Predicting protein complexes in biosynthetic gene clusters

Yoshitaka Moriwaki[1,2,†], Taro Shiraishi[3], Yohei Katsuyama[2,4], Kenichi Matsuda[5], Toyoyuki Ose[6], Atsushi Minami[7], Hideaki Oikawa[8], Tomohisa Kuzuyama[2,4], Ryuichiro Ishitani[1], Tohru Terada[2,4]

[1]Department of Computational Drug Discovery and Design, Medical Research Laboratory, Institute of Integrated Research, Institute of Science Tokyo, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan.

[2]Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.

[3]Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo-ku, 606-8502 Kyoto, Japan.

[4]Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan

[5]Faculty of Pharmaceutical Sciences, Hokkaido University, kita-ku, Sapporo, 060-0812, Japan.

[6]Faculty of Advanced Life Science, Hokkaido University, kita-ku, Sapporo, 060-0810, Japan.

[7]Department of Chemistry, Institute of Science Tokyo, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8551, Japan.

[8]Division of Chemistry, Faculty of Science, Hokkaido University, Sapporo, 060-0810, Japan.

ABSTRACT: Biosynthetic gene clusters (BGCs) are contiguous genomic regions that encode diverse, non-homologous proteins required for the production of specific natural products. Their genetic diversity underlies the structural complexity of these compounds, and their biosynthetic pathways still remain to be clarified in many cases. The biosynthetic mechanisms rely not only on substrate specificity between proteins and ligands, but also on protein–protein interactions that mediate transport of intermediates, regulation of activity, and structural stabilization. However, sequence-based functional predictions have had limited success for many uncharacterized proteins within BGCs. To address this challenge, we built a high-throughput complex prediction pipeline by replacing AlphaFold3's multiple sequence alignment generation with a faster MMSeqs2. We systematically screened 487,828 protein pairs derived from 2,437 BGCs registered in the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database and predicted 15,438 heteromeric interactions with an ipTM ≥ 0.6. Among them, 1,390 protein pairs exhibited structural homology with an RMSD ≤ 2.0 Å. These predictions highlight intriguing molecular mechanisms involving proteins previously annotated as uncharacterized or potentially dysfunctional. Furthermore, we provide these results in a reusable format and as a protein interaction network map to facilitate future experimental validation by researchers.

INTRODUCTION

Secondary metabolites are natural products produced by organisms that are not directly involved in their growth, development, or reproduction. Unlike primary metabolites, such as amino acids, nucleotides, and carbohydrates, secondary metabolites often play specialized roles in ecological interactions—such as defense against predators[2], competition with other organisms[3], or symbiotic relationships[4]. These compounds are typically synthesized by biosynthetic gene clusters (BGCs), which are co-located sets of genes that collectively encode enzymes required to produce a specific metabolite and auxiliary proteins such as transporter and regulatory proteins[5,6]. Secondary metabolites contain a wide variety of chemical structures and classes, including non-ribosomal peptides (NRPs), polyketides (PKs), ribosomally synthesized and post-translationally modified peptides (RiPPs), terpenes, and saccharides, which have attracted significant interest from both enzymology and organic chemistry.

Early studies on secondary metabolites were limited in scope due to the large number of host organisms that could not be cultured under laboratory conditions. However, advances in genetic engineering and the establishment of heterologous expression systems have enabled the expression of a wider range of proteins. Since the 2000s, progress in sequencing technologies, such as next-generation sequencing and metagenomic analysis, along with the development of bioinformatics tools and the expansion of genomic databases, has significantly advanced research on secondary metabolites derived from uncultivable microorganisms. For example, antiSMASH[7-14] is a widely used bioinformatics tool for the identification, annotation, and analysis of BGCs in microbial genomes. It also classifies clusters into known types by comparing them to curated reference datasets, including the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database[15]. DeepBGC[16] incorporates deep learning and natural language processing strategies and incorporates genomic position dependency and order information not considered in hidden Markov model (HMM)-based tools such as ClusterFinder[17] and antiSMASH. This improved the identification of BGCs in bacterial genomes and contributed to the discovery of novel BGCs encoding natural products that could not be detected by existing methods.

In addition to these sequence-based analyses, highly accurate prediction of protein tertiary structures, as exemplified by AlphaFold 2 (AF2)[18], has become available since 2021. The computational

and biological breakthrough significantly expanded the availability of protein structural information, which had previously been limited to around 200,000 experimental structures. AlphaFold can generate accurate models of protein monomers and complexes by predicting inter-residue distances based on co-evolution signals concealed in the multiple sequence alignments (MSA) for the query sequence(s), provided that more than roughly 30–100 effective homologous sequences are available in the target sequence database. The predicted structures have been already used as search models of molecular replacement method of for X-ray crystallography[19,20], as well as initial structures for molecular dynamics (MD) simulations and quantum mechanics/molecular mechanics (QM/MM) calculations[21]. Furthermore, AlphaFold provides valuable predictions about protein–protein[22] and protein–ligand[23] interactions that are difficult to obtain from sequence-based bioinformatics tools alone. One of the applications of the former prediction is the discovery and experimental validation of the protein Tmem81, which is conserved across vertebrates during fertilization of sperm and egg[24].

Formation of protein complexes also plays a crucial role in BGCs. Notably, protein–protein interactions observed in carrier proteins[25], *trans*-type acyltransferase (AT) polyketide synthase (PKS)[26], *cis*-type AT PKS[27], type I/II PKS, and fatty acid synthase[28,29] have been extensively studied and discussed as key components in the biosynthetic pathways. In homomultimer complexes, residues that are distant in the primary and tertiary structures assemble at the multimer interface to form a single binding pocket[30]. Additionally, some heteromeric complexes influence the stability, specificity, and catalytic activity[31,32]. Because these interactions underlie the remarkable structural and chemical diversity of the products, many enzymes in these systems have been structurally characterized to elucidate their molecular basis. Nevertheless, most BGCs still harbor numerous proteins of unknown function, often annotated simply as "hypothetical protein" in MIBiG, which represents a major challenge for the comprehensive understanding of their biosynthetic pathways.

Here, we report a comprehensive search of all pairwise complex predictions among proteins encoded within each BGCs using AlphaFold3 (AF3)[23]. These computational analyses enabled highly sensitive detection of potential homo- and heteromultimer formation that were previously difficult to predict using conventional sequence-based bioinformatics tools. This approach led to the identification of novel pairs of proteins that are likely to act cooperatively. Furthermore, our results suggest that numerous structurally homologous heterocomplexes exist in BGCs and that many of them involve proteins previously annotated as having unknown functions. Finally, we visualized the protein–protein interaction network maps to provide a comprehensive overview in the BGCs. Taken together, these findings offer novel and significant insights into previously uncharacterized proteins in BGCs, thereby providing valuable contributions to both experimental and computational researcher communities engaged in the investigation of secondary metabolites.

MATERIALS AND METHODS

Structure Prediction Workflow.

Input amino acid sequences. We selected the proteins derived from BGCs for comprehensive protein-protein complex prediction based on the following criteria. First, we focused on 2,437 "active" BGC accession IDs deposited in MIBiG version 4.0 (2024-11-15) that are currently maintained and updated. IDs marked as "retired" or "pending" were excluded from analysis. Second, among the proteins encoded in these BGCs, those with more than 1,950 amino acids, which are commonly found in PKS- or NRPS-modular type proteins, were excluded from the prediction, as they can exceed the computational limits of modeling using AF3. The amino acid sequences were obtained from the GBK format files provided by MIBiG. Using the amino acid sequences that meet the criteria, we performed exhaustive predictions of homo- and heterodimeric complexes for each BGC.

MSA generation and structure prediction. Despite recent advances, protein structure prediction with AF3 still relies on sufficient number of multiple sequence alignments (MSAs) for the query sequences. To achieve both high accuracy and high-throughput prediction, we replaced the default MSA generation pipeline using HMMER 3[33] implemented in AF3 with a faster alternative: the *colabfold_search* command in LocalColabFold[34]. MMseqs2 version 15-6f452[35] was used to build two sequence databases, Uniref30_2302 and colabfold_envdb_202108 (released on July 31, 2023), which can be downloaded from https://colabfold.mmseqs.com/ and generate MSAs for the query sequences. The obtained MSAs have been shown to be sufficient for producing highly accurate predicted models with AF2 or ColabFold in most cases.[34] In addition to unpaired MSAs, we also obtained paired MSAs to further improve the accuracy of complex modeling. The protein template search against PDB100 database was not performed, as it is known to improve monomeric but not complex prediction accuracy. The MSA generation was performed on the SuperComputer Flow Type III subsystem at Nagoya University. This system comprises 16 Intel Xeon Platinum 8280M processors (2.7 GHz, 28 cores each) and 24 TiB of RAM per node. Each query took approximately 1–2 minutes on the calculation node. The obtained a3m-formatted MSA files were added to the "pairedMsa" and "UnpairedMsa" keys of the AF3 input JSON file using the *msatojson* command of the alphafold3_tools software that we developed and made available on our GitHub repository: https://github.com/cddlab/alphafold3_tools/. The "modelSeeds" value in the AF3 input JSON file was set to 1 and five models were generated for each prediction. For dimer with a total number of residues below 1500, the complex structure prediction was performed using the SuperComputer Flow Type II subsystem with NVIDIA GPU V100; for those with 1500 to 2000 residues, our in-house NVIDIA GPU RTX4090; and for those with 2000–3900 residues, NVIDIA GPU H100 (VRAM 94GB). The Predicted Aligned Error (PAE) of the predicted structures was visualized using the *paeplot* command of alphafold3_tools.

