

Comprehensive profiling of antibiotic resistance genes in diverse environments and novel function discovery

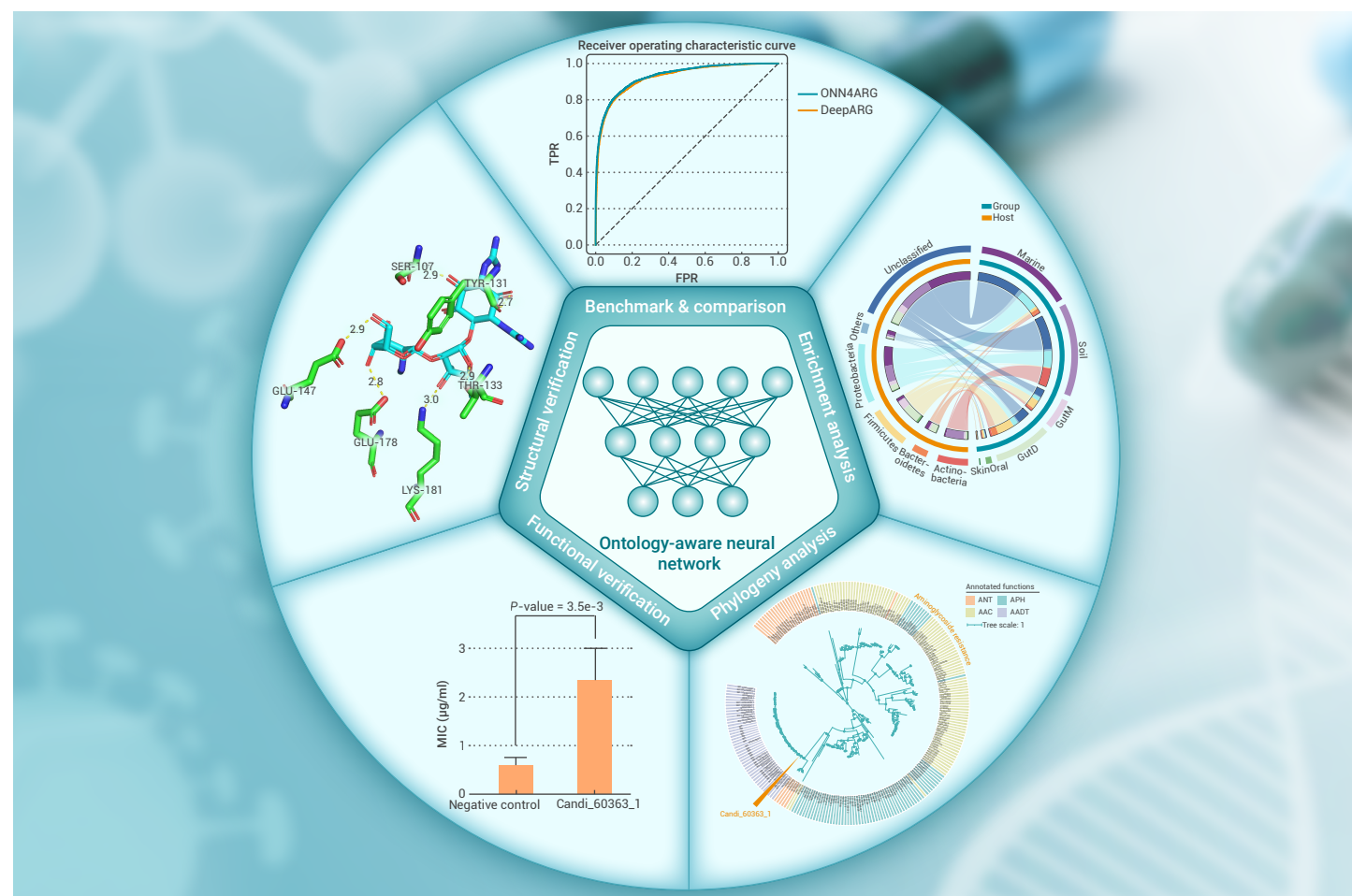
Yuguo Zha,¹ Cheng Chen,² Qihong Jiao,² Xiaomei Zeng,^{1*} Xuefeng Cui,^{2*} and Kang Ning^{1,*}

*Correspondence: xmzeng@hust.edu.cn (X.Z.); xfcui@email.sdu.edu.cn (X.C.); ningkang@hust.edu.cn (K.N.)

Received: November 5, 2023; Accepted: January 23, 2024; Published Online: February 28, 2024; <https://doi.org/10.59717/j.innn-life.2024.100054>

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- ONN4ARG is an ontology-aware neural network model for antibiotic resistance gene (ARG) prediction.
- ONN4ARG detects remotely homologous ARGs and generates a comprehensive set of ARGs with high fidelity.
- ONN4ARG enables comprehensive enrichment analysis of ARGs among diverse hosts and environments.
- Computational work and wet-lab experimental validation together validated a novel streptomycin ARG.



