Deciphering Biosynthetic Gene Clusters with a Context-aware Protein Language Model

Zixin Kang[1,#], Haohong Zhang[1,#], Chaoqin Liang[2], Ronghua Yang[3], Ying Ye[2,*], Hong Bai[1,*], Yonghui Zhang[2,*] , Kang Ning[1,*]

[1]*MOE Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China*

[2]*Hubei Key Laboratory of Natural Medicinal Chemistry and Resource Evaluation, School of Pharmacy, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei, China*

[3]*Dovetree synbio Ltd., 215 Qingnian street, Shenyang, China*

[#] Equal contribution.

[*] Correspondence author.

E-mail: ying_ye@hust.edu.cn (Ye Y). baihong@hust.edu.cn (Bai H).

zhangyh@mails.tjmu.edu.cn (Zhang Y). ningkang@hust.edu.cn (Ning K).

Highlights

- CoreFinder is a context-aware deep learning framework leveraging protein language model to predict gene functions and associated metabolites classes within biosynthetic gene clusters (BGCs).

- CoreFinder deciphers BGCs without sequence alignment and manually defined rules. It achieved precision of 0.945 (842/891) and recall of 0.821 (842/1,025) for core gene annotation in over 700 experimentally validated fungal BGCs.

- Built on CoreFinder, we introduced an end-to-end deep learning-based workflow for BGC screening and deciphering, which is two orders of magnitude faster than current methods. Applied to 256 fungal genomes spanning 197 taxa, CoreFinder identified 6,414 core genes within 4,585 BGCs.

- CoreFinder uncovered an ancient NRPS gene family existed before the divergence of *Fusarium* and *Aspergillus*, underscoring evolutionary retention of biosynthetic tools in fungi for ecological adaptation.

Abstract

Microbial secondary metabolites, synthesized by biosynthetic gene clusters (BGCs), play critical roles in ecological interactions and offer vast potential for biotechnological and pharmaceutical applications. Despite advances in computational BGC detection, current methods face challenges, including time-consuming sequence alignments, dependence on known homologs and manually defined rules, limiting its robustness and generalizability. To address these, we present CoreFinder, a deep learning framework that integrates protein language models (pLMs) and genomic contexts to predict product class and decipher gene functions within BGCs without alignment. CoreFinder demonstrated higher precision of 0.945 (842/891) and recall of 0.821 (842/1,025) than antiSMASH for core gene annotation in over 700 experimentally validated fungal BGCs. Built on CoreFinder, we introduced an end-to-end scalable workflow for BGC screening and deciphering, which is about 240 times faster than antiSMASH. Applied to 256 genomes spanning 197 taxa, CoreFinder identified 6,414 core genes within 4,585 BGCs. Further analysis indicates that a non-ribosomal peptide synthetase (NRPS) family likely existed prior to the divergence of *Fusarium* and *Aspergillus* and evolved into function-specific gene clusters. These findings emphasize the potential of CoreFinder as a powerful tool for accelerating natural product discovery and driving innovation in synthetic biology by unlocking novel biosynthetic pathways for biotechnological and pharmaceutical advancements.

Introduction

In complex ecosystems, microbiome produce a variety of secondary metabolites, such as antibiotics, antifungals and siderophores, to meditate microbe-microbe and microbe -host interactions [1, 2]. These bioactive natural products are mainly synthesized by Biosynthetic Gene Clusters (BGCs), which are genomically co-localized genes [3]. While BGCs in microbiome share the common purpose of producing chemical diversity, their biosynthetic strategies diverge significantly. For instance, bacterial polyketide synthases (PKSs) typically employ modular assembly-line mechanisms, whereas fungal PKSs often use iterative catalytic processes, highlighting the distinct biosynthetic logic of fungi [4].

Over the past decades, many computational methods have been developed to unravel BGCs and predict product class [5-12]. Early approaches, like ClusterFinder, used Hidden Markov Model (HMM) to assess BGC likelihood via Pfam domains [11], while antiSMASH employed profile Hidden Markov Model (pHMM) [13] to identify BGCs across known secondary metabolite classes [10]. MetaBGC [12] adapted segmented pHMM algorithm to detect new BGCs for metagenomic reads. These tools have broadened BGC catalogs [14-16], yet their dependence on annotated sequences, high-quality multiple sequence alignments (MSAs), or detectable Pfam domains, combined with their highly time-consuming nature, restricts their accuracy and generalization, especially for fungal BGCs with divergent features. Previous study has also documented limitations in antiSMASH, which was reported to misclassify core genes encoding biosynthetic enzyme as other genes [17], contributing to elevated false-negative rates in BGC identification. Besides, advances in sequencing technologies and reconstruction of metagenomic data, have enabled the generation of vast numbers of metagenome-assembled genomes (MAGs), which are increasingly used to discover novel BGCs and analyze the biosynthetic diversity of microbial communities. However, conventional methods require several months to mine BGCs across thousands of MAGs, creating a bottleneck in large-scale studies. Deep learning approaches present a transformative opportunity to address these challenges by directly learning sequence-function relationships and identifying functional homologies between proteins without reliance on sequence alignment [18]. DeepBGC [9] leveraged a Bidirectional Long Short-term Memory (BiLSTM) RNN and a word embedding neural network to identify BGCs in bacterial genome. Our previous work, BGC-Prophet [19], introduced a transformer-based method for high-throughput screening, achieving rapid identification of BGCs. However, these methods remain limited in their ability to functionally annotate genes within identified BGCs, underscoring the need for integrative approaches for fast and accurate deciphering of BGCs.

Identifying the gene functions and positioning core genes represent a vital subtask in BGC identification. The core (biosynthetic-core) gene within a BGC typically

encodes a fundamental biosynthetic enzyme that is responsible for the initial metabolic step in microbial secondary metabolite synthesis [20]. Tailoring (biosynthetic-additional) genes, transport genes and regulatory genes within clusters collectively contribute to the modification, transportation and activation of BGCs [21]. Therefore, accurate classification of BGC gene components facilitates streamlined wet-lab experimental validation. Moreover, uncovering the functional roles of gene within BGCs is important for elucidating gene interactions and networks that drive complex microbial activities, and identifying core genes can pave the way for pharmaceutical biosynthesis and promising synthetic biology applications, such as biosynthetic pathway design. The current methods rely on sequence alignment and defined rules to annotate genes, which limits its ability to align remote homologs or unknown genes. Alignment-free approaches, which learn latent patterns among sequences could overcome these limitations, yet a robust alignment-free model for deciphering BGC remains undeveloped.

With the advent of large language models (LLMs), models trained on massive biological data have been widely applied across tasks such as sequence analysis, function annotation and structure prediction [22]. The protein language model (pLM) [18, 23-26] is trained to learn the patterns in protein sequences, establishing complex mapping relationships between sequence and function by capturing biological co-evolution information from millions of proteins. One notable example is ESM [27], a protein language model pretrained on 65 million proteins. This model has demonstrated the ability to encode the remote homology and the alignment within a protein family [28]. However, pLMs treat each protein as an isolated entity, overlooking the genomic contexts shaped by evolutionary processes, including regulation, transport and modification, which function cooperatively within BGCs. To learn contextual representations of genes and close the gap between genomic context and gene function, Hwang et al. constructed gLM, a genomic language model, by leveraging pLM embeddings as input [29]. This transformer-based language model captured biologically meaningful information from diverse metagenomic data in MGnify [30] during pre-training. Therefore, gLM can be effectively applied to BGC analysis, particularly those with limited homology.

Fungi stand out among microbial producers due to their remarkable biosynthetic richness and ecological significance. Their BGCs encode specialized enzymes and produce bioactive compounds, which drive nutrient cycling, sustain plant symbiosis, and offer pharmaceutical potential [31-33]. Despite this, the vast biosynthetic diversity of fungal BGCs remains underexplored, limiting potential application in biotechnology, agriculture, and medicine [3, 32, 34]. Beyond the experimentally validated BGCs cataloged in databases like MIBiG [35] and FunBGCs [36], many fungal BGCs lack effective functional characterization. This gap stems from the time-intensive and costly nature of wet-lab validation, compounded by the unique properties of fungal BGCs, such as remote homology and complex genomic contexts, all of which make them challenging to analyze [37].

To address these, we introduce CoreFinder, a context-aware deep learning framework aimed at predicting product class and deciphering gene functions within BGCs by leveraging protein language model and integrating genomic contexts [27, 38]. In benchmark evaluations against established methods, CoreFinder demonstrated higher accuracy and recall in overall product prediction and gene annotation. Across over 700 experimentally validated fungal BGCs, CoreFinder successfully identified 842 out of 1,025 core genes with higher precision of 0.945, outperforming antiSMASH. Built on CoreFinder, we proposed an end-to-end deep learning-based workflow for BGC screening and deciphering, which is two orders of magnitude faster than antiSMASH. Applied to 256 fungal genomes spanning 197 taxa, CoreFinder identified 6,414 core genes within 4,585 BGCs. Besides, wet-lab experiment validated a remote homolog (share high structure similarity but low sequence identity with known homologs) discovered in *Aspergillus oryzae*. Further analysis reveals that a non-ribosomal peptide synthetase (NRPS) core gene family might exist before the divergence of *Fusarium* and *Aspergillus*, subsequently evolving into function-specific biosynthetic gene clusters in these two species [39]. These findings highlight CoreFinder' s potential as a fast and sensitive tool for BGC deciphering, advancing our understanding of secondary metabolite production, and accelerating discovery of natural product.

Results

Overview of CoreFinder

We introduced CoreFinder, a transformer-based method that simultaneously predict BGC gene functions and product classes through attention-driven alignment of genomic features. Unlike current methods that require additional training phases or fine-tuning for two prediction tasks separately, CoreFinder integrates product prediction into its function classification framework. This bidirectional linkage between genes and biosynthetic outputs enables interpretable characterization of unannotated BGCs. The pipeline of CoreFinder involves four main steps (**Fig. 1, Methods**):

Data Preparation: The dataset for training comprised 26,520 genes from 2,502 experimentally validated BGCs (MIBiG Version 3.1). The MIBiG dataset is divided into training-set, validation-set and test-set according to 8:1:1. The external validation leveraged two experimentally validated, phylogenetically distinct (bacteria and fungi) BGC datasets to evaluate model' s accuracy and generalization capacity. Additionally, fungal genomes from FungiDB served as a dataset for detecting and deciphering novel BGCs (**Fig. 1a**).

Context-free Protein Embedding: Genes are encoded using context-free embeddings

generated by the ESM2 model with 650 million parameters. Product class embeddings were initialized with zero within the same latent space dimensionality as ESM2. Each BGC was represented as a structured sequence of vectors, starting with the product embedding followed by gene embeddings (**Fig. 1b**).

Context-aware Gene and Product Prediction: The classifier comprises a 5-layer transformer with parameters initialized from gLM to capture long-range genomic syntax (**Supplementary Fig. 1**). During training, the attention mechanism deciphered gene functional roles and aggregated information toward the initial product embedding position in each sequence. A single-layer neural network then performed dual predictions across 12 categories, including 7 product classes (NRPS, PKS, RiPPs, etc.) and 5 gene functional types (Biosynthetic, Transport, Regulatory, etc.). (**Fig. 1c**).

Experimental Validation: Novel BGCs discovered by CoreFinder were prioritized for experimental validation to verify their biosynthetic roles, demonstrating the model' s practical utility (**Fig. 1d**).

