High-throughput prediction of protein–protein interactions uncovers hidden molecular networks 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, Kyoto, 606-8502, 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 proteins responsible for natural product biosynthesis. These proteins collectively produce various secondary metabolites with complex chemical structure, including antibiotics and mycotoxins, yet the complete biosynthetic pathways have been experimentally resolved for only a limited number of compounds. Protein–protein interactions within BGCs have recently been recognized as key determinants of intermediate transfer, enzymatic regulation, and structural stability. However, many BGCs still contain proteins of unknown function that cannot be predicted by conventional sequence-based bioinformatics tools, hindering a comprehensive understanding of their biosynthetic pathways. 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 interesting molecular mechanisms involving proteins previously annotated as "uncharacterized" or "potentially dysfunctional". Our analysis further showed that the ipSAE metric can distinguish correct heterocomplex pairs when multiple functionally homologous proteins are present within a BGC. Overall, our computational analysis revealed molecular interaction networks among proteins encoded by each BGC and identified enzyme complexes that are likely functional only when assembled. These predicted complexes may represent previously unrecognized links in their biosynthetic pathways. The complete results are available in a reusable format at https://doi.org/10.5281/zenodo.17451667 to support future experimental validation.

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

48 interactions—such as defense against predators[1], competition with other organisms[2], or symbiotic
49 relationships[3]. These compounds are typically produced by biosynthetic gene clusters (BGCs), which
50 are co-located sets of genes that collectively encode enzymes required to produce a specific
51 metabolite and auxiliary proteins such as transporter and regulatory proteins[4, 5]. Secondary metabolites
52 contain a wide variety of chemical structures and classes, including non-ribosomal peptides (NRPs),
53 polyketides (PKs), ribosomally synthesized and post-translationally modified peptides (RiPPs),
54 terpenes, and saccharides, which have attracted significant interest from both enzymology and organic
55 chemistry. Since the 2000s, progress in sequencing technologies, such as next-generation sequencing
56 and metagenomic analysis, along with the development of bioinformatics tools and the expansion of
57 genomic databases, has significantly advanced research on secondary metabolites derived from
58 uncultivable microorganisms. For example, antiSMASH[6, 7, 8, 9, 10, 11, 12, 13] is a widely used
59 bioinformatics tool for the identification, annotation, and analysis of BGCs in microbial genomes. It
60 also classifies clusters into known types by comparing them to curated reference datasets, including
61 the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database[14]. DeepBGC[15]
62 incorporates deep learning and natural language processing strategies to capture genomic positional
63 dependencies and order information that are not considered in hidden Markov model (HMM)-based
64 tools such as ClusterFinder[16] and antiSMASH. This improved the identification of BGCs in bacterial
65 genomes and contributed to the discovery of novel BGCs encoding natural products that could not be
66 detected by experimental methods.

67 Proteins can expand their functional capabilities by forming complexes with other proteins or
68 ligands. As in other types of gene clusters, the formation of protein complexes plays a crucial role in
69 BGCs. Notably, protein–protein interactions observed in carrier proteins[17], *trans*-type acyltransferase
70 (AT) polyketide synthase (PKS)[18], *cis*-type AT PKS[19], type I/II PKS, and fatty acid synthase[20, 21] have
71 been extensively studied and discussed. These interactions are key components of the biosynthetic
72 machinery. In homomultimer complexes, residues that are distant in the primary and tertiary
73 structures assemble at the multimer interface to form a single binding pocket[22]. Additionally, some
74 heteromeric complexes influence the stability, specificity, and catalytic activity[23, 24]. Because these

interactions underlie the remarkable structural and chemical diversity of the products, many enzymes in these systems have been structurally characterized to clarify their molecular basis to date. Nevertheless, most BGCs still harbor numerous proteins of unknown function, often annotated simply as "hypothetical protein" in MIBiG. Sequence-based bioinformatics tools generally cannot predict the function of these proteins unless homologous proteins with known functions have already been characterized. This limitation represents a major challenge for the comprehensive understanding of their biosynthetic pathways.

Highly accurate prediction of protein tertiary structures, as exemplified by AlphaFold2 (AF2)[25], has significantly advanced structural bioinformatics. Structures predicted with high confidence are often sufficiently accurate to be used as search models of molecular replacement method of for X-ray crystallography[26, 27], as well as initial structures for molecular dynamics (MD) simulations and quantum mechanics/molecular mechanics (QM/MM) calculations[28]. Furthermore, AlphaFold provides valuable predictions about protein–protein[29] interactions that are difficult to obtain from sequence-based bioinformatics tools alone. Notably, AlphaFold predicts inter-residue distances based on co-evolutionary signals concealed in the multiple sequence alignments (MSA) for the query sequence(s). When more than roughly 30–100 effective homologous sequences are available in sequence database, this strategy enables accurate prediction without relying on experimentally determined homologous structures or prior knowledge of protein–protein interactions. One application of protein complex prediction is the discovery and experimental validation of the protein Tmem81, which is conserved across vertebrates during fertilization of sperm and egg but whose function could not be inferred prior to the study[30].

Here, we report a comprehensive search of all pairwise complex predictions among proteins encoded within each BGC using AlphaFold3 (AF3)[31]. 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.

Results

High-throughput complex prediction pipeline

We developed a computational pipeline for the systematic prediction of protein–protein complexes encoded within BGCs (**Fig. 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.

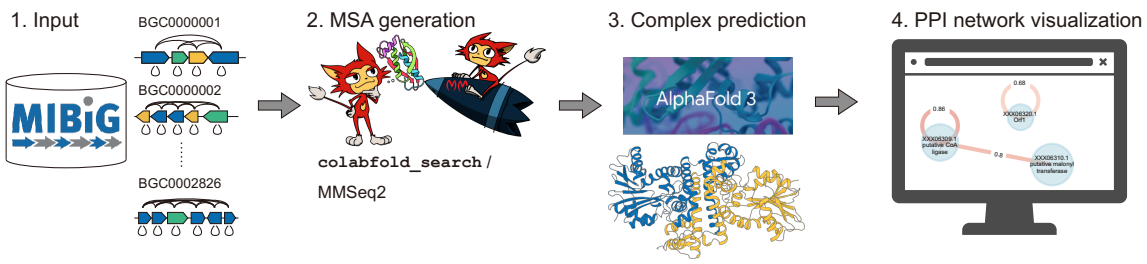


Fig. 1: computational workflow for protein complex prediction. 1. Input amino acid sequences. List all pairs of self or other protein sequences encoded in each BGC accession ID, excluding those with a length more than 1950 residues. 2. Generate paired and unpaired MSAs for all the protein pairs using 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.

