Bioinformatic Mining of Novel Lipopeptides Enabled by Daptomycin Cs Domain and Structural Modeling

Hao Zhu[1]*

[1] *School of Pharmacy, Nanjing University of Chinese Medicine, Nanjing, China*

*Corresponding author

zhuhao@njucm.edu.cn*

Abstract

Amidst the escalating antibiotic resistance crisis, lipopeptides emerge as promising antibiotic candidates due to their unique amphipathic structures. This study established a systematic bioinformatic platform using the condensation starter (Cs) domain of daptomycin as a molecular probe. Sequence similarity network (SSN) analysis identified 613 potential lipopeptide biosynthetic gene clusters (BGCs), with 432 (70.5%) originating from *Streptomyces* species. Subsequent integration of antiSMASH boundary prediction and evolutionary analysis prioritized 37 candidate BGCs harboring multiple post-modification modules. Five novel BGC types were ultimately selected based on Cs domain homology (<40% identity) and modification complexity. AlphaFold3 modeling revealed that WP_386473946.1 possess distinctive loop architecture, an expanded catalytic cavity, and a unique His59-Asp91 structural unit. Crucially, glycine residues adjacent to the conserved HHxxDG motif in the active pocket provide targets for substrate recognition and rational engineering. This work delivers structure-guided genomic resources and molecular blueprints for accelerated discovery of anti-drug-resistant lipopeptides.

Keywords: lipopeptides, Cs domain, bioinformatic platform, sequence similarity network.

Introduction

Antimicrobial resistance (AMR) has emerged as a critical global public health threat in the 21st century. World Health Organization (WHO) data reveal that drug-resistant bacterial infections already cause approximately 1.27 million deaths annually worldwide. This figure is projected to exceed 10 million deaths (Fig. 1A) by 2050 without effective interventions, accompanied by economic losses surpassing 100 trillion US dollars[1]. This crisis stems from a dual imbalance, the

stagnation of drug development and the relentless evolution of pathogens. Over the past two decades, only 33 novel antibiotics have been approved (Fig. 1B). Metagenomic studies revealing that 23.78% of identified genes are multidrug-resistant genes underscore the rapid evolution and dissemination of antibiotic resistance. As the development rate of new antimicrobial agents lags far behind the adaptive evolution of pathogens[2], the development of novel anti-resistance drugs based on innovative molecular mechanisms[3] has become a strategic imperative in the biomedical field.

Lipopeptide represent a promising breakthrough in combating drug-resistant bacteria due to their unique membrane-targeting mechanisms[4]. Their amphipathic architecture, featuring a peptide backbone linked via an amide bond to a fatty acid chain, enables disruption of bacterial cell membranes or inhibition of critical physiological processes[5–7]. For instance, daptomycin exerts its bactericidal effects by targeting the membrane of Gram-positive bacteria in the presence of Ca^{2+}, perturbing lipid organization and inhibiting peptidoglycan biosynthesis[8,9], while globomycin competitively inhibits signal peptidase to block outer membrane synthesis towards Gram-negative pathogens[10]. Despite broad-spectrum potential of lipopeptides, daptomycin has been the only calcium-dependent antibiotic approved for clinical use in the past two decades[1], revealing the inefficiency of traditional phenotype-based screening strategies and highlighting the urgent need for structure-guided and systematic discovery platforms. This imperative necessitates comprehensive investigation into lipopeptide biosynthesis pathways. The bioassembly of lipopeptides hinges on the modular machinery of non-ribosomal peptide synthetases (NRPS) and the substrate specificity of condensation starter (Cs) domains. NRPS enzymes operate via an assembly-line process, precisely orchestrating amino acid activation (A domain), thiolation (T domain), and condensation (C

domain), while relying on tailoring enzymes to incorporate non-proteinogenic units[11,12] (e.g., methylated glutamate[13]). Crucially, the Cs domain acts as the gateway for fatty acid chain incorporation, catalyzing polyketides acylation through fatty acyl-AMP ligase (FAAL)[14], fatty acyl CoA ligase (FACL)[15], or polykedtide synthase (PKS)-associated pathways[11]. The architecture of the fatty acyl chain serves as the primary structural determinant governing the bioactivity of lipopeptides. Prioritizing Cs in mining pipelines therefore accelerates the discovery of novel bioactive lipopeptides. Genomic mining of uncultured soil microbes by the Sean Brady group revealed macolacin[16] and cilagicin[17] as potent anti-MRSA antibiotics, which paves a new path for combating the increasingly serious challenge of antibiotic resistance.