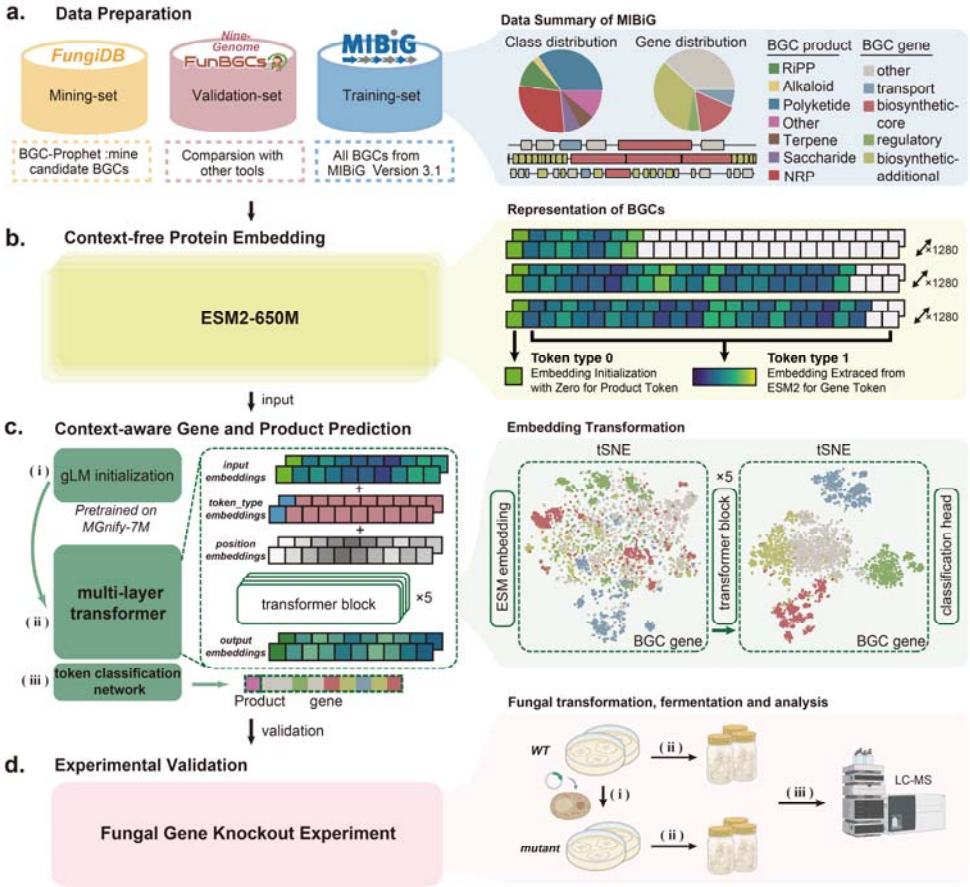


Figure 1. CoreFinder training strategy and workflow. a. Training set constructed from all samples in MIBiG; external manually curated fungal BGCs used as validation set for comparison with antiSMASH. BGC-Prophet was used to screen candidate BGCs in the genome, followed by CoreFinder for further analysis. The detailed picture on the right shows the data summary of MIBiG Version 3.1. **b.**

Embedding generated by ESM2-650M was employed for gene representation, while the class representation was initialized randomly. Gene vacancies were encoded as an all-zero vector and masked. **c.** (i) The transformer was initialized with gLM pretrained on MGnify-7M before training. (ii) Before input to the transformer blocks, the embeddings generated from the previous step was added with position embeddings and token-type embeddings. The transformer is a 5-layer encoder with ten attention heads each layer. During training, the model learns the map from genes to products using attention mechanism. The detailed picture on the right shows tSNE visualization of BGC gene embeddings before and after transformation. (iii) The single-layer classification network predicts product classes and gene functions from embeddings generated by the transformer. **d.** A wet-lab experiment was conducted to validate the novel BGC identified exclusively by CoreFinder. (i) fungal transformation; (ii) fungal fermentation; (iii) LC-MS analysis of products.

Model evaluation and comparison

CoreFinder demonstrated superior performance in both gene function classification (AUROC = 0.990, AUPR = 0.986) and product prediction (AUROC = 0.959, AUPR = 0.869) on the MIBiG test-set (see Methods). Notably, CoreFinder surpassed DeepBGC [9] and BGC-Prophet [19] in both AUC and recall for product prediction (see Methods) (**Fig. 2a, b, Supplementary Fig. 2**).

Unlike DeepBGC and BGC-Prophet, which focus on binary classification of BGCs, CoreFinder is designed for detailed gene function annotation and core gene prediction, complicating direct benchmarking with these tools. To enable evaluation on an external dataset, we redefined core (biosynthetic-core) gene as BGC and all other genes as non-BGC genes during CoreFinder inference, potentially reducing its F1 score. Nevertheless, in a benchmark using nine microbial genomes with manually annotated BGCs [11], CoreFinder outperformed DeepBGC, achieving higher accuracy and a markedly lower false positive rate, despite not being optimized for BGC detection (**Fig. 2c, Supplementary Fig. 3**). Besides, BGC-Prophet, our prior model for BGC detection in genomes, attained the highest accuracy and F1 scores in this task.

Leveraging CoreFinder's strength, we proposed an end-to-end scalable workflow ('BGC-Prophet+CoreFinder') combining BGC-Prophet for initial BGC screening with CoreFinder for detailed annotation and product prediction (**Table 1, Fig. 2d**). To evaluate the computational speed of the model, we compared it with antiSMASH, the most commonly used model. We mined the TPMC-MAG (Tibetan Plateau Microbial Catalog - Metagenome-assembled Genomes) resources [40] in our local server. Using a single NVIDIA RTX 3090 GPU, our method was time-expectable and 2 orders of magnitude faster than antiSMASH (**Fig. 2d)**.

To assess our protein encoding strategy and model architecture, we further trained two

Random Forest (RF) models, RF-Gene for gene function annotation task and RF-Production product prediction task, on the same MIBiG training-set (see Methods). We compared ESM2-650M embeddings with one-hot embeddings. ESM2 embeddings significantly outperformed one-hot encodings in both tasks (**Supplementary Fig. 4**), reflecting their richer biological and functional content. Notably, ESM2 embeddings exhibit a significant difference between core gene similarity and overall gene similarity (P = 2.96e-48), underscoring ESM2's capacity to discern biologically relevant distinctions in core genes, whereas one-hot embeddings showed no such difference (P = 0.51) (**Supplementary Fig. 5a, c**). While RF-Gene matched CoreFinder' s performance in gene function prediction, RF-Product's average AUROC for product classification was only 0.791, even with ESM2 embeddings. This disparity underscores the advantage of CoreFinder's transformer-based architecture over traditional machine learning models in leveraging genomic features for product inference.

To explore CoreFinder's attention pattern for interpretability, we analyzed attention scores during inference on the MIBiG inner test set. A Spearman correlation analysis identified several attention heads with strong positive correlations (P > 0.40) to core gene positions, indicating their role in capturing core gene information (**Fig. 2e**, marked by green dot box). In addition, we collected all attention scores from gene tokens to product tokens and categorized them by gene function to evaluate their contributions to product prediction (**Fig. 2f**). Core genes contribute the most, followed by tailoring (biosynthetic-additional) genes and transport genes. These findings reveal that CoreFinder autonomously aligns core gene embeddings with product embeddings by leveraging its multi-head attention mechanism, without prior biological knowledge. Interestingly, transport genes showed a notable influence on product prediction, which is consistent with established evidence of their roles in product characteristics, especially in siderophore-like BGCs [41]. This analysis underscores CoreFinder' s ability to elucidate functional roles of genes within BGCs, enhancing our understanding of biosynthetic processes.

Method	BGC Detection	Gene Annotation	Product Prediction	Computational Speed
antiSMASH	✓	✓	✓	Slow
DeepBGC	✓	✗	✓	Medium
BGC-Prophet	✓	✗	✓	Fast
CoreFinder	✗	✓	✓	Fast
BGC-Prophet+CoreFinder	✓	✓	✓	Fast

Table 1: Comparison of BGC Analysis Methods. This table summarizes the capabilities and computational efficiency of various methods designed for BGC analysis. 'BGC Detection' refers to the ability to identify BGCs within genomic sequences, 'Gene Annotation' denotes the provision of functional annotations for genes within BGCs, 'Product Prediction' indicates the capacity to predict the chemical product synthesized by the BGC, and 'Computational Speed' provides a qualitative

assessment of processing speed, categorized as Fast, Medium, or Slow. The entry 'BGC-Prophet+CoreFinder' represents a combined workflow where BGC-Prophet performs BGC detection, followed by CoreFinder for gene annotation and product prediction. (Fast:$10^{-1}\sim10^{1}$s per genome; Medium: $10^{1}\sim10^{2}$s per genome; Slow: $10^{2}\sim10^{3}$s per genome.)

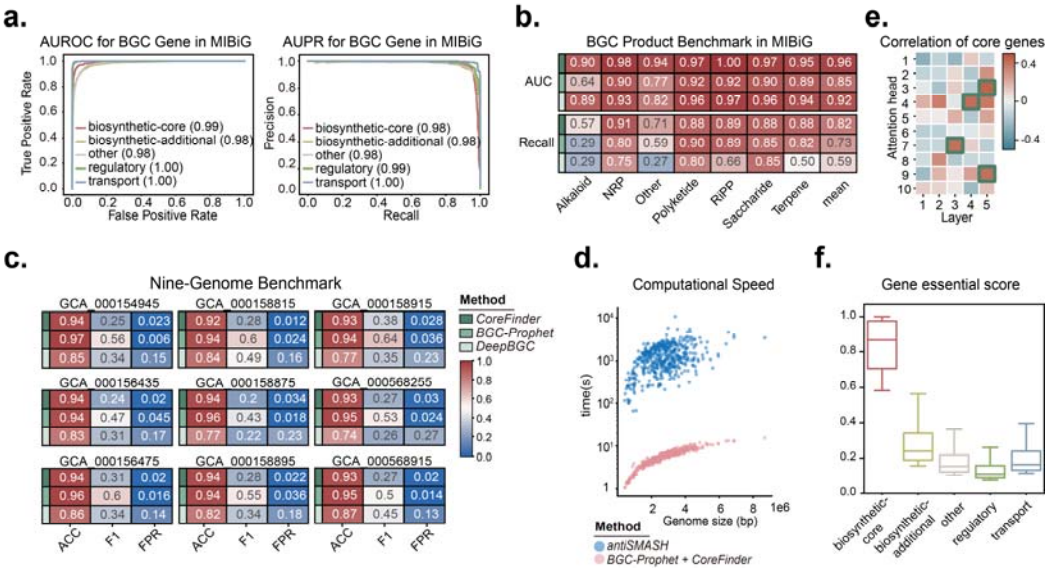


Figure 2. Evaluation of CoreFinder, comparison with other methods and interpretability analysis. a. Receiver operating characteristic curve (ROC) of CoreFinder in gene function classification on MIBiG test-set. Precision-recall curve (PR) of CoreFinder in gene function classification on MIBiG test-set. **b.** Evaluation metrics (AUC, Recall) comparing performance of CoreFinder, BGC-Prophet and DeepBGC in BGC product prediction on MIBiG test-set. **c.** Evaluation metrics comparing performance of CoreFinder, BGC-Prophet, and DeepBGC in BGC gene detection across an external nine-genome dataset that was used in previous studies. Higher accuracy and F_1-scores, along with lower false positive rate (FPR) values, indicate better performance. **d.** The computational speed of identifying and deciphering BGCs by different methods. We used antiSMASH and our workflow ('BGC-Prophet+CoreFinder') to unravel BGCs from TPMC-MAGs [40] in our local server to evaluate the computational efficiency of the two models. BGC-Prophet was used for BGC detection, and CoreFinder was used for deciphering. Our model is 2 to 3 orders of magnitude more computationally efficient than antiSMASH. **e.** Spearman correlation between core genes and attention scores for each attention head. Redder color corresponds to stronger correlation with core genes. Green dot box marks attention head with Spearman correlation higher than 0.40. **f.** Gene essential scores on MIBiG test-set, where higher scores indicate a greater gene contribution to product inference. The top 5% of gene-product scores from each gene category were collected and displayed.

Identification of core genes within experimentally validated BGCs

While most training data for BGC prediction models derive from bacterial BGCs, fungal BGCs present unique complexities due to their biosynthetic diversity, necessitating robust benchmarks to evaluate model generalization across organisms. Here, we compare CoreFinder and antiSMASH in the task of identifying gene roles, especially core genes in experimentally validated BGCs.

To rigorously evaluate CoreFinder' s ability to identify core genes, we used FunBGCs [17], a manually curated dataset of 742 experimentally validated fungal BGCs. This dataset provides a gold standard, with annotations rooted in wet-lab evidence rather than computational predictions alone (**Table S1**). There are 1,025 experimentally validated core genes within fungal BGCs. CoreFinder outperformed antiSMASH, achieving an accuracy of 0.956, precision of 0.945 (842/891), recall of 0.821 (842/1,025), and F1 score of 0.879 in detecting these core genes **(Fig. 3a, Fig. 3b)**. CoreFinder can recall more terpene-related core genes than antiSMASH, like geranylgeranyl pyrophosphate synthase and DMATS-type prenyltransferase (**Table S1**). CoreFinder correctly identified 842 out of 1,025 core genes, demonstrating its superior capability to capture a higher number of core genes while minimizing false positives compared to antiSMASH. Besides, many tailoring (biosynthetic-additional) genes misclassified by antiSMASH were annotated correctly by CoreFinder. For instance, cytochrome P450 monooxygenase contributing to a broad range of biological functions in versatile metabolism [42, 43], were frequently misclassified by antiSMASH [44], with 159 out of 733 labeled as 'other' or 'biosynthetic-core'. In contrast, CoreFinder accurately identified 731 out of 733 cytochrome P450 monooxygenases (**Table S1**). Notably, it failed to detect some core genes responding for non-typical class product synthase, like phosphoenolpyruvate mutase and isocyanide synthase. Besides, CoreFinder and antiSMASH are both unable to identify UbiA-type prenyltransferase and terpene cyclase due to the difference of gene annotations in MIBiG **(Fig. 3b)**. Overall, CoreFinder predicted 851 core genes, of which 787 were experimentally confirmed, exhibiting higher sensitivity and lower false positives in deciphering functional roles with BGCs.