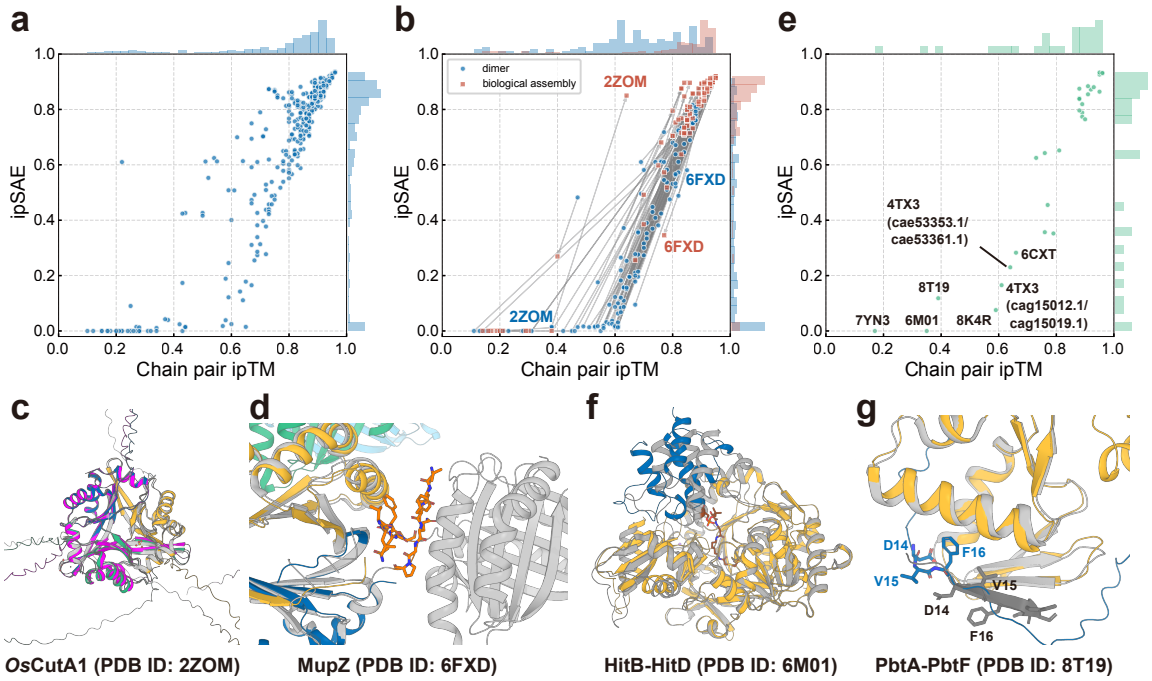
Highly accurate structure prediction using AF3 requires the generation of multiple sequence alignments (MSAs) for the input sequences. However, comprehensive prediction using the AF3 Web server is not feasible due to its usage limits (currently 30 runs per day). Moreover, the official AF3 inference code available on GitHub is also impractical for the high-throughput workflows, mainly because the MSA generation process using HMMER 3[32] is slow and cannot be efficiently parallelized. To address these issues, we used *colabfold_search* from LocalColabFold[33] in conjunction with MMseqs2[34] 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 was performed using the *msatojson* command, which is part of our *alphafold3_tools* toolkit. Using this pipeline, we generated MSAs in 1–2 minutes per complex on a server with 768 GB of RAM or more and efficiently distributed the structure inference tasks for a total of 487,828 protein pairs across multiple GPU-equipped computers.

Validation of the predicted complexes

We subsequently compared the predicted complexes with the experimentally determined protein complexes deposited in Protein Data Bank (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 multimerization interface and thereby lead to inaccurate measurements.

For the protein structures registered in 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%) (**Fig. 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 (**Fig. 2b**). A representative example is CutA1 from *Oryza sativa* (PDB ID: 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 (**Fig. 2c**). 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 Fig. 1**). Despite this, ipSAE remained high, underscoring its reliability as an evaluation metric in such cases.



168 **Fig. 2: Validation of the predicted complexes. a** Scatter plot of chain-level ipTM and ipSAE values for predicted
169 homodimers (n = 408). Marginal histograms for each axis are displayed above and to the right. **b** Differences in
170 ipTM and ipSAE between the predicted homodimer (blue) and the biological assembly (red). Only biological
171 assemblies forming homomeric complexes larger than dimers and computable on our GPUs are included (n =
172 135). **c** Superimposition of the predicted *Os*CutA1 trimer (GenBank ID: BAT10611.1; blue, yellow, and green)
173 and dimer (purple) with the crystal structure (PDB ID: 2ZOM). **d** Predicted MupZ homotetramer (colored) and
174 its crystal structure (white). The artificial N-terminal peptide (orange) is shown in stick representation. **e** Scatter
175 plot of ipTM and ipSAE values for predicted heterodimers. **f** Predicted HitB (GenBank ID: BAR73008.1)–HitD
176 (BAR73010.1) and its crystal structure (white and gray). The cross-linking probe in the crystal structure is shown
177 in brown stick representation. **g** Superimposition of the predicted PbtA (AGY49587.1; blue)–PbtF (AGY49592.1;
178 yellow) complex with the crystal structure (PDB ID: 8T19; white and gray). Residues of PbtA at the interface are
179 shown in stick representation.

180

181 Nevertheless, some of the predicted complexes did not exhibit high ipTM/ipSAE values even
182 when using the stoichiometries of biological assemblies. Closer inspection revealed that several of
183 these cases represented artifacts, either due to misassignments of stoichiometry inferred from the
184 crystal packing using PDBePISA[35], or the inclusion of artificial sequence tags that induced unnatural
185 oligomerization. The former example is MxaA in *Stigmatella aurantiaca* (GenBank ID:
186 AAK57184.1; PDB ID: 4U7W), which is experimentally validated as a monomer[36]. The latter
187 example is MupZ from *Pseudomonas fluorescens* (**Fig. 2d**; GenBank ID: AFD33556.1; PDB ID:
188 6FXD), which is registered as a homo-tetramer due to C-terminal His_6-tag–induced tetramerization,
189 although the native protein forms a dimer.[37] We summarized the proteins exhibiting low ipSAE and
190 ipTM values in **Supplementary Table 1**. These observations highlight the challenges of constructing
191 fully reliable datasets of multimeric assemblies from PDB. Nonetheless, our findings suggest that an
192 ipTM threshold of 0.6 is effective for screening proteins for the potential of homooligomerization.

193 Validation of the predicted heteromeric complexes

194 For the 30 heteromeric complexes, we observed similar ipTM and ipSAE correlations to those
195 in the homomeric complexes. Among these, 6M01[38], 7YN3[39], 8K4R[40], 8T19 (DOI:
196 10.2210/pdb8T19/pdb), and 4TX3[41] showed ipTM < 0.6 (**Fig. 2e**). We further investigated the cause
197 of these low ipTM values. First, in PDB entries of 7YN3, 8K4R (VinM-VinL), and 6M01 (**Fig. 2f**), a

198 chemical reagent to induce covalent cross-linking between the two proteins was used to obtain the
199 crystal structure of the heterodimeric complexes. Biologically, one of the proteins in these two PDB
200 entries is an acyl carrier protein (ACP), which is known to form a transient complex with the other
201 protein. Thus, it is speculated that the ipTM does not yield high values for such complexes.