Validation of the predicted complex structures.

Metrics. To identify biologically plausible protein–protein interactions in the AF3-predicted complexes, we employed two metrics: the interface predicted TM-score (ipTM) and the interaction prediction score from aligned errors (ipSAE)[36]. The ipTM measures the accuracy of the predicted relative arrangement of subunits. Values ≥ 0.8 indicate confident and high-quality predictions, < 0.6 suggest failure, and 0.6–0.8 represent an uncertain range.[22,23] The ipTM values are provided in the AF3 output JSON files. The ipSAE is a modified scoring function that evaluates the quality of predicted interactions by focusing on residue pairs likely to participate in the interface, rather than the entire sequence.

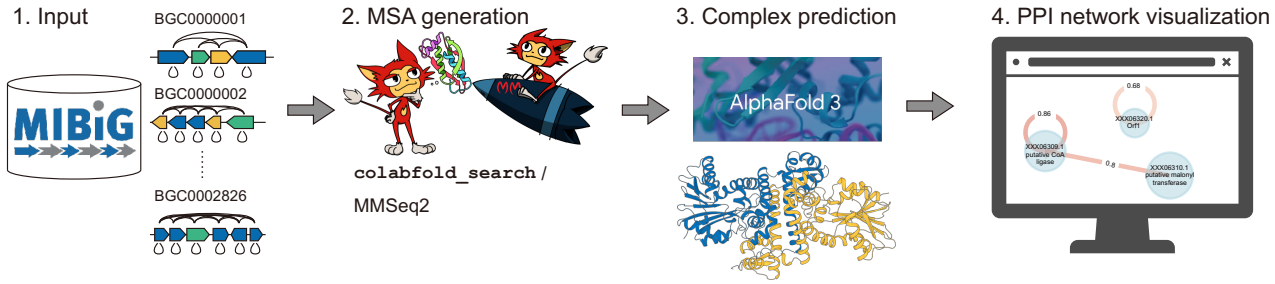

Figure 1 Computational workflow. 1. Input amino acid sequences. List all pairs of self or other protein sequences encoded in each BGC accession ID, excluding those with a length more than 1950 residues. 2. Generate paired and unpaired MSAs for all the protein pairs using MMseqs2 and colabfold_search. 3. Predict the dimeric complex using AlphaFold3. The input JSON files were created with alphafold3_tools. 4. Visualize a protein-protein interaction (PPI) map for each BGC accession ID. Output JSON files containing ipTM and ipSAE metrics were also generated.

This metric can be computed by applying the ipSAE script[36] to the AF3 output JSON file with both the PAE cutoff (pae_cutoff) and Cα-Cα distance cutoff (dist_cutoff) set to 10. We did not use other metrics such as pDock[37], pDockQ2[38], or LIS[39] in this analysis, although they serve similar purposes in evaluating protein–protein interactions.

Validation datasets and complex analysis. We evaluated both ipTM and ipSAE metrics for the predicted homo- and heterodimeric structures by validating them against experimentally determined complexes. For proteins encoded in the BGCs listed in MIBiG version 4.0 and also registered in the Protein Data Bank (PDB), we retrieved their biologically relevant complex structures (also referred as "biological assemblies") and stoichiometry information were retrieved from Protein Data Bank (PDB) as of November 30, 2024. To minimize the impact of amino acid substitutions on protein oligomerization, the sequence identity threshold was set to 95%. Additionally, for complexes whose biological assemblies contained more than two chains and did not exceed 3,900 amino acid residues in total, we performed further structure predictions with the stoichiometries to obtain the corresponding ipTM and ipSAE metrics.

For the obtained heterodimeric structures, the RMSD of the backbone heavy atoms (N, Cα, C, and O) between the two subunits was calculated using Gemmi version 0.7.3 (https://github.com/project-gemmi/gemmi) [40], with the maximum number of outlier rejection cycles set to 5. To prevent calculation errors caused by very short sequences (e.g., precursor peptides), RMSD values were not computed for proteins with fewer than 20 residues.

Network visualization

The protein–protein networks obtained for the 2437 BGCs were drawn using NetworkX version 3.5. Each protein was represented as a node, and edges were drawn between heterodimers that showed ipTM > 0.55 in the AF3 structural prediction. Proteins that formed homodimers at the same threshold were represented with self-loop edges. Then, the network maps were output in SVG image format and made available on a web server. Additionally, the JSON files and Python codes used for generating these network maps will be published as supplementary files, allowing users to regenerate them in their own environments using arbitrary thresholds.

RESULTS

High-throughput complex prediction pipeline. We developed a computational pipeline for the systematic prediction of protein–protein complexes encoded within BGCs (**Figure 1**). The pipeline takes as input the GenBank (.gbk) files associated with BGC accession IDs registered in MIBiG. Users may also provide a GenBank files generated by antiSMASH or FASTA-formatted protein sequence files for unregistered BGCs. Here, we focused on 2437 BGCs that are labeled as "active" in MIBiG and targeted proteins with up to 1,950 amino acid residues. For each BGC, complexes were predicted for all possible protein combinations, including self-pairs. Specifically, for a BGC containing N qualifying proteins, we performed $N(N-1)/2$ predictions for heteromeric pairs and N additional predictions for homomeric pairs, resulting in a total of $N(N+1)/2$ predictions per BGC.

Highly accurate structure prediction using AF3 requires the generation of multiple sequence alignments (MSAs) for the input sequences. However, comprehensive prediction using the AF3 Web server is not feasible due to its usage limits (currently 30 trials per day). Moreover, the official AF3 inference code available on GitHub is also impractical for the high-throughput workflows because the MSA generation process using HMMER 3 (http://hmmer.org) [33] is slow. To address these issues, we used *colabfold_search* from LocalColabFold[34] in conjunction with MMseqs2[35] to rapidly generate MSAs. The resultant MSAs included both unpaired MSAs for individual proteins and paired MSAs generated by concatenating sequences from the same organism. These were assigned to the "unpairedMsa" and "pairedMsa" keys, respectively, in the AF3 input JSON file. This conversion process was achieved using *msatojson* command developed as part of our toolkit of *alphafold3_tools*. Using this pipeline, we prepared the input JSON files for subsequent complex structure prediction on 487,828 protein pairs.

Validation of the predicted homomeric complexes. We subsequently compared the predicted complexes with the experimentally determined protein complexes deposited in the PDB to evaluate the validity. As of November 30, 2024, a total of 555 homomeric and 30 heteromeric protein entries were found in PDB to exhibit over 95% sequence identity with proteins registered in MIBiG ver. 4.0. Among the 555 homomeric proteins, 408 were annotated as homodimers, and 135 as homotrimers or higher-order assemblies that could be modeled on our GPUs. The remaining 12 entries could not be predicted due to their extremely large complex sizes. A high sequence similarity threshold of 95% was applied to minimize the effects of amino acid differences, insertions, or deletions that could alter the composition of residues at the