To further validate CoreFinder's predictions against experimental ground truth, we examined cases where predictions where CoreFinder and antiSMASH prediction overlapped and diverged. In the overlapped set of 807 core genes, 767 were experimentally validated correct **(Fig. 3c)**, yielding a higher precision of 0.950 compared the predictions of individual models. The union of predictions (combining the overlapped and diverged sets) achieved a higher recall of 0.844 (865/1,025) **(Fig. 3a, Fig. 3b)**. In summary, CoreFinder could enhance precision and improve coverage of core genes in gene function prediction when applied with antiSMASH. CoreFinder's superior performance is further exemplified by its ability to detect remote homolog missed by antiSMASH. For instance, in FBGC00213, antiSMASH

misclassified a polyketide synthase as a tailoring (biosynthetic-additional) gene, while CoreFinder correctly identified it as a core gene, corroborated by FunBGCs experimental validation.

The core gene was quarried in NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi, accessed February 2025), and aligned with a homolog with low sequence similarity (SI: 39.35%) and high structure similarity (TM-score: 0.508) (**Fig. 3d, Supplementary Fig. 6**). This highlights CoreFinder' s robustness in pinpointing key biosynthetic enzymes within BGCs and capturing remote homologs sensitively. To conclude, CoreFinder's predictions align exceptionally well with experimental evidence, establishing it as a powerful tool for fungal core gene discovery with validated real-world utility.

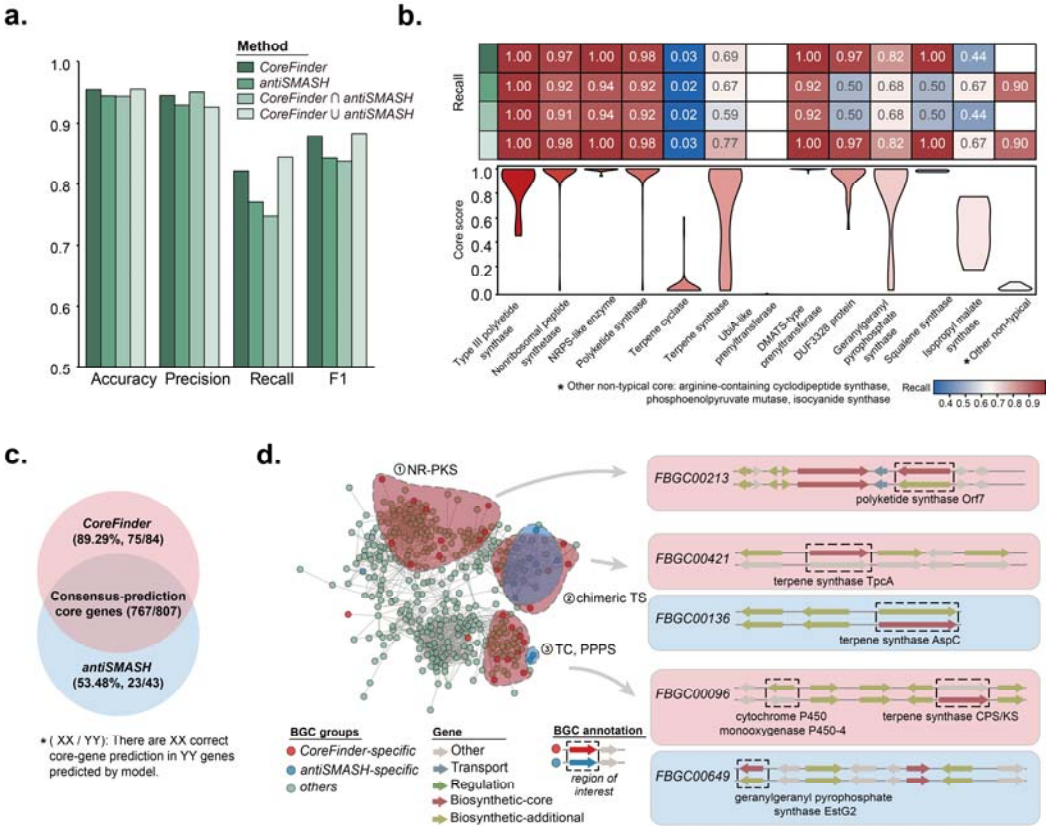


Figure 3. Comparative analysis of CoreFinder and antiSMASH in detecting core genes within fungal BGCs. The two models reached a consensus on 807 predicted core genes within BGCs, of which 767 were experimentally validated as correct. a. Performance metrics of four strategy (CoreFinder only, antiSMASH only, CoreFinder ∩ antiSMASH, CoreFinder ∪ antiSMASH) in detecting core genes within experimentally validated fungal BGCs. CoreFinder outperforms antiSMASH in accuracy, precision, recall, and F1 score, demonstrating its superior ability to identify a greater number of core genes within BGCs while minimizing false positives. When the results of the two models are intersected, the precision is the highest. When the results of the two models are combined, the recall rate and F1 will be higher, and

more core genes can be covered. **b.** Recall of antiSMASH and CoreFinder on different core genes and confidence score distribution of CoreFinder for different core genes. **c.** Venn diagram showing the overlap and accuracy of the core gene prediction results of CoreFinder and antiSMASH. Of the 84 core genes uniquely identified by CoreFinder, 75 were accurate. Of the 43 core genes uniquely identified by antiSMASH, 23 were accurate. In total, CoreFinder detected 842 out of 1,025 (Recall: 0.8215) experimentally validated core genes, while antiSMASH detected 790 (Recall: 0.7707) core genes. **d.** The main network constructed by BGC similarity ($dss \geq 0.35$) calculated by BiG-SCAPE. BGCs containing core genes only identified by CoreFinder (antiSMASH) were colored in red (blue), and others were colored in green. The region containing BGCs of interest was highlighted in shading with same color. CoreFinder could discovered some remote polyketide synthase, and both CoreFinder and antiSMASH would miss some terpene synthase.

Comprehensive BGC discovery and analysis across fungal genomes

To showcase CoreFinder' s ability to uncover novel insights into fungal BGCs, we analyzed 256 fungal genomes from 197 species across major phyla (e.g., *Ascomycota* and *Basidiomycota*) from FungiDB [45]. Leveraging the 'BGC-Prophet + CoreFinder' workflow, we identified 6,414 core genes within 4,585 BGCs **(Fig. 4a)**.

CoreFinder revealed a diverse array of BGCs, many of which were previously uncharacterized. Among the 6,414 core genes, 2,223 were annotated as 'hypothetical' or 'unspecified' by Gene Ontology annotation, underscoring the extent of unexplored biosynthetic potential in fungi (**Table S2**). BGCs associated with NRPS and PKS were particularly enriched in *Aspergillus* and *Fusarium*, two genera frequently linked to plant infections. Few BGC was identified in *Saccharomyces* and *Pichiomycetes*, consisting with existing research (**Fig. 4a**) [46, 47]. To validate CoreFinder' s predictions, we focused on *Aspergillus oryzae*, a well-studied model organism [48, 49], selecting a NRP BGC where CoreFinder identified one additional core gene, termed *Gene 363*, encoding a terpene synthase, which antiSMASH failed to detect (**Fig. 4b**). A homolog search found an isoprenoid synthase [50, 51] with high structure similarity and moderate sequence identity with *Gene 363*. This gene was uniquely identified by CoreFinder within the *AO-NSPID1 cluster 8.1*. Though novel BGCs may be particularly prevalent among uncultivated microorganisms, our findings suggest they may also be overlooked in well-characterized species [5].

For experimental validation, we knocked out *Gene 363* in *A. oryzae* and compared wild-type (WT) and mutant strains, confirming its functional role as a core gene. PCR analysis show that mutant strains were positive for 5F (~1,400 bp) and 3F (~1,650 bp) fragments (absent in WT) and negative for the RT (~850 bp) fragment (present in WT), verifying successful disruption of *Gene 363* in *AO-NSPID1 cluster 8.1* (**Supplementary Fig. 7, Supplementary Fig. 8**). Liquid chromatography-mass spectrometry (LC-MS) analysis of organic extracts from WT and mutant cultures

revealed WT-specific peaks at m/z 379.2486, indicating *Gene 363* contributes to the biosynthesis of terpene-related products. These differences were observed across different types of culture media (**Fig. 4c** for rice medium, **Supplementary Fig. 9** for CYP20 medium). The distinct metabolic profiles between WT and mutant strains confirm *Gene 363*'s biosynthetic role. Wet-lab validation, combined with protein alignment, underscores CoreFinder' s capability to detect terpene BGCs and identify remote homologs across fungal genomes.

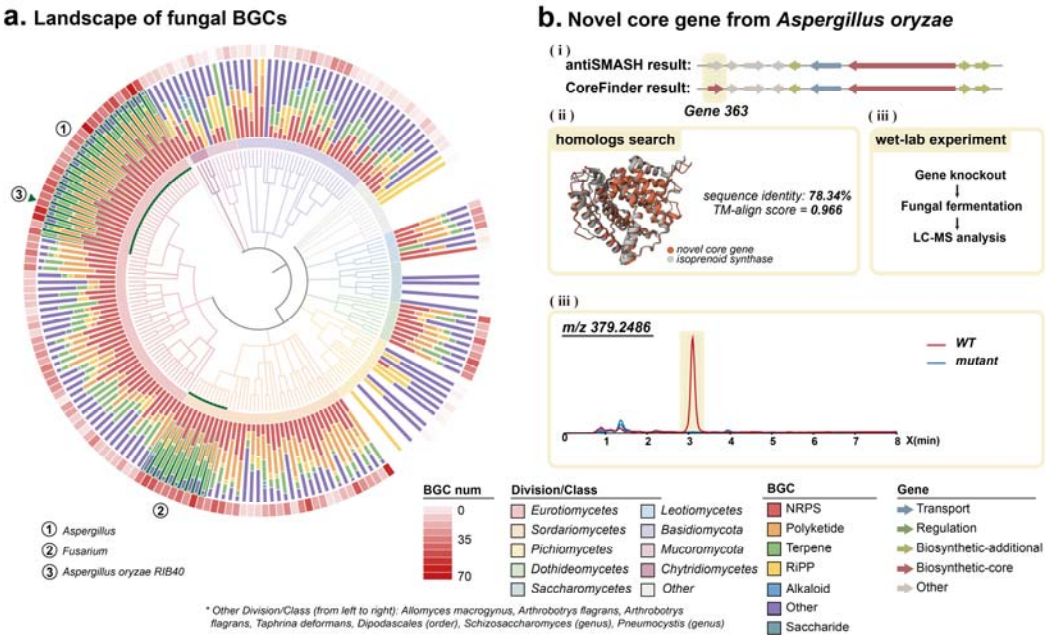


Figure 4. The global distribution of fungal BGCs and identification of a novel core gene. a. Phylogenomic tree constructed at the order level, encompassing 197 fungal species from the FungiDB database. Average numbers of biosynthetic gene clusters (BGCs) and core genes per genome are shown for each order. Inner rings depict BGC distribution across categories: NRP, polyketide, terpene, RiPP, saccharide, alkaloid, and others. The outer ring illustrates the distribution of BGC counts across the 197 species. **b.** (i) Comparison of CoreFinder and antiSMASH annotations for an NRP BGC identified from Aspergillus oryzae. (ii) CoreFinder identified an additional core gene, a terpene synthase, exhibiting high structural similarity (TM-score = 0.966) and moderate sequence identity (78.34%) to a known isoprenoid synthase. (iii) LC-MS analysis of metabolites extracted from rice cultures of WT and mutant A. oryzae strains lacking the novel core gene. Extracted ion chromatograms (EICs) at m/z 379.2486 reveal WT-specific peaks (highlighted in yellow), absent in the mutant, indicating products of the terpene synthase.

Phytotoxin BGCs in *Fusarium* Species

From seventeen *Fusarium* genomes, a diverse genus associated with plant and microbial interactions with significant implications for human health, animal welfare, and food security, we identified 625 biosynthetic gene clusters (BGCs) and 862 core

genes. Phytotoxin synthases, notably AM-toxin (AMT) and HC-toxin (HCT) synthases, were significantly enriched in *Fusarium*, making it an ideal model for studying BGCs linked to these compounds **(Supplementary Fig. 10)**. Here, we utilized CoreFinder to explore homologous core genes and reconstruct BGC families, focusing on AMT- and HCT-related clusters to demonstrate its analytical capabilities across fungal genomes.

To elucidate evolutionary relationships among phytotoxin BGCs, we calculated core genes' sequence identity [52, 53], comparing annotated AMT and HCT core genes to all 862 CoreFinder-identified core genes. This analysis revealed 115 core genes with high sequence identity to known phytotoxin synthases. Given the modular nature of NRPS genes and their capacity to generate diverse products [54-57], we mapped these core genes to their respective BGCs and clustered them into 59 distinct GCCs (gene cluster clans) by BiG-SCAPE [52], including three conserved AMT-related GCCs (GCC 540, GCC 1058, GCC 1703) and four conserved HCT-related GCCs (GCC 21, GCC 101, GCC 792, GCC 455), suggesting a variety of biosynthetic pathways. Phytotoxin-related GCCs were predominantly enrich in *Fusarium*, and two conserved HCT-related GCCs were found exclusive to *Coccidioides* (**Supplementary Fig. 10a, Fig. 5a**). Besides, GCC 1058 and GCC 1703, two conserved AMT-related GCCs in different *Fusarium* species, shared high GCF (gene cluster family) similarity and core gene structure similarity (**Supplementary Fig. 10b, Fig. 5a**). Furthermore, the analysis revealed that GCC 540, an AMT-related BGC family in *Fusarium*, exhibited substantial core gene similarity with GCC 12, a siderophore-related BGC family in *Aspergillus* (**Fig. 5a, b**).