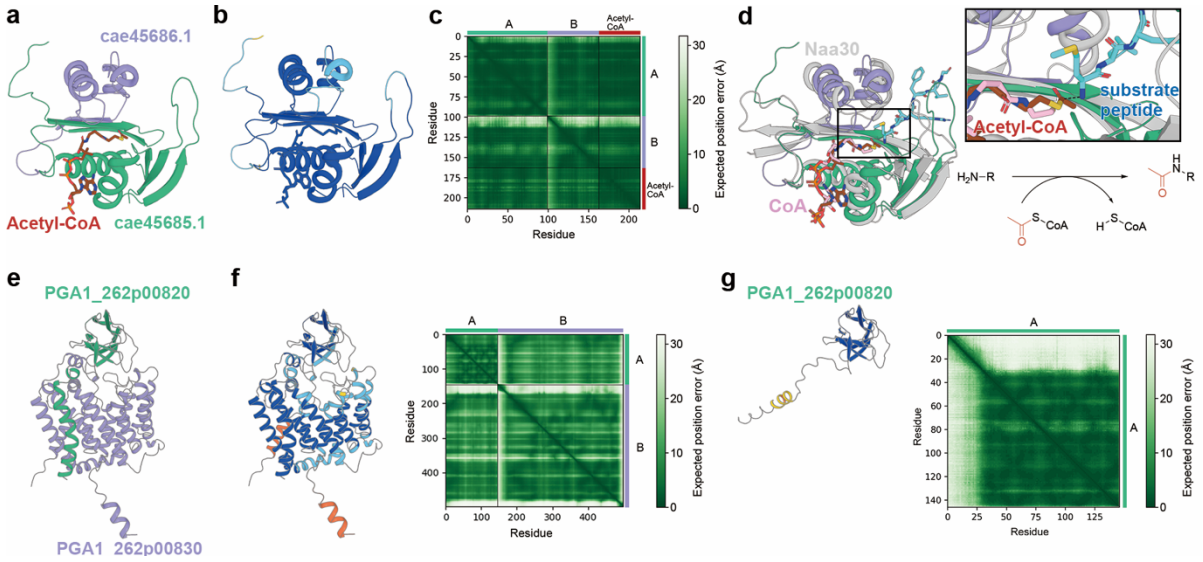
202 PDB ID 8T19 represents the complex structure of PbtA, a RiPP precursor peptide, and its
203 modifying enzyme PbtF from the GE2270 BGC (BGC0001155). The ipTM and ipSAE values were
204 0.39 and 0.1, respectively. Interestingly, AF3 predicted that the peptide binds to the β-strand region of
205 PbtF; however, the predicted binding position was shifted by two residues compared with the crystal
206 structure (**Fig. 2g**). We further investigated NisB from *Lactococcus lactis subsp. lactis* and StrB from
207 *Streptococcus thermophilus* LMD-9 (BGC0001209)[42] to gain insights into proteins that bind RiPP
208 precursor peptides. The crystal structure of NisB in complex with its precursor peptide NisA was not
209 included in the validation dataset because these proteins were fused within the same chain (PDB ID:
210 4WD9)[43] or the peptide is chemically modified (PDB ID: 6M7Y)[44]. StrB shares 94.98% sequence
211 identity with its homologous protein SuiB from *Streptococcus suis* (PDB ID: 5V1T)[45]. The predicted
212 complex structures for both matched well with their corresponding crystal structures, showing no
213 residue shift and high ipTM and ipSAE values (**Supplementary Fig. 2**). However, because of the
214 extremely limited number of available structures or experimental data on formation of RiPP precursor
215 peptide–enzyme complexes, a statistical evaluation for them remains difficult.

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

220 Novel heterocomplexes

221 Among 487,828 predicted protein pairs, 15,438 heterodimers exhibited ipTM ≥ 0.6, and 3,754
222 showed ipTM ≥ 0.8. These include complexes formed by previously uncharacterized proteins. Here,
223 we can hypothesize the functions of these proteins based on the predicted models. For example, a

224 heterodimer of cae45685.1 and cae45686.1 (**Fig. 3a–c**) found in borrelidin BGC from *Streptomyces*
225 *parvulus* (MIBiG accession ID: BGC0000031)[46, 47, 48] is annotated as "hypothetical protein" in
226 MIBiG, and to our knowledge, no studies have been reported for these proteins. BLASTp[49] search
227 predicted both proteins have partial sequence similarity with GCN5-related *N*-acetyltransferase
228 (GNAT). Structure-based searches against PDB100 database using the Foldseek webserver[50] indicated
229 that the predicted complex structure resembles the single subunit Naa30 of the heterotrimeric NatC
230 complex (PDB ID: 6YGD, chain A)[51], which catalyzes N-terminal acetylation (Nt-acetylation) of
231 numerous eukaryotic proteins. Superimposition of the complex model on the Naa30 structure revealed
232 that a binding pocket for the cofactor acetyl-CoA is formed between the cae45685.1–cae45686.1
233 interface, with the acetyl group of acetyl-CoA being exposed to the peptide-binding site of Naa30
234 (**Fig. 3d**). This suggests that this heterocomplex may also catalyze Nt-acetylation of its substrate.
235 Recently, *Streptomyces mutabilis sp*. MII, which is thought to possess a BGC similar to
236 BGC0000031, has been reported to produce *N*-acetylborrelidin B in addition to borrelidin
237 (**Supplementary Fig. 3a**)[52]. Given that borrelidin B, a known precursor of borrelidin, has an
238 aminomethyl group at C12 carbon, it is plausible that the heterocomplex binds borrelidin B at a
239 peptide-binding site analogous to that of Naa30 and catalyzes its acetylation to yield *N*-
240 acetylborrelidin B (**Supplementary Fig. 3b and 3c**).



241

242 **Fig. 3: Examples of predicted complex models. a** cae45685.1 (green)–cae45686.1 (purple)–Acetyl-CoA
243 (brown) trimeric complex model. **b** The predicted complex colored by pLDDT (orange, 0–50; yellow, 50–70;

244 cyan, 70–90; and blue, 90–100). **c** The predicted aligned error (PAE) matrix with chain coloring of Panel **a** on the
245 side bars. The dashed black lines indicate the chain boundaries. **d** Superimposition of the Naa30 crystal structure
246 (white) on the predicted model shown in Panel **a**. The N-terminal region of major capsid protein as a substrate
247 peptide (Uniprot ID: P32503) is shown as cyan stick. **e** PGA1_262p00820 (green)–PGA1_262p00830 (purple)
248 complex model found in BGC0000932. **f** pLDDT-colored predicted PGA1_262p00820–PGA1_262p00830
249 complex. The PAE matrix with chain coloring of Panel **f** is shown on right panel. **g** Predicted PGA1_262p00820
250 monomer structure colored by pLDDT. The right panel shows the PAE matrix. These data were obtained from the
251 AlphaFold Protein Structure Database[53] (UniProt ID: I7E6A3).

252

253 We also identified several Domain Unknown Function (DUF) proteins whose function appears
254 to arise only upon complex formation. For example, PGA1_262p00820 (Genbank ID: AFO93364.1)
255 in tropodithietic acid BGC from *Phaeobacter inhibens* DSM 17395 (MIBiG accession ID:
256 BGC0000932)[54] was annotated as a DUF4399 domain–containing protein, a family for which
257 functional characterization remains scarce. Our analysis predicted that this protein forms a complex
258 with PGA1_262p00830 (AFO93365.1), a DoxX-family protein, with an ipTM of 0.82 and an ipSAE
259 of 0.64 (**Fig. 3e**). Notably, the N-terminal region of PGA1_262p00820 was predicted to be disordered
260 as a monomer; however, in the heterodimer, this region adopted an α-helical conformation with high
261 pLDDT and formed an interaction interface with a helix of PGA1_262p00830 (**Fig. 3f and g**). Given
262 that DoxX-family proteins are localized to the cell membrane, PGA1_262p00820 is expected to be
263 embedded with PGA1_262p00830 in the cell membrane. Thus, the two proteins may act
264 cooperatively as membrane-bound enzymes or transporters.

265 Structurally homologous heterocomplexes

266 By examining the predicted heterodimers, we identified 952 pairs (4.9%) in which the
267 structural RMSD between constituent proteins was ≤1.0 Å, and 1390 pairs (7.1%) with RMSD ≤ 2.0
268 Å (**Fig. 4a**). The median ipTM and ipSAE values for the protein pairs with RMSD ≤2.0 Å were 0.82
269 and 0.71, respectively, indicating that these predictions were made with high confidence. This
270 observation suggests that structurally homologous heterodimers are present across BGCs and
271 contribute to the biosynthesis. A well-characterized and popular example is a ketosynthase (KS)–
272 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 ID: 1TQY)[55]. 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 (**Fig. 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.