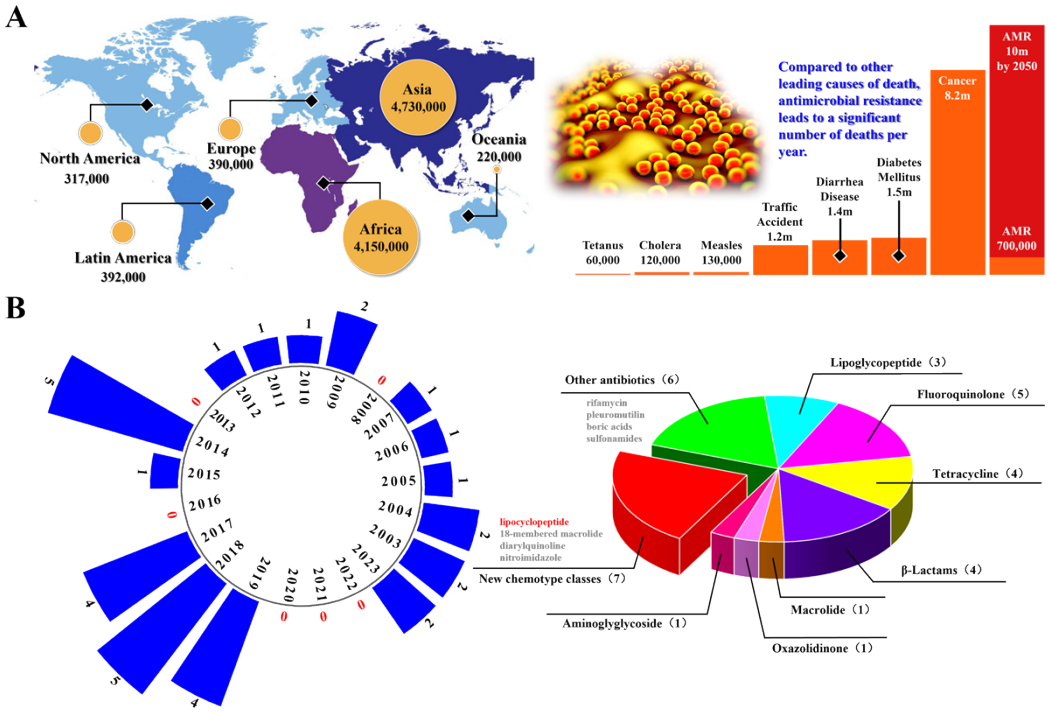

Fig. 1. The proliferation rate of drug-resistant pathogens substantially outpaces the development rate of novel antibiotics.

(A) Global incidence of antimicrobial resistance (AMR) infections and projections across six continents[1]; (B) FDA-approved antibiotics in the past 20 years[1].

We employed the Cs domain of daptomycin as a molecular probe for genomic mining in this work, establishing a systematic platform for lipopeptide antibiotic discovery. Leveraging sequence similarity network (SSN) analysis centered on the daptomycin Cs domain, we identified 37 novel candidate biosynthetic gene clusters (BGCs) encoding multiple tailoring enzymes from a pool of 613 potential BGCs. Subsequent BGC evolutionary classification and AlphaFold3-based structural modeling pinpointed a novel Cs domain (WP_386473946.1) within *Streptomyces zaomyceticus*, possessing an extended loop region beyond the V-shaped catalytic pocket conformation. This integrated strategy provides a molecular blueprint for designing next-generation lipopeptides with enhanced membrane penetration and anti-resistance properties. which may serve as an innovative solution to the global AMR crisis.