To further investigate whether this similarity reflect horizontal gene transfer [58-60] or convergent evolution of NRPS BGCs in *Fusarium* and *Aspergillus*, we analyzed the evolutionary timeline and structural similarity of NRPS core genes in these two genera. The evolutionary timeline indicates that these two genera diverged approximately 352 million years ago during the Paleozoic, an Era marked by extensive forest coverage. Despite this ancient divergence, their NRPS core genes display remarkable structural similarity, a feature not observed in other co-evolving fungal species (**Supplementary Fig. 11**). We inferred that this NRPS gene family likely exists before their divergence, with both genera retaining it for synthesizing plant infection-related metabolites, while other species lost it over time may due to its reliance on other fungal relatives [47, 61]. Differences in genomic context [62-64] and substrate specificity [65], driven by subtle amino acid variations, resulted in the NRPS gene's divergent product profiles (**Supplementary Fig. 12**) [55, 66].

Our analysis of *Fusarium* genomes has unveiled a diverse spectrum of phytotoxin BGC and suggest the birth and evolution of an NRPS family within *Aspergillus* and *Fusarium* [46, 47, 67]. CoreFinder effectively identified homologous core genes, highlighting *Fusarium*'s role in plant-fungal pathogen interactions. This study enhances our understanding of fungal pathogenicity and offers insights for potential

agricultural interventions targeting these biosynthetic pathways.

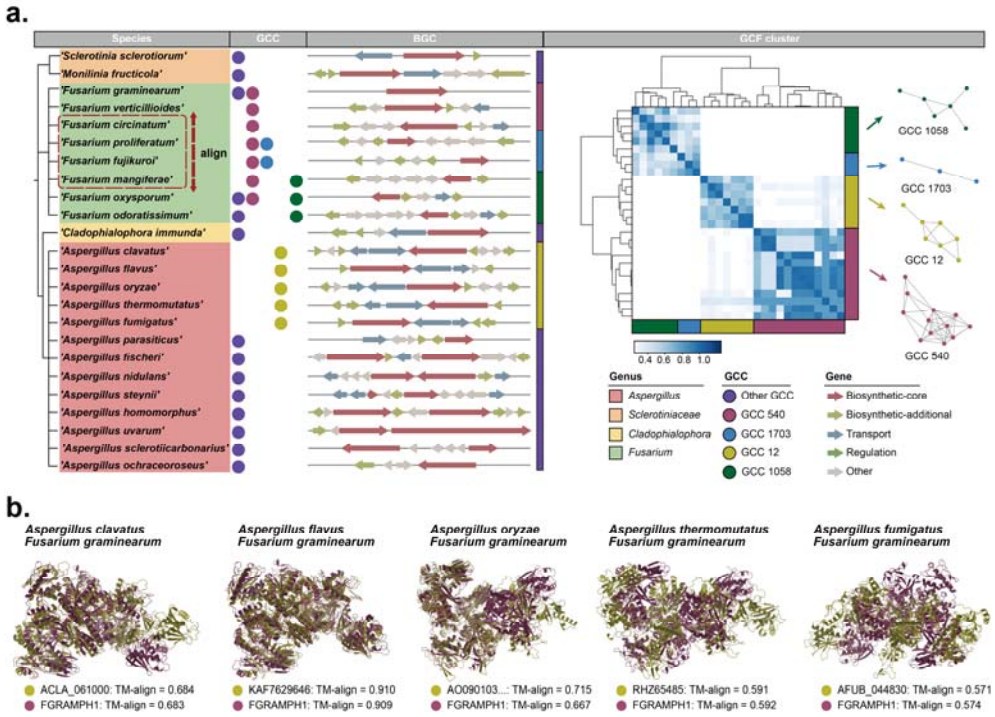


Figure 5. Phylogeny reconstruction of AM-toxin BGCs. a. All core genes annotated as 'AM-toxin synthase' were utilized as reference sequences for alignment against the complete set of core genes identified in the FungiDB database. The reference AM-toxin BGCs segregated into two distinct gene cluster clades: GCC 540 and GCC 1703. The aligned core genes predominantly distributed across three gene cluster clades: GCC 540, GCC1058, and GCC 12. A cluster map on the right shows the similarity between AM-toxin BGCs. **b.** The protein structures of core genes from GCC 540 and GCC 12 were predicted by AlphaFold3 Server, followed by structure alignment using TM-align.

Discussion

Traditional *in silico* BGC detection methods suffer from time-consuming sequence alignment, manually defined-rules, heavy reliance on known homologous sequences, and limited accuracy and sensitivity. In this study, CoreFinder, a context-aware deep learning framework that integrates a protein language model with genomic information to predict gene functions and product types of BGCs, was introduced. Unlike traditional approaches that rely on sequence alignment and manually defined rules, CoreFinder leverages multi-scale architectures and attention mechanisms to captures both global genomic patterns and local sequence features, accelerating the accurate resolution of BGCs.

CoreFinder demonstrated exceptional performance in both BGC product prediction and gene function annotation tasks. CoreFinder achieves higher overall accuracy and

recall in product prediction than DeepBGC. Besides, CoreFinder successfully identified 842 across 1,025 experimentally confirmed core genes with higher precision of 0.945, surpassing antiSMASH, which identified 790 core genes with precision of 0.929. CoreFinder achieved higher sensitivity in identifying terpene synthase in fungal genomes, an area where antiSMASH often struggles due to its reliance on predefined signatures. Integrating antiSMASH with CoreFinder could also lead to enhanced recall rates. Besides, model interpretability analysis reveals that CoreFinder autonomously captured core gene features and functional effects of different genes on product synthesis based on the muti-head attention mechanism without any biology knowledge.

Our end-to-end, scalable workflow, BGC-Prophet+CoreFinder, which integrates BGC-Prophet for initial BGC screening with CoreFinder for accurate annotation, also has advantage of speed: it outperforms antiSMASH by two orders of magnitude, enabling the mining and resolution of BGCs from 100,000 MAGs on a single GPU in just one week, significantly accelerating the natural product discovery from metagenome samples.

To explore the biosynthetic diversity in fungi, we applied the workflow on a broad-spectrum of fungi genomes, and identified 6,414 core genes within 4,585 BGCs across 197 taxa from 256 fungal genomes. Notebly, we revealed the enrichment of NRPS and PKS in *Fusarium* and *Aspergillus*, two genera of significant ecological and agricultural importance. An NRPS gene family involved in plant-pathogen interactions existed before the divergence of *Fusarium* and *Aspergillus*, but was evolved into distinct gene clusters and lay different functional roles. This finding provides new insights into the origin and evolution of NRPS BGCs and fungal biosynthetic machinery. Besides, CoreFinder identified a novel terpene synthase from *Aspergillus oryzae* that was overlooked by antiSMASH and subsequently validated experimentally. It shares high structure similarity and low sequence identity with known synthase, highlighting CoreFinder' s ability to capture remote homolog and uncover novel biosynthetic pathways with robust experimental corroboration.

Although CoreFinder outperformed other commonly used models, limitations warrant consideration. The reliance on ESM2-650M demands substantial computational resources, which limits CoreFinder' s accessibility. To mitigate this, we developed a user-friendly BGC gene annotation tool and released it on Google Colab. Future work could focus on optimizing computational efficiency to enhance accessibility. Additionally, further investigation into CoreFinder' s capacity to identify novel biosynthetic pathways and elucidate the evolutionary diversity of gene clusters across fungal genomes could open new avenues for natural product discovery and biotechnological applications [68]. Moreover, the embeddings generated by CoreFinder remain unexplored. Their analysis could yield valuable insights for clustering BGCs and predicting the chemical structures of natural products, further enhancing its utility.

In conclusion, CoreFinder enhances BGC analysis with high accuracy, robust recall, and notably interpretability. Additionally, the 'BGC-Prophet+CoreFinder' scalable workflow speeds up mining and deciphering of BGCs by two orders of magnitude. Its application to fungal genomes has provided novel insights and experimentally validated discoveries, demonstrating its potential to advance our understanding of microbial secondary metabolism and its ecological and biotechnological significance. By elucidating the functional roles of genes within BGCs, CoreFinder paves the way for the design and engineering of bespoke biosynthetic pathways. This capability can drive the sustainable production of high-value compounds such as novel antibiotics and antineoplastic drugs, offering a powerful platform for synthetic biology applications.

Code availability

The code for CoreFinder is available at https://github.com/HUST-NingKang-Lab/CoreFinder. The example colab notebook for BGC gene annotation and product prediction is available at https://colab.research.google.com/drive/1gjYhrrkbxqJv_YWfhZS6kK5zA1Bsl-1G?usp=sharing. The CoreFinder model's weights and its basic usage is available at https://huggingface.co/KangHuggingface/CoreFinder.

Methods

Data collection

We collected three datasets in our study: data for training, data for validation and data for mining.

Data for training. A total of 2,499 BGCs from MIBiG Version 3.1 were collected. Each sample was labeled with both product type and gene function. BGCs with multiple products were duplicated and labeled with a single product to ensure each sample was associated with only one product, resulting in 2994 samples. Of these, 2,394 BGCs were allocated for model training (MIBiG training-set), 300 for internal validation (MIBiG validation-set), and 300 for model test (MIBiG test-set).

Date for validation. The nine bacteria genomes containing 291 annotated biosynthetic gene clusters (BGCs) from Cimermancic et *al.* [11] were utilized for model evaluation with DeepBGC and BGC-Prophet. Additionally, FunBGCs, a manually curated dataset of fungal BGCs, [17] served as a benchmark in core gene discovery task.

Data for mining. A total of 256 genomes representing 197 fungal species from FungiDB latest version (68), as reported by Basenko et al. [18], were analyzed to identify putative fungal BGCs and biosynthetic core genes. Each gene in the genomes was annotated with Gene Ontology (GO) terms, and the results obtained from CoreFinder were cross-referenced with these annotations to identify additional potential core genes.

Establishment of CoreFinder

Model architecture. We developed the CoreFinder model based on RoBERTa architecture aligned with gLM. CoreFinder consisted of 8 layers, each with a hidden size of 1280 and 10 attention heads per layer, utilizing relative positional embeddings ('relative_key_query') [69]. Each transformer layer comprises of a self-attention layer and a fully connected layer, allowing the model to capture complex dependencies within genomic data. The classification head is implemented as a single-layer feed-forward neural network, with an output layer dimension that corresponds to the number of gene and product categories to be predicted. Specifically, the gene categories include: Biosynthetic-core (core), Biosynthetic-additional (tailoring), Transport, Regulation, and Other. The product categories consist of: Non-Ribosomal Peptide (NRP), Ribosomally Synthesized and Post-Translationally Modified Peptide (RiPP), Alkaloid, Polyketide, Saccharide, Terpene, and Other.

Each gene in the samples was characterized by a 1280-dimensional embedding generated by ESM2-650M, and mean pooling was applied to the token embeddings to generate a single, fixed-length vector representation for each gene. Product embeddings were initialized with zero as a CLS token, serving as a representative embedding for product classification and concatenated with the gene embeddings to form a unified representation of the BGCs. To distinguish the tokens of products and genes, the 'token_type_id' of products is set to 0 and the 'token_type_id 'of genes is set to 1. Given the fixed input length requirement of transformer models, we set input length to 128 to ensure adequate coverage of BGC size and each position in the sequence is represented by a 1280-dimensional vector. The input representation for a single BGC is formally defined as:

$$X_i = [P_i | G_{i,1} | G_{i,2} | \dots | G_{i,k}] \in \mathbb{R}^{n \times (1+k)}$$

where X_i denotes a single BGC, P_i represents the product embedding vector, $G_{i,j} \in \mathbb{R}^{n \times 1}$ represents the embedding vector of the j-th gene in the BGC embedding, $n = 1280$ is the dimensionality of each embedding vector and k is the number of genes in the BGC, with padding or truncation applied as needed to match the fixed input length. Before input to the transformer block, X was added to the relative position embeddings and token type embeddings.

The self-attention mechanism used in RoBERTa adheres to the scaled dot-product attention formulation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q (queries), K (keys), and V (values) are linear projections of the input, and d_k is the dimensionality of the keys. This formulation allows the model to capture contextual relationships among genes and product embedding within the sequence.

Training details. In CoreFinder, the Adam optimizer was applied with an initial learning rate of 5×10^{-5}. The model parameters were initialized with gLM. Training was set to run for a maximum of 100 epochs, employing an early stopping strategy to retain the best-performing model. For hyperparameter selection, we conducted a grid search in the internal validation-set and selected a structure comprising 5 layers and 10 attention heads (**Supplementary Fig. 1**). The modeling framework was implemented in PyTorch, utilizing the HuggingFace Transformers library for model configuration and training.