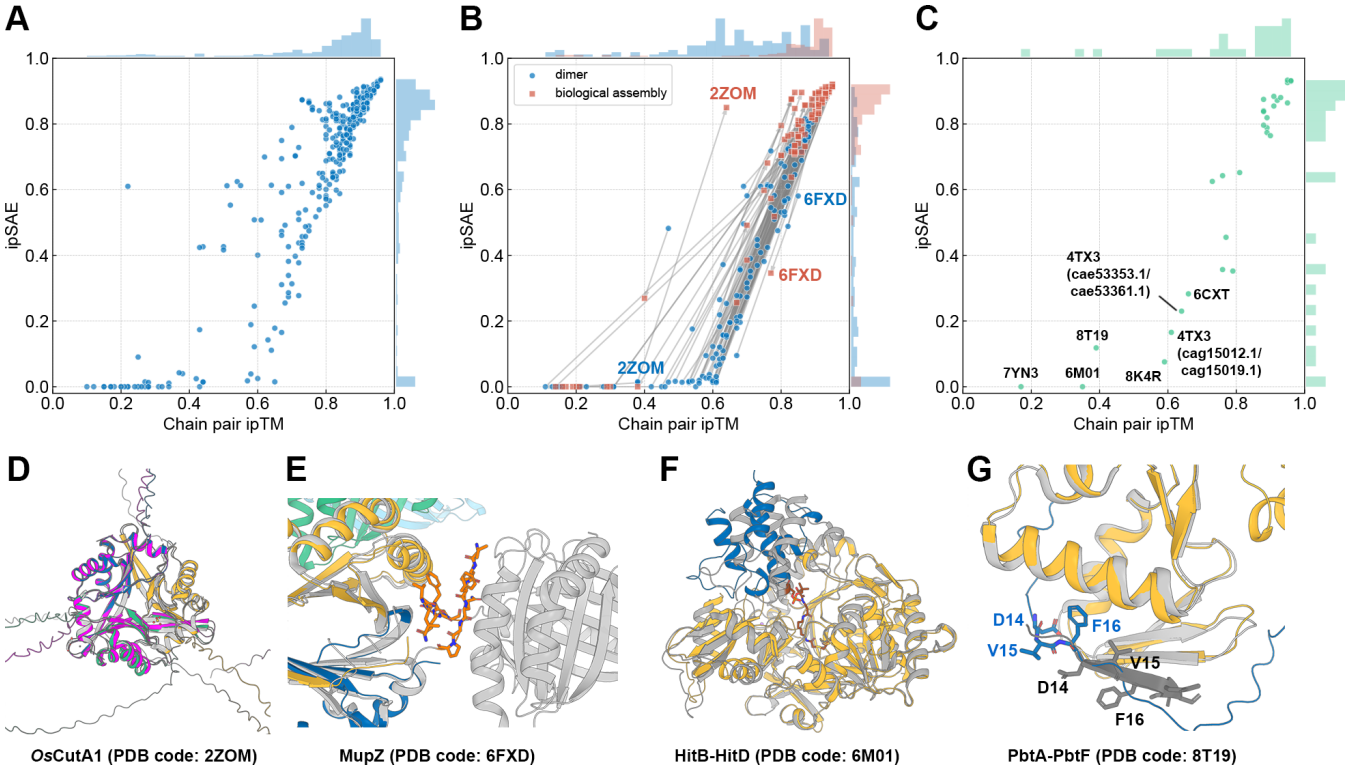

Figure 2. Validation of the predicted complexes. (A) Scatter plot of chain-level ipTM and ipSAE values for predicted homodimers (n = 408). Marginal histograms for each axis are displayed above and to the right. (B) Differences in ipTM and ipSAE between the predicted homodimer (blue) and the biological assembly (red). Only biological assemblies forming homomeric complexes larger than dimers and computable on our GPUs are included (n = 135). (C) Scatter plot of ipTM and ipSAE values for predicted heterodimers. (D) Superimposition of the predicted *Os*CutA1 trimer (GenBank ID: BAT10611.1; blue, yellow, and green) and dimer (purple) with the crystal structure (PDB code: 2ZOM). (E) Predicted MupZ homotetramer (colored) and its crystal structure (white). An artificial N-terminal peptide (orange) is shown in stick representation. (F) Predicted HitB (GenBank ID: BAR73008.1)–HitD (BAR73010.1) and its crystal structure (white and gray). The cross-linking probe in the crystal structure is shown in brown stick representation. (G) Superimposition of the predicted PbtA (AGY49587.1; blue)–PbtF (AGY49592.1; yellow) complex with the crystal structure (PDB code: 8T19; white and gray). Residues of PbtA at the interface are shown in stick representation.

multimerization interface and thereby lead to inaccurate measurements.

For the protein structures registered in the PDB as homodimers, our predicted dimeric models showed ipTM ≥ 0.8 for 294/408 (72.1%) and ipTM ≥ 0.6 for 358/408 (87.7%) (**Figure 2A**). The ipSAE metric, which represents a corrected ipTM calculated using only interchain residue pairs with reliable PAE scores, correlated well with ipTM above the 0.6 threshold, but approached zero below this cutoff. We next examined the 135 proteins annotated as forming higher-order homomeric complexes. When predicted as dimers, these proteins frequently yielded ipTM values around 0.6. However, when re-predicted with their stoichiometries of biological assemblies registered in PDB, the ipTM values substantially improved, exceeding 0.8 in most cases (**Figure 2B**). A representative example is CutA1 from *Oryza sativa* (PDB code: 2ZOM). In the dimeric prediction, the substructure was nearly identical to the crystal structure but exhibited low confidence scores (ipTM = 0.38; ipSAE = 0.001). In contrast, the trimeric prediction, corresponding to the biological assembly, closely reproduced the experimental structure, with improved ipTM and ipSAE values of 0.64 and 0.84, respectively (**Figure 2D**). These results highlight the limitations of using ipTM or ipSAE thresholds for dimer prediction, especially when the inter-subunit interface is small. Additionally, the ipTM and ipSAE metrics were further improved to 0.93 and 0.86, respectively, when its N-terminal long transit peptide region (residues 1–64) was not included in the prediction. This finding also indicated that the long-disordered region lowered ipTM (**Supplementary Figure S1**). Despite this, ipSAE remained high, underscoring its reliability as an evaluation metric in such cases.

Nevertheless, some of the predicted complexes did not exhibit high ipTM/ipSAE values even when using the stoichiometries of biological assemblies. Closer inspection revealed that several of these cases represented artifacts, either due to misassignments of stoichiometry inferred from the crystal packing using PDBePISA[41], or the inclusion of artificial sequence tags that induced unnatural oligomerization. The former example is MxaA in *Stigmatella aurantiaca* (GenBank ID: AAK57184.1; PDB code: 4U7W), which is experimentally validated as a monomer[42]. The latter example is MupZ from *Pseudomonas fluorescens* (GenBank ID: AFD33556.1; PDB code: 6FXD), which is registered as a homotetramer due to C-terminal His_6-tag–induced tetramerization, although the native protein forms a dimer.[43] We summarized the proteins exhibiting low ipSAE and ipTM values in **Supplementary Table 1**. These observations highlight the challenges of constructing fully reliable datasets of multimeric assemblies from the PDB. Nonetheless, our findings suggest that an ipTM threshold of 0.6 is effective for screening proteins for the potential of homooligomerization.

Validation of the predicted heteromeric complexes. For the 30 heteromeric complexes, we observed similar ipTM and ipSAE

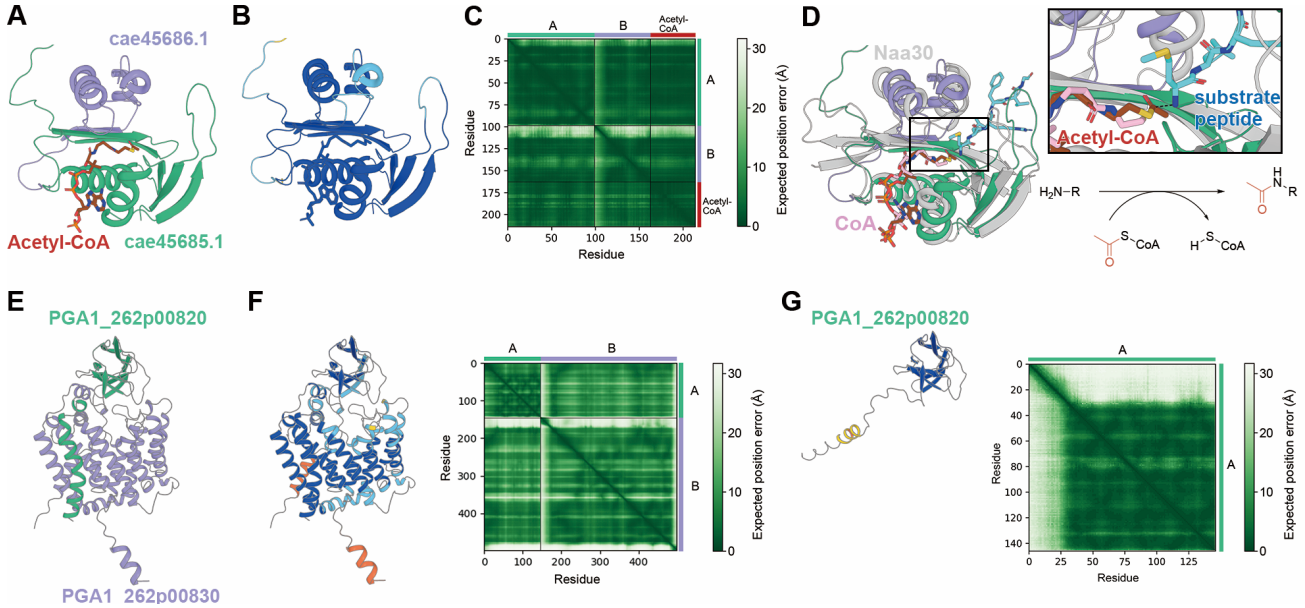

Figure 3. Examples of predicted complex models. (A) cae45685.1 (green)–cae45686.1 (purple)–Acetyl-CoA (brown) trimeric complex model. (B) The predicted complex colored by pLDDT (orange, 0–50; yellow, 50–70; cyan, 70–90; and blue, 90–100). (C) The predicted aligned error (PAE) matrix with chain coloring of (A) on the side bars. The dashed black lines indicate the chain boundaries. (D) Superimposition of the Naa30 crystal structure (white) on the predicted model shown in Panel A. The N-terminal region of major capsid protein as a substrate peptide (Uniprot ID: P32503) is shown as cyan stick. (E) PGA1_262p00820 (green)–PGA1_262p00830 (purple) complex model found in BGC0000932. (F) pLDDT-colored predicted PGA1_262p00820 –PGA1_262p00830 complex. The PAE matrix with chain coloring of (F) is shown on right panel. (G) Predicted PGA1_262p00820 monomer structure colored by pLDDT. The right panel shows the PAE matrix. These data were obtained from the AlphaFold Protein Structure Database[1] (UniProt ID: I7E6A3).

correlations to those in the homomeric complexes. Among these, 6M01[44], 7YN3[45], 8K4R[46], 8T19 (DOI: 10.2210/pdb8T19/pdb), and 4TX3[47] showed ipTM < 0.6. We further investigated the cause of these low ipTM values. First, in PDB entries of 7YN3, 8K4R (VinM-VinL), and 6M01, a chemical reagent to induce covalent cross-linking between the two proteins was used to obtain the crystal structure of the heterodimeric complexes. Biologically, one of the proteins in these two PDB entries is an acyl carrier protein (ACP), which is known to form a transient complex with the other protein. Thus, it is speculated that the ipTM does not yield high values for such complexes.