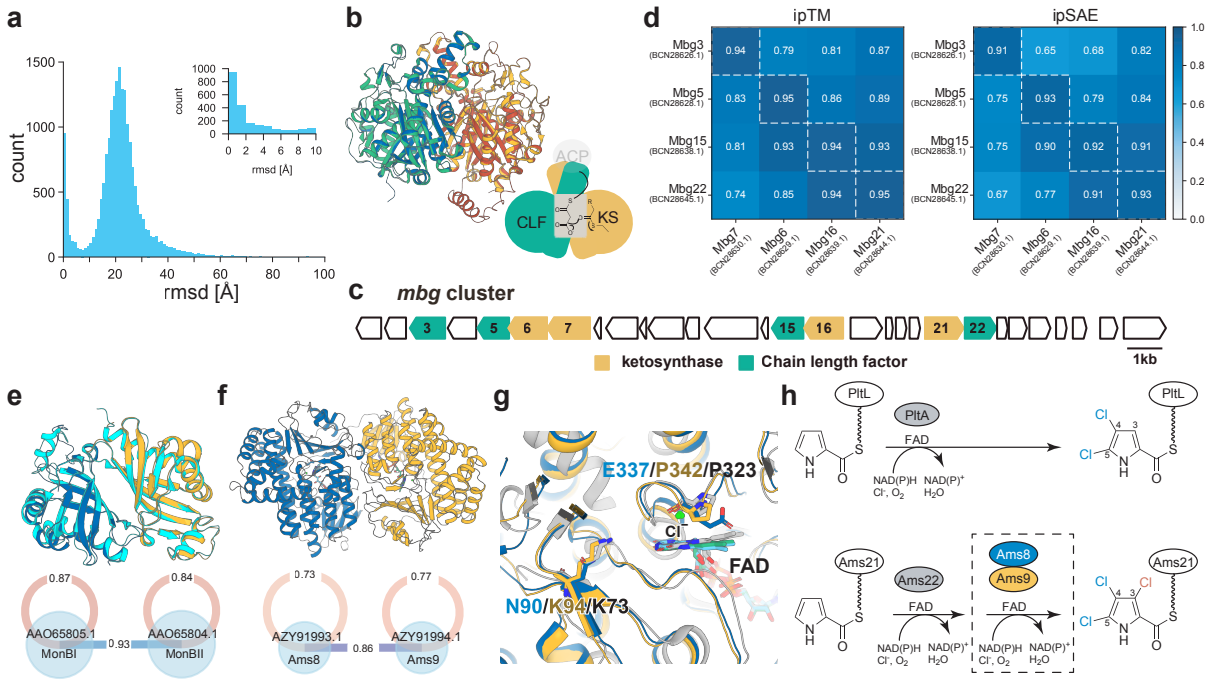


Fig. 4: Predicted structurally homologous complexes. a Histogram of structural RMSD values between monomeric units of predicted heterodimer pairs. An inset shows a magnified view of the 0–10 Å range. **b** Superimposition of the predicted complex structure of actinorhodin polyketide β-ketoacyl synthase alpha (KS) and beta (CLF) subunits (blue and red) on the crystal structure (PDB ID: 1TQY; green and yellow).

291 KS is shown in blue and green, and CLF in red and yellow. **c** *mbg* cluster. Only the genes encoding KS and
292 CLF are depicted. **d** Computed ipTM and ipSAE values for each complex prediction of the four KS
293 (horizontal) and four CLF (vertical) subunits in *mbg* cluster. **e** Superimposition of the predicted MonBI
294 (blue)–MonBII (yellow) heterodimer on the crystal structure of the MonBI homodimer (PDB ID: 3WMD;
295 cyan). The computed ipTM values for each homo- and hetero-dimer are shown below. **f** Predicted Ams8
296 (blue)–Ams9 (yellow) complex. **g** Close-up view of the FAD binding site in the predicted structures of
297 Ams8 (blue) and Ams9 (yellow). The pltA crystal structure (PDB ID: 5DBJ) is shown in gray. Residues
298 presumed to be important for chlorination activity are shown in stick representation. **h** Proposed
299 trichlorination mechanism in the armeniaspirol BGC (bottom). Formation of the dichloropyrrole moiety
300 catalyzed by PltA during pyoluteorin biosynthesis is shown on the top for comparison.

301

302 Subsequently, we examined whether ipTM and ipSAE could identify correct KS–CLF pairs
303 within a BGC cluster encoding multiple KS and CLF proteins. In our previous work, we
304 experimentally validated the correct pairing in *mbg* cluster (BGC0002441; **Fig. 4c**) as Mbg7–Mbg3,
305 Mbg6–Mbg5, Mbg16–Mbg15, and Mbg21–Mbg22 [56]. Our all-by-all complex predictions for the 4 × 4
306 combinations showed that both ipTM and ipSAE values reached their maximum values for the
307 experimentally validated pairs (**Fig. 4d**). Notably, ipSAE yielded lower values for incorrect pairings
308 than ipTM, indicating greater discriminatory power. A similar trend was observed in the Ysf cluster
309 (BGC0002547)[57], which encodes three KSs and three CLFs, where the highest values again
310 corresponded to the experimentally confirmed pairings (**Supplementary Fig. 4**). Collectively, these
311 results demonstrate that the ipTM and ipSAE metrics provide reliable indicators for identifying true
312 KS–CLF pairs and predicting their released polyketide intermediates, even when the corresponding
313 genes are not located adjacent within the BGC.

314 Our comprehensive predictions further identified 381 protein pairs that are likely to form both
315 homodimers and heterodimers, meeting the criteria of ipTM ≥ 0.6 and ipSAE ≥ 0.6. A representative
316 example is the pair MonBI and MonBII from the monensin BGC (BGC0000100). Although the
317 crystal structure of the MonBI homodimer has been determined in our previous work (PDB ID:

318 3WMD)[58], biological assays have shown that MonBI also forms a heterodimer with MonBII, and this
319 complex displays enzymatic activity. AF3 predictions indicated that MonBI and MonBII form
320 homodimers with ipTM values of 0.87 and 0.84, respectively, while they form a heterodimer with a
321 higher ipTM value of 0.93 (**Fig. 4e**). The predicted MonBI–MonBII heterodimer structure closely
322 matched the MonBI homodimer crystal structure, with an RMSD of 1.0 Å. This result suggests that
323 dimer formation between these two proteins is competitive, and that the heterodimer is preferentially
324 formed.

325 Proteins predicted to form both homo- and hetero-dimers were also observed among other
326 functional proteins. We highlight three flavin-dependent halogenase-like proteins—Ams8, Ams9, and
327 Ams22—from the armeniaspirol BGC (BGC0002022)[59] as an example. Within the BGC, Ams21 and
328 Ams22 share sequence similarities of 43% and 65% with PltL and PltA, respectively, from the
329 characterized pyoluteorin BGC (BGC0000127), suggesting that Ams21 functions as an ACP and
330 Ams22 as a flavin-dependent halogenase. Moreover, the structure of the active site of Ams22 is
331 highly similar to that of PltA (PDB ID: 5DBJ)[60], which catalyzes the 4,5-dichlorination of the pyrrolyl
332 moiety covalently attached to the phosphopantetheinyl arm of PltL, strongly suggesting that Ams22
333 catalyzes the same reaction (**Supplementary Fig. 5**). However, unlike in the pyoluteorin BGC, the
334 armeniaspirol BGC was reported to produce a 3,4,5-trichlorinated pyrollyl intermediate during the
335 biosynthetic process[59]. Thus, the remaining halogenase-like enzymes, Ams8 and Ams9, are
336 hypothesized to catalyze chlorination at the 3-position of the pyrrolyl moiety, although the mechanism
337 remains unclear[59].