Model evaluation implementation

Retraining of DeepBGC and BGC-Prophet. To evaluate the product classification performance of DeepBGC and BGC-Prophet. DeepBGC and BGC-Prophet were trained on the merge of CoreFinder' s training-set and inner validation-set, and evaluated on the CoreFinder' s inner test-set.

Classification performance. To assess classification performance of CoreFinder and the Random Forest model (RF), we computed areas under the receiver operating characteristic curve (AUROC) for both gene classification and product prediction tasks using test data derived from MIBiG training-set. The ROC curve and AUROC value were calculated using scikit-learn methods. The RF was retrained separately for gene classification and product prediction, and ROC curve and AUROC score were calculated separately to assess its individual performance on each task.

In the external nine-genome benchmark, we compared the performance of CoreFinder, DeepBGC, and BGC-Prophet for their ability to accurately detect BGC genes. The evaluation metrics included accuracy, F_1-score, precision, recall and false positive rate (FPR):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives predicted, respectively.

Model interpretability. To assess the interpretability of the CoreFinder model and its ability to capture core gene features, we focused on analyzing the model's attention mechanisms. Specifically, we extracted attention weights from all attention heads (n=50) by running inference on all sequences in the MIBiG test set. To evaluate the model's capacity to capture core gene features, each sequence from the test set was represented as a binary vector of BGC length, where a value of 0 indicates a non-core gene and 1 indicates a core gene. We calculated the Spearman correlation between the attention weights and the presence of core genes using the scikit-learn library. We further explored the gene contribution to product inference by examining the attention weights between gene embeddings and product embeddings. Specifically, attention weights were extracted where product embeddings served as Queries (Q) and gene embeddings acted as Keys (K), and these attention scores are referred to as gene-product scores. To highlight the roles of different genes in product inference, the top 5% of gene-product scores from each gene category were collected and displayed.

BGC reconstruction

The latest version of all protein sequences and corresponding chromosomal DNA were downloaded from fungiDB (https://fungidb.org/fungidb/app/downloads). CoreFinder and BGC-Prophet were employed and cross-validated to prioritize high-confidence BGCs from FungiDB. For model inference, genes were grouped into batches based on their genomic location, ensuring that genes from different chromosomes were not combined into a single batch. Following inference, the protein sequences and corresponding DNA of all BGCs were retrieved and reconstructed into GenBank format files. BiG-SCAPE [52] was employed to analyze the Gene Cluster Families (GCFs) and Gene Cluster Contexts (GCCs) for CoreFinder-identified BGCs. DSS value (Domain Sequence Similarity) was used to measure the similarity between two BGCs as follows:

$$DSS = 1 - \left(\frac{c_1}{\sum_{d \in \alpha} max\left(N_A^d, N_B^d\right)} \sum sd(d_A, d_B) + \frac{c_2}{\sum_{d \in n\alpha} max\left(N_A^d, N_B^d\right)} \sum sd(d_A, d_B) \right)$$

Where α represents the anchor domains, a list of domains can be given a special weight, such as those in NRPS or PKS, and $n\alpha$ accounts for the remaining domains;

d are all distinct domain types in the BGC pair, and N^d donates the number of copies of domain d in the BGC. sd computes the sequence dissimilarity between domain d in BGC A and B by taking all copies of the domain and summing the complement of their sequence identities. c_1 and c_2 are weighting coefficients that adjust the contributions of anchor and non-anchor domains, respectively.

Homologs search and structure prediction

To identify homologous proteins, we used DIAMOND [70] with default parameters, querying against the core gene database from CoreFinder. This search detected sequences similar to the query proteins, including both closely related and remote homologs. To identify homologs of novel core gene discovered by CoreFinder, we queried the gene in NCBI database and performed a BLASTp search to detect similar genes with known functions. For protein structure prediction, AlphaFold3 Server [71] generated the tertiary structures of selected proteins. We used TM-align [72] to compare and align protein structures, computing the TM-score to evaluate structural similarity as follows:

$$TM-score\left(P_{target}, P_{source}\right) = \frac{1}{L_{target}} \sum_{i}^{L_{alignment}} \frac{1}{1+\left(\frac{x_i}{x_0 \cdot L_{target}}\right)^2}$$

Where P_{source} is the source protein aligned to target protein P_{target}. L_{target} is the length of target protein, and $L_{alignment}$ is the number of paired residues. x_i is the distance between the i-th paired residues. All procedures followed default settings.

DNA extraction

Aspergillus oryzae NSPID1 (AO-NSPID1) was obtained from our laboratory for the following gene deletion and validation. AO-NSPID1 glycerol stock was retrieved from -80°C and thawed. A 40 μL aliquot was transferred onto a PDA + UU agar plate using a sterile pipette tip. The sample was evenly spread across the surface with a sterile disposable spreader, and the plate was sealed, labeled, and inverted for incubation at 30°C for 5 days. Fungal biomass was dried and transferred to a 2 mL tube with 400 μL LETS Buffer and 3 steel beads, then homogenized at 70 Hz for 60 seconds. After a 5-minute rest, 300 μL LETS Buffer was added, followed by a 5-minute incubation. PCI extraction was performed by adding 700 μL of the lower phase, shaking for 20 seconds, and centrifuging at 13,000 rpm for 10 minutes at 4°C. The aqueous phase was transferred to a new tube, mixed with 1 mL pre-cooled 95% ethanol, and centrifuged at 13,000 rpm for 10 minutes. The pellet was washed with 300 μL 75% ethanol, air-dried, and DNA was eluted with 30 μL ddH2O at 65°C. The genomic DNA was quantified, diluted to 50–100 ng/μL, and stored at -20°C.

Gene amplification

PCR procedure. Recombinant plasmids were constructed using SnapGene software to insert the target gene into a suitable plasmid. Primers for each fragment were designed, including forward (FW) and reverse (RV) primers. The primers were approximately 20 bp in length, with an annealing temperature of around 55°C and homologous arm lengths of 20 bp. The PCR reaction mixture was prepared with the final addition of 2× Phanta Flash Master Mix. The Master Mix was stored at -20°C after use. The PCR amplification program began with pre-denaturation at 98°C for 10 seconds, followed by 34 cycles of denaturation at 98°C for 10 seconds, annealing at the primer-specific melting temperature for 5 seconds, and extension at 72°C for 5 seconds per kilobase. A final extension at 72°C for 2 minutes completed the process, with the products stored at 4°C.

DNA Agarose Gel Electrophoresis. For DNA separation, 0.6 g of high-purity, low electro-osmotic agarose was mixed with 60 mL of TAE buffer (30 mL for smaller volumes). The solution was microwaved for 2 minutes to dissolve, followed by an additional 30 seconds of heating until visible bubbles formed. After slight cooling, a red fluorescent DNA dye was added. The agarose solution was poured into the gel mold, and combs were inserted to create wells. The gel was allowed to solidify for approximately 20 minutes, after which the combs were carefully removed. The gel was placed into an electrophoresis chamber, ensuring that the TAE buffer completely covered the gel. The wells were oriented towards the negative electrode. PCR products were mixed with 6× Loading Buffer, and 30 μL of the mixture was loaded into each well along with 5 μL of DL 5000 DNA Marker. Electrophoresis was performed at 120 V (not exceeding 140 V, 200 mA) for approximately 25 minutes. The gel was then visualized under UV light using a gel imaging system, with the image captured after adjusting the brightness and contrast for clarity.

Gene knockout experiment

Fungal preparation. The germinated Aspergillus oryzae culture was filtered through an iron sieve, washed with 200 mL of water, pressed dry, and transferred to a 50 mL centrifuge tube. Yatalase (120 mg) was dissolved in 20 mL of Solution 0 (final concentration: 60 mg/mL), filtered through a 0.22 μm membrane, and added to the fungal biomass. The mixture was incubated at 30°C and 120 rpm for approximately 60 minutes, with protoplast formation monitored at 50 minutes. After 100 minutes, the suspension was filtered through a glass funnel, kept on ice, and centrifuged at 1000 g, 4°C, for 5 minutes. The supernatant was discarded, and the pellet was gently resuspended in 10 mL of 0.8 M NaCl. Protoplasts were counted, with each experimental group containing 4 × 10□ protoplasts. The suspension was centrifuged at 900 g, 4°C, for 5 minutes, the supernatant was discarded, and the pellet was resuspended in Solution II and III (4:1, 100 μL per group).

Fungal transformation. The DNA solution was prepared by dissolving at least 10 μg of DNA in 50 μL of sterile water and mixing with 50 μL of 2× Solution II. A 100 μL aliquot of the protoplast suspension was added to the DNA solution and incubated on ice for 20 minutes. Solution III (50 μL) was then added, followed by incubation at 25°C for 20 minutes. Subsequently, 2 mL of Solution III was added in steps, gently mixed, and incubated at room temperature for 5 minutes. The mixture was then diluted with 4 mL of Solution II. A 600 μL aliquot of the transformed protoplast suspension was plated onto each CD plate, with 10 plates per group, and incubated at 30°C overnight. The following day, 5 mL of Top Agar was overlaid onto each plate. Once solidified, the plates were sealed and incubated upside down at 30°C. Transformant growth was assessed on the third day. To screen positive transformants, secondary screening plates (8 mL each) were prepared and transformants were picked using a sterile pipette tip, transferred to the plates, labeled, and incubated at 30°C.

Gene knockout validation. To obtain fungal biomass, MPY liquid medium was prepared, sterilized at 121°C for 30 minutes, and distributed into sterile 40 mL vials, each containing 5 mL. Biomass was acquired from the rescreening plates and added to the vials. A control was also prepared by adding wild-type Aspergillus oryzae spores to another vial. The vials were sealed with breathable sealing film, labeled, and incubated at 30°C and 220 rpm for 3-5 days. Genomic DNA was extracted from each transformant. Primers were designed for PCR validation of each knockout strain. For the 8.1-5F, 8.1-3F, and 8.1-RT fragments, the respective i5F-FW, i5F-RV, i3F-FW, i3F-RV, iRT-FW, and iRT-RV primers were used. PCR amplification was performed using the extracted genomic DNA as templates. The PCR program included initial denaturation at 98°C, followed by 34 cycles of denaturation, annealing, and extension, with final extension at 72°C. Validation of knockouts was confirmed based on the presence or absence of PCR amplicons.

Fungi Fermentation

Transformant fermentation. CYA, CYA20, and YES small-scale fermentation was conducted [73]. The respective liquid media were prepared, sterilized at 121°C for 30 minutes, and distributed into sterile 40 mL vials, each containing 5 mL. Spore suspensions were added to the vials, with a control vial containing wild-type Aspergillus oryzae spores. The vials were sealed with breathable sealing film and incubated at 30°C and 220 rpm for 5 days.

Rice small-scale fermentation. For the rice small-scale fermentation, 2 g of rice and 1.8 mL of water were added to 40 mL vials, sealed with breathable sealing film, and sterilized at 121°C for 30 minutes. After cooling, spore suspensions were added to the vials, with a control vial containing wild-type Aspergillus oryzae spores. The rice was loosened with a sterile pipette to ensure homogeneity. The vials were sealed with breathable sealing film and incubated at 30°C for 12 days.

Extraction of fermentation products. Fermentation products from liquid and solid media were extracted using industrial-grade methanol overnight. The methanol extracts were concentrated by rotary evaporation, reconstituted in chromatographic-grade methanol, and stored at -20°C.

LC-MS Analysis of Fermentation Products

LC-MS analysis of crude extracts was performed using an Eclipse Plus C18 column (2.1 × 100 mm, 3.5 μm). The mobile phase consisted of CH_3CN and H_2O (both containing 0.1% HCOOH). The CH3CN concentration was increased linearly from 5% to 98% over 11 minutes and maintained at 100% for 2 minutes. The flow rate was 0.5 mL/min, and the column temperature was set at 25°C.

Data visualization

The Python packages Matplotlib [74] and Seaborn [75] were used for data visualization, while iTOL [76] was employed for tree visualization.

Acknowledge

This work was partially supported by the National Key R&D Program of China (Grant No. 2023YFA1800900, 2021YFA0910500 and 2018YFC0910502) and the National Natural Science Foundation of China (Grant Nos. 32071465, 31871334, 81827901, 22277035).

Author contribution

K.N., Y.Z. and H.B. conceived and proposed the idea. Z.K., H.Z. and C.L. performed the experiments and analyzed the data. Z.K. and H.Z. visualized the data. Z.K., H.Z., Y.Y., Y.Z., K.N. and H.B. contributed to editing and proof-reading the manuscript. R.Y. joined the discussions. All authors read and approved the final manuscript.