PDB code 8T19 represents the complex structure of PbtA, a RiPP precursor peptide, and its modifying enzyme PbtF from the GE2270 BGC (BGC0001155). The ipTM and ipSAE values were 0.39 and 0.1, respectively. Interestingly, AF3 predicted that the peptide binds to the β-strand region of PbtF; however, the predicted binding position was shifted by two residues compared with the crystal structure. We further investigated NisB from *Lactococcus lactis subsp. lactis* and StrB from *Streptococcus thermophilus* LMD-9 (BGC0001209)[48] to gain insights into proteins that bind RiPP precursor peptides. The crystal structure of NisB in complex with its precursor peptide NisA was not included in the validation dataset because these proteins were fused within the same chain (PDB code: 4WD9)[49] or the peptide is chemically modified (PDB code: 6M7Y)[50]. StrB shares 94.98% sequence identity with its homologous protein SuiB from *Streptococcus suis* (PDB code: 5V1T)[51]. The predicted complex structures for both matched well with their corresponding crystal structures, showing no residue shift and high ipTM and ipSAE values (**Supplementary Figure S2**). However, because of the extremely limited number of available structures or experimental data on formation of RiPP precursor peptide–enzyme complexes, a statistical evaluation for them remains difficult.

Taken together, these results indicate that AF3 successfully predicted complex formation (ipTM ≥ 0.6) in 26 out of the 30 heteromeric cases. However, there are limitations in its predictive capabilities, particularly for transient complexes or those with small binding interfaces, such as those formed by ACPs or precursor peptides.

Novel heterocomplexes. Among 487,828 predicted protein pairs, 15,438 heterodimers exhibited ipTM ≥ 0.6, and 3,754 showed ipTM ≥ 0.8. These include complexes formed by previously uncharacterized proteins. Here, we can hypothesize the functions of these proteins based on the predicted models. For example, a heterodimer of cae45685.1 and cae45686.1 (**Figure 3A–C**) found in borrelidin BGC from *Streptomyces parvulus* (MIBiG accession ID: BGC0000031)[52-54] is annotated as "hypothetical protein" in MIBiG, and to our knowledge, no studies have been reported for these proteins. BLASTp[55] search predicted both proteins have partial sequence similarity with GCN5-related *N*-acetyltransferase (GNAT). Structure-based searches against PDB100 database using the Foldseek webserver[56] indicated that the predicted complex structure resembles the single subunit Naa30 of the heterotrimeric NatC complex (PDB code: 6YGD, chain A)[57], which catalyzes N-terminal acetylation (Nt-acetylation) of numerous eukaryotic proteins. Superimposition of the complex model on the Naa30 structure revealed that a binding pocket for the cofactor acetyl-CoA is formed between the cae45685.1–cae45686.1 interface, with the acetyl group of acetyl-CoA being exposed to the peptide-binding site of Naa30 (**Figure 3D**). This suggests that this heterocomplex may also catalyze Nt-acetylation of its substrate. Recently, *Streptomyces mutabilis sp.* MII, which is thought to possess a BGC similar to

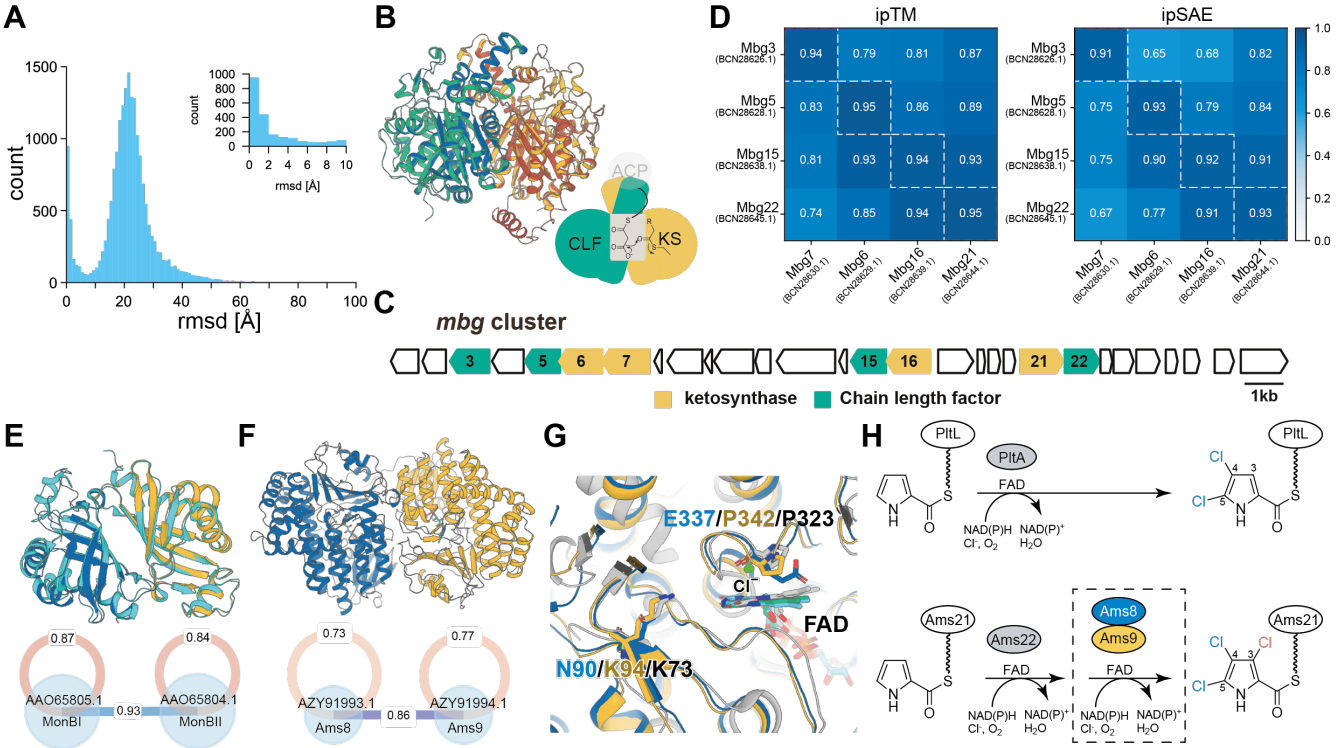

Figure 4. Predicted structurally homologous complexes. (A) Histogram of structural RMSD values between monomeric units of predicted heterodimer pairs. An inset shows a magnified view of the 0–10 Å range. (B) Superimposition of the predicted complex structure of actinorhodin polyketide β-ketoacyl synthase alpha (KS) and beta (CLF) subunits (blue and red) on the crystal structure (PDB code: 1TQY; green and yellow). KS is shown in blue and green, and CLF in red and yellow. (C) *mbg* cluster. Only the genes encoding KS and CLF are depicted. (D) Computed ipTM and ipSAE values for each complex prediction of the four KS (horizontal) and four CLF (vertical) subunits in *mbg* cluster. (E) Superimposition of the predicted MonBI (blue)–MonBII (yellow) heterodimer on the crystal structure of the MonBI homodimer (PDB code: 3WMD; cyan). The computed ipTM values for each homo- and hetero-dimer are shown below. (F) Predicted Ams8 (blue)–Ams9 (yellow) complex. (G) Close-up view of the FAD binding site in the predicted structures of Ams8 (blue) and Ams9 (yellow). The pltA crystal structure (PDB code: 5DBJ) is shown in gray. Residues presumed to be important for chlorination activity are shown in stick representation. (H) Proposed trichlorination mechanism in the armeniaspirol BGC (bottom). Formation of the dichloropyrrole moiety catalyzed by PltA during pyoluteorin biosynthesis is shown on the top for comparison.

BGC0000031, has been reported to produce *N*-acetylborrelidin B in addition to borrelidin (**Supplementary Figure S3A**)[58]. Given that borrelidin B, a known precursor of borrelidin, has an aminomethyl group at C12 carbon, it is plausible that the heterocomplex binds borrelidin B at a peptide-binding site analogous to that of Naa30 and catalyzes its acetylation to yield *N*-acetylborrelidin B (**Supplementary Figure S3B and C**).

We also identified several Domain Unknown Function (DUF) proteins whose function appears to arise only upon complex formation. For example, PGA1_262p00820 (Genbank ID: AFO93364.1) in tropodithietic acid BGC from *Phaeobacter inhibens* DSM 17395 (MIBiG accession ID: BGC0000932)[59] was annotated as a DUF4399 domain–containing protein, a family for which functional characterization remains scarce. Our analysis predicted that this protein forms a complex with PGA1_262p00830 (AFO93365.1), a DoxX-family protein, with an ipTM of 0.82 and an ipSAE of 0.64 (**Figure 3E**). Notably, the N-terminal region of PGA1_262p00820 was predicted to be disordered as a monomer; however, in the heterodimer, this region adopted an α-helical conformation with high pLDDT and formed an interaction interface with a helix of PGA1_262p00830 (**Figure 3F and G**). Given that DoxX-family proteins are localized to the cell membrane, PGA1_262p00820 is expected to be embedded with PGA1_262p00830 in the cell membrane. Thus, the two proteins may act cooperatively as membrane-bound enzymes or transporters.