338 Our structure prediction pipeline showed that Ams8 and Ams9 may form homodimers with
339 ipTM values of 0.77 and 0.73, respectively, whereas their heterodimer displayed a higher ipTM value
340 of 0.86 (**Fig. 4f**). Intriguingly, the predicted structure of Ams8 revealed that E337 is positioned at a
341 site that would inhibit the binding of flavin adenine dinucleotide (FAD), an essential cofactor for
342 halogenation. In addition, the conserved lysine residue required to mediate the reaction between the
343 substrate and hypochlorous acid[61, 62, 63] generated from FAD, molecular oxygen, and chloride ions—is
344 replaced by N90 in Ams8. (**Fig. 4g**). These unusual residue substitutions strongly suggest that Ams8

345 lacks halogenase activity. By contrast, Ams9 and Ams22 retain these conserved residues. However,
346 gene deletion of *ams8* completely abolished the biosynthetic process of armeniaspirol, whereas
347 deletion of either *ams9* or *ams22* reduced its production by 100-fold but did not lead to complete
348 loss[59]. This indicates that Ams8 is more essential for the biosynthetic pathway. Overall, these findings
349 suggest that Ams9 functions as the enzyme catalyzing trichlorination of 4,5-dichloropyrrolyl–*S*-
350 Ams22, while Ams8 acts as an auxiliary protein that supports the activity of Ams9 through complex
351 formation (**Fig. 4h**).

352 Visualization of the protein–protein interaction network map

353 To facilitate the use of our computational results by a broad range of researchers, we created a
354 web site that visualizes protein–protein network diagrams for the 2437 BGCs (http://vivace.bi.a.u-
355 tokyo.ac.jp/network_in_bgc/publish.html). On this site, users can filter the data by MIBiG accession
356 ID, taxonomy, class, or compound names, and view both the network diagrams and representative
357 chemical structures of the products (**Fig. 5a**). Proteins are represented as nodes, and pairs of proteins
358 predicted to form homodimers or heterodimers are connected by edges when they exceed a certain
359 threshold (**Fig. 5b**). The JSON files and Python scripts used to generate these network diagrams are
360 available in the supplementary files, allowing users to perform their own analyses.

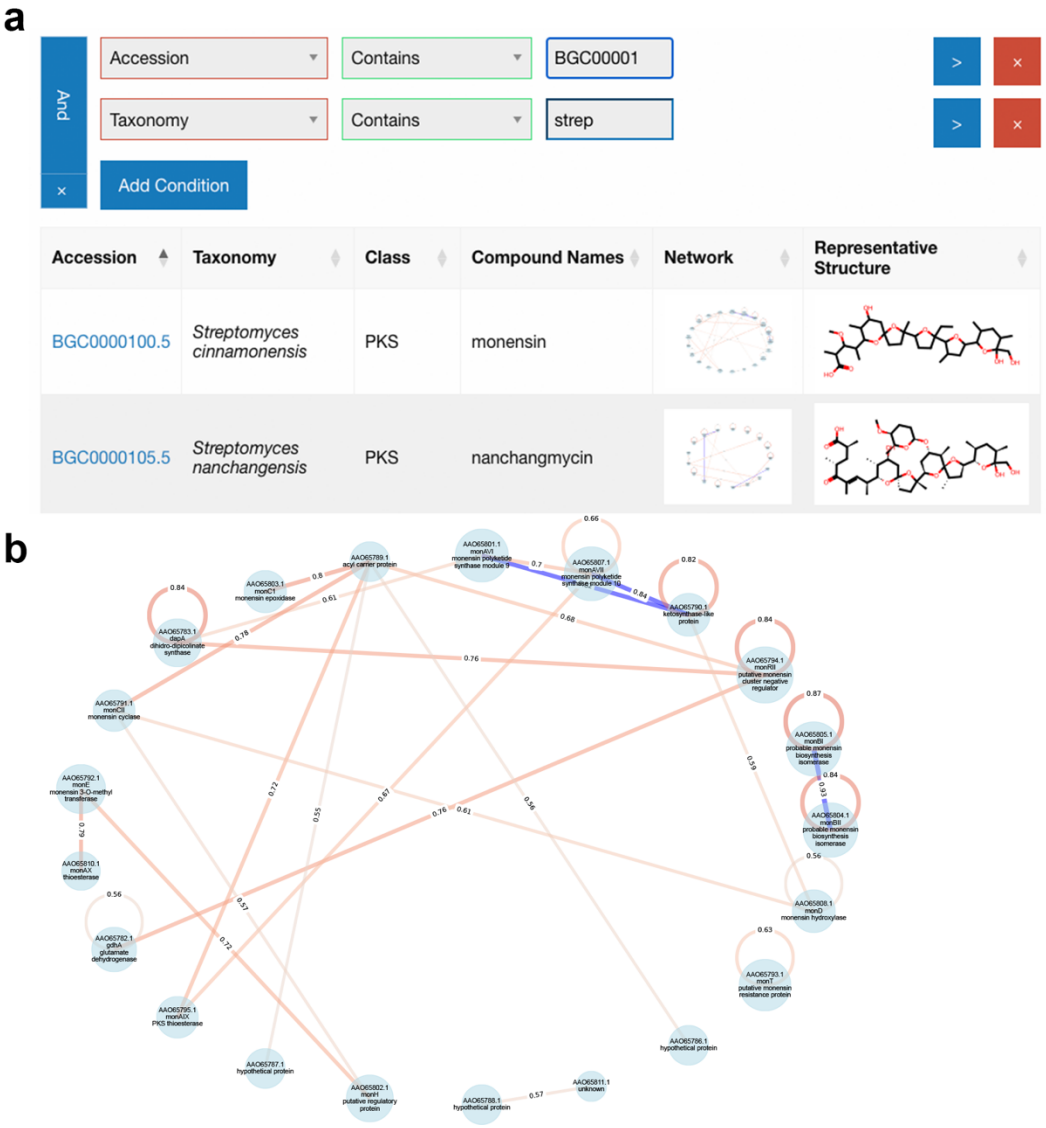


Fig. 5: Visualization of the protein–protein interaction (PPI) network map. a The interface on the website. **b** The PPI network map in the monensin BGC (BGC0000100). The nodes represent proteins encoded within the BGC. Edges indicate predicted interactions, with labels showing the corresponding ipTM 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).

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

372 showed that BGCs contain a considerable proportion of protein homo- or hetero- complexes whose
373 function emerges only upon complex formation, as well as structurally homologous heterocomplexes.

374 Although our pipeline and network maps provide valuable insights into the interactome in
375 BGCs, they have three limitations. First, AF2 (including AlphaFold-Multimer) and AF3 often fail to
376 accurately predict weak or transient protein complexes. This limitation arises because their algorithms
377 infer three-dimensional protein structures primarily from coevolutionary signals in the MSA derived
378 from individual amino acid sequence(s), and inter-chain interactions were not considered during their
379 training processes[31]. Carrier proteins, such as ACPs in PKSs and peptidyl carrier proteins (PCPs) in
380 NRPSs, exemplify this limitation. These proteins tether biosynthetic intermediates via a thioester
381 bond to the thiol group of a phosphopantetheinyl arm attached to a conserved serine residue. They
382 then interact with multiple catalytic domains in a specific yet transient manner to promote stepwise
383 transformations of the substrate. Although identifying their partner enzymes is crucial for studying
384 biosynthetic pathways, improving their prediction accuracy remains an important technical challenge.
385 Similarly, predicting the binding position of precursor peptides in RiPPs is also difficult for the same
386 reason.

387 Second, our computational pipeline did not perform complex predictions for proteins larger
388 than 1,950 residues due to computational limitations. Such large proteins are typically found in type I
389 PKS or NRPS systems, where multiple catalytic domains are linked within a single polypeptide to
390 form the core scaffold of the product. While these modular proteins mainly assemble the product
391 through their intramolecular domain–domain interactions, some tailoring enzymes within the same
392 BGC can modify intermediates tethered to the modular proteins, thereby expanding the chemical
393 diversity of the final products.[64, 65] Prior prediction of these interactions could help identify novel
394 tailoring enzymes responsible for introducing unusual functional groups; however, such predictions
395 were not fully performed in this study.

