiBiG.
(XLSX)

S3 Table. Subtype prediction by NRPS Motif Finder and annotation by antiSMASH v6 for C domains in 16,820 bacterial genomes.
(XLSX)

S4 Table. Subtype prediction by NRPS Motif Finder and annotation by antiSMASH v6 for C domains in 2,505 fungal genomes.
(XLSX)

S5 Table. Amino acid composition and conservation in three potential motifs.
(XLSX)

S6 Table. All available structures of AMP-binding-domain-containing proteins from the PDB database.
(XLSX)

S7 Table. Structural effects of point mutations in G-motif G409 of FmqC estimated by Missense3D.
(XLSX)

S8 Table. 2,636 A domains sequences information.
(XLSX)

S9 Table. Bacterial and fungus genome information used in this study.
(XLSX)

S10 Table. The definition of conserved motifs in NRPS domains and the positions in reference sequences.
(XLSX)

S1 File. C domain subtype reference HMM files.
(ZIP)

S2 File. C domain subtype reference sequences (MSA).
(ZIP)

S3 File. The result file of the phylogenetic tree of C domain and E domain by IQ-TREE.
(ZIP)

S4 File. NRPS Motif Finder online version code (Python).
(ZIP)

S5 File. NRPS Motif Finder Matlab version code.
(ZIP)

S1 Text. Table A, Table B, **Table C**. Table A. Product yields in different strains by HPLC/MS analysis. Table B. Fungal plasmids and strains used in this study. Table C. PCR primers used in this study.
(DOCX)

Acknowledgments

We thank Haonan Zheng for helpful discussions. We thank Stefan Günther and Paul F Zierep for providing detailed A domains dataset in SeMPI 2.0. We also appreciated the tech team at the BDA Informatics Suite (www.bdainformatics.org) for the online platform.

Author Contributions

Conceptualization: Zhiyuan Li.

Data curation: Ruolin He, Yuanzhe Shao.

Formal analysis: Ruolin He, Chen Song.

Funding acquisition: Wen-Bing Yin, Zhiyuan Li.

Investigation: Wen-Bing Yin, Zhiyuan Li.

Methodology: Ruolin He.

Project administration: Wen-Bing Yin.

Software: Yuanzhe Shao, Long Qian.

Supervision: Zhiyuan Li.

Validation: Ruolin He, Jinyu Zhang.

Visualization: Ruolin He, Jinyu Zhang, Yuanzhe Shao, Shaohua Gu, Long Qian.

Writing – original draft: Ruolin He, Jinyu Zhang.

Writing – review & editing: Shaohua Gu, Wen-Bing Yin, Zhiyuan Li.

References

1. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proceedings of the National Academy of Sciences. 2014; 111(25):9259–64. https://doi.org/10.1073/pnas.1401734111 PMID: 24927540

2. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. Trends in Microbiology. 2017; 25(4):280–92. Epub 2017/01/01. https://doi.org/10.1016/j.tim.2016.12.002 PMID: 28038926.

3. Wei B, Du AQ, Zhou ZY, Lai C, Yu WC, Yu JB, et al. An atlas of bacterial secondary metabolite biosynthesis gene clusters. Environmental Microbiology. 2021; 23(11):6981–92. Epub 2021/09/08. https://doi.org/10.1111/1462-2920.15761 PMID: 34490968.

4. Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Guimarães DO, de Frias UA, et al. Global biogeographic sampling of bacterial secondary metabolism. eLife. 2015; 4:e05048. https://doi.org/10.7554/eLife.05048 PMID: 25599565

5. Edgar RC. High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. bioRxiv. 2022.

6. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Research. 2020; 49(D1):D490–D7. https://doi.org/10.1093/nar/gkaa812 PMID: 33010170

7. González O, Ortíz-Castro R, Díaz-Pérez C, Díaz-Pérez AL, Magaña-Dueñas V, López-Bucio J, et al. Non-ribosomal Peptide Synthases from Pseudomonas aeruginosa Play a Role in Cyclodipeptide Biosynthesis, Quorum-Sensing Regulation, and Root Development in a Plant Host. Microbial ecology. 2017; 73(3):616–29. Epub 2016/12/03. https://doi.org/10.1007/s00248-016-0896-4 PMID: 27900439.

8. Morgan GL, Kretsch AM, Santa Maria KC, Weeks SJ, Li B. Specificity of Nonribosomal Peptide Synthetases in the Biosynthesis of the Pseudomonas virulence factor. Biochemistry. 2019; 58(52):5249–54. Epub 2019/06/28. https://doi.org/10.1021/acs.biochem.9b00360 PMID: 31243997; PubMed Central PMCID: PMC7179087.

9. Thomas MD, Langston-Unkefer PJ, Uchytil TF, Durbin RD. Inhibition of Glutamine Synthetase from Pea by Tabtoxinine-β-lactam. Plant Physiology. 1983; 71(4):912–5. https://doi.org/10.1104/pp.71.4.912 PMID: 16662928

10. Lee B-N, Kroken S, Chou DYT, Robbertse B, Yoder OC, Turgeon BG. Functional Analysis of All Nonribosomal Peptide Synthetases in *Cochliobolus heterostrophus* Reveals a Factor, NPS6, Involved in Virulence and Resistance to Oxidative Stress. Eukaryotic Cell. 2005; 4(3):545–55. https://doi.org/10.1128/EC.4.3.545–555.2005

11. Cociancich S, Pesic A, Petras D, Uhlmann S, Kretz J, Schubert V, et al. The gyrase inhibitor albicidin consists of p-aminobenzoic acids and cyanoalanine. Nature Chemical Biology. 2015; 11(3):195–7. https://doi.org/10.1038/nchembio.1734 PMID: 25599532

12. Bhatt K, Machado H, Osório NS, Sousa J, Cardoso F, Magalhães C, et al. A Nonribosomal Peptide Synthase Gene Driving Virulence in Mycobacterium tuberculosis. mSphere. 2018; 3(5). Epub 2018/11/02. https://doi.org/10.1128/mSphere.00352-18 PMID: 30381350; PubMed Central PMCID: PMC6211224.

13. Zan J, Li Z, Tianero MD, Davis J, Hill RT, Donia MS. A microbial factory for defensive kahalalides in a tripartite marine symbiosis. Science. 2019; 364(6445). https://doi.org/10.1126/science.aaw6732 PMID: 31196985.

14. Visca P, Imperi F, Lamont IL. Pyoverdine siderophores: from biogenesis to biosignificance. Trends in Microbiology. 2007; 15(1):22–30. https://doi.org/10.1016/j.tim.2006.11.004 PMID: 17118662

15. Süssmuth RD, Mainz A. Nonribosomal Peptide Synthesis—Principles and Prospects. Angewandte Chemie International Edition. 2017; 56(14):3770–821. https://doi.org/10.1002/anie.201609079 PMID: 28323366

16. Pacheco-Moreno A, Stefanato FL, Ford JJ, Trippel C, Uszkoreit S, Ferrafiat L, et al. Pan-genome analysis identifies intersecting roles for Pseudomonas specialized metabolites in potato pathogen inhibition. eLife. 2021; 10:e71900. https://doi.org/10.7554/eLife.71900 PMID: 34792466

17. Du L, Sánchez C, Chen M, Edwards DJ, Shen B. The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. Chemistry & biology. 2000; 7(8):623–42. Epub 2000/10/26. https://doi.org/10.1016/s1074-5521(00)00011-9 PMID: 11048953.

18. Dreyfuss M, Härri E, Hofmann H, Kobel H, Pache W, Tscherter H. Cyclosporin A and C. European J Appl Microbiol. 1976; 3(2):125–33. https://doi.org/10.1007/BF00928431

19. Actinomycin Hollstein U. Chemistry and mechanism of action. Chemical Reviews. 1974; 74(6):625–52. https://doi.org/10.1021/cr60292a002

20. Fleming A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. British journal of experimental pathology. 1929; 10(3):226.

21. Queener SW. Molecular biology of penicillin and cephalosporin biosynthesis. Antimicrobial agents and chemotherapy. 1990; 34(6):943–8. Epub 1990/06/01. https://doi.org/10.1128/AAC.34.6.943 PMID: 2203307; PubMed Central PMCID: PMC171734.