Comprehensive profiling of antibiotic resistance genes in diverse environments and novel function discovery

Yuguo Zha,¹ Cheng Chen,² Qihong Jiao,² Xiaomei Zeng,^{1,*} Xuefeng Cui,^{2,*} and Kang Ning^{1,*}

¹Key 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, China

²School of Computer Science and Technology, Shandong University, Qingdao 266237, China

*Correspondence: xmzeng@hust.edu.cn (X.Z.); xfcui@email.sdu.edu.cn (X.C.); ningkang@hust.edu.cn (K.N.)

Received: November 5, 2023; Accepted: January 23, 2024; Published Online: February 28, 2024; <https://doi.org/10.59717/j.xinn-life.2024.100054>

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Zha Y., Chen C., Jiao Q., et al., (2024). Comprehensive profiling of antibiotic resistance genes in diverse environments and novel function discovery. *The Innovation Life* **2**(1): 100054.

Antibiotic resistance genes (ARGs) have emerged in pathogens and are arousing worldwide concern, and accurately identifying unknown ARGs is a formidable challenge in studying the generation and spread of antibiotic resistance in diverse environments. Current methods can identify known ARGs but have limited utility for the discovery of novel ARGs, thus rendering the profiling of ARGs incomprehensive. Here, we developed ONN4ARG, an ontology-aware deep learning approach for comprehensive ARG discovery. Systematic evaluation revealed that ONN4ARG outperforms previous methods in terms of efficiency, accuracy, and comprehensiveness. Experiments using 200 million microbial genes collected from 815 metagenomic samples from diverse environments or hosts have resulted in 120,726 candidate ARGs, of which more than 20% are not yet present in public databases. The comprehensive set of ARGs revealed environment-specific and host-specific patterns. The wet-lab functional validation together with structural investigation have validated a novel streptomycin resistance gene from oral microbiome samples, confirming ONN4ARG's ability to discover novel functions. In summary, ONN4ARG enables comprehensive ARG discovery toward a grand view of ARGs worldwide.

INTRODUCTION

With the development of metagenomics and next-generation sequencing, many new microbial genes have been discovered, but different kinds of “unknowns” remain.¹ The functional diversity of microbiomes has not been fully explored, and approximately 40% of microbial gene functions have yet to be discovered.² A typical example is the antibiotic resistance gene (ARG), which is an urgent and growing threat to public health.³ In the past few decades, problems caused by antibiotic resistance have drawn the public's attention.⁴ Antibiotic resistance data are an ever-expanding data source, with many new ARG families being discovered in recent years.^{5,6} The discovery of resistance genes in diverse environments offers possibilities for early surveillance, action to reduce transmission, gene-based diagnostics, and improved treatment.⁷

Existing annotated ARGs have been curated manually or automatically for decades. Presently, there are 4,661 annotated ARGs in the CARD^{5,6} (v3.2.5, released in September 2022), 3,131 in the ResFinder⁸ (as of December 2022), and 2,476 in the SwissProt database⁹ (as of December 2022). Current ARG databases are far from complete: although no ARG database contains more than 4,000 well-annotated ARGs, NCBI non-redundant database searches yielded more than 9,000 putative genes annotated with “antibiotic resistance” as of December 2022. Therefore, we deemed that there is a large gap between the genes annotated in ARG databases and the possible ARGs that already exist in general databases, not to mention ARGs that are not yet annotated.

Many ARG prediction tools have been proposed in the past few years.^{8,10-20} These tools can generally be divided into two approaches. One approach is sequence-alignment, such as BLAST,²¹ USEARCH,²² and Diamond,²³ which use homologous genes to annotate unclassified genes. The other approach includes deep learning methods, such as DeepARG¹² and HMD-ARG,¹⁶ which use neural network models to predict and annotate ARGs. Several limitations still preclude comprehensive profiling of ARGs. A more comprehensive set of ARGs could be roughly defined as having more ARGs in type and number with fewer false-positive entries, regardless of the homology with known

ARGs, and many of these ARGs could be experimentally validated. Based on this definition, existing tools fall short in comprehensive profiling of ARGs. First, existing tools are limited to a few types of ARGs because the datasets used for building models are specialized. For example, HMD-ARG¹⁶ identifies only 15 types of resistance genes, and PATRIC¹³ is limited to identifying ARGs encoding resistance to carbapenem, methicillin, and beta-lactam antibiotics. Second, existing tools fail to discover novel ARGs, which usually lack homology to known sequences in reference databases. For instance, the *VraR* gene, which confers resistance to vancomycin, has a sequence identity of only 24% with the homolog from the CARD.¹² Therefore, there is an urgent need for a new approach to address these limitations.