396 Third, our method may fail to correctly predict the complex formation of proteins with
397 potentially conformational changes. Although talaromyolides *tlx* BGC from *Talaromyces*

398 *purpureogenus* has not been registered in the current MIBiG dataset yet, the TlxJ–TlxI heterodimer
399 crystal structure (PDB ID: 7VBQ)[66] is notable case to discuss the issue (**Supplementary Fig. 6a**).
400 While sequence analysis suggests that both TlxJ (GenBank ID: BDC03472.1) and TlxI (BDC03471.1)
401 were nonheme iron oxygenases and the RMSD between the two proteins was 1.6 Å, TlxI is suggested
402 to be enzymatically dysfunctional as oxidases due to the lack of an arginine residue for α-
403 ketoglutarate (α-KG) binding and serve as an auxiliary protein to assist in TlxJ expression enhancing
404 its catalytic activity[66]. The gel filtration analysis and the X-ray crystallography indicated that TlxI is
405 monomeric (PDB ID: 7VBR).[66] However, while TlxJ–TlxI was correctly predicted as a
406 heterocomplex with an ipTM of 0.91 (**Supplementary Fig. 6b**), TlxI was predicted to form a
407 homodimer with an ipTM of 0.85 (**Supplementary Fig. 6c**). Interestingly, the crystal structures
408 revealed that TlxI undergoes a large conformational change when forming a heterocomplex with TlxJ.
409 Yet the predicted TlxI structure was close to the conformation observed in the heterocomplex crystal
410 structure in both monomeric and homomeric forms (**Supplementary Fig. 6c and d**). This observation
411 is consistent with the caveat that AF3 does not adequately predict proteins that undergo
412 conformational changes[31]. By contrast, AF3 successfully predicted the homodimer of MonBI with an
413 ipTM of 0.87 and the heterodimer with MonBII with a higher ipTM of 0.93 (**Fig. 4e**). These results
414 suggest that MonBI does not undergo significant conformational changes upon dimerization. To our
415 knowledge, MSA subsampling with AF2 or ColabFold[67] is one of the best strategies for predicting
416 such conformational changes in advance. However, we did not apply this approach in this study
417 because of its high computational cost.

418 Although several technical challenges remain in accurately predicting protein complexes, our
419 comprehensive approach uncovered numerous plausible complexes hidden within BGCs. Notably, the
420 structurally homologous heterodimers identified in this study may serve as intriguing targets for future
421 research, particularly in the context of microbial evolution. According to the genome streamlining
422 theory, natural selection in prokaryotes tends to eliminate genetic redundancy and nonessential
423 sequences, leading to the evolution of highly compact and efficient genomes[68, 69]. Thus, the
424 structurally homologous complexes are unlikely to be mere products of gene duplication but are

425 instead expected to possess distinct functional roles such as regulation or assistance in protein
426 expression. Examples of the former function have been reported in both homo-[70, 71] and hetero-
427 complexes[66], and their characterization is relatively straightforward. In contrast, the latter function can
428 only be uncovered through co-expression of a specific protein combination, and its underlying
429 molecular mechanisms remain largely unexplored. Among the reported cases, MonBI–MonBII[58] and
430 LstA–LstB[72] become functional only when co-expressed. Co-expression of LxmK and LxmY has
431 been shown not only to improve stability but also to increase yield.[73] Interestingly, TalA from
432 *Talaromyces pinophilus* LD-7[74] and Lsd19 from *Streptomyces lasaliensis*[75, 76, 77, 78] have been
433 identified as naturally occurring chimeric enzymes, in which two structurally homologous protein
434 domains are fused into a single polypeptide chain. Such protein (or domain) pairs may exist to expand
435 the range of catalytic functions that can be achieved by a single enzyme alone.

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

446 Methods

447 Structure Prediction Workflow

448 Input amino acid sequences

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

458 MSA generation and structure prediction

459 Despite recent advances, protein structure prediction with AF3 still relies on sufficient number
460 of multiple sequence alignments (MSAs) for the query sequences. To achieve both high accuracy and
461 high-throughput prediction, we replaced the default MSA generation pipeline using HMMER 3[32]
462 implemented in AF3 with a faster alternative: the *colabfold_search* command in LocalColabFold[33].
463 MMseqs2 version 15-6f452[34] was used to build two sequence databases, Uniref30_2302 and
464 colabfold_envdb_202108 (released on July 31, 2023), which can be downloaded from
465 https://colabfold.mmseqs.com/ and generate MSAs for the query sequences. The obtained MSAs have
466 been shown to be sufficient for producing highly accurate predicted models with AF2 or ColabFold in
467 most cases.[33] In addition to unpaired MSAs, we also obtained paired MSAs to further improve the
468 accuracy of complex modeling. The protein template search against PDB100 database was not
469 performed, as it is known to improve monomeric but not complex prediction accuracy. The MSA
470 generation was performed on the SuperComputer Flow Type III subsystem at Nagoya University.
471 This system comprises 16 Intel Xeon Platinum 8280M processors (2.7 GHz, 28 cores each) and 24
472 TiB of RAM per node. Each query took approximately 1–2 minutes on the calculation node. The
473 obtained a3m-formatted MSA files were added to the "pairedMsa" and "unpairedMsa" keys of the
474 AF3 input JSON file using the *msatojson* command of the alphafold3_tools software that we
475 developed and made available on our GitHub repository: https://github.com/cddlab/alphafold3_tools/.

476 The "modelSeeds" value in the AF3 input JSON file was set to 1 and five models were generated for
477 each prediction. For dimer with a total number of residues below 1500, the complex structure
478 prediction was performed using the SuperComputer Flow Type II subsystem with NVIDIA GPU
479 V100; for those with 1500 to 2000 residues, our in-house NVIDIA GPU RTX4090; and for those with
480 2000–3900 residues, NVIDIA GPU H100 (VRAM 94GB). The Predicted Aligned Error (PAE) of the
481 predicted structures was visualized using the *paeplot* command of alphafold3_tools.

482 Validation of the predicted complex structures

483 Metrics

484 To identify biologically plausible protein–protein interactions in the AF3-predicted
485 complexes, we employed two metrics: the interface predicted TM-score (ipTM) and the interaction
486 prediction score from aligned errors (ipSAE)[79]. The ipTM measures the accuracy of the predicted
487 relative arrangement of subunits. Values ≥ 0.8 indicate confident and high-quality predictions, < 0.6
488 suggest failure, and 0.6–0.8 represent an uncertain range.[29, 31] The ipTM values are provided in the
489 AF3 output JSON files. The ipSAE is a modified scoring function that evaluates the quality of
490 predicted interactions by focusing on residue pairs likely to participate in the interface, rather than the
491 entire sequence. This metric can be computed by applying the ipSAE script[79] to the AF3 output JSON
492 file with both the PAE cutoff (pae_cutoff) and Cα-Cα distance cutoff (dist_cutoff) set to 10. We did
493 not use other metrics such as pDock[80], pDockQ2[81], or LIS[82] in this analysis, although they serve
494 similar purposes in evaluating protein–protein interactions.

495 Validation datasets and complex analysis

496 We evaluated both ipTM and ipSAE metrics for the predicted homo- and heterodimeric
497 structures by validating them against experimentally determined complexes. For proteins encoded in
498 the BGCs listed in MIBiG version 4.0 and also registered in the Protein Data Bank (PDB), we
499 retrieved their biologically relevant complex structures (also referred as "biological assemblies") and
500 stoichiometry information were retrieved from Protein Data Bank (PDB) as of November 30, 2024.
501 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) [83], 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.