References

1. Donia, M.S., et al., *A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics.* Cell, 2014. **158**(6): p. 1402-1414.
2. Crits-Christoph, A., et al., *Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis.* Nature, 2018. **558**(7710): p. 440-444.
3. Robey, M.T., et al., *An interpreted atlas of biosynthetic gene clusters from*

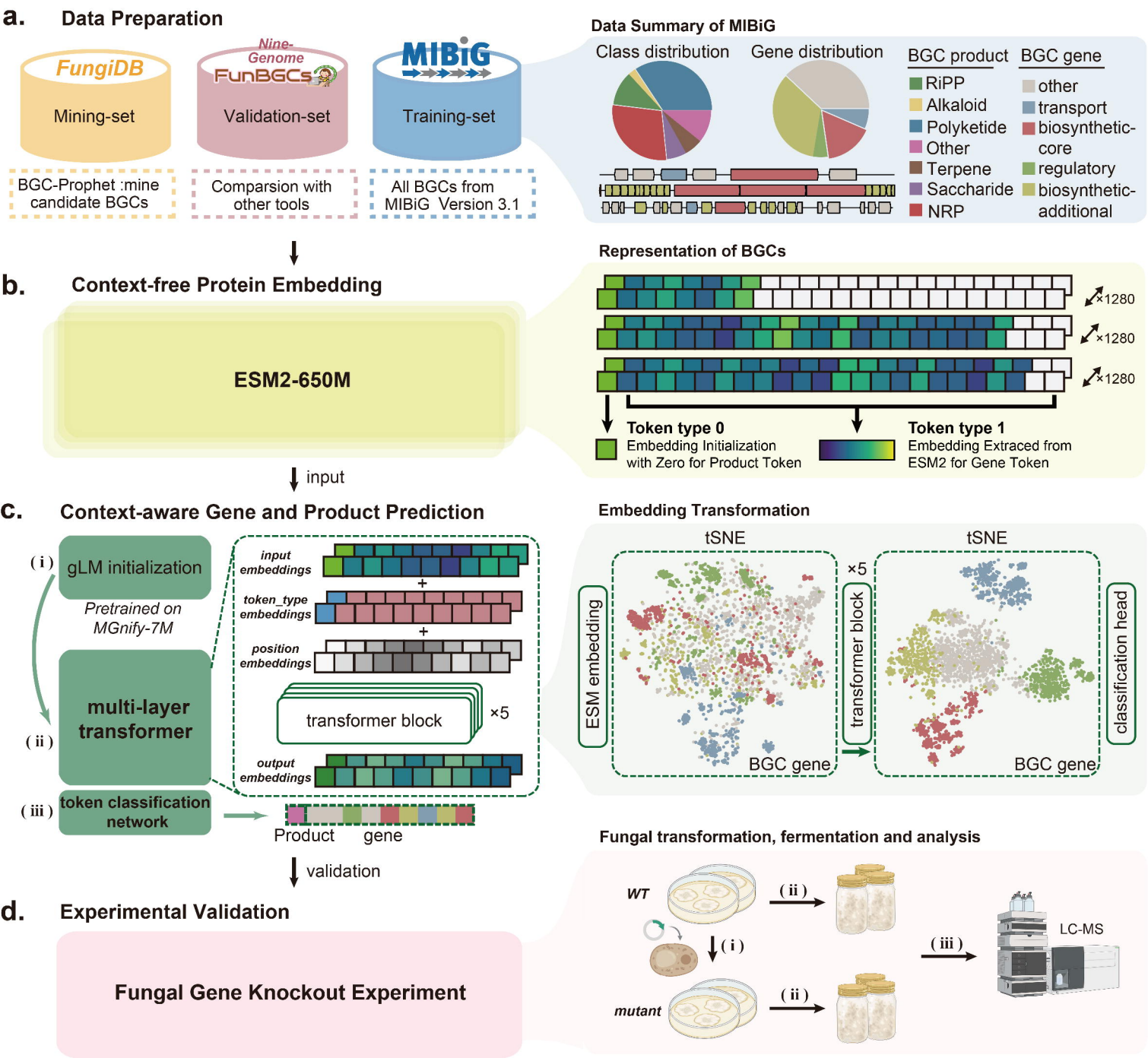
1,000 fungal genomes. Proc Natl Acad Sci U S A, 2021. **118**(19).
4. Campbell, C.D. and J.C. Vederas, *Biosynthesis of lovastatin and related metabolites formed by fungal iterative PKS enzymes*. Biopolymers, 2010. **93**(9): p. 755-63.
5. Medema, M.H. and M.A. Fischbach, *Computational approaches to natural product discovery.* Nat Chem Biol, 2015. **11**(9): p. 639-48.
6. Rios-Martinez, C., et al., *Deep self-supervised learning for biosynthetic gene cluster detection and product classification.* PLoS Comput Biol, 2023. **19**(5): p. e1011162.
7. Mullowney, M.W., et al., *Artificial intelligence for natural product drug discovery.* Nat Rev Drug Discov, 2023. **22**(11): p. 895-916.
8. Carroll, L.M., et al., *Accurate de novo identification of biosynthetic gene clusters with GECCO.* BioRxiv, 2021: p. 2021.05. 03.442509.
9. Hannigan, G.D., et al., *A deep learning genome-mining strategy for biosynthetic gene cluster prediction.* Nucleic Acids Res, 2019. **47**(18): p. e110.
10. Medema, M.H., 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**(Web Server issue): p. W339-46.
11. Cimermancic, P., et al., *Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters*. Cell, 2014. **158**(2): p. 412-421.
12. Sugimoto, Y., et al., *A metagenomic strategy for harnessing the chemical repertoire of the human microbiome.* Science, 2019. **366**(6471).
13. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
14. Kautsar, S.A., et al., *BiG-FAM: the biosynthetic gene cluster families database*. Nucleic Acids Res, 2021. **49**(D1): p. D490-D497.
15. Blin, K., et al., *The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more.* Nucleic Acids Res, 2024. **52**(D1): p. D586-D589.
16. Palaniappan, K., et al., *IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase.* Nucleic Acids Res, 2020. **48**(D1): p. D422-D430.
17. Tang, J. and Y. Matsuda, *Discovery of fungal onoceroid triterpenoids through domainless enzyme-targeted global genome mining*. Nat Commun, 2024. **15**(1): p. 4312.
18. Flamholz, Z.N., S.J. Biller, and L. Kelly, *Large language models improve annotation of prokaryotic viral proteins.* Nature Microbiology, 2024. **9**(2): p. 537-549.
19. Lai, Q., et al., *Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model.* Nucleic Acids Research, 2025. **53**(7).
20. Khaldi, N., et al., *SMURF: Genomic mapping of fungal secondary metabolite*

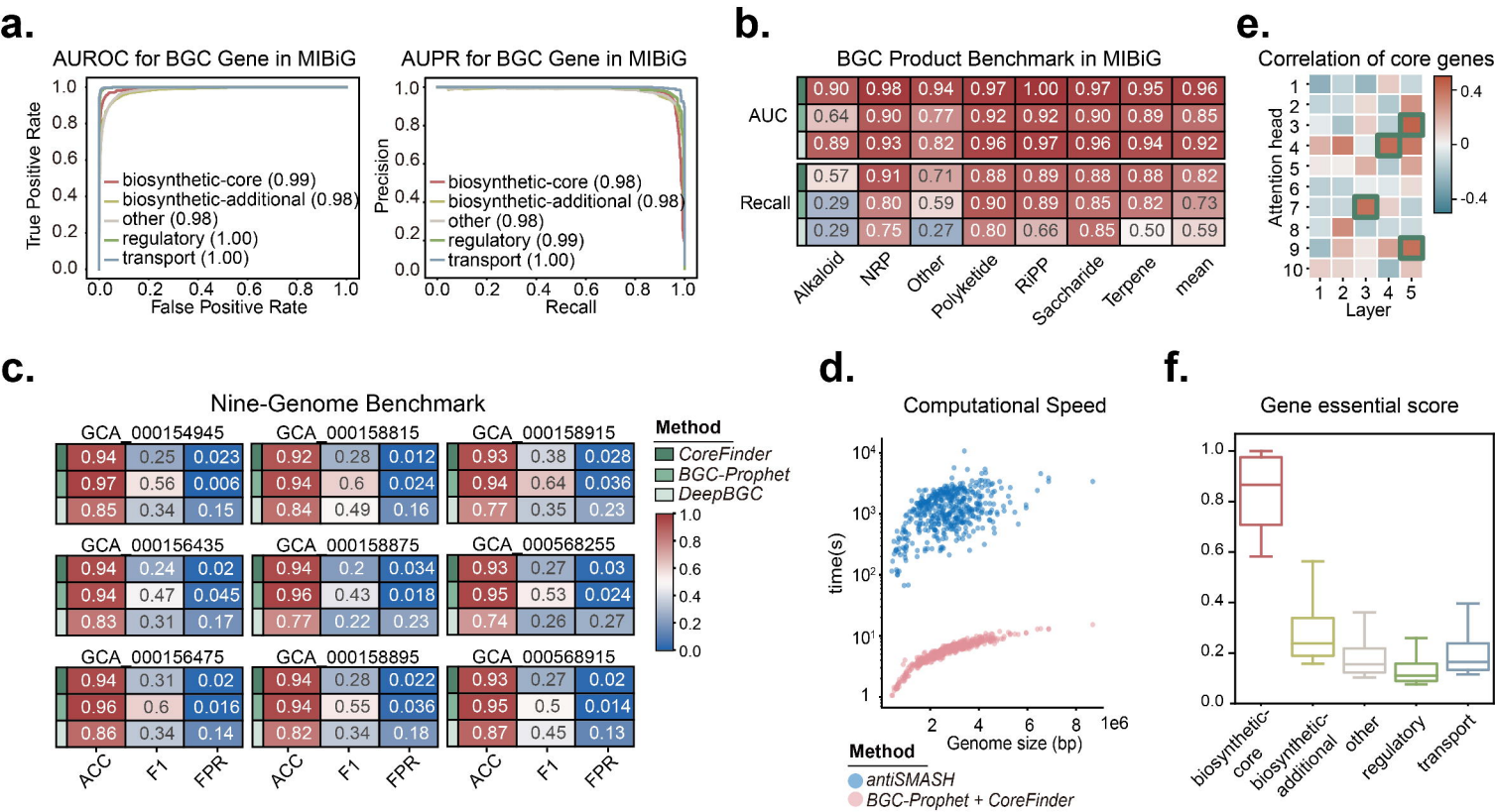
clusters. Fungal Genet Biol, 2010. **47**(9): p. 736-41.
21. Kwon, M.J., et al., *Beyond the Biosynthetic Gene Cluster Paradigm: Genome-Wide Coexpression Networks Connect Clustered and Unclustered Transcription Factors to Secondary Metabolic Pathways.* Microbiol Spectr, 2021. **9**(2): p. e0089821.
22. Li, Q., et al., *Progress and opportunities of foundation models in bioinformatics*. Brief Bioinform, 2024. **25**(6).
23. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
24. Ferruz, N., S. Schmidt, and B. Höcker, *ProtGPT2 is a deep unsupervised language model for protein design.* Nature communications, 2022. **13**(1): p. 4348.
25. Madani, A., et al., *Large language models generate functional protein sequences across diverse families*. Nature Biotechnology, 2023. **41**(8): p. 1099-1106.
26. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* nature, 2021. **596**(7873): p. 583-589.
27. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
28. Hong, L., et al., *Fast, sensitive detection of protein homologs using deep dense retrieval.* Nat Biotechnol, 2024.
29. Hwang, Y., et al., *Genomic language model predicts protein co-regulation and function.* Nat Commun, 2024. **15**(1): p. 2880.
30. Richardson, L., et al., *MGnify: the microbiome sequence data analysis resource in 2023.* Nucleic Acids Res, 2023. **51**(D1): p. D753-D759.
31. Brundrett, M.C. and L. Tedersoo, *Evolutionary history of mycorrhizal symbioses and global host plant diversity.* New Phytol, 2018. **220**(4): p. 1108-1115.
32. Case, N.T., et al., *Fungal impacts on Earth's ecosystems.* Nature, 2025. **638**(8049): p. 49-57.
33. Keller, N.P., *Fungal secondary metabolism: regulation, function and drug discovery.* Nature Reviews Microbiology, 2019. **17**(3): p. 167-180.
34. Harvey, A.L., R. Edrada-Ebel, and R.J. Quinn, *The re-emergence of natural products for drug discovery in the genomics era.* Nat Rev Drug Discov, 2015. **14**(2): p. 111-29.
35. Terlouw, B.R., et al., *MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters.* Nucleic Acids Res, 2023. **51**(D1): p. D603-D610.
36. Tang, J. and Y. Matsuda, *Discovery of fungal onoceroid triterpenoids through domainless enzyme-targeted global genome mining.* Nature Communications, 2024. **15**(1): p. 4312.
37. Kjaerbolling, I., et al., *A comparative genomics study of 23 Aspergillus species from section Flavi.* Nat Commun, 2020. **11**(1): p. 1106.
38. Vaswani, A., *Attention is all you need.* Advances in Neural Information