Structurally homologous heterocomplexes. By examining the predicted heterodimers, we identified 952 pairs (4.9%) in which the structural RMSD between constituent proteins was ≤1.0 Å, and 1390 pairs (7.1%) with RMSD ≤ 2.0 Å (**Figure 4A**). The median ipTM and ipSAE values for the protein pairs with RMSD ≤2.0 Å were 0.82 and 0.71, respectively, indicating that these predictions were made with high confidence. This observation suggests that structurally homologous heterodimers are present across BGCs and contribute to the biosynthesis. A well-characterized and popular example is a ketosynthase (KS)–chain length factor (CLF) complex, which constitutes the central catalytic machinery in Type II PKSs. In this complex, the KS subunit catalyzes the decarboxylation of the malonyl unit tethered to a phosphopantetheinyl arm of an ACP, and the condensation of the resulting enolate anion with a growing polyketide chain attached to its catalytic cysteine. The CLF subunit is a KS homolog lacking catalytic residues, but the internal tunnel formed by the complex plays a critical role in determining the polyketide chain length. The actinorhodin polyketide beta-ketoacyl synthase alpha (GenBank ID: CAC44200.1) and beta (CAC44201.1) subunits of BGC0000194 have been identified as KS and CLF, respectively, and their heterocomplex structure was determined (PDB code: 1TQY)[60]. The protein pair was predicted to form a complex with high ipTM and ipSAE values of 0.96 and 0.93, respectively, and closely matched the crystal structure (**Figure 4B**). Interestingly, despite the KS and CLF subunits sharing

the same fold with an RMSD of 1.02 Å, the homodimeric prediction for KS yielded relatively low scores (ipTM = 0.65, ipSAE = 0.38), and CLF showed almost no signal (ipTM = 0.20, ipSAE = 0.0). These findings indicate that AF3 is capable of correctly identifying heterocomplex formation independent of overall structural similarity between monomers.

Subsequently, we examined whether ipTM and ipSAE could identify correct KS–CLF pairs within a BGC cluster encoding multiple KS and CLF proteins. In our previous work, we experimentally validated the correct pairing in *mbg* cluster (BGC0002441; **Figure 4C**) as Mbg7–Mbg3, Mbg6–Mbg5, Mbg16–Mbg15, and Mbg21–Mbg22 [61]. Our all-by-all complex predictions for the 4 × 4 combinations showed that both ipTM and ipSAE values reached their maximum values for the experimentally validated pairs (**Figure 4D**). Notably, ipSAE yielded lower values for incorrect pairings than ipTM, indicating greater discriminatory power. A similar trend was observed in the Ysf cluster (BGC0002547)[62], which encodes three KSs and three CLFs, where the highest values again corresponded to the experimentally confirmed pairings (**Supplementary Figure S4**). Collectively, these results demonstrate that the ipTM and ipSAE metrics provide reliable indicators for identifying true KS–CLF pairs and predicting their released polyketide intermediates, even when the corresponding genes are not located adjacent within the BGC.

Our comprehensive predictions further identified 381 protein pairs that are likely to form both homodimers and heterodimers, meeting the criteria of ipTM ≥ 0.6 and ipSAE ≥ 0.6. A representative example is the pair MonBI and MonBII from the monensin BGC (BGC0000100). Although the crystal structure of the MonBI homodimer has been determined in our previous work (PDB code: 3WMD)[63], biological assays have shown that MonBI also forms a heterodimer with MonBII, and this complex displays enzymatic activity. AF3 predictions indicated that MonBI and MonBII form homodimers with ipTM values of 0.87 and 0.84, respectively, while they form a heterodimer with a higher ipTM value of 0.93 (**Figure 4E**). The predicted MonBI–MonBII heterodimer structure closely matched the MonBI homodimer crystal structure, with an RMSD of 1.0 Å. This result suggests that dimer formation between these two proteins is competitive, and that the heterodimer is preferentially formed.

Proteins predicted to form both homo- and hetero-dimers were also observed among other functional proteins. We highlight three flavin-dependent halogenase-like proteins—Ams8, Ams9, and Ams22—from the armeniaspirol BGC (BGC0002022)[64] as an example. Within the BGC, Ams21 and Ams22 share sequence similarities of 43% and 65% with PltL and PltA, respectively, from the characterized pyoluteorin BGC (BGC0000127), suggesting that Ams21 functions as an ACP and Ams22 as a flavin-dependent halogenase. Moreover, the structure of the active site of Ams22 is highly similar to that of PltA (PDB code: 5DBJ)[65], which catalyzes the 4,5-dichlorination of the pyrrolyl moiety covalently attached to the phosphopantetheinyl arm of PltL, strongly suggesting that Ams22 catalyzes the same reaction (**Supplementary Figure S5**). However, unlike in the pyoluteorin BGC, the armeniaspirol BGC was reported to produce a 3,4,5-trichlorinated pyrollyl intermediate during the biosynthetic process[64]. Thus, the remaining halogenase-like enzymes, Ams8 and Ams9, are hypothesized to catalyze chlorination at the 3-position of the pyrrolyl moiety, although the mechanism remains unclear[64].

Our structure prediction pipeline showed that Ams8 and Ams9 may form homodimers with ipTM values of 0.77 and 0.73, respectively, whereas their heterodimer displayed a higher ipTM value of 0.86 (**Figure 4F**). Intriguingly, the predicted structure of Ams8 revealed that E337 is positioned at a site that would inhibit the binding of flavin adenine dinucleotide (FAD), an essential cofactor for halogenation. In addition, the conserved lysine residue required to mediate the reaction between the substrate and hypochlorous acid[66-68]—generated from FAD, molecular oxygen, and chloride ions—is replaced by N90 in Ams8. (**Figure 4G**). These unusual residue substitutions strongly suggest that Ams8 lacks halogenase activity. By contrast, Ams9 and Ams22 retain these conserved residues. However, gene deletion of *ams8* completely abolished the biosynthetic process of armeniaspirol, whereas deletion of either *ams9* or *ams22* reduced its production by 100-fold but did not lead to complete loss[64]. This indicates that Ams8 is more essential for the biosynthetic pathway. Overall, these findings suggest that Ams9 functions as the enzyme catalyzing trichlorination of 4,5-dichloropyrrolyl–*S*-Ams22, while Ams8 acts as an auxiliary protein that supports the activity of Ams9 through complex formation (**Figure 4H**).

Visualization of the protein–protein interaction network map. To facilitate the use of our computational results by a broad range of researchers, we created a web site that visualizes protein–protein network diagrams for the 2437 BGCs. On this site, users can filter the data by MIBiG accession ID, taxonomy, class, or compound names, and view both the network diagrams and representative chemical structures of the products (**Figure 5A**). Proteins are represented as nodes, and pairs of proteins predicted to form homodimers or heterodimers are connected by edges when they exceed a certain threshold (**Figure 5B**). The JSON files and Python scripts used to generate these network diagrams are available in the supplementary files, allowing users to perform their own analyses.

DISCUSSION

One of the core challenges in the field of secondary metabolism and biosynthesis research is to elucidate the origins and biosynthetic mechanisms of the vast repertoire of natural products. In this study, we established a high-throughput complex prediction pipeline based on AF3 and demonstrated that it enables comprehensive detection of protein–protein interactions within BGCs. We further showed that BGCs contain a considerable proportion of protein homo- or hetero- complexes whose function emerges only upon complex formation, as well as structurally homologous heterocomplexes.

Although our pipeline and network maps provide valuable insights into the interactome in BGCs, they have three limitations. First, AF2 (including AlphaFold-Multimer) and AF3 often fail to accurately predict weak or transient protein complexes. This limitation arises because their algorithms infer three-dimensional protein structures primarily from coevolutionary signals in the MSA derived from individual amino acid sequence(s), and inter-chain interactions were not considered during their training processes[23]. Carrier proteins, such as ACPs in PKSs and peptidyl carrier proteins (PCPs) in NRPSs, exemplify this limitation. These proteins tether biosynthetic intermediates via a thioester bond to the thiol group of a phosphopantetheinyl arm attached to a conserved serine residue. They then interact with multiple catalytic domains in a specific yet transient manner to promote stepwise transformations of the substrate. Although identifying their partner enzymes is crucial for