Materials and Methods

Bibliometric Analysis

This study employed a systematic bibliometric approach to conduct a visual analysis of lipopeptide antibiotic research within the biomedical field. Literature retrieval was performed using the Web of Science Core Collection database. The search strategy was defined as: Topic sentences (TS) = ("lipopeptide" AND "antibiotic" AND "bio*"), with the publication date range restricted to January 2021 through December 2025. The retrieved records were exported in batches as plain text files (Supplementary Materials, download_1.txt to download_5.txt). Following import into CiteSpace (version 6.3.R1) and removal of duplicate entries, the time slicing parameter ("Year Per Slice") was configured as 1-year intervals spanning from January 2021 to December 2025. Utilizing the keyword co-occurrence network analysis mode, a knowledge map was constructed with the g-

index set to k = 20.

Construction of Sequence Similarity Network (SSN) and Cluster Analysis

BGCs of characterized lipopeptide antibiotics[18] were systematically retrieved from the MIBiG database[19] (https://mibig.secondarymetabolites.org/). Representative amino acid sequences (known.fasta) were downloaded and integrated with candidate sequences identified in this study through mining of the daptomycin Cs domain. Following redundancy removal, a non-redundant dataset of Cs sequences (Cs.fasta) was compiled. This dataset was submitted to the EFI-EST platform[20] (https://efi.igb.illinois.edu/efi-est/) for construction of a SSN using default parameters. Within the SSN, nodes represent individual protein sequences, and edge weights are defined by pairwise sequence similarity scores. The inherent clustering algorithm of the platform was employed to analyze network topology, partitioning clusters at a filter value threshold of 80% to merge low-confidence edges. Final network data (node attribute tables and edge lists) were imported into Cytoscape (v3.10.3) for visualization of key functional clusters and core hub proteins.

Boundary Prediction and Functional Annotation of Gene Clusters

BGC boundary prediction and functional annotation were performed using the antiSMASH (v8.0) online platform[21] (https://antismash.secondarymetabolites.org/#!/start) based on the candidate protein dataset (Supplementary Materials, id_protein_50k.fasta). BGCs containing tailoring enzymes (e.g., P450s, methyltransferases, and oxidoreductases) with high abundance of modification modules were selected. Following deduplication, a non-redundant dataset of candidate BGCs (Supplementary Materials, CS_selection.fasta) was constructed for downstream analyses.

Sequence Alignment and Phylogenetic Tree Construction

Multiple sequence alignment of the non-redundant candidate BGC dataset (Supplementary Materials, CS_selection.fasta) was performed using MEGA11 (v11.0) with the ClustalW algorithm to compute sequence similarity matrices. The alignment results were visualized through the ESpript 3.0 online platform[22] (https://espript.ibcp.fr/ESPript/ESPript/index.php), generating annotated alignment maps with conserved domain markers. Subsequently, the aligned sequences were subjected to maximum likelihood (ML) phylogenetic reconstruction implemented in MEGA11. The resulting tree was topologically refined and annotated using the iTOL (v6) online tool[23] (https://itol.embl.de/), with adjustments to branch length scaling and application of color-coded node labels to enhance visual interpretation of functional clusters and evolutionary relationships.

Prediction of Three-Dimensional Models

The three-dimensional structure of the Cs domain in target protein WP_386473946.1 was predicted using the AlphaFold3 web server[24] (https://alphafoldserver.com/welcome). Models were filtered based on per-residue plDDT confidence scores (0–100 scale), retaining regions with high-confidence plDDT values for downstream analysis. Selected models were visualized and structurally refined with default parameters in PyMOL 3.1.3, including surface rendering. To investigate the evolutionary conservation of Cs domains, the predicted model was systematically aligned with experimentally determined Cs structures from seven characterized lipopeptide antibiotics: A54145, daptomycin, enduracidin, globomycin, rakicidin, taromycin A, and telomycin. Structural superposition of Cs backbones was performed using PyMOL.

Results and discussion

Bibliometric Analysis of Lipopeptide Antibiotics

Based on bibliometric analysis of 903 publications retrieved from the Web of Science Core Collection database, a keyword co-occurrence network was constructed using CiteSpace (Fig. 2). The network comprised 253 keywords connected by 345 links, yielding a network density of 0.0108. Analysis of high-frequency keywords (Table 1) revealed that "biological control" (frequency: 72, centrality: 0.06) and "biocontrol" (frequency: 59, centrality: 0.08) constituted the core thematic foci of lipopeptide antibiotic research over the past five years, reflecting a prevailing research emphasis on biological control applications. However, although "biosynthesis" ranked 12th among high-frequency terms with 54 occurrences, its centrality metric of zero indicates a significant knowledge gap in research on biosynthetic assembly. This research deficit stands in stark contrast to the escalating global burden of multidrug-resistant (MDR) infections and the worsening AMR crisis across six continents. WHO data indicate that AMR-associated infections currently cause millions of deaths annually, with projections suggesting AMR may surpass cancer as a global mortality threat by 2050 (Fig. 1A).