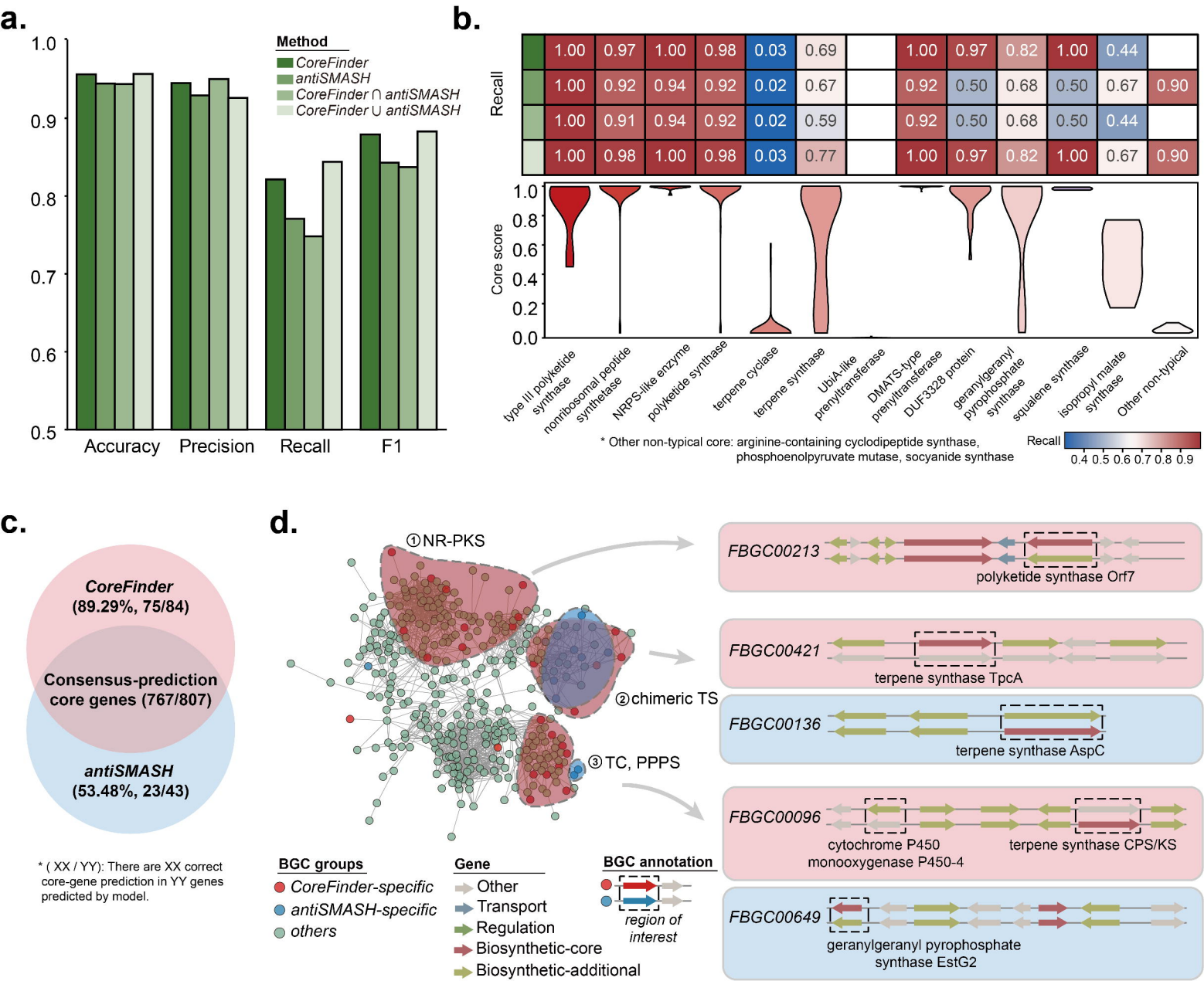
Processing Systems, 2017.
39. Dodds, P.N. and J.P. Rathjen, *Plant immunity: towards an integrated view of plant-pathogen interactions.* Nat Rev Genet, 2010. **11**(8): p. 539-48.
40. Cheng, M., et al., *A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau.* Nat Commun, 2024. **15**(1): p. 1438.
41. Crits-Christoph, A., et al., *Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity.* Genome Research, 2021. **31**(2): p. 239-250.
42. Urlacher, V.B. and M. Girhard, *Cytochrome P450 monooxygenases: an update on perspectives for synthetic application.* Trends in biotechnology, 2012. **30**(1): p. 26-36.
43. Chen, W., et al., *Fungal cytochrome p450 monooxygenases: their distribution, structure, functions, family expansion, and evolutionary origin.* Genome biology and evolution, 2014. **6**(7): p. 1620-1634.
44. Cong, Z., et al., *NPtagM: A Tailoring Enzyme Genome Mining Toolkit and Its Application in Terpenoid P450s from Phytopathogenic Fungi.* Journal of Agricultural and Food Chemistry, 2024. **72**(49): p. 27225-27234.
45. Basenko, E.Y., et al., *What is new in FungiDB: a web-based bioinformatics platform for omics-scale data analysis for fungal and oomycete species.* Genetics, 2024. **227**(1): p. iyae035.
46. Rokas, A., et al., *Biosynthetic gene clusters and the evolution of fungal chemodiversity.* Nat Prod Rep, 2020. **37**(7): p. 868-878.
47. Rokas, A., J.H. Wisecaver, and A.L. Lind, *The birth, evolution and death of metabolic gene clusters in fungi.* Nat Rev Microbiol, 2018. **16**(12): p. 731-744.
48. Machida, M., et al., *Genome sequencing and analysis of Aspergillus oryzae.* Nature, 2005. **438**(7071): p. 1157-61.
49. Gasch, A.P., B.A. Payseur, and J.E. Pool, *The Power of Natural Variation for Model Organism Biology.* Trends Genet, 2016. **32**(3): p. 147-154.
50. Lange, B.M., et al., *Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes.* Proc Natl Acad Sci U S A, 2000. **97**(24): p. 13172-7.
51. González-Hernández, R.A., et al., *Overview of fungal terpene synthases and their regulation.* World Journal of Microbiology and Biotechnology, 2023. **39**(7): p. 194.
52. Navarro-Munoz, J.C., et al., *A computational framework to explore large-scale biosynthetic diversity.* Nat Chem Biol, 2020. **16**(1): p. 60-68.
53. Selem-Mojica, N., et al., *EvoMining reveals the origin and fate of natural product biosynthetic enzymes.* Microb Genom, 2019. **5**(12).
54. Stachelhaus, T. and M.A. Marahiel, *Modular structure of genes encoding multifunctional peptide synthetases required for non-ribosomal peptide synthesis.* FEMS Microbiol Lett, 1995. **125**(1): p. 3-14.
55. Jenner, M., et al., *Elucidating the molecular programming of a nonlinear non-ribosomal peptide synthetase responsible for fungal siderophore biosynthesis.* Nat Commun, 2023. **14**(1): p. 2832.

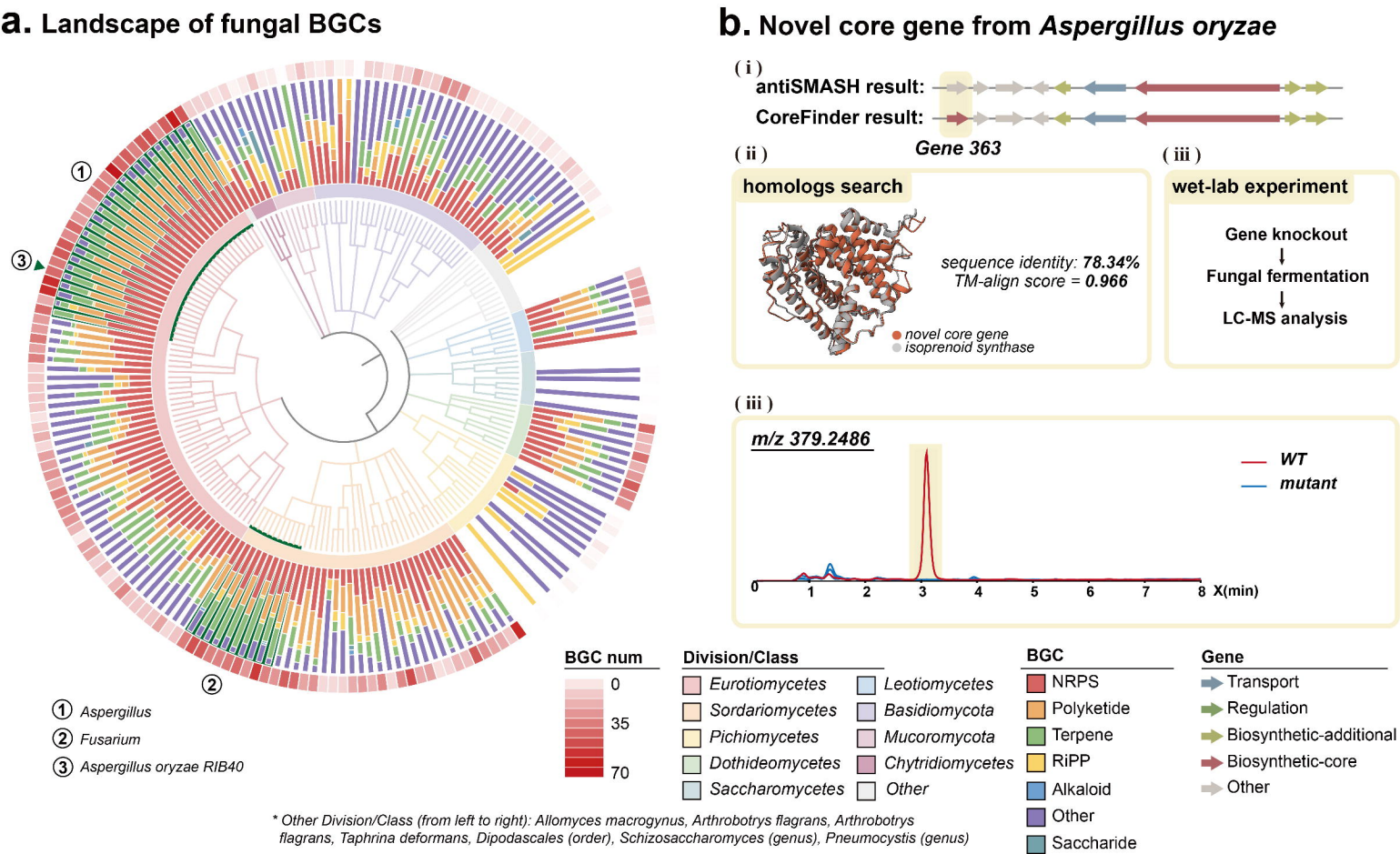
56. Jaremko, M.J., et al., *Type II non-ribosomal peptide synthetase proteins: structure, mechanism, and protein-protein interactions.* Nat Prod Rep, 2020. **37**(3): p. 355-379.
57. Hur, G.H., C.R. Vickery, and M.D. Burkart, *Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology.* Nat Prod Rep, 2012. **29**(10): p. 1074-98.
58. Jensen, P.R., *Natural Products and the Gene Cluster Revolution.* Trends Microbiol, 2016. **24**(12): p. 968-977.
59. Wu, J., et al., *Systematic characterization of horizontally transferred biosynthetic gene clusters in the human gut microbiota using HTBGC□Finder.* iMetaOmics, 2025: p. e62.
60. Slot, J.C. and A. Rokas, *Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi.* Current biology, 2011. **21**(2): p. 134-139.
61. Richards, T.A. and N.J. Talbot, *Horizontal gene transfer in osmotrophs: playing with public goods.* Nature Reviews Microbiology, 2013. **11**(10): p. 720-727.
62. Kountz, D.J. and E.P. Balskus, *Leveraging Microbial Genomes and Genomic Context for Chemical Discovery.* Acc Chem Res, 2021. **54**(13): p. 2788-2797.
63. Salamzade, R. and L.R. Kalan, *Context matters: assessing the impacts of genomic background and ecology on microbial biosynthetic gene cluster evolution.* mSystems, 2025: p. e0153824.
64. Chase, A.B., et al., *Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites.* mBio, 2021. **12**(6): p. e0270021.
65. Stachelhaus, T., H.D. Mootz, and M.A. Marahiel, *The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases.* Chem Biol, 1999. **6**(8): p. 493-505.
66. Witte, T.E., et al., *Accessory chromosome-acquired secondary metabolism in plant pathogenic fungi: the evolution of biotrophs into host-specific pathogens.* Frontiers in Microbiology, 2021. **12**: p. 664276.
67. Medema, M.H., et al., *A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis.* PLoS Comput Biol, 2014. **10**(12): p. e1004016.
68. Alkhalaf, L.M., et al., *Thoughts for the future.* nature chemical biology, 2025. **21**(1): p. 6-15.
69. Huang, Z., et al., *Improve transformer models with better relative position embeddings.* arXiv preprint arXiv:2009.13658, 2020.
70. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
71. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3.* Nature, 2024. **630**(8016): p. 493-500.
72. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic acids research, 2005. **33**(7): p. 2302-2309.