Here, we propose an ontology-aware deep learning approach, ONN4ARG, which allows comprehensive identification of ARGs. Systematic evaluation of reference datasets revealed that the ONN4ARG model outperforms state-of-the-art models such as DeepARG, especially for the detection of remotely homologous ARGs. Experiments based on more than 200 million candidate microbial genes collected from 815 samples in various environments have resulted in 120,726 candidate ARGs, of which more than 20% are not yet present in public databases. Our experiments confirmed that ARGs are both environment-specific and host-specific, as exemplified by the enrichment of rifamycin resistance genes in Actinobacteria and in the soil environment. A case study of tens of experimentally validated ARGs also verified the ability of ONN4ARG to discover novel ARGs. We also validated a novel streptomycin ARG from oral microbiome samples via wet-lab experiments. In summary, ONN4ARG enables comprehensive ARG discovery, which provides a relatively complete picture of the prevalence of ARGs and leads to a comprehensive view of ARGs worldwide.

MATERIALS AND METHODS

Datasets

The ARGs used in this study for model training and testing were obtained from the Comprehensive Antibiotic Resistance Database (CARD) v3.0.3.^{5,6} We also used protein sequences from the UniProt (SwissProt and TrEMBL) database, which contains an ARG dataset (ONN4ARG-DB) and a non-ARG dataset, to expand our training dataset. First, genes with ARG annotations were collected from the CARD (2,587 ARGs) and SwissProt (2,261 ARGs) databases. Notably, to remove redundancy, the genes collected from the CARD and SwissProt were clustered with a sequence identity threshold of 90% using the MMseqs2 tool (version 10), and representative genes were retained. Then, their close homologs (sequence identity > 90% and coverage > 98%) were collected from TrEMBL (23,728 homologous genes). These annotated and homologous ARGs made up our ARG dataset. Finally, redundant genes with identical sequences were filtered out. As a result, our ARG dataset, namely, ONN4ARG-DB, contained 28,396 ARGs. The non-ARG dataset was constructed from non-ARGs that had relatively low sequence similarity to ARGs (sequence identity < 90% and bit-scores < alignment lengths) without resistance annotation (by text mining and manual curation) in SwissProt. Notably, to remove redundancy, the non-ARGs collected from SwissProt were subjected to clustering with a sequence identity threshold of 90% using the MMseqs2 tool (version 10), and 17,937 representative non-ARGs were retained.

The gene clustering of the 681 newly added ARGs in CARD v3.1.3 was performed using the MMseqs2 tool (version 10), with an identity of 90% and