Table 1. Top 15 keywords in lipopeptide antibiotic research.

Rank	Frequency	Centrality	Keywords
1	114	0.01	identification
2	99	0.02	bacillus subtilis
3	98	0.03	subtilis
4	88	0.00	resistance
5	87	0.10	lipopeptides
6	77	0.02	staphylococcus aureus
7	75	0.00	surfactin

8	74	0.11	antibiotics
9	72	0.06	biological control
10	59	0.08	biocontrol
11	54	0.14	vancomycin
12	54	0.00	biosynthesis
13	53	0.00	antibiotic resistance
14	47	0.08	mechanisms
15	46	0.10	strains

Critically, the pipeline for novel lipopeptide antibiotics lags far behind the accelerating pace of the rapidly evolving resistance mechanisms. Statistical data reveal that only one lipopeptide, daptomycin, was approved by the U.S. Food and Drug Administration (FDA) among 33 novel antibiotics approved over the past two decades (Fig. 1B). The lag in elucidating biosynthetic pathways and optimizing metabolic engineering may represent a critical bottleneck restricting the development of the lipopeptide class. With a constraint acutely exacerbating against the backdrop of rapidly rising AMR-associated disease incidence. Consequently, intensifying research on lipopeptide antibiotic biosynthesis is not merely pivotal for overcoming technical barriers, but an urgent imperative to address the global public health crisis. The antimicrobial potency of lipopeptide antibiotics is primarily governed by the structural architecture of their lipid tails[25]. As the Cs domain catalyzes the pivotal incorporation of these lipid moieties, targeted mining of Cs domain homologs constitutes a rational strategy for the accelerated discovery of novel bioactive lipopeptides.

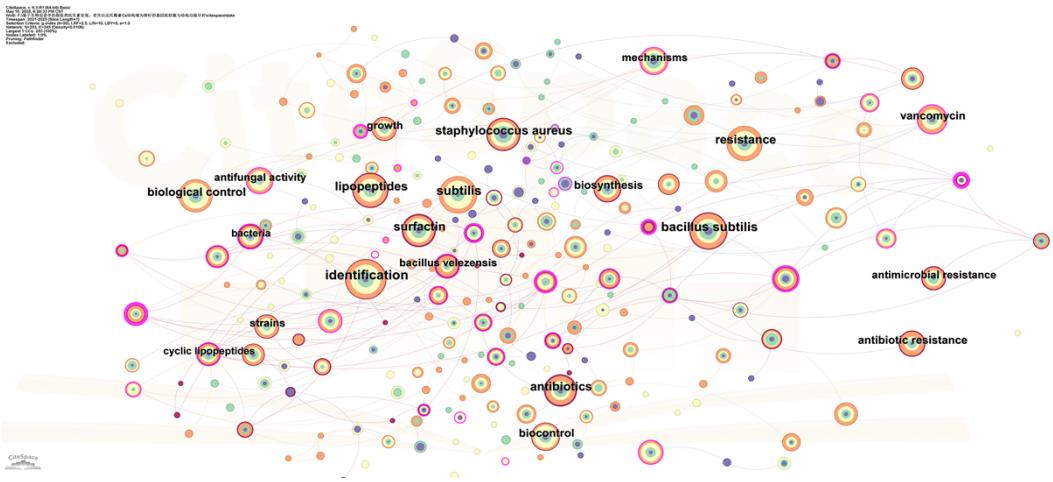

Fig. 2. Keyword co-occurrence network for lipopeptide antibiotics research.