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. The data used for the plots in **Fig. 2** are provided as a Supplementary Excel file.

Code availability

The python scripts used for the preprocessing, evaluation, and visualization are publicly available at https://github.com/cddlab/bgccomplexbuilder. The *msatojson*, *paeplot*, and other utility

525 commands to process the input and output files of AlphaFold3 can be downloaded as
526 *alphafold3_tools* (https://github.com/cddlab/alphafold3_tools).

527 # References

528 1. Zhang, S., Mukherji, R., Chowdhury, S., Reimer, L. & Stallforth, P. Lipopeptide-mediated
529 bacterial interaction enables cooperative predator defense. *Proc. Natl. Acad. Sci. USA* **118**,
530 e2013759118 (2021).
531 2. Chevrette, M. et al. Microbiome composition modulates secondary metabolism in a
532 multispecies bacterial community. *Proc. Natl. Acad. Sci. USA* **119**, e2212930119 (2022).
533 3. Modolon, F., Barno, A., Villela, H. & Peixoto, R. Ecological and biotechnological importance
534 of secondary metabolites produced by coral-associated bacteria. *J. Appl. Microbiol.* **129**,
535 1441–1457 (2020).
536 4. Malpartida, F. & Hopwood, D. A. Physical and genetic characterisation of the gene cluster for
537 the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2). *Mol. Genet. Genomics* **205**, 66–
538 73 (1986).
539 5. Stanzak, R., Matsushima, P., Baltz, R. & Rao, R. Cloning and Expression in *Streptomyces*
540 *lividans* of Clustered Erythromycin Biosynthesis Genes from *Streptomyces erythreus*.
541 *Bio/Technology* **4**, 229–232 (1986).
542 6. Medema, M. et al. antiSMASH: rapid identification, annotation and analysis of secondary
543 metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic*
544 *Acids Res.* **39**, W339–W346 (2011).
545 7. Blin, K. et al. antiSMASH 2.0—a versatile platform for genome mining of secondary
546 metabolite producers. *Nucleic Acids Res.* **41**, W204–W212 (2013).
547 8. Weber, T. et al. antiSMASH 3.0-a comprehensive resource for the genome mining of
548 biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237–W243 (2015).
549 9. Blin, K. et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster
550 boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).

10. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

11. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).

12. Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).

13. Blin, K. et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res.* **53**, W32–W38 (2025).

14. Zdouc, M. et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res.* **53**, D678–D690 (2025).

15. Hannigan, G. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, (2019).

16. Cimermancic, P. et al. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **158**, 412–421 (2014).

17. Gulick, A. & Aldrich, C. Trapping interactions between catalytic domains and carrier proteins of modular biosynthetic enzymes with chemical probes. *Nat. Prod. Rep.* **35**, 1156–1184 (2018).

18. Kosol, S., Jenner, M., Lewandowski, J. & Challis, G. Protein–protein interactions in *trans*-AT polyketide synthases. *Nat. Prod. Rep.* **35**, 1097–1109 (2018).

19. Dodge, G., Maloney, F. & Smith, J. Protein–protein interactions in "*cis*-AT" polyketide synthases. *Nat. Prod. Rep.* **35**, 1082–1096 (2018).

20. Herbst, D., Townsend, C. & Maier, T. The architectures of iterative type I PKS and FAS. *Nat. Prod. Rep.* **35**, 1046–1069 (2018).

21. Chen, A., Re, R. & Burkart, M. Type II fatty acid and polyketide synthases: deciphering protein–protein and protein–substrate interactions. *Nat. Prod. Rep.* **35**, 1029–1045 (2018).

22. Mori, T. et al. Molecular basis for the diversification of lincosamide biosynthesis by pyridoxal phosphate-dependent enzymes. *Nat. Chem.* **17**, 256–264 (2025).

578 23. Sikandar, A. & Koehnke, J. The role of protein-protein interactions in the biosynthesis of
579 ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **36**,
580 1576–1588 (2019).
581 24. Rattray, D. & Foster, L. Dynamics of protein complex components. *Curr. Opin. Chem. Biol.*
582 **48**, 81–85 (2019).
583 25. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
584 583–589 (2021).
585 26. McCoy, A., Sammito, M. & Read, R. Implications of *AlphaFold2* for crystallographic
586 phasing by molecular replacement. *Acta Crystallographica Section D: Structural Biology* **78**,
587 1–13 (2022).
588 27. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. *Nat.*
589 *Struct. Mol. Biol.* **29**, 1056–1067 (2022).
590 28. Homma, M. et al. Insights into stereoselective ring formation in canonical strigolactone:
591 Discovery of a dirigent domain-containing enzyme catalyzing orobanchol synthesis. *Proc.*
592 *Natl. Acad. Sci. USA* **121**, e2313683121 (2024).
593 29. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*,
594 2021.2010.2004.463034 (2022).
595 30. Deneke, V. et al. A conserved fertilization complex bridges sperm and egg in vertebrates. *Cell*
596 **187**, 7066–7078.e7022 (2024).
597 31. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold
598 3. *Nature* **630**, 493–500 (2024).
599 32. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comp. Biol.* **7**, e1002195 (2011).
600 33. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–
601 682 (2022).
602 34. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
603 analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
604 35. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J.*
605 *Mol. Biol.* **372**, 774–797 (2007).

606 36. Barajas, J. et al. Comprehensive Structural and Biochemical Analysis of the Terminal
607 Myxalamid Reductase Domain for the Engineered Production of Primary Alcohols. *Chem.*
608 *Biol.* **22**, 1018–1029 (2015).
609 37. Wang, L. et al. A Rieske oxygenase/epoxide hydrolase-catalysed reaction cascade creates
610 oxygen heterocycles in mupirocin biosynthesis. *Nat. Catal.* **1**, 968–976 (2018).
611 38. Miyanaga, A., Kurihara, S., Chisuga, T., Kudo, F. & Eguchi, T. Structural Characterization of
612 Complex of Adenylation Domain and Carrier Protein by Using Pantetheine Cross-Linking
613 Probe. *ACS Chem. Biol.* **15**, 1808–1812 (2020).
614 39. Mori, T. et al. Molecular basis for carrier protein-dependent amide bond formation in the
615 biosynthesis of lincosamide antibiotics. *Nat. Catal.* **6**, 531–542 (2023).
616 40. Miyanaga, A. et al. Structural Basis of Amide-Forming Adenylation Enzyme VinM in
617 Vicenistatin Biosynthesis. *ACS Chem. Biol.* **18**, 2343–2348 (2023).
618 41. Haslinger, K., Peschke, M., Brieke, C., Maximowitsch, E. & Cryle, M. X-domain of peptide
619 synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature* **521**, 105–109
620 (2015).
621 42. Schramma, K. R., Bushin, L. B. & Seyedsayamdost, M. R. Structure and biosynthesis of a
622 macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat. Chem.*
623 **7**, 431–437 (2015).
624 43. Ortega, M. A. et al. Structure and mechanism of the tRNA-dependent lantibiotic dehydratase
625 NisB. *Nature* **517**, 509–512 (2015).
626 44. Bothwell, I. R. et al. Characterization of glutamyl-tRNA-dependent dehydratases using
627 nonreactive substrate mimics. *Proc. Natl. Acad. Sci. USA* **116**, 17245–17250 (2019).
628 45. Davis, K. M. et al. Structures of the peptide-modifying radical SAM enzyme SuiB elucidate
629 the basis of substrate recognition. *Proc. Natl. Acad. Sci. USA* **114**, 10420–10425 (2017).
630 46. Olano, C. et al. Evidence from engineered gene fusions for the repeated use of a module in a
631 modular polyketide synthase. *Chem. Commun.*, 2780–2782 (2003).
632 47. Olano, C. et al. Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces*
633 *parvulus* Tü4055: insights into nitrile formation. *Mol. Microbiol.* **52**, 1745–1756 (2004).