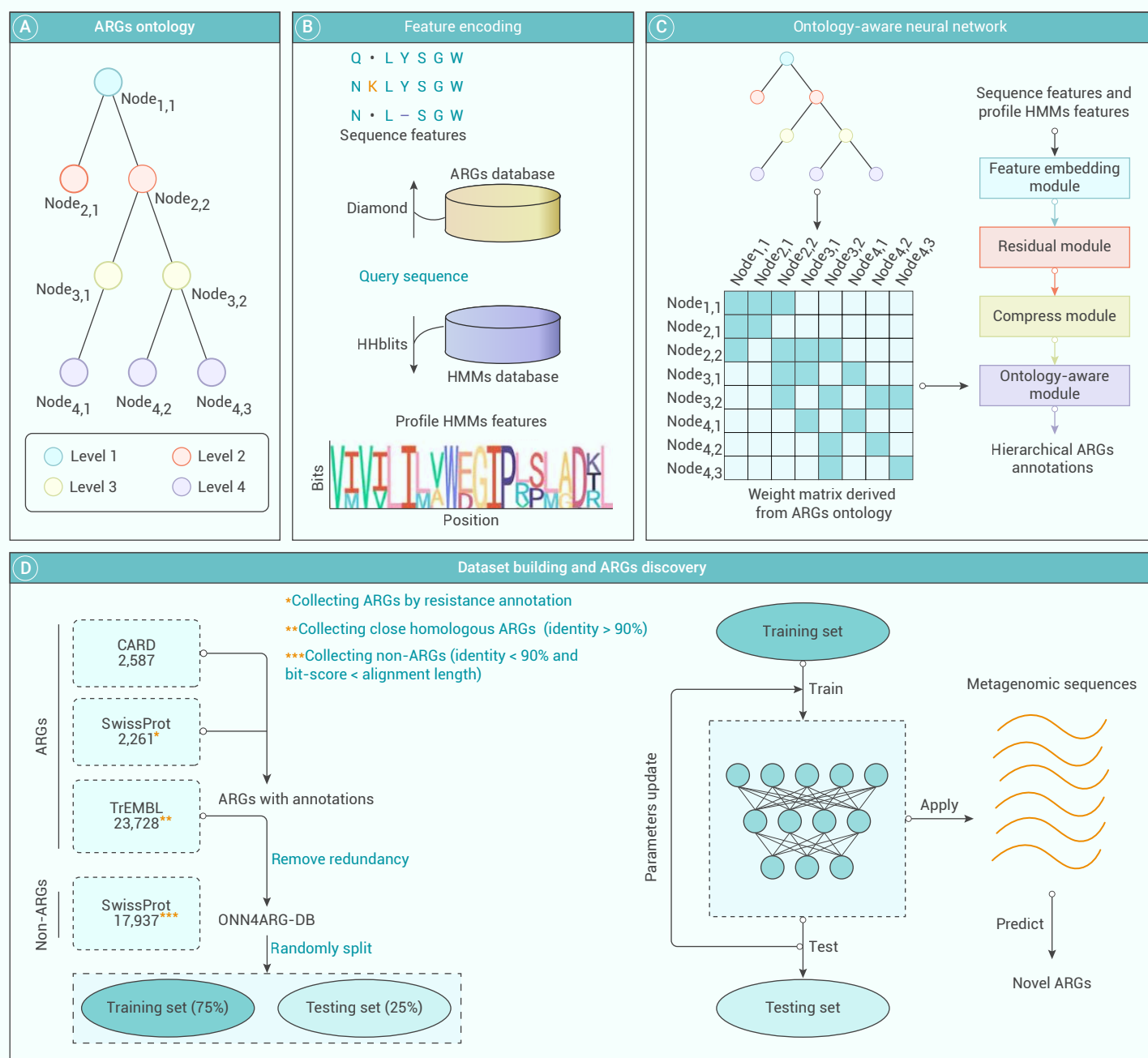


Figure 1. The ONN4ARG model employs an ontology-aware neural network and shows potential for novel ARG discovery in metagenomic data (A) The antibiotic resistance gene ontology term contains four levels. The root (first level) is a single node, namely, "arg". There are 1, 2, 34, and 277 nodes from the first level to the fourth level, respectively. (B) The feature encoding procedure of the ONN4ARG model. The sequence alignment features and profile HMM features are encoded by calling Diamond and HHblits. (C) The architecture of the ontology-aware neural network can be described in four functional modules, namely, the feature embedding module, residual module, compress module and ontology-aware module. The ontology-aware module is a partially connected layer that encourages annotation predictions satisfying the ontology rules. Specifically, weights between nodes with relationships (e.g., parent and child) satisfying the ontology rules are saved in the partially connected layer, and weights between irrelevant nodes are masked. (D) Building the dataset for training and testing and applying the ONN4ARG model to metagenomic samples to discover candidate novel ARGs.

coverage of 98%. The ResFinder dataset was obtained in June 2022 from https://bitbucket.org/genomicepidemiology/resfinder_db/src/master/.

Antibiotic resistance ontology

The ARG ontology contains a list of types of antibiotic resistance (drug class) and was organized into a four-level tree structure (a graph containing no cycles), in which higher-level resistance types cover lower-level resistance types (Figure 1A). There are 1, 2, 34, and 277 nodes (terms) from the first level to the fourth level, respectively. For instance, the root (first level) is a single node; namely, "arg", there are "beta-lactam" and "non-beta-lactam" in the second level, "acridine dye" and "aminocoumarin" in the third level, and "acriflavine" and "clorobiocin" in the fourth level.

ONN4ARG framework

ONN4ARG model. Considering a query gene q represented by its protein sequence, as well as its potential resistance categories represented by the antibiotic resistance ontology O , to predict resistance categories $\hat{y}_q$ of the query gene q , we employed an ontology-aware neural network to learn a mapping M from a set of genes $b \in S$ to their resistance categories $\hat{y}_b = (y_b^1, y_b^2, y_b^3, y_b^4)$. Here, S is the set of genes (i.e., ONN4ARG-DB), and y_b^i is the resistance category for gene b in the first level of the antibiotic resistance ontology. Then, we applied M to q to determine the potential resistance categories of the query gene.