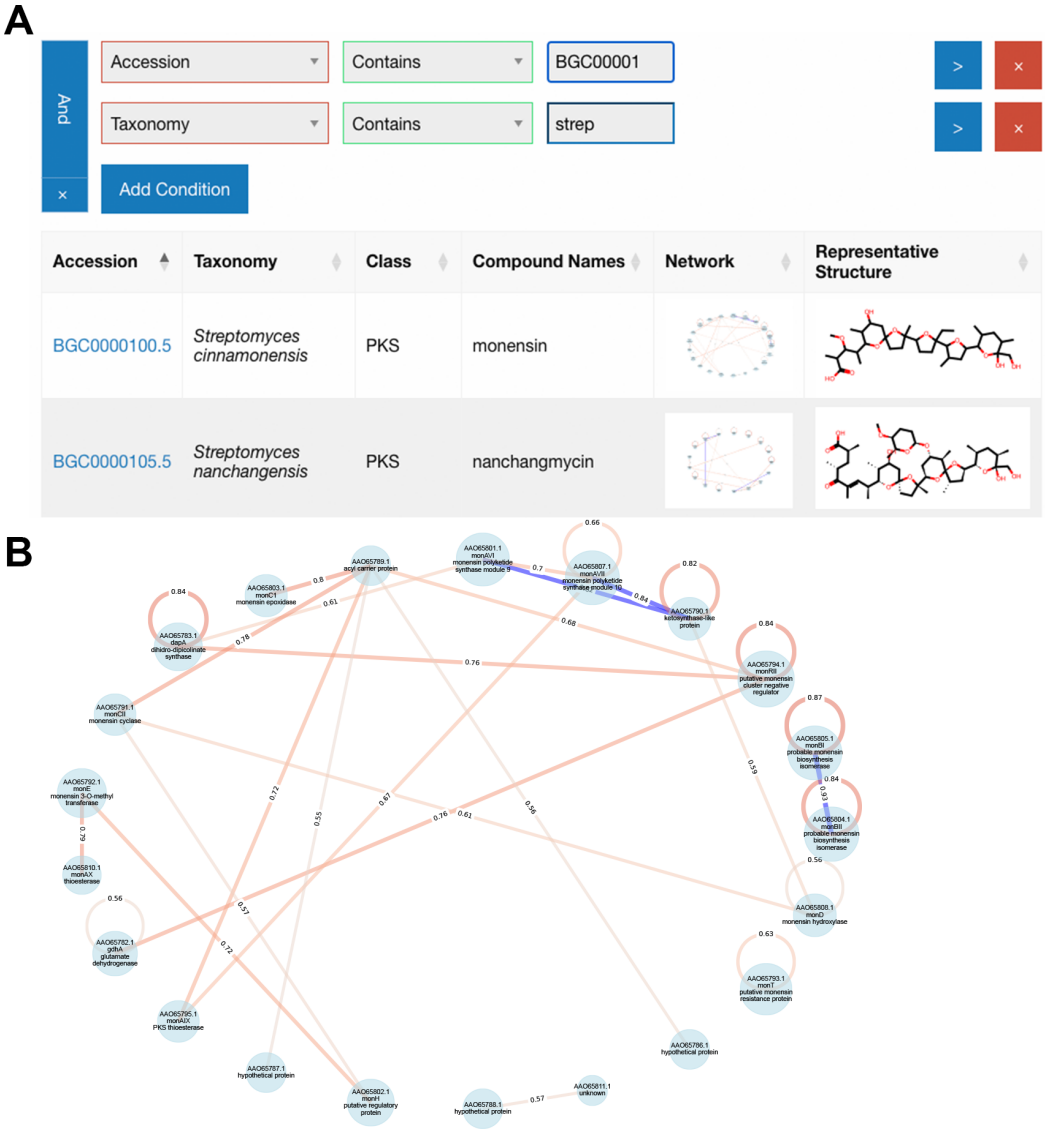

Figure 5. Visualization of the protein–protein interaction (PPI) network map. (A) The interface on the website. (B) The PPI network map in the monensin BGC (BGC0000100). The nodes represent proteins encoded within the BGC. Edges indicate predicted interactions, with labels showing the corresponding ipTM values. Blue edges denote structurally homologous interactions. Only nodes connected by edges above the ipTM threshold are displayed (ipTM > 0.55 was applied on the website).

studying biosynthetic pathways, improving their prediction accuracy remains an important technical challenge. Similarly, predicting the binding position of precursor peptides in RiPPs is also difficult for the same reason.

Second, our computational pipeline did not perform complex predictions for proteins larger than 1,950 residues due to computational limitations. Such large proteins are typically found in type I PKS or NRPS systems, where multiple catalytic domains are linked within a single polypeptide to form the core scaffold of the product. While these modular proteins mainly assemble the product through their intramolecular domain–domain interactions, some tailoring enzymes within the same BGC can modify intermediates tethered to the modular proteins, thereby expanding the chemical diversity of the final products.[69,70] Prior prediction of these interactions could help identify novel tailoring enzymes responsible for introducing unusual functional groups; however, such predictions were not fully performed in this study.

Third, our method may fail to correctly predict the complex formation of proteins with potentially conformational changes. Although talaromyolides *tlx* BGC from *Talaromyces purpureogenus* has not been registered in the current MIBiG dataset yet, the TlxJ–TlxI heterodimer crystal structure (PDB code: 7VBQ)[71] is notable case to discuss the issue (**Supplementary Figure S6A**). While sequence analysis suggests that both TlxJ (GenBank ID: BDC03472.1) and TlxI (BDC03471.1) were nonheme iron oxygenases and the RMSD between the two proteins was 1.6 Å, TlxI is suggested to be enzymatically dysfunctional as oxidases due to the lack of an arginine residue for α-ketoglutarate (α-KG) binding and serve as an auxiliary protein to assist in TlxJ expression enhancing its catalytic activity[71]. The gel filtration analysis and the X-ray crystallography indicated that TlxI is monomeric (PDB code: 7VBR).[71] However, while TlxJ–TlxI was correctly predicted as a heterocomplex with an ipTM of 0.91 (**Supplementary Figure S6B**), TlxI was predicted to form a homodimer with an ipTM of 0.85 (**Supplementary Figure S6C**). Interestingly, the crystal structures

revealed that TlxI undergoes a large conformational change when forming a heterocomplex with TlxJ. Yet the predicted TlxI structure was close to the conformation observed in the heterocomplex crystal structure in both monomeric and homomeric forms (**Supplementary Figure S6C and D**). This observation is consistent with the caveat that AF3 does not adequately predict proteins that undergo conformational changes[23]. By contrast, AF3 successfully predicted the homodimer of MonBI with an ipTM of 0.87 and the heterodimer with MonBII with a higher ipTM of 0.93 (**Figure 4E**). These results suggest that MonBI does not undergo significant conformational changes upon dimerization. To our knowledge, MSA subsampling with AF2 or ColabFold[72] is one of the best strategies for predicting such conformational changes in advance. However, we did not apply this approach in this study because of its high computational cost.

Although several technical challenges remain in accurately predicting protein complexes, our comprehensive approach uncovered numerous plausible complexes hidden within BGCs. Notably, the structurally homologous heterodimers identified in this study may serve as intriguing targets for future research, particularly in the context of microbial evolution. According to the genome streamlining theory, natural selection in prokaryotes tends to eliminate genetic redundancy and nonessential sequences, leading to the evolution of highly compact and efficient genomes[73,74]. Thus, the structurally homologous complexes are unlikely to be mere products of gene duplication but are instead expected to possess distinct functional roles such as regulation or assistance in protein expression. Examples of the former function have been reported in both homo-[75,76] and hetero-complexes[71], and their characterization is relatively straightforward. In contrast, the latter function can only be uncovered through co-expression of a specific protein combination, and its underlying molecular mechanisms remain largely unexplored. Among the reported cases, MonBI–MonBII[63] and LstA–LstB[77] become functional only when co-expressed. Co-expression of LxmK and LxmY has been shown not only to improve stability but also to increase yield.[78] Interestingly, TalA from *Talaromyces pinophilus* LD-7[79] and Lsd19 from *Streptomyces lasaliensis*[80-83] have been identified as naturally occurring chimeric enzymes, in which two structurally homologous protein domains are fused into a single polypeptide chain. Such protein (or domain) pairs may exist to expand the range of catalytic functions that can be achieved by a single enzyme alone.

Currently, sequence-based prediction tools such as DeepBGC and ClusterFinder are widely used to identify BGCs that produce novel natural products. In contrast, structure-based approaches have the potential to elucidate the molecular mechanisms underlying these products based on the relationship between protein function and tertiary structure. The development of prediction tools for protein structures and their interaction networks is a key step toward unraveling enzymatic mechanisms and entire metabolic pathways, even in uncultivable microorganisms. Our results provide a comprehensive overview of the constituent proteins within BGCs and highlight intriguing protein complexes for future investigation. These bottom-up approaches will accelerate identification of protein pairs that should be co-expressed or co-analyzed and deepen our biochemical understanding, ultimately providing insights essential for the rational engineering of enzymes.

ASSOCIATED CONTENT

Data Availability

The predicted protein–protein interaction network and JSON files underlying this article are available in Zenodo at https://doi.org/10.5281/zenodo.17451667.

AUTHOR INFORMATION

Corresponding Author

moriwaki.yoshitaka@tmd.ac.jp

Author Contributions (CRediT Taxonomy)

Yoshitaka Moriwaki (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Funding acquisition [equal], Investigation [lead], Methodology[lead], Resources [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]),
Taro Shiraishi (Conceptualization [equal], Investigation [supporting], Resources [equal], Validation [equal], Writing—review & editing [equal]),
Yohei Katsuyama (Conceptualization [equal], Investigation [supporting], Validation [equal], Writing—review & editing [equal]),
Kenichi Matsuda (Conceptualization [equal], Investigation [supporting], Writing—review & editing [equal]),
Toyoyuki Ose (Writing—review & editing [supporting]),
Atsushi Minami (Writing—review & editing [supporting]),
Hideaki Oikawa (Writing—review & editing [supporting]),
Tomohisa Kuzuyama (Funding acquisition [lead], Project administration [lead], Writing—review & editing [equal]),
Ryuichiro Ishitani (Funding acquisition [equal], Resources [equal], Supervision [equal], Writing—review & editing [equal]),
Tohru Terada (Funding acquisition [equal], Project administration [lead], Resources [equal], Supervision [equal], Writing—review & editing [equal])

ACKNOWLEDGMENT

We thank Dr Naruki Yoshikawa for a fruitful discussion.