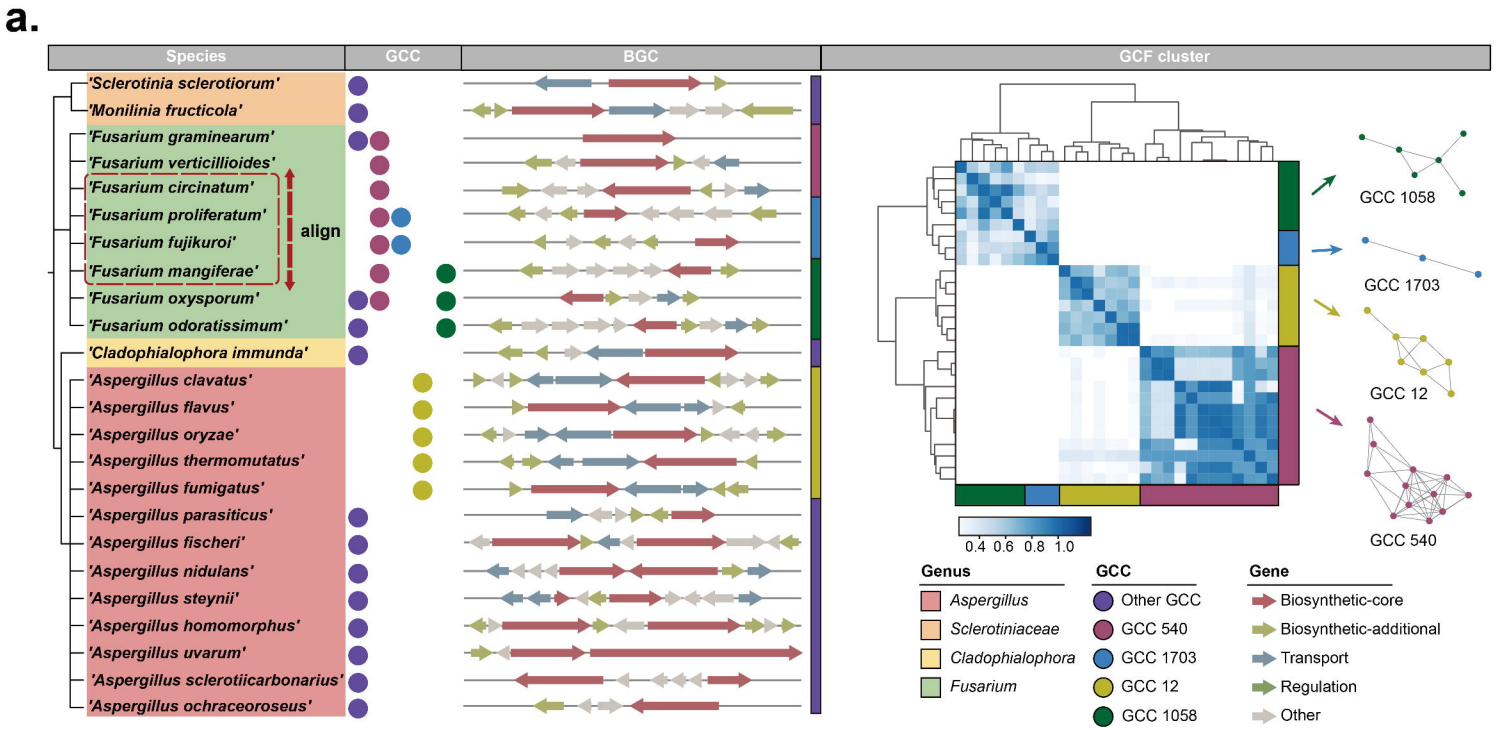
73. Kan, E., et al., *Discovery of the 2, 4'□Dihydroxy□3'□methoxypropiophenone Biosynthesis Genes in Aspergillus oryzae*. ChemBioChem, 2021. **22**(1): p. 203-211.
74. Barrett, P., et al. *matplotlib--A Portable Python Plotting Package*. in *Astronomical data analysis software and systems XIV*. 2005.
75. Waskom, M.L., *Seaborn: statistical data visualization*. Journal of Open Source Software, 2021. **6**(60): p. 3021.
76. Letunic, I. and P. Bork, *Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool*. Nucleic acids research, 2024. **52**(W1): p. W78-W82.

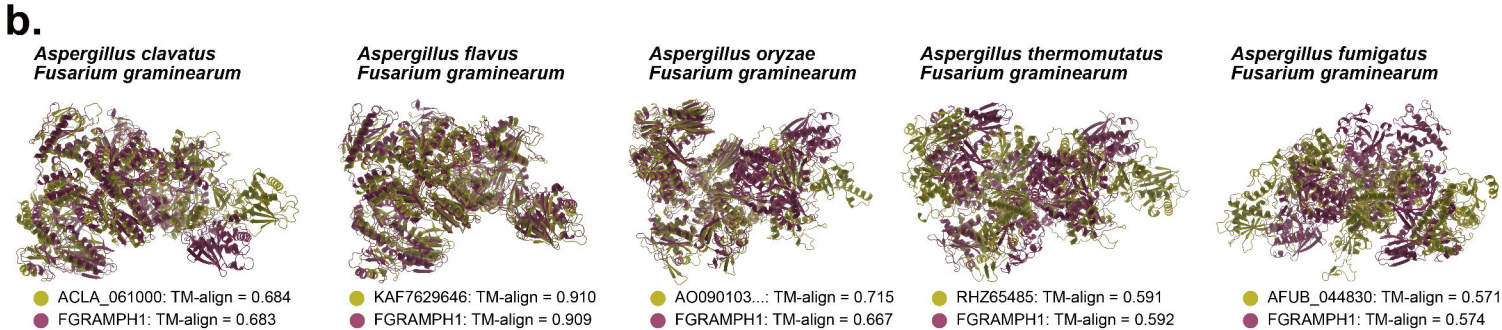

a.
Data Preparation
FungiDB
Mining-set
Nine-Genome FunBGCs
Validation-set
MIBiG
Training-set
BGC-Prophet :mine candidate BGCs
Comparsion with other tools
All BGCs from MIBiG Version 3.1
Data Summary of MIBiG
Class distribution
Gene distribution
BGC product
RiPP
Alkaloid
Polyketide
Other
Terpene
Saccharide
NRP
BGC gene
other
transport
biosynthetic-core
regulatory
biosynthetic-additional
b.
Context-free Protein Embedding
ESM2-650M
input
Representation of BGCs
×1280
×1280
×1280
Token type 0
Embedding Initialization with Zero for Product Token
Token type 1
Embedding Extraced from ESM2 for Gene Token
c.
Context-aware Gene and Product Prediction
(i)
gLM initialization
Pretrained on MGnify-7M
(ii)
multi-layer transformer
(iii)
token classification network
input embeddings
token_type embeddings
position embeddings
transformer block
×5
output embeddings
Product
gene
validation
Embedding Transformation
tSNE
tSNE
ESM embedding
×5
transformer block
classification head
BGC gene
BGC gene
d.
Experimental Validation
Fungal Gene Knockout Experiment
Fungal transformation, fermentation and analysis
WT
mutant
(i)
(ii)
(ii)
(iii)
LC-MS


a.
AUROC for BGC Gene in MIBiG
True Positive Rate
False Positive Rate
biosynthetic-core (0.99)
biosynthetic-additional (0.98)
other (0.98)
regulatory (1.00)
transport (1.00)
AUPR for BGC Gene in MIBiG
Precision
Recall
biosynthetic-core (0.98)
biosynthetic-additional (0.98)
other (0.98)
regulatory (0.99)
transport (1.00)
b.
BGC Product Benchmark in MIBiG
AUC
Recall
Alkaloid
NRP
Other
Polyketide
RiPP
Saccharide
Terpene
mean
e.
Correlation of core genes
Attention head
Layer
c.
Nine-Genome Benchmark
GCA_000154945
GCA_000158815
GCA_000158915
GCA_000156435
GCA_000158875
GCA_000568255
GCA_000156475
GCA_000158895
GCA_000568915
ACC
F1
FPR
Method
CoreFinder
BGC-Prophet
DeepBGC
d.
Computational Speed
time(s)
Genome size (bp)
1e6
Method
antiSMASH
BGC-Prophet + CoreFinder
f.
Gene essential score
biosynthetic-core
biosynthetic-additional
other
regulatory
transport


a.
Method
CoreFinder
antiSMASH
CoreFinder ∩ antiSMASH
CoreFinder ∪ antiSMASH
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
Precision
Recall
F1
b.
Recall
1.00 0.97 1.00 0.98 0.03 0.69 1.00 0.97 0.82 1.00 0.44
1.00 0.92 0.94 0.92 0.02 0.67 0.92 0.50 0.68 0.50 0.67 0.90
1.00 0.91 0.94 0.92 0.02 0.59 0.92 0.50 0.68 0.50 0.44
1.00 0.98 1.00 0.98 0.03 0.77 1.00 0.97 0.82 1.00 0.67 0.90
Core score
type III polyketide synthase
nonribosomal peptide synthetase
NRPS-like enzyme
polyketide synthase
terpene cyclase
terpene synthase
UbiA-like prenyltransferase
DMATS-type prenyltransferase
DUF3328 protein
geranylgeranyl pyrophosphate synthase
squalene synthase
isopropyl malate synthase
Other non-typical
* Other non-typical core: arginine-containing cyclodipeptide synthase, phosphoenolpyruvate mutase, socyanide synthase
Recall
0.4 0.5 0.6 0.7 0.8 0.9
c.
CoreFinder (89.29%, 75/84)
Consensus-prediction core genes (767/807)
antiSMASH (53.48%, 23/43)
* (XX / YY): There are XX correct core-gene prediction in YY genes predicted by model.
d.
①NR-PKS
②chimeric TS
③TC, PPPS
BGC groups
CoreFinder-specific
antiSMASH-specific
others
Gene
Other
Transport
Regulation
Biosynthetic-core
Biosynthetic-additional
BGC annotation
region of interest
FBGC00213
polyketide synthase Orf7
FBGC00421
terpene synthase TpcA
FBGC00136
terpene synthase AspC
FBGC00096
cytochrome P450 monooxygenase P450-4
terpene synthase CPS/KS
FBGC00649
geranylgeranyl pyrophosphate synthase EstG2


a. Landscape of fungal BGCs
b. Novel core gene from Aspergillus oryzae
(i)
antiSMASH result:
CoreFinder result:
Gene 363
(ii)
homologs search
sequence identity: 78.34%
TM-align score = 0.966
novel core gene
isoprenoid synthase
(iii)
wet-lab experiment
Gene knockout
Fungal fermentation
LC-MS analysis
(iii)
m/z 379.2486
WT
mutant
0
1
2
3
4
5
6
7
8
X(min)
① Aspergillus
② Fusarium
③ Aspergillus oryzae RIB40
BGC num
0
35
70
Division/Class
Eurotiomycetes
Sordariomycetes
Pichiomycetes
Dothideomycetes
Saccharomycetes
Leotiomycetes
Basidiomycota
Mucoromycota
Chytridiomycetes
Other
BGC
NRPS
Polyketide
Terpene
RiPP
Alkaloid
Other
Saccharide
Gene
Transport
Regulation
Biosynthetic-additional
Biosynthetic-core
Other
* Other Division/Class (from left to right): Allomyces macrogynus, Arthrobotrys flagrans, Arthrobotrys flagrans, Taphrina deformans, Dipodascales (order), Schizosaccharomyces (genus), Pneumocystis (genus)


a.
Species
GCC
BGC
GCF cluster
'Sclerotinia sclerotiorum'
'Monilinia fructicola'
'Fusarium graminearum'
'Fusarium verticillioides'
'Fusarium circinatum'
'Fusarium proliferatum'
'Fusarium fujikuroi'
'Fusarium mangiferae'
'Fusarium oxysporum'
'Fusarium odoratissimum'
'Cladophialophora immunda'
'Aspergillus clavatus'
'Aspergillus flavus'
'Aspergillus oryzae'
'Aspergillus thermomutatus'
'Aspergillus fumigatus'
'Aspergillus parasiticus'
'Aspergillus fischeri'
'Aspergillus nidulans'
'Aspergillus steynii'
'Aspergillus homomorphus'
'Aspergillus uvarum'
'Aspergillus sclerotiicarbonarius'
'Aspergillus ochraceoroseus'
align
GCC 1058
GCC 1703
GCC 12
GCC 540
0.4 0.6 0.8 1.0
Genus
Aspergillus
Sclerotiniaceae
Cladophialophora
Fusarium
GCC
Other GCC
GCC 540
GCC 1703
GCC 12
GCC 1058
Gene
Biosynthetic-core
Biosynthetic-additional
Transport
Regulation
Other


b.
Aspergillus clavatus
Fusarium graminearum
ACLA_061000: TM-align = 0.684
FGRAMPH1: TM-align = 0.683
Aspergillus flavus
Fusarium graminearum
KAF7629646: TM-align = 0.910
FGRAMPH1: TM-align = 0.909
Aspergillus oryzae
Fusarium graminearum
AO090103...: TM-align = 0.715
FGRAMPH1: TM-align = 0.667
Aspergillus thermomutatus
Fusarium graminearum
RHZ65485: TM-align = 0.591
FGRAMPH1: TM-align = 0.592
Aspergillus fumigatus
Fusarium graminearum
AFUB_044830: TM-align = 0.571
FGRAMPH1: TM-align = 0.574