$$\hat{y}_q = (y_q^i)_{1 \leq i \leq 4} = M(q)$$

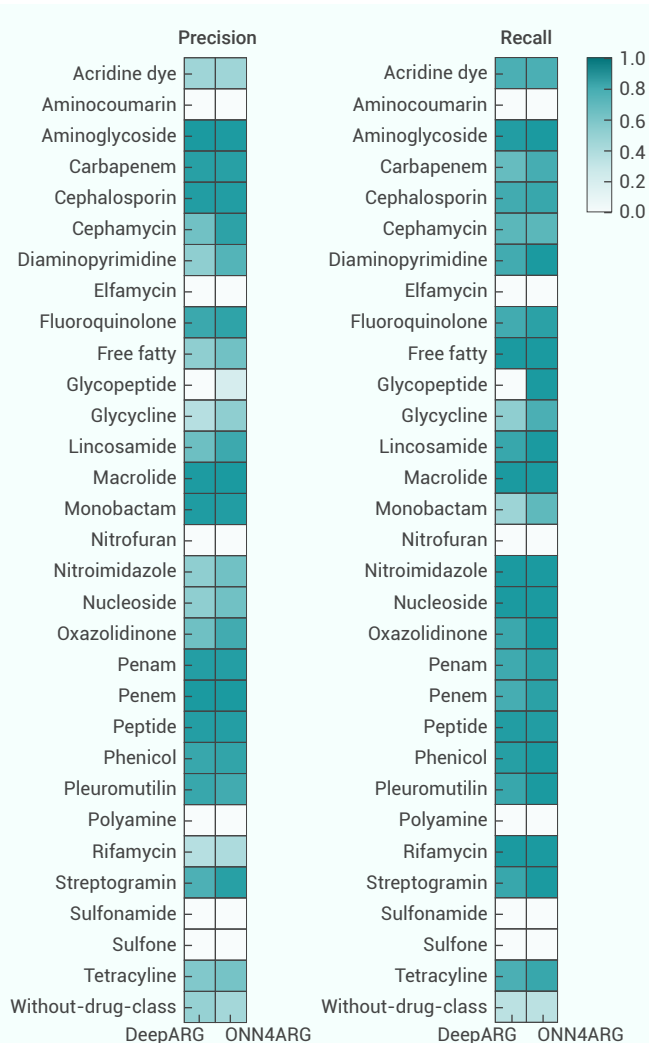


Figure 2. ONN4ARG outperformed DeepARG in terms of precision and recall The precision and recall of DeepARG and ONN4ARG for ARG classification for each resistance type at the second ARG ontology level. The masking threshold of the testing set was 0.4 (details of the masking threshold are provided in the Methods).

Feature encoding. The task of feature encoding is to abstract the homologous signal of a query gene. ONN4ARG takes homologous signals (e.g., identity, e-value, bit-score) between the protein sequence of the query gene and protein sequences and profiles (i.e., position-specific scoring matrix) of genes in ONN4ARG-DB (referred to as source genes) as features. The homologous signal abstraction procedure is as follows. First, a protein sequence library of the base genes was constructed by using the “makedb” function of Diamond software. The protein sequences of the query genes and base genes were subsequently aligned by using the “BLASTP” function of the Diamond program, version 0.9.10 (Figure 1B). Second, profile hidden Markov models (HMMs) of the source genes were generated by using the “HHblits” function of HH-suite3 software (version 3.2.0). The protein sequences of the query genes and profile HMMs of the source genes were subsequently aligned by using the “HHblits” function of HH-suite3 software (Figure 1B). Third, these homologous signals were normalized (i.e., divided by alignment length) and saved as numeric vectors for the input of the ONN4ARG model. The length of the numeric vector input to the ONN4ARG model is determined based on the number of sequences and profile HMMs in the ONN4ARG-DB. Specifically, the lengths of the numeric vectors of the sequence features are 25,868 and 9,564 for the profile HMM features.

Architecture of the ontology-aware neural network. PyTorch version 1.7.1 was used for generating the ONN model. The architecture of the ontology-aware neural network can be described by four functional modules, namely, the feature embedding module, residual module, compression

module and ontology-aware module (Figure S1).

First, the feature embedding module comprises two fully connected layers with group normalization⁵⁰ and GELU activation.⁵¹ This module accepts the flattened features (e.g., sequence alignment features) and outputs a highly informative embedding vector of size 1024. Second, the residual module is used to improve the generalization ability and avoid degradation of the neural network model. It comprises four self-defined residual layers similar to the ResNet proposed by He et al.⁵² Third, the compression module is a simple fully connected layer with group normalization and GELU activation, which is used to change the size of the embedding vector. For example, the output embedding vector of the residual layer is 1,024, and it is compressed to the size of 193 (the number of terms in the ARG ontology). Finally, the ontology-aware module is a partially connected layer that encourages annotation predictions satisfying the hierarchical structure of the ARG ontology. Specifically, weights between terms with relationships (e.g., parent and child) are preserved in the partially connected layer, and weights between irrelevant terms are masked to zero. The ontology-aware layer accepts an embedding vector V_{pre} . Then, the “Layer Norm” and “GELU” functions are applied to the vector V_{pre} . Finally, the partially connected layer is applied to generate a vector V_{post} . The output of the ontology-aware layer is $V_{pre} + V_{post}$.

Training and testing. We performed 4-fold cross-validation in the systematic evaluation of the ONN4ARG model. For each fold, we divided the ONN4ARG-DB into a training set and a testing set. The training set included 75% of the genes randomly selected from the ONN4ARG-DB, whereas the remaining 25% of the genes were selected as the testing set. We created a binary label vector for each protein sequence. If a protein sequence is annotated with a resistance type from the ontology, then we assign 1 to the type's position in the binary label vector. Otherwise, we assign 0.