Funding Sources

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan [Grant-in-Aid for Transformative Research Areas, 22H05126 to Y.M. and T.T., 22H05120 to T.K. and T.S., 22H05130 to Y.K., 22H05128 to K.M., 25H01575 to A.M., and 25H02250 to R.I.; Grant-in-Aid for Scientific Research (JSPS KAKENHI), 25K01890 to Y.M. and 19H04634 to T.O.]. Additional support was provided by the Medical Research Center Initiative for High Depth Omics at the Institute of Science Tokyo, Nanken-Kyoten at the Institute of Science Tokyo 2025, and the Multilayered Stress Diseases project (JPMXP1323015483) at the Institute of Science Tokyo. This work utilized computational resources from the supercomputer "Flow" provided by the Information Technology Center at Nagoya University through the HPCI System Research Project (Project IDs: jh240001 and jh250002) and the TSUBAME 4.0 supercomputer at the Institute of Science Tokyo.

Conflict of interest

None declared.

ABBREVIATIONS

BGC, biosynthetic gene cluster; ipTM, interface predicted template modeling; AF3, AlphaFold3; RiPP(s), ribosomally synthesized and post-translationally modified peptide(s); ACP, acyl-carrier protein.

REFERENCES

1. Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-

sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–D444. https://doi.org/10.1093/nar/gkab1061

2. Zhang S, Mukherji R, Chowdhury S *et al.* Lipopeptide-mediated bacterial interaction enables cooperative predator defense. *Proc Natl Acad Sci USA* 2021;**118**:e2013759118. https://doi.org/10.1073/pnas.2013759118

3. Chevrette M, Thomas C, Hurley A *et al.* Microbiome composition modulates secondary metabolism in a multispecies bacterial community. *Proc Natl Acad Sci USA* 2022;**119**:e2212930119. https://doi.org/10.1073/pnas.2212930119

4. Modolon F, Barno A, Villela H *et al.* Ecological and biotechnological importance of secondary metabolites produced by coral-associated bacteria. *J Appl Microbiol* 2020;**129**:1441–1457. https://doi.org/10.1111/jam.14766

5. Malpartida F and Hopwood D A. Physical and genetic characterisation of the gene cluster for the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2). *Mol Genet Genomics* 1986;**205**:66–73.

6. Stanzak R, Matsushima P, Baltz R *et al.* Cloning and Expression in *Streptomyces lividans* of Clustered Erythromycin Biosynthesis Genes from *Streptomyces erythreus*. *Bio/Technology* 1986;**4**:229–232. https://doi.org/10.1038/nbt0386-229

7. Medema M, Blin K, Cimermancic P *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011;**39**:W339–W346. https://doi.org/10.1093/nar/gkr466

8. Blin K, Medema M, Kazempour D *et al.* antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* 2013;**41**:W204–W212. https://doi.org/10.1093/nar/gkt449

9. Weber T, Blin K, Duddela S *et al.* antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 2015;**43**:W237–W243. https://doi.org/10.1093/nar/gkv437

10. Blin K, Wolf T, Chevrette M *et al.* antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* 2017;**45**:W36–W41. https://doi.org/10.1093/nar/gkx319

11. Blin K, Shaw S, Steinke K *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;**47**:W81–W87. https://doi.org/10.1093/nar/gkz310

12. Blin K, Shaw S, Kloosterman A *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 2021;**49**:W29–W35. https://doi.org/10.1093/nar/gkab335

13. Blin K, Shaw S, Augustijn H *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* 2023;**51**:W46–W50. https://doi.org/10.1093/nar/gkad344

14. Blin K, Shaw S, Vader L *et al.* antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res* 2025;**53**:W32–W38. https://doi.org/10.1093/nar/gkaf334

15. Zdouc M, Blin K, Louwen N *et al.* MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* 2025;**53**:D678–D690. https://doi.org/10.1093/nar/gkae1115

16. Hannigan G, Prihoda D, Palicka A *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;**47**:https://doi.org/10.1093/nar/gkz654

17. Cimermancic P, Medema M, Claesen J *et al.* Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* 2014;**158**:412–421. https://doi.org/10.1016/j.cell.2014.06.034

18. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–589. https://doi.org/10.1038/s41586-021-03819-2

19. McCoy A, Sammito M and Read R. Implications of *AlphaFold2* for crystallographic phasing by molecular replacement. *Acta Crystallographica Section D: Structural Biology* 2022;**78**:1–13. https://doi.org/10.1107/S2059798321012122

20. Akdel M, Pires D, Pardo E *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 2022;**29**:1056–1067. https://doi.org/10.1038/s41594-022-00849-w

21. Homma M, Wakabayashi T, Moriwaki Y *et al.* Insights into stereoselective ring formation in canonical strigolactone: Discovery of a dirigent domain-containing enzyme catalyzing orobanchol synthesis. *Proc Natl Acad Sci USA* 2024;**121**:e2313683121. https://doi.org/10.1073/pnas.2313683121

22. Evans R, O'Neill M, Pritzel A *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2022;2021.2010.2004.463034. https://doi.org/10.1101/2021.10.04.463034

23. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500. https://doi.org/10.1038/s41586-024-07487-w

24. Deneke V, Blaha A, Lu Y *et al.* A conserved fertilization complex bridges sperm and egg in vertebrates. *Cell* 2024;**187**:7066–7078.e7022. https://doi.org/10.1016/j.cell.2024.09.035

25. Gulick A and Aldrich C. Trapping interactions between catalytic domains and carrier proteins of modular biosynthetic enzymes with chemical probes. *Nat Prod Rep* 2018;**35**:1156–1184. https://doi.org/10.1039/c8np00044a

26. Kosol S, Jenner M, Lewandowski J *et al.* Protein–protein interactions in *trans*-AT polyketide synthases. *Nat Prod Rep* 2018;**35**:1097–1109. https://doi.org/10.1039/c8np00066b

27. Dodge G, Maloney F and Smith J. Protein–protein interactions in "*cis*-AT" polyketide synthases. *Nat Prod Rep* 2018;**35**:1082–1096. https://doi.org/10.1039/c8np00058a

28. Herbst D, Townsend C and Maier T. The architectures of iterative type I PKS and FAS. *Nat Prod Rep* 2018;**35**:1046–1069. https://doi.org/10.1039/c8np00039e

29. Chen A, Re R and Burkart M. Type II fatty acid and polyketide synthases: deciphering protein–protein and protein–substrate interactions. *Nat Prod Rep* 2018;**35**:1029–1045. https://doi.org/10.1039/c8np00040a

30. Mori T, Moriwaki Y, Sakurada K *et al.* Molecular basis for the diversification of lincosamide biosynthesis by pyridoxal phosphate-dependent enzymes. *Nat Chem* 2025;**17**:256–264. https://doi.org/10.1038/s41557-024-01687-7

31. Sikandar A and Koehnke J. The role of protein-protein interactions in the biosynthesis of ribosomally synthesized and post-translationally modified peptides. *Nat Prod Rep* 2019;**36**:1576–1588. https://doi.org/10.1039/c8np00064f

32. Rattray D and Foster L. Dynamics of protein complex components. *Curr Opin Chem Biol* 2019;**48**:81–85. https://doi.org/10.1016/j.cbpa.2018.11.003

33. Eddy S R. Accelerated Profile HMM Searches. *PLoS Comp Biol* 2011;**7**:e1002195. https://doi.org/10.1371/journal.pcbi.1002195

34. Mirdita M, Schütze K, Moriwaki Y *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**:679–682. https://doi.org/10.1038/s41592-022-01488-1

35. Steinegger M and Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–1028. https://doi.org/10.1038/nbt.3988

36. Dunbrack R L. *Rēs ipSAE loquunt*: What's wrong with AlphaFold's ipTM score and how to fix it. *bioRxiv* 2025;2025.2002.2010.637595. https://doi.org/10.1101/2025.02.10.637595

37. Bryant P, Pozzati G and Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* 2022;**13**:1265. https://doi.org/10.1038/s41467-022-28865-w

38. Zhu W, Shenoy A, Kundrotas P *et al.* Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* 2023;**39**:btad424. https://doi.org/10.1093/bioinformatics/btad424

39. Kim A-R, Hu Y, Comjean A *et al.* Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer. *bioRxiv* 2024;2024.2002.2019.580970. https://doi.org/10.1101/2024.02.19.580970

40. Wojdyr M. GEMMI: A library for structural biology. *J Open Source Softw* 2022;7:4200. https://doi.org/10.21105/joss.04200

41. Krissinel E and Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 2007;**372**:774–797. https://doi.org/10.1016/j.jmb.2007.05.022

42. Barajas J, Phelan R, Schaub A *et al.* Comprehensive Structural and Biochemical Analysis of the Terminal Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem Biol* 2015;**22**:1018–1029. https://doi.org/10.1016/j.chembiol.2015.06.022

43. Wang L, Parnell A, Williams C *et al.* A Rieske oxygenase/epoxide hydrolase-catalysed reaction cascade creates oxygen heterocycles in mupirocin biosynthesis. *Nat Catal* 2018;**1**:968–976. https://doi.org/10.1038/s41929-018-0183-5

44. Miyanaga A, Kurihara S, Chisuga T *et al.* Structural Characterization of Complex of Adenylation Domain and Carrier Protein by Using Pantetheine Cross-Linking Probe. *ACS Chem Biol* 2020;**15**:1808–1812. https://doi.org/10.1021/acschembio.0c00403

45. Mori T, Kadlcik S, Lyu S *et al.* Molecular basis for carrier protein-dependent amide bond formation in the biosynthesis of lincosamide antibiotics. *Nat Catal* 2023;**6**:531–542. https://doi.org/10.1038/s41929-023-00971-y