Screening of Putative Lipopeptide BGCs

Using the Cs domain of daptomycin (Fig. 3A) as a bioinformatic probe, we retrieved (>)5,000 homologous sequences from the NCBI non-redundant protein database through systematic homology searches. SSN analysis enabled clustering and redundancy reduction of these sequences (Fig. 3B), ultimately identifying 613 putative lipopeptide BGCs. Strikingly, 432 BGCs (70.5% of total candidates) (Fig. 3B, Supplmentary Materials) originated from *Streptomyces* species, indicating this genus dominates lipopeptide biosynthetic capacity. These results establish a valuable genetic repository for mining novel lipopeptide natural products and elucidating their biosynthetic mechanisms.

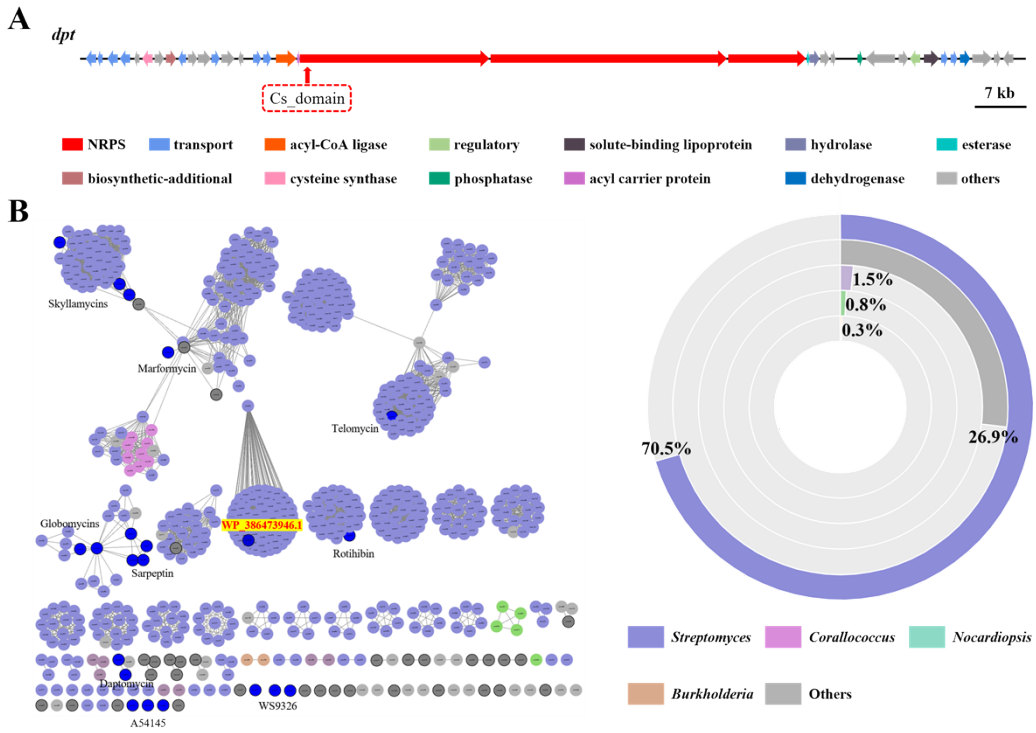

Fig. 3. Genomic mining using daptomycin's Cs domain as a probe within the NCBI database.

(A) The BGC of daptomycin named *dpt*; (B) SSN of daptomycin Cs and its 612 homologues.

Evolutionary Classification of Candidate BGCs

Boundary prediction and functional annotation of target gene clusters via the antiSMASH platform enabled selection of 37 candidate BGCs harboring multiple post-assembly modification modules. Integration with 26 characterized Cs domains facilitated phylogenetic tree construction through multiple sequence alignment (Fig. 4A), revealing evolutionary relationships and potential functional divergence between candidate and characterized BGCs. Application of dual screening criteria of Cs sequence identity (<40%) and complexity of modification modules prioritized five distinct BGC categories exhibiting unique structural features and high diversification potential (Fig. 4B). The low sequence homology (<40% identity) strongly suggests these clusters encode novel lipopeptide scaffolds, while their complex modification machinery indicates capacity to produce

structurally diverse lipopeptides with unprecedented bioactivities. These BGCs may constitute high-value targets for heterologous expression and metabolic mining.