Masking threshold. To simulate remotely homologous ARGs in our experiments, homologous signals between the query protein and its close homologs with sequence identities greater than a certain threshold were masked as zeros (i.e., no signals). For instance, when the masking threshold of the testing set was 0.4, homologous signals between the query protein (in the testing set) and its close homologs (in the training set) with sequence identities greater than 40% were masked as zeros. Occasionally, all homologs were masked for a query protein, and such query proteins were removed during testing. For example, if query X had two homologs, M and N , and assuming that the identity of M is 0.45 and the identity of N is 0.95, when the masking threshold of the testing set equaled 0.9, homologous signals between query X and homolog N were masked as zeros. When the masking threshold of the testing set equaled 0.4, query X was removed during testing. In the cross-validation experiment, 2,444, 103, and 17 homolog queries were removed from the testing datasets at masking thresholds of 0.4, 0.7, and 0.9, respectively.

Other methods

We used Diamond (version 0.9.0)²³ as the sequence-alignment tool for comparison. We used the same training and testing sets as in the ONN4ARG model to evaluate the sequence-alignment method. We searched for queries in the testing set against the training set. The target with the highest identity was defined as the closest homologous gene for each query. Then, we compared whether the actual annotation of the query was consistent with the annotation of its closest homologous gene to evaluate the performance. DeepARG¹² is a newly developed tool that applies a plain neural network (e.g., five fully connected layers) to predict ARGs. Here, we reconstructed the DeepARG model with PyTorch v1.7.1 by using the same architecture as the original DeepARG model and used the same training and testing sets as in the ONN4ARG model to train and test the DeepARG model. For queries in the testing set, we used the reconstructed DeepARG model to predict ARG annotations and compared whether the actual annotations were consistent with the predicted annotations to evaluate the performance.

Performance measures

To assess the performance of the ONN4ARG model and other methods, we used an accuracy metric with the following formula:

Table 1. Accuracy comparison of sequence-alignment, DeepARG and ONN4ARG based on different masking thresholds in the testing set.

Masking threshold	No. ARG genes	No. non-ARG genes	Sequence-alignment ^a	DeepARG	ONN4ARG-RO ^b	ONN4ARG
0.4	4916	2240	37.16%	89.85%	80.98%	94.26%
0.7	6401	3096	71.56%	97.29%	81.18%	98.50%
0.9	6487	3096	77.95%	99.22%	81.39%	99.10%
1.0	6504	3096	89.78%	99.20%	81.25%	98.95%
Overall	-	-	69.11%	96.39%	81.20%	97.70%

Note: The first column indicates different masking threshold of testing set; ^a, Sequence-alignment is based on Diamond; ^b, ONN4ARG model with random ontologies, the random ontologies are generated by randomly shuffling the terms on the lowest level of antibiotic resistance ontology; "-", not reported.

$$Accuracy = \frac{N_{corp}}{N_{pred}}$$

where N_{corp} is the number of correct predictions and N_{pred} is the total number of predictions. Notably, a prediction was defined as correct if and only if all ARG annotations (including ancestor annotations from ARG ontology) were correctly predicted.

Furthermore, we used precision, recall, F1, AUROC, and AUPRC measures to assess the performance of the ONN4ARG model and other methods for each antibiotic resistance type:

$$Precision(f) = \frac{TP(f)}{TP(f) + FP(f)}$$

$$Recall(f) = \frac{TP(f)}{TP(f) + FN(f)}$$

$$F1 = \frac{2 \times Precision(f) \times Recall(f)}{Precision(f) + Recall(f)}$$

$$TPR(f) = \frac{TP(f)}{TP(f) + FP(f)}$$

$$FPR(f) = \frac{FP(f)}{FP(f) + TN(f)}$$

where f represents one resistance type, $TP(f)$ is the number of true positive predictions of resistance type f , $FP(f)$ is the number of false positive predictions of resistance type f , $TN(f)$ is the number of true negative predictions of resistance type f , and $FN(f)$ is the number of false negative predictions of resistance type f . The AUROC is the area under the TPR - FPR curve, and the AUPRC is the area under the $Precision$ - $Recall$ curve.

Taxonomic annotation

The Kraken2 (version 2.1.2)⁵³ program with default parameters was used to identify the hosts of the gene contigs. Then, each ARG predicted by ONN4ARG was annotated according to the host of its gene contigs. The length range of the contigs containing ARGs was between 928 and 152,372 bases.