46. Miyanaga A, Nagata K, Nakajima J *et al.* Structural Basis of Amide-Forming Adenylation Enzyme VinM in Vicenistatin Biosynthesis. *ACS Chem Biol* 2023;**18**:2343–2348. https://doi.org/10.1021/acschembio.3c00517

47. Haslinger K, Peschke M, Brieke C *et al.* X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature* 2015;**521**:105–109. https://doi.org/10.1038/nature14141

48. Schramma K R, Bushin L B and Seyedsayamdost M R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat Chem* 2015;**7**:431–437. https://doi.org/10.1038/nchem.2237

49. Ortega M A, Hao Y, Zhang Q *et al.* Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* 2015;**517**:509–512. https://doi.org/10.1038/nature13888

50. Bothwell I R, Cogan D P, Kim T *et al.* Characterization of glutamyl-tRNA-dependent dehydratases using nonreactive substrate mimics. *Proc Natl Acad Sci USA* 2019;**116**:17245–17250. https://doi.org/10.1073/pnas.1905240116

51. Davis K M, Schramma K R, Hansen W A *et al.* Structures of the peptide-modifying radical SAM enzyme SuiB elucidate the basis of substrate recognition. *Proc Natl Acad Sci USA* 2017;**114**:10420–10425. https://doi.org/10.1073/pnas.1703663114

52. Olano C, Wilkinson B, Moss S *et al.* Evidence from engineered gene fusions for the repeated use of a module in a modular polyketide synthase. *Chem Commun* 2003;2780–2782. https://doi.org/10.1039/b310648a

53. Olano C, Moss S, Braña A *et al.* Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: insights into nitrile formation. *Mol Microbiol* 2004;**52**:1745–1756. https://doi.org/10.1111/j.1365-2958.2004.04090.x

54. Olano C, Wilkinson B, Sánchez C *et al.* Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: Cluster analysis and assignment of functions. *Chem Biol* 2004;**11**:87–97. https://doi.org/10.1016/j.chembiol.2003.12.018

55. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+ : architecture and applications. *BMC Bioinformatics* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421

56. van Kempen M, Kim S S, Tumescheit C *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;**42**:243–246 https://doi.org/10.1038/s41587-023-01773-0

57. Grunwald S, Hopf L, Bock-Bierbaum T *et al.* Divergent architecture of the heterotrimeric NatC complex explains N-terminal acetylation of cognate substrates. *Nat Commun* 2020;**11**:5506. https://doi.org/10.1038/s41467-020-19321-8

58. Hamed A, Abdel-Razek A, Frese M *et al.* *N*-Acetylborrelidin B: a new bioactive metabolite from *Streptomyces mutabilis* sp MII. *Z Naturforsch, C: Biosci* 2018;**73**:49–57. https://doi.org/10.1515/znc-2017-0140

59. Brock N, Nikolay A and Dickschat J. Biosynthesis of the antibiotic tropodithietic acid by the marine bacterium *Phaeobacter inhibens*. *Chem Commun* 2014;**50**:5487–5489. https://doi.org/10.1039/c4cc01924e

60. Keatinge-Clay A, Maltby D, Medzihradszky K *et al.* An antibiotic factory caught in action. *Nat Struct Mol Biol* 2004;**11**:888–893. https://doi.org/10.1038/nsmb808

61. Zhang J, Yuzawa S, Thong W *et al.* Reconstitution of a Highly Reducing Type II PKS System Reveals 6π-Electrocyclization Is Required for *o*-Dialkylbenzene Biosynthesis. *J Am Chem Soc* 2021;**143**:2962–2969. https://doi.org/10.1021/jacs.0c13378

62. Deng Z, Liu J, Li T *et al.* An Unusual Type II Polyketide Synthase System Involved in Cinnamoyl Lipid Biosynthesis. *Angew Chem Int Ed* 2021;**60**:153–158. https://doi.org/10.1002/anie.202007777

63. Minami A, Ose T, Sato K *et al.* Allosteric Regulation of Epoxide Opening Cascades by a Pair of Epoxide Hydrolases in Monensin Biosynthesis. *ACS Chem Biol* 2014;**9**:562–569. https://doi.org/10.1021/cb4006485

64. Fu C, Xie F, Hoffmann J *et al.* Armeniaspirol Antibiotic Biosynthesis: Chlorination and Oxidative Dechlorination Steps Affording Spiro[4.4]non-8-ene. *ChemBioChem* 2019;**20**:764–769. https://doi.org/10.1002/cbic.201800791

65. Pang A, Garneau-Tsodikova S and Tsodikov O. Crystal structure of halogenase PltA from the pyoluteorin biosynthetic pathway. *J Struct Chem* 2015;**192**:349–357. https://doi.org/10.1016/j.jsb.2015.09.013

66. Barker R D, Yu Y, De Maria L *et al.* Mechanism of Action of Flavin-Dependent Halogenases. *ACS Catal* 2022;**12**:15352–15360. https://doi.org/10.1021/acscatal.2c05231

67. Yeh E, Blasiak L C, Koglin A *et al.* Chlorination by a long-lived intermediate in the mechanism of flavin-dependent halogenases. *Biochemistry* 2007;**46**:1284–1292. https://doi.org/10.1021/bi0621213

68. Flecks S, Patallo E P, Zhu X *et al.* New insights into the mechanism of enzymatic chlorination of tryptophan. *Angew Chem Int Ed* 2008;**47**:9533–9536. https://doi.org/10.1002/anie.200802466

69. Ames B, Nguyen C, Bruegger J *et al.* Crystal structure and biochemical studies of the trans-acting polyketide enoyl reductase LovC from lovastatin biosynthesis. *Proc Natl Acad Sci USA* 2012;**109**:11144–11149. https://doi.org/10.1073/pnas.1113029109

70. Matthews M, Krest C, Barr E *et al.* Substrate-Triggered Formation and Remarkable Stability of the C-H Bond-Cleaving Chloroferryl Intermediate in the Aliphatic Halogenase, SyrB2. *Biochemistry* 2009;**48**:4331–4343. https://doi.org/10.1021/bi900109z

71. Li X, Awakawa T, Mori T *et al.* Heterodimeric Non-heme Iron Enzymes in Fungal Meroterpenoid Biosynthesis. *J Am Chem Soc* 2021;**143**:21425–21432. https://doi.org/10.1021/jacs.1c11548

72. del Alamo D, Sala D, McHaourab H S *et al.* Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* 2022;**11**:https://doi.org/10.7554/eLife.75751

73. Giovannoni S, Tripp H, Givan S *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 2005;**309**:1242–1245. https://doi.org/10.1126/science.1114057

74. Giovannoni S, Thrash J and Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014;**8**:1553–1565. https://doi.org/10.1038/ismej.2014.60

75. Kotaka M, Ren J, Lockyer M *et al.* Structures of R- and T-state *Escherichia coli* aspartokinase III: Mechanisms of the allosteric transition and inhibition by lysine. *J Biol Chem* 2006;**281**:31544–31552. https://doi.org/10.1074/jbc.M605886200

76. Mattevi A, Valentini G, Rizzi M *et al.* Crystal structure of *Escherichia coli* pyruvate kinase type I: molecular basis of the allosteric transition. *Structure* 1995;**3**:729–741. https://doi.org/10.1016/S0969-2126(01)00207-6

77. Zhang D, Zhang F and Liu W. A KAS-III Heterodimer in Lipstatin Biosynthesis Nondecarboxylatively Condenses C_8 and C_{14} Fatty Acyl-CoA Substrates by a Variable Mechanism during the Establishment of a C_{22} Aliphatic Skeleton. *J Am Chem Soc* 2019;**141**:3993–4001. https://doi.org/10.1021/jacs.8b12843

78. Randall G, Grant-Mackie E, Chunkath S *et al.* A Stable Dehydratase Complex Catalyzes the Formation of Dehydrated Amino Acids in a Class V Lanthipeptide. *ACS Chem Biol* 2024;**19**:2548–2556. https://doi.org/10.1021/acschembio.4c00637

79. Ren M, Li Z, Wang Z *et al.* Antiviral Chlorinated Drimane Meroterpenoids from the Fungus *Talaromyces pinophilus* LD-7 and Their Biosynthetic Pathway. *J Nat Prod* 2024;**87**:2034–2044. https://doi.org/10.1021/acs.jnatprod.4c00539

80. Hotta K, Chen X, Paton R *et al.* Enzymatic catalysis of anti-Baldwin ring closure in polyether biosynthesis. *Nature* 2012;**483**:355–358. https://doi.org/10.1038/nature10865

81. Minami A, Oguri H, Watanabe K *et al.* Biosynthetic machinery of ionophore polyether lasalocid: enzymatic construction of polyether skeleton. *Curr Opin Chem Biol* 2013;**17**:555–561. https://doi.org/10.1016/j.cbpa.2013.06.004

82. Migita A, Watanabe M, Hirose Y *et al.* Identification of a Gene Cluster of Polyether Antibiotic Lasalocid from *Streptomyces lasaliensis*. *Biosci Biotechnol Biochem* 2009;**73**:169–176. https://doi.org/10.1271/bbb.80631

83. Wong F, Hotta K, Chen X *et al.* Epoxide Hydrolase-Lasalocid A Structure Provides Mechanistic Insight into Polyether Natural Product Biosynthesis. *J Am Chem Soc* 2015;**137**:86–89. https://doi.org/10.1021/ja511374k