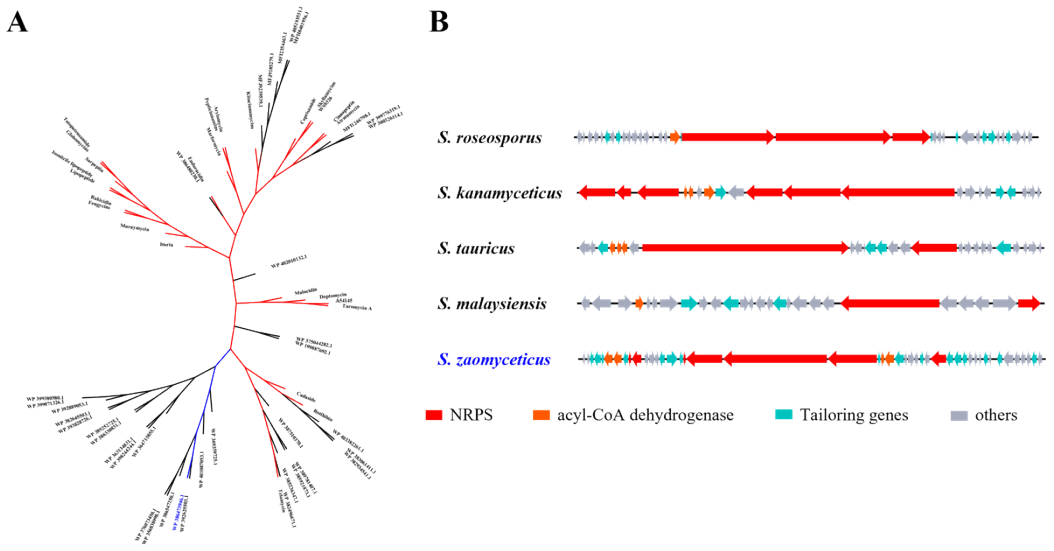

Fig. 4. Phylogenetic analysis of distinct Cs domains.

(A) Phylogeny tree of 66 Cs domains. Red branches represent characterized Cs (n = 26), black branches denote uncharacterized Cs with potential novel functions (n = 36), and blue branche corresponds to WP_386473946.1; (B) Five phylogenetically discrete BGCs classes, each demonstrating signature structural architectures and significant diversification capacity. *S. zaomyceticus* possesses WP_386473946.1.

Structure-Guided Functional Analysis of the Cs Domain

Leveraging the AlphaFold3 protein structure prediction platform, we pioneered a high-confidence 3D structural model of the Cs domain WP_386473946.1 (Fig. 5A). Comparative structural analysis revealed a distinct V-shaped catalytic cleft topology with significantly larger cavity volume than homologous structures. Crucially, a characteristic His59-Asp91 secondary structural element (Fig. 5B) forms unique spatial arrangements likely modulating substrate-binding

specificity and catalytic microenvironment. Further multiple sequence alignment identified two conserved glycine residues (Gly168/Gly170) precisely positioned adjacent to the HHxxDG functional motif within the active pocket (Fig. 5C). This discovery not only provides structural rationale for substrate recognition mechanisms but also establishes critical engineering targets for rational enzyme redesign.