Acquired and intrinsic ARG annotation

Mobile genetic elements (MGEs) were explored to distinguish acquired and intrinsic ARGs. Specifically, contigs of predicted ARGs from the metagenomic samples were searched against the ACLAME⁴⁴ (<http://aclame.ulb.ac.be/>) database for plasmid annotation. The contigs containing predicted ARGs were searched for plasmid-like sequences against the ACLAME database using Diamond blastx with an identity of $\geq 90\%$. The predicted ARGs in contigs that aligned to the plasmids in the ACLAME database were classified as acquired ARGs, and ARGs in contigs not aligned to the plasmids in the ACLAME database were classified as intrinsic ARGs.

Phylogenetic tree

The protein sequences most closely related to Candi_60363_1 were

collected via "BLASTP" with default parameters from the NCBI non-redundant protein database. The retrieved proteins, Candi_60363_1 and all aminoglycoside resistance proteins from ResFinder⁸ (https://bitbucket.org/genomic-epidemiology/resfinder_db/src/master, last update Jun 2022), were aligned with ClustalW. The phylogenetic tree was constructed with MEGA⁵⁴ (v10) using the maximum likelihood algorithm with default parameters. The Interactive Tree of Life (iTOL v6) online tool⁵⁵ was used to construct the phylogenetic tree for display.

Protein model and docking

The Rosetta⁵⁶ was used to predict the protein structure via ab initio protein folding (<http://rosetta.bakerlab.org/>). The top five protein pockets were generated for docking calculations with Surface Topography of proteins⁵⁷ (CASTp). We used the Cambridge Structure Database⁵⁸ to generate streptomycin conformers. The 3D protein-ligand complexes were obtained from AutoDock Vina.⁵⁹

ARG candidate gene expression plasmid construction and expression verification

The candidate resistance gene Candi_60363_1 and the positive control resistance gene AHE40557.1 were synthesized and subcloned into pUC19 vector, replacing the *lacZ'* gene. The recombinant plasmids were subsequently transformed into *E. coli* BL21 (DE3). The expression of resistance genes was induced by Isopropyl β -D-1-thiogalactopyranoside (IPTG) and verified by quantitative Real-time PCR (qRT-PCR) assay. Briefly, bacteria were grown in LB supplemented with ampicillin (100 μ g/ml) to an OD₆₀₀ of 0.5-0.6 by incubation at 37°C with 220 rpm agitation, and the bacterial cultures continued to grow until the OD₆₀₀ reached 1.0 by the addition or absence of 1 mM IPTG. The cells were harvested, and total RNA was extracted using a Bacterial RNA Extraction Kit (Vazyme Biotech). RNA reverse transcription was performed by using a HiScript[®] II Q Select RT SuperMix for qPCR kit (Vazyme Biotech). qRT-PCR was performed by using SYBR Green Master Mix-High ROX Premixed (Vazyme Biotech) in a Stepone Plus system (Applied Biosystems). The *ldh* gene was used as an internal control in all reactions. The relative fold changes were determined using the $2^{-\Delta\Delta Ct}$ method, in which *ldh* was used for normalization.

MIC determination

The minimal inhibitory concentrations (MICs) of the antibiotics for the strains containing resistance genes were determined using E-tests (three replicates). Single colonies of the strains were incubated in 3 ml Mueller-Hinton (MH) medium with the addition of 100 μ g/ml ampicillin at 35 °C for 4 hours, and the cells equal to 1.5×10^8 cells/ml were spread on MH agar plates with the addition of 100 μ g/ml ampicillin and 1 mM IPTG, and streptomycin MIC Test Strips (Liofilchem[®]) were put in the middle of the plates. The plates were incubated at 35°C for 18-24 hours, after which the MICs were determined. The strain containing the empty vector was used as a negative control.

Statistical test

The normality of the data was verified by the Shapiro-Wilk test and Levene's test, and the ARG abundance data were distributed in a Gaussian manner



Figure 3. The length, abundance and novelty of the predicted ARGs varied among the diverse environments (A) The proportion of predicted ARGs for different protein sequence lengths. (B) The abundance ratio of predicted ARGs among diverse environments. The abundance ratio was defined as the number of ARGs divided by the number of total genes. (C) The proportion of predicted ARGs for different sequence identities among diverse environments. (D) Number of genes in the ONN4ARG-DB (left), predicted homologous ARGs (middle), and predicted novel ARGs (right) for various resistance types. The horizontal axis indicates the logarithmic number of genes, and the vertical axis indicates different antibiotic resistance types. We collected metagenomic samples from several published studies; these samples were mainly from "marine," "soil," and "human" environments. Human-associated samples consisted of two gut groups (one group from Madagascar, i.e., GutM; the other group from Denmark, i.e., GutD), one oral group, and one skin group (both oral and skin groups were from the HMP project).

with unequal variance. Thus, a statistical test of the enrichment analysis was performed utilizing Welch's *t* test (one-tailed) at a significance level of 0.005.⁶⁰ For all the tests, when the associated *P* value is lower than the significance level, one should reject the null hypothesis *H*₀ (ARGs are not enriched in the environment or host) and accept the alternative hypothesis *H*_a (ARGs are enriched in the environment or host).