48. Olano, C. et al. Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: Cluster analysis and assignment of functions. *Chem. Biol.* **11**, 87–97 (2004).

49. Camacho, C. et al. BLAST+ : architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

50. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

51. Grunwald, S. et al. Divergent architecture of the heterotrimeric NatC complex explains N-terminal acetylation of cognate substrates. *Nat. Commun.* **11**, 5506 (2020).

52. Hamed, A. et al. *N*-Acetylborrelidin B: a new bioactive metabolite from *Streptomyces mutabilis* sp MII. *Z. Naturforsch., C: Biosci.* **73**, 49–57 (2018).

53. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

54. Brock, N., Nikolay, A. & Dickschat, J. Biosynthesis of the antibiotic tropodithietic acid by the marine bacterium *Phaeobacter inhibens*. *Chem. Commun.* **50**, 5487–5489 (2014).

55. Keatinge-Clay, A., Maltby, D., Medzihradszky, K., Khosla, C. & Stroud, R. An antibiotic factory caught in action. *Nat. Struct. Mol. Biol.* **11**, 888–893 (2004).

56. Zhang, J. et al. Reconstitution of a Highly Reducing Type II PKS System Reveals 6π-Electrocyclization Is Required for *o*-Dialkylbenzene Biosynthesis. *J. Am. Chem. Soc.* **143**, 2962–2969 (2021).

57. Deng, Z. et al. An Unusual Type II Polyketide Synthase System Involved in Cinnamoyl Lipid Biosynthesis. *Angew. Chem. Int. Ed.* **60**, 153–158 (2021).

58. Minami, A. et al. Allosteric Regulation of Epoxide Opening Cascades by a Pair of Epoxide Hydrolases in Monensin Biosynthesis. *ACS Chem. Biol.* **9**, 562–569 (2014).

59. Fu, C. et al. Armeniaspirol Antibiotic Biosynthesis: Chlorination and Oxidative Dechlorination Steps Affording Spiro[4.4]non-8-ene. *ChemBioChem* **20**, 764–769 (2019).

661 60. Pang, A., Garneau-Tsodikova, S. & Tsodikov, O. Crystal structure of halogenase PltA from
662 the pyoluteorin biosynthetic pathway. *J. Struct. Chem.* **192**, 349–357 (2015).
663 61. Barker, R. D., Yu, Y., De Maria, L., Johannissen, L. O. & Scrutton, N. S. Mechanism of
664 Action of Flavin-Dependent Halogenases. *ACS Catal.* **12**, 15352–15360 (2022).
665 62. Yeh, E., Blasiak, L. C., Koglin, A., Drennan, C. L. & Walsh, C. T. Chlorination by a long-
666 lived intermediate in the mechanism of flavin-dependent halogenases. *Biochemistry* **46**,
667 1284–1292 (2007).
668 63. Flecks, S. et al. New insights into the mechanism of enzymatic chlorination of tryptophan.
669 *Angew. Chem. Int. Ed.* **47**, 9533–9536 (2008).
670 64. Ames, B. et al. Crystal structure and biochemical studies of the trans-acting polyketide enoyl
671 reductase LovC from lovastatin biosynthesis. *Proc. Natl. Acad. Sci. USA* **109**, 11144–11149
672 (2012).
673 65. Matthews, M. et al. Substrate-Triggered Formation and Remarkable Stability of the C-H
674 Bond-Cleaving Chloroferryl Intermediate in the Aliphatic Halogenase, SyrB2. *Biochemistry*
675 **48**, 4331–4343 (2009).
676 66. Li, X. et al. Heterodimeric Non-heme Iron Enzymes in Fungal Meroterpenoid Biosynthesis. *J.*
677 *Am. Chem. Soc.* **143**, 21425–21432 (2021).
678 67. del Alamo, D., Sala, D., McHaourab, H. S. & Meiler, J. Sampling alternative conformational
679 states of transporters and receptors with AlphaFold2. *eLife* **11**, (2022).
680 68. Giovannoni, S. et al. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**,
681 1242–1245 (2005).
682 69. Giovannoni, S., Thrash, J. & Temperton, B. Implications of streamlining theory for microbial
683 ecology. *ISME J.* **8**, 1553–1565 (2014).
684 70. Kotaka, M., Ren, J., Lockyer, M., Hawkins, A. & Stammers, D. Structures of R- and T-state
685 *Escherichia coli* aspartokinase III: Mechanisms of the allosteric transition and inhibition by
686 lysine. *J. Biol. Chem.* **281**, 31544–31552 (2006).
687 71. Mattevi, A. et al. Crystal structure of *Escherichia coli* pyruvate kinase type I: molecular basis
688 of the allosteric transition. *Structure* **3**, 729–741 (1995).

689 72. Zhang, D., Zhang, F. & Liu, W. A KAS-III Heterodimer in Lipstatin Biosynthesis
690 Nondecarboxylatively Condenses C_8 and C_{14} Fatty Acyl-CoA Substrates by a Variable
691 Mechanism during the Establishment of a C_{22} Aliphatic Skeleton. *J. Am. Chem. Soc.* **141**,
692 3993–4001 (2019).
693 73. Randall, G. et al. A Stable Dehydratase Complex Catalyzes the Formation of Dehydrated
694 Amino Acids in a Class V Lanthipeptide. *ACS Chem. Biol.* **19**, 2548–2556 (2024).
695 74. Ren, M. et al. Antiviral Chlorinated Drimane Meroterpenoids from the Fungus *Talaromyces*
696 *pinophilus* LD-7 and Their Biosynthetic Pathway. *J. Nat. Prod.* **87**, 2034–2044 (2024).
697 75. Hotta, K. et al. Enzymatic catalysis of anti-Baldwin ring closure in polyether biosynthesis.
698 *Nature* **483**, 355–358 (2012).
699 76. Minami, A., Oguri, H., Watanabe, K. & Oikawa, H. Biosynthetic machinery of ionophore
700 polyether lasalocid: enzymatic construction of polyether skeleton. *Curr. Opin. Chem. Biol.*
701 **17**, 555–561 (2013).
702 77. Migita, A. et al. Identification of a Gene Cluster of Polyether Antibiotic Lasalocid from
703 *Streptomyces lasaliensis*. *Biosci. Biotechnol. Biochem.* **73**, 169–176 (2009).
704 78. Wong, F. et al. Epoxide Hydrolase-Lasalocid A Structure Provides Mechanistic Insight into
705 Polyether Natural Product Biosynthesis. *J. Am. Chem. Soc.* **137**, 86–89 (2015).
706 79. Dunbrack, R. L. *Rēs ipSAE loquunt*: What's wrong with AlphaFold's ipTM score and how to
707 fix it. *bioRxiv*, 2025.2002.2010.637595 (2025).
708 80. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions
709 using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
710 81. Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. Evaluation of AlphaFold-Multimer
711 prediction on multi-chain protein complexes. *Bioinformatics* **39**, btad424 (2023).
712 82. Kim, A.-R. et al. Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer.
713 *bioRxiv*, 2024.2002.2019.580970 (2024).
714 83. Wojdyr, M. GEMMI: A library for structural biology. *J. Open Source Softw.* **7**, 4200 (2022).

715 # Acknowledgements

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.

We thank Dr Naruki Yoshikawa for a fruitful discussion.

Author information

Contributions

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 [support-ing], 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]

Corresponding authors

Correspondence to Yoshitaka Moriwaki.

Ethics declarations

Competing interests

The authors declare no competing interests.