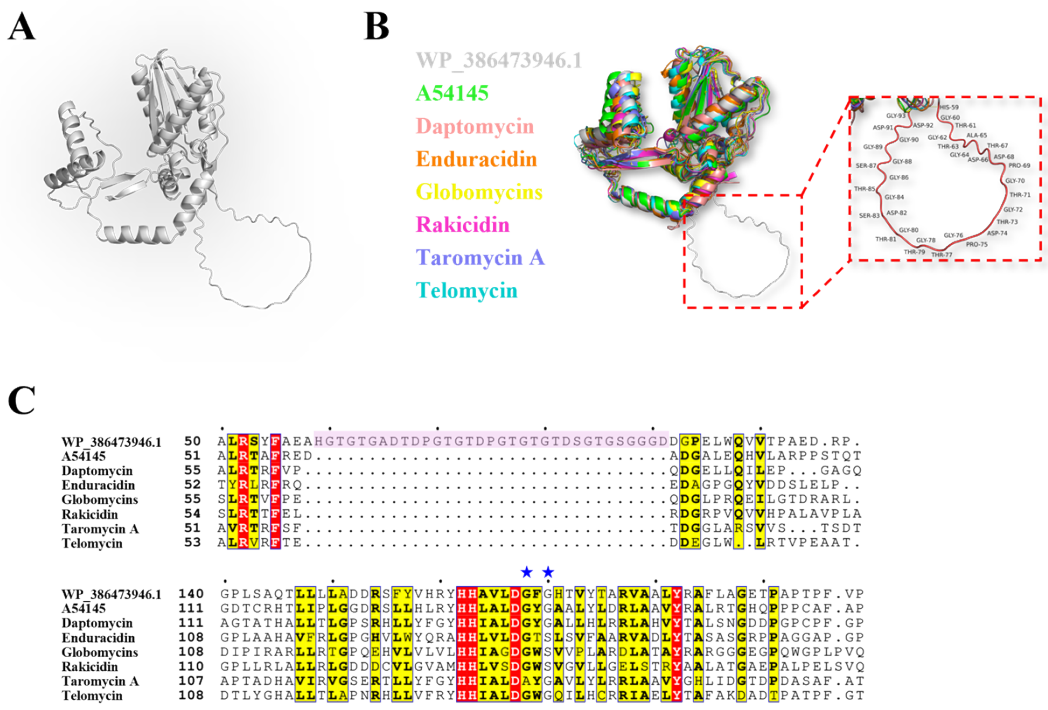

Fig. 5. Structural prediction and sequence alignment of WP_386473946.1.

(A) Three-dimensional structural modeling of WP_386473946.1 using AlphaFold3 (https://alphafoldserver.com/welcome); (B) Overlaid plot of 7 modeled known Cs and WP_386473946.1; (C) Sequence alignment of Cs of 7 known lipopeptides and WP_386473946.1.

Conclusions

This study established a bioinformatics-driven strategy for the rapid discovery of lipopeptides, resulting in a high-value resource library of *Streptomyces* lipopeptide BGCs. From this library, we identified BGCs predicted to encode novel bioactive lipopeptides. Furthermore, three-dimensional

structural predictions of candidate Cs domains directly inform engineering strategies to modulate substrate scope and selectivity. Collectively, this Cs domain-centric mining strategy provides an accelerated pathway for the discovery of lipopeptides with novel bioactivities.

Acknowledgments

We acknowledge Dr. Wei. Bai at Nanjing University of Chinese Medicine, for the technical assistance in using SSN for the analysis of lipopeptides biosynthetic gene clusters.

Supplementary Information

The following supporting information can be downloaded at: https://zenodo.org/records/15669838.

Funding

This research received no external funding.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its Supplementary Materials. Portions of the data are also available in the publicly accessible databases of the NCBI (https://www.ncbi.nlm.nih.gov/).

Conflicts of Interest

The author declares no conflicts of interest.

References

1. Qiao W, Wang L, Luo Y, Yang T. Synthetic Approaches and Therapeutic Applications of FDA-Approved Antibacterial Agents: A Comprehensive Review from 2003 to 2023. Eur J Med Chem. 2025; 117267–117267. doi:10.1016/j.ejmech.2025.117267

2. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399: 629–655. doi:10.1016/s0140-6736(21)02724-0
3. Kersten RD, Yang Y-L, Xu Y, Cimermancic P, Nam S-J, Fenical W, et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011;7: 794–802. doi:10.1038/nchembio.684
4. Yang J, Bahreman A, Daudey G, Bussmann J, Olsthoorn RCL, Kros A. Drug Delivery via Cell Membrane Fusion Using Lipopeptide Modified Liposomes. ACS Cent Sci. 2016;2: 621–630. doi:10.1021/acscentsci.6b00172
5. Sreedharan SM, Rishi N, Singh R. Microbial lipopeptides: Properties, mechanics and engineering for novel lipopeptides. Microbiol Res. 2023;271: 127363. doi:10.1016/j.micres.2023.127363
6. Pattelli ON, Valdivia EM, Beyersdorf MS, Regan CS, Rivas M, Hebert KA, et al. A Lipopeptidomimetic of Transcriptional Activation Domains Selectively Disrupts the Coactivator Med25 Protein–Protein Interactions. Angew Chem Int Ed. 2024;63: e202400781–e202400781. doi:10.1002/anie.202400781
7. Vicente-Garcia C, Colomer I. Lipopeptides as tools in catalysis, supramolecular, materials and medicinal chemistry. Nat Rev Chem. 2023;7: 710–731. doi:10.1038/s41570-023-00532-8
8. Müller A, Wenzel M, Strahl H, Grein F, Saaki TNV, Kohl B, et al. Daptomycin inhibits cell envelope synthesis by interfering with fluid membrane microdomains. Proc Natl Acad Sci USA. 2016;113: E7077–E7086. doi:10.1073/pnas.1611173113