RESULTS

The ONN4ARG model employs an ontology-aware neural network for ARG identification and classification

To address the large gap between the genes annotated in ARG databases and the possible ARGs that already exist in general databases, along with the ARGs that have not yet been annotated, we propose ONN4ARG, an ontology-

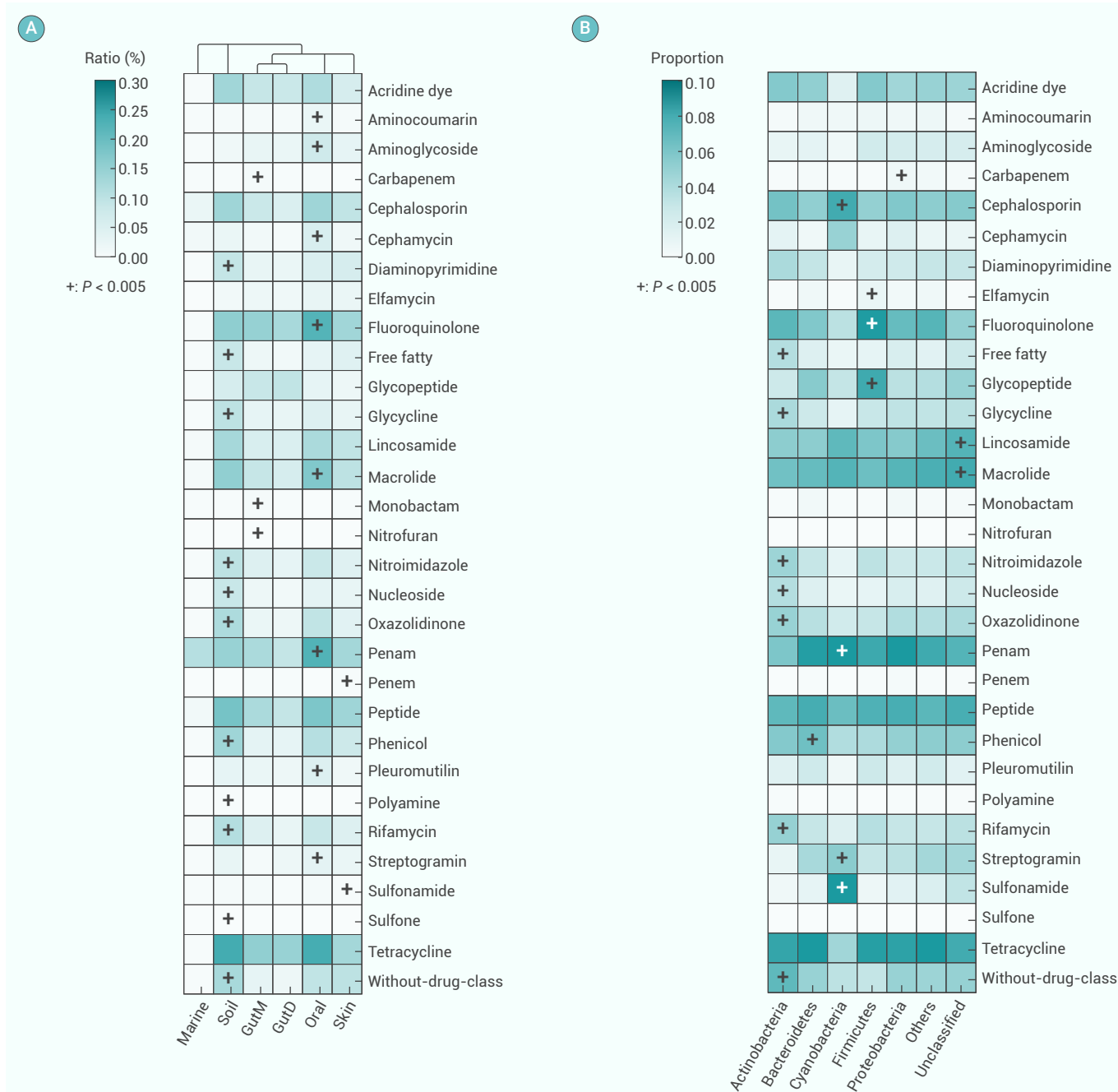


Figure 4. The predicted ARGs exhibited environment-specific and host-specific enrichment patterns (A) Relative abundance and enrichment of ARGs among diverse environments. The abundance ratio was defined as the number of ARGs divided by the number of total genes. (B) Proportion and enrichment of ARGs among diverse hosts. The colors indicate the proportion of ARGs for each phylum and resistance type. The results for the five most abundant phyla that carry