22. Yim G, Thaker MN, Koteva K, Wright G. Glycopeptide antibiotic biosynthesis. The Journal of Antibiotics. 2014; 67(1):31–41. https://doi.org/10.1038/ja.2013.117 PMID: 24220108

23. Brown AS, Calcott MJ, Owen JG, Ackerley DF. Structural, functional and evolutionary perspectives on effective re-engineering of non-ribosomal peptide synthetase assembly lines. Natural product reports. 2018; 35(11):1210–28. https://doi.org/10.1039/c8np00036k PMID: 30069573

24. Hacker C, Cai X, Kegler C, Zhao L, Weickhmann AK, Wurm JP, et al. Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. Nature Communications. 2018; 9(1):4366. https://doi.org/10.1038/s41467-018-06712-1 PMID: 30341296

25. Gao X, Haynes SW, Ames BD, Wang P, Vien LP, Walsh CT, et al. Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. Nature Chemical Biology. 2012; 8(10):823–30. https://doi.org/10.1038/nchembio.1047 PMID: 22902615

26. Bozhüyük KAJ, Fleischhacker F, Linck A, Wesche F, Tietze A, Niesert C-P, et al. De novo design and engineering of non-ribosomal peptide synthetases. Nature Chemistry. 2018; 10(3):275–81. https://doi.org/10.1038/nchem.2890 PMID: 29461518

27. Bozhüyük KAJ, Linck A, Tietze A, Kranz J, Wesche F, Nowak S, et al. Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. Nature Chemistry. 2019; 11(7):653–61. https://doi.org/10.1038/s41557-019-0276-z PMID: 31182822

28. Calcott MJ, Owen JG, Ackerley DF. Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. Nature Communications. 2020; 11(1):4554. https://doi.org/10.1038/s41467-020-18365-0 PMID: 32917865

29. Stachelhaus T, Mootz HD, Marahiel MA. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & Biology. 1999; 6(8):493–505. https://doi.org/10.1016/S1074-5521(99)80082-9 WOS:000084001200004. PMID: 10421756

30. Eppelmann K, Stachelhaus T, Marahiel MA. Exploitation of the Selectivity-Conferring Code of Nonribosomal Peptide Synthetases for the Rational Design of Novel Peptide Antibiotics. Biochemistry. 2002; 41(30):9718–26. https://doi.org/10.1021/bi0259406 PMID: 12135394

31. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS computational biology. 2014; 10(12):e1004016. Epub 2014/12/05. https://doi.org/10.1371/journal.pcbi.1004016 PMID: 25474254; PubMed Central PMCID: PMC4256081 following competing interests: MAF is on the scientific advisory board of Warp Drive Bio.

32. Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Research. 2019; 48(D1):D422–D30. https://doi.org/10.1093/nar/gkz932 PMID: 31665416

33. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019; 48(D1): D454–D8. https://doi.org/10.1093/nar/gkz882 PMID: 31612915

34. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology. 2015; 11(9):639–48. https://doi.org/10.1038/nchembio.1884 PMID: 26284671

35. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 2021; 49(W1):W29–W35. https://doi.org/10.1093/nar/gkab335 PMID: 33978755

36. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Research. 2015; 43 (20):9645–62. Epub 2015/10/08. https://doi.org/10.1093/nar/gkv1012 PMID: 26442528; PubMed Central PMCID: PMC4787774.

37. Zierep PF, Ceci AT, Dobrusin I, Rockwell-Kollmann SC, Günther S. SeMPI 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening in Natural Product Databases. Metabolites. 2020; 11(1). Epub 2021/01/02. https://doi.org/10.3390/metabo11010013 PMID: 33383692; PubMed Central PMCID: PMC7823522.

38. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2020; 49(D1):D412–D9. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078

39. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997; 28(3):405–20. Epub 1997/07/01. https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l PMID: 9223186.

40. Park HB, Perez CE, Barber KW, Rinehart J, Crawford JM. Genome mining unearths a hybrid nonribosomal peptide synthetase-like-pteridine synthase biosynthetic gene cluster. eLife. 2017; 6:e25229. https://doi.org/10.7554/eLife.25229 PMID: 28431213

41. Rottig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Research. 2011; 39(Web Server issue):W362–7. Epub 2011/05/12. https://doi.org/10.1093/nar/gkr323 PMID: 21558170; PubMed Central PMCID: PMC3125756.

42. Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. Journal of molecular biology. 2007; 368 (5):1500–17. Epub 2007/04/03. https://doi.org/10.1016/j.jmb.2007.02.099 PMID: 17400247.

43. Khayatt BI, Overmars L, Siezen RJ, Francke C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. PLoS One. 2013; 8(4):e62136. Epub 2013/05/03. https://doi.org/10.1371/journal.pone.0062136 PMID: 23637983; PubMed Central PMCID: PMC3630128.

44. Prieto C, Garcia-Estrada C, Lorenzana D, Martin JF. NRPSsp: non-ribosomal peptide synthase substrate predictor. Bioinformatics. 2012; 28(3):426–7. Epub 2011/12/02. https://doi.org/10.1093/bioinformatics/btr659 PMID: 22130593.

45. Challis GL, Ravel J, Townsend CA. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chemistry & Biology. 2000; 7(3):211–24. https://doi.org/10.1016/S1074-5521(00)00091-0 PMID: 10712928

46. Lee TV, Johnson LJ, Johnson RD, Koulman A, Lane GA, Lott JS, et al. Structure of a Eukaryotic Nonribosomal Peptide Synthetase Adenylation Domain That Activates a Large Hydroxamate Amino Acid in Siderophore Biosynthesis*. Journal of Biological Chemistry. 2010; 285(4):2415–27. https://doi.org/10.1074/jbc.M109.071324 PMID: 19923209

47. Li Y, Zhang W, Zhang H, Tian W, Wu L, Wang S, et al. Structural Basis of a Broadly Selective Acyltransferase from the Polyketide Synthase of Splenocin. Angewandte Chemie International Edition. 2018; 57(20):5823–7. https://doi.org/10.1002/anie.201802805 PMID: 29536601

48. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nature Reviews Genetics. 2013; 14(4):249–61. https://doi.org/10.1038/nrg3414 PMID: 23458856

49. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proceedings of the National Academy of Sciences. 1993; 90(15):7176–80. Epub 1993/08/01. https://doi.org/10.1073/pnas.90.15.7176 PMID: 8346232; PubMed Central PMCID: PMC47099.

50. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences. 2012; 109(26):10340–5. Epub 2012/06/14. https://doi.org/10.1073/pnas.1207864109 PMID: 22691493; PubMed Central PMCID: PMC3387073.

51. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28 (2):184–90. Epub 2011/11/22. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153.

52. Teşileanu T, Colwell LJ, Leibler S. Protein sectors: statistical coupling analysis versus conservation. PLoS computational biology. 2015; 11(2):e1004091. Epub 2015/02/28. https://doi.org/10.1371/journal.pcbi.1004091 PMID: 25723535; PubMed Central PMCID: PMC4344308.

53. Rivoire O, Reynolds KA, Ranganathan R. Evolution-Based Functional Decomposition of Proteins. PLoS computational biology. 2016; 12(6):e1004817. Epub 2016/06/03. https://doi.org/10.1371/journal.pcbi.1004817 PMID: 27254668; PubMed Central PMCID: PMC4890866.

54. Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, et al. An interdomain sector mediating allostery in Hsp70 molecular chaperones. Molecular systems biology. 2010; 6:414. Epub 2010/09/25. https://doi.org/10.1038/msb.2010.65 PMID: 20865007; PubMed Central PMCID: PMC2964120.

55. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009; 138(4):774–86. Epub 2009/08/26. https://doi.org/10.1016/j.cell.2009.07.038 PMID: 19703402; PubMed Central PMCID: PMC3210731.

56. Reynolds KA, Russ WP, Socolich M, Ranganathan R. Chapter Ten—Evolution-Based Design of Proteins. In: Keating AE, editor. Methods in Enzymology. 523: Academic Press; 2013. p. 213–35.

57. Conti E, Stachelhaus T, Marahiel MA, Brick P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. The EMBO journal. 1997; 16(14):4174–83. Epub 1997/07/16. https://doi.org/10.1093/emboj/16.14.4174 PMID: 9250661; PubMed Central PMCID: PMC1170043.

58. Tanovic A, Samel SA, Essen LO, Marahiel MA. Crystal structure of the termination module of a nonribosomal peptide synthetase. Science. 2008; 321(5889):659–63. Epub 2008/06/28. https://doi.org/10.1126/science.1159850 PMID: 18583577.

59. Samel SA, Czodrowski P, Essen LO. Structure of the epimerization domain of tyrocidine synthetase A. Acta Crystallographica Section D: Biological Crystallography. 2014; 70(Pt 5):1442–52. Epub 2014/05/13. https://doi.org/10.1107/S1399004714004398 PMID: 24816112.

60. Konz D, Marahiel MA. How do peptide synthetases generate structural diversity? Chemistry & Biology. 1999; 6(2):R39–R48. https://doi.org/10.1016/S1074-5521(99)80002-7 WOS:000082564800002. PMID: 10021423

61. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evolutionary Biology. 2007; 7(1):78. https://doi.org/10.1186/1471-2148-7-78 PMID: 17506888

62. Bloudoff K, Fage CD, Marahiel MA, Schmeing TM. Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. Proceedings of the National Academy of Sciences. 2017; 114(1):95–100. Epub 2016/12/21. https://doi.org/10.1073/pnas.1614191114 PMID: 27994138; PubMed Central PMCID: PMC5224394.

63. Reimer JM, Eivaskhani M, Harb I, Guarné A, Weigt M, Schmeing TM. Structures of a dimodular nonribosomal peptide synthetase reveal conformational flexibility. Science. 2019; 366(6466):eaaw4388. https://doi.org/10.1126/science.aaw4388 PMID: 31699907

64. Linne U, Doekel S, Marahiel MA. Portability of Epimerization Domain and Role of Peptidyl Carrier Protein on Epimerization Activity in Nonribosomal Peptide Synthetases. Biochemistry. 2001; 40 (51):15824–34. https://doi.org/10.1021/bi011595t PMID: 11747460

65. Wheadon MJ, Townsend CA. Evolutionary and functional analysis of an NRPS condensation domain integrates β-lactam, D-amino acid, and dehydroamino acid synthesis. Proceedings of the National Academy of Sciences. 2021; 118(17):e2026017118. https://doi.org/10.1073/pnas.2026017118 PMID: 33893237

66. Reitz ZL, Hardy CD, Suk J, Bouvet J, Butler A. Genomic analysis of siderophore β-hydroxylases reveals divergent stereocontrol and expands the condensation domain family. Proceedings of the National Academy of Sciences. 2019; 116(40):19805–14. https://doi.org/10.1073/pnas.1903161116 PMID: 31527229

67. Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, et al. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. Journal of Biological Chemistry. 2022; 298(10):102480. https://doi.org/10.1016/j.jbc.2022.102480 PMID: 36108739

68. Hai Y, Jenner M, Tang Y. Fungal siderophore biosynthesis catalysed by an iterative nonribosomal peptide synthetase. Chemical Science. 2020; 11(42):11525–30. https://doi.org/10.1039/d0sc03627g PMID: 34094397

69. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research. 2019; 48(D1):D265–D8. https://doi.org/10.1093/nar/gkz991 PMID: 31777944

70. Koketsu K, Watanabe K, Suda H, Oguri H, Oikawa H. Reconstruction of the saframycin core scaffold defines dual Pictet-Spengler mechanisms. Nature Chemical Biology. 2010; 6(6):408–10. https://doi.org/10.1038/nchembio.365 PMID: 20453862

71. Patteson JB, Fortinez CM, Putz AT, Rodriguez-Rivas J, Bryant LH, III, Adhikari K, et al. Structure and Function of a Dehydrating Condensation Domain in Nonribosomal Peptide Biosynthesis. Journal of the American Chemical Society. 2022; 144(31):14057–70. https://doi.org/10.1021/jacs.1c13404 PMID: 35895935

72. Zaleta-Rivera K, Xu C, Yu F, Butchko RAE, Proctor RH, Hidalgo-Lara ME, et al. A Bidomain Nonribosomal Peptide Synthetase Encoded by FUM14 Catalyzes the Formation of Tricarballylic Esters in the Biosynthesis of Fumonisins. Biochemistry. 2006; 45(8):2561–9. https://doi.org/10.1021/bi052085s PMID: 16489749

73. Lin S, Huang T, Horsman GP, Huang S-X, Guo X, Shen B. Specificity of the Ester Bond Forming Condensation Enzyme SgcC5 in C-1027 Biosynthesis. Organic Letters. 2012; 14(9):2300–3. https://doi.org/10.1021/ol300720s PMID: 22519717

74. Eddy SR. Accelerated Profile HMM Searches. PLOS Computational Biology. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

75. Keatinge-Clay AT. The structures of type I polyketide synthases. Natural Product Reports. 2012; 29 (10):1050–73. https://doi.org/10.1039/c2np20019h PMID: 22858605

76. Raymond KN, Dertz EA, Kim SS. Enterobactin: An archetype for microbial iron transport. Proceedings of the National Academy of Sciences. 2003; 100(7):3584–8. https://doi.org/10.1073/pnas.0630018100 PMID: 12655062

77. May JJ, Wendrich TM, Marahiel MA. The dhb Operon of Bacillus subtilisEncodes the Biosynthetic Template for the Catecholic Siderophore 2,3-Dihydroxybenzoate-Glycine-Threonine Trimeric Ester Bacillibactin*. Journal of Biological Chemistry. 2001; 276(10):7209–17. https://doi.org/10.1074/jbc.M009140200 PMID: 11112781

78. Reimmann C, Patel HM, Serino L, Barone M, Walsh CT, Haas D. Essential PchG-dependent reduction in pyochelin biosynthesis of Pseudomonas aeruginosa. Journal of Bacteriology. 2001; 183(3):813–20. Epub 2001/02/24. https://doi.org/10.1128/JB.183.3.813-820.2001 PMID: 11208777; PubMed Central PMCID: PMC94946.

79. Koumoutsi A, Chen XH, Henne A, Liesegang H, Hitzeroth G, Franke P, et al. Structural and functional characterization of gene clusters directing nonribosomal synthesis of bioactive cyclic lipopeptides in Bacillus amyloliquefaciens strain FZB42. Journal of Bacteriology. 2004; 186(4):1084–96. Epub 2004/02/06. https://doi.org/10.1128/JB.186.4.1084-1096.2004 PMID: 14762003; PubMed Central PMCID: PMC344220.

80. Veith B, Herzberg C, Steckel S, Feesche J, Maurer KH, Ehrenreich P, et al. The complete genome sequence of Bacillus licheniformis DSM13, an organism with great industrial potential. Journal of Molecular Microbiology and Biotechnology. 2004; 7(4):204–11. Epub 2004/09/24. https://doi.org/10.1159/000079829 PMID: 15383718.

81. Pourmasoumi F, De S, Peng H, Trottmann F, Hertweck C, Kries H. Proof-Reading Thioesterase Boosts Activity of Engineered Nonribosomal Peptide Synthetase. ACS Chemical Biology. 2022; 17 (9):2382–8. https://doi.org/10.1021/acschembio.2c00341 PMID: 36044980

82. Bushley KE, Ripoll DR, Turgeon BG. Module evolution and substrate specificity of fungal nonribosomal peptide synthetases involved in siderophore biosynthesis. BMC Evolutionary Biology. 2008; 8 (1):328. https://doi.org/10.1186/1471-2148-8-328 PMID: 19055762

83. Johnson LJ, Koulman A, Christensen M, Lane GA, Fraser K, Forester N, et al. An Extracellular Siderophore Is Required to Maintain the Mutualistic Interaction of Epichloë festucae with Lolium perenne. PLOS Pathogens. 2013; 9(5):e1003332. https://doi.org/10.1371/journal.ppat.1003332 PMID: 23658520

84. Wight WD, Labuda R, Walton JD. Conservation of the genes for HC-toxin biosynthesis in Alternaria jesenskae. BMC Microbiology. 2013; 13(1):165. https://doi.org/10.1186/1471-2180-13-165 PMID: 23865912

85. Li W, Fan A, Wang L, Zhang P, Liu Z, An Z, et al. Asperphenamate biosynthesis reveals a novel two-module NRPS system to synthesize amino acid esters in fungi. Chemical Science. 2018; 9(9):2589–94. https://doi.org/10.1039/c7sc02396k PMID: 29719714

86. Robey MT, Caesar LK, Drott MT, Keller NP, Kelleher NL. An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes. Proceedings of the National Academy of Sciences. 2021; 118 (19):e2020230118. https://doi.org/10.1073/pnas.2020230118 PMID: 33941694

87. Haese A, Schubert M, Herrmann M, Zocher R. Molecular characterization of the enniatin synthetase gene encoding a multifunctional enzyme catalysing N-methyldepsipeptide formation in Fusarium scirpi. Molecular Microbiology. 1993; 7(6):905–14. Epub 1993/03/01. https://doi.org/10.1111/j.1365-2958.1993.tb01181.x PMID: 8483420.

88. Xu Y, Orozco R, Kithsiri Wijeratne EM, Espinosa-Artiles P, Leslie Gunatilaka AA, Patricia Stock S, et al. Biosynthesis of the cyclooligomer depsipeptide bassianolide, an insecticidal virulence factor of Beauveria bassiana. Fungal Genetics and Biology. 2009; 46(5):353–64. Epub 2009/03/17. https://doi.org/10.1016/j.fgb.2009.03.001 PMID: 19285149.

89. Xu Y, Orozco R, Wijeratne EM, Gunatilaka AA, Stock SP, Molnár I. Biosynthesis of the cyclooligomer depsipeptide beauvericin, a virulence factor of the entomopathogenic fungus Beauveria bassiana. Chemistry & Biology. 2008; 15(9):898–907. Epub 2008/09/23. https://doi.org/10.1016/j.chembiol.2008.07.011 PMID: 18804027.

90. Zhang L, Wang C, Chen K, Zhong W, Xu Y, Molnár I. Engineering the biosynthesis of fungal nonribosomal peptides. Natural Product Reports. 2023; 40(1):62–88. https://doi.org/10.1039/d2np00036a PMID: 35796260

91. Steiniger C, Hoffmann S, Mainz A, Kaiser M, Voigt K, Meyer V, et al. Harnessing fungal nonribosomal cyclodepsipeptide synthetases for mechanistic insights and tailored engineering. Chemical Science. 2017; 8(11):7834–43. https://doi.org/10.1039/c7sc03093b PMID: 29163920

92. Fierro F, García-Estrada C, Castillo NI, Rodríguez R, Velasco-Conde T, Martín JF. Transcriptional and bioinformatic analysis of the 56.8 kb DNA region amplified in tandem repeats containing the penicillin gene cluster in Penicillium chrysogenum. Fungal Genetics and Biology. 2006; 43(9):618–29. Epub 2006/05/23. https://doi.org/10.1016/j.fgb.2006.03.001 PMID: 16713314.

93. Ames BD, Haynes SW, Gao X, Evans BS, Kelleher NL, Tang Y, et al. Complexity Generation in Fungal Peptidyl Alkaloid Biosynthesis: Oxidation of Fumiquinazoline A to the Heptacyclic Hemiaminal Fumiquinazoline C by the Flavoenzyme Af12070 from Aspergillus fumigatus. Biochemistry. 2011; 50 (40):8756–69. https://doi.org/10.1021/bi201302w PMID: 21899262

94. Lim FY, Ames B, Walsh CT, Keller NP. Co-ordination between BrlA regulation and secretion of the oxidoreductase FmqD directs selective accumulation of fumiquinazoline C to conidial tissues in Aspergillus fumigatus. Cellular microbiology. 2014; 16(8):1267–83. Epub 2014/03/13. https://doi.org/10.1111/cmi.12284 PMID: 24612080; PubMed Central PMCID: PMC4114987.

95. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? Journal

of Molecular Biology. 2019; 431(11):2197–212. Epub 2019/04/18. https://doi.org/10.1016/j.jmb.2019.04.009 PMID: 30995449; PubMed Central PMCID: PMC6544567.

96. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 2015; 10(6):845–58. https://doi.org/10.1038/nprot.2015.053 PMID: 25950237

97. Burgard C, Zaburannyi N, Nadmid S, Maier J, Jenke-Kodama H, Luxenburger E, et al. Genomics-Guided Exploitation of Lipopeptide Diversity in Myxobacteria. ACS Chemical Biology. 2017; 12 (3):779–86. Epub 2017/01/28. https://doi.org/10.1021/acschembio.6b00953 PMID: 28128551.

98. Yan F, Burgard C, Popoff A, Zaburannyi N, Zipf G, Maier J, et al. Synthetic biology approaches and combinatorial biosynthesis towards heterologous lipopeptide production. Chemical Science. 2018; 9 (38):7510–9. Epub 2018/10/16. https://doi.org/10.1039/c8sc02046a PMID: 30319751; PubMed Central PMCID: PMC6180311.

99. Weber T, Baumgartner R, Renner C, Marahiel MA, Holak TA. Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. Structure. 2000; 8(4):407–18. https://doi.org/10.1016/s0969-2126(00)00120-9 PMID: 10801488

100. Izoré T, Cryle MJ. The many faces and important roles of protein-protein interactions during non-ribosomal peptide synthesis. Natural product reports. 2018; 35(11):1120–39. Epub 2018/09/13. https://doi.org/10.1039/c8np00038g PMID: 30207358.

101. Crüsemann M, Kohlhaas C, Piel J. Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. Chemical Science. 2013; 4(3):1041–5. https://doi.org/10.1039/C2SC21722H

102. Kries H, Niquille DL, Hilvert D. A Subdomain Swap Strategy for Reengineering Nonribosomal Peptides. Chemistry & Biology. 2015; 22(5):640–8. https://doi.org/10.1016/j.chembiol.2015.04.015 PMID: 26000750

103. Thong WL, Zhang Y, Zhuo Y, Robins KJ, Fyans JK, Herbert AJ, et al. Gene editing enables rapid engineering of complex antibiotic assembly lines. Nature Communications. 2021; 12(1):6872. https://doi.org/10.1038/s41467-021-27139-1 PMID: 34824225

104. Lehninger AL, Nelson DL, Cox MM, Cox MM. Lehninger principles of biochemistry: Macmillan; 2005.

105. Yan BX, Sun YQ. Glycine residues provide flexibility for enzyme active sites. Journal of Biological Chemistry. 1997; 272(6):3190–4. https://doi.org/10.1074/jbc.272.6.3190 PMID: 9013553

106. OKONIEWSKA M TANAKA T, YADA RY. The pepsin residue glycine-76 contributes to active-site loop flexibility and participates in catalysis. Biochemical Journal. 2000; 349(1):169–77. https://doi.org/10.1042/bj3490169

107. Xu Y, Feng L, Jeffrey PD, Shi Y, Morel FMM. Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. Nature. 2008; 452(7183):56–61. https://doi.org/10.1038/nature06636 PMID: 18322527

108. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins: Structure, Function, and Bioinformatics. 1992; 12(4):345–64. https://doi.org/10.1002/prot.340120407.

109. Nwafor J, Salguero C, Welcome F, Durmus S, Glasser RN, Zimmer M, et al. Why Are Gly31, Gly33, and Gly35 Highly Conserved in All Fluorescent Proteins? Biochemistry. 2021. Epub 2021/11/23. https://doi.org/10.1021/acs.biochem.1c00587 PMID: 34806355.

110. Koglin A, Mofid MR, Löhr F, Schäfer B, Rogov VV, Blum M-M, et al. Conformational Switches Modulate Protein Interactions in Peptide Antibiotic Synthetases. Science. 2006; 312(5771):273–6. https://doi.org/10.1126/science.1122928 PMID: 16614225

111. Izoré T, Candace Ho YT, Kaczmarski JA, Gavriilidou A, Chow KH, Steer DL, et al. Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. Nature Communications. 2021; 12(1):2511. https://doi.org/10.1038/s41467-021-22623-0 PMID: 33947858

112. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999; 10(1):37–48. Epub 1999/01/15. PMID: 9888278.

113. Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2020; 49(D1):D480–D9. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286

114. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016; 17(1):132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

115. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. Journal of Genetics and Genomics. 2021; 48(9):844–50. https://doi.org/10.1016/j.jgg.2021.03.006 PMID: 34001434

116. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;

7:539. Epub 2011/10/13. https://doi.org/10.1038/msb.2011.75 PMID: 21988835; PubMed Central PMCID: PMC3261699.

117. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2014; 32 (1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

118. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 2017; 14(6):587–9. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

119. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Research. 2016; 44(W1):W344–W50. https://doi.org/10.1093/nar/gkw408 PMID: 27166375

120. Agnieszka Golicz PVT, Madeira Fábio, Martin David M. A., Procter James B.and Geoffrey J. AACon: A Fast Amino Acid Conservation Calculation Service 2018. Available from: http://www.compbio.dundee.ac.uk/aacon/.

121. Baccile JA, Spraker JE, Le HH, Brandenburger E, Gomez C, Bok JW, et al. Plant-like biosynthesis of isoquinoline alkaloids in Aspergillus fumigatus. Nature chemical biology. 2016; 12(6):419–24. Epub 2016/04/12. https://doi.org/10.1038/nchembio.2061 PMID: 27065235; PubMed Central PMCID: PMC5049701.

122. Szewczyk E, Nayak T, Oakley CE, Edgerton H, Xiong Y, Taheri-Talesh N, et al. Fusion PCR and gene targeting in Aspergillus nidulans. Nature Protocols. 2006; 1(6):3111–20. https://doi.org/10.1038/nprot.2006.405 PMID: 17406574

123. Bok JW, Keller NP. LaeA, a regulator of secondary metabolism in Aspergillus spp. Eukaryotic Cell. 2004; 3(2):527–35. Epub 2004/04/13. https://doi.org/10.1128/EC.3.2.527-535.2004 PMID: 15075281; PubMed Central PMCID: PMC387652.