9. Baltz RH, Miao V, Wrigley SK. Natural products to drugs: daptomycin and related lipopeptide antibiotics. Nat Prod Rep. 2005;22: 717. doi:10.1039/b416648p

10. Olatunji S, Yu X, Bailey J, Huang C-Y, Zapotoczna M, Bowen K, et al. Structures of lipoprotein signal peptidase II from Staphylococcus aureus complexed with antibiotics globomycin and myxovirescin. Nat Commun. 2020;11. doi:10.1038/s41467-019-13724-y

11. Zhang S, Chen Y, Zhu J, Lu Q, Cryle MJ, Zhang Y, et al. Structural diversity, biosynthesis, and biological functions of lipopeptides from*Streptomyces*. Nat Prod Rep. 2023;40: 557–594. doi:10.1039/d2np00044j

12. Kadaikunnan S, Vijayaragavan P, Rathi MA, Balamuralikrishnan B, Alharbi NS, Gopalakrishnan VK, et al. Antibacterial and biofilm disruptive nonribosomal lipopeptides from Streptomyces parvulus against multidrug-resistant bacterial infections. J Infect Public Health. 2024;17: 450–456. doi:10.1016/j.jiph.2024.01.001

13. Lombardo G, Dorm BC, Salvay AG, Franzi L, Gaffney ML, Peredo Camio JB, et al. Novel chitosan-based hydrogels as promising wound dressing materials with advanced properties. Int J Biol Macromol. 2024;279: 135423. doi:10.1016/j.ijbiomac.2024.135423

14. Liong A, Leão PN. Fatty acyl-AMP ligases in bacterial natural product biosynthesis. Nat Prod Rep. 2025. doi:10.1039/d4np00073k

15. Bhattacharya A, Cho CJ, Brea RJ, Devaraj NK. Expression of Fatty Acyl-CoA Ligase Drives One-Pot *De Novo* Synthesis of Membrane-Bound Vesicles in a Cell-Free Transcription-Translation System. J Am Chem Soc. 2021;143: 11235–11242. doi:10.1021/jacs.1c05394

16. Wang X, Ning X, Zhao Q, Kang Q, Bai L. Improved PKS Gene Expression With Strong

Endogenous Promoter Resulted in Geldanamycin Yield Increase. Biotechnol J. 2017;12. doi:10.1002/biot.201700321

17. Wang Z, Koirala B, Hernandez Y, Zimmerman M, Park S, Perlin DS, et al. A naturally inspired antibiotic to target multidrug-resistant pathogens. Nature. 2022;601: 606–611. doi:10.1038/s41586-021-04264-x

18. Yan F, Burgard C, Popoff A, Zaburannyi N, Zipf G, Maier J, et al. Synthetic biology approaches and combinatorial biosynthesis towards heterologous lipopeptide production. Chem Sci. 2018;9: 7510–7519. doi:10.1039/c8sc02046a

19. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res. 2024. doi:10.1093/nar/gkae1115

20. Oberg N, Zallot R, Gerlt JA. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. J Mol Biol. 2023;435: 168018–168018. doi:10.1016/j.jmb.2023.168018

21. Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, Augustijn HE, et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Res. 2025. doi:10.1093/nar/gkaf334

22. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 2014;42: W320–W324. doi:10.1093/nar/gku316

23. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024;52: W78–W82. doi:10.1093/nar/gkae268

24. Roy R, Al-Hashimi HM. AlphaFold3 takes a step toward decoding molecular behavior and biological computation. Nat Struct Mol Biol. 2024. doi:10.1038/s41594-024-01350-2

25. Bai X, Zhong L, Chen H, Liu Y, Shi X, Wang X, et al. Changing the initiation unit of nonribosomal peptide synthetases to access underexplored biosynthetic potential. In: bioRxiv. Cold Spring Harbor Laboratory; 2025. doi:10.1101/2025.05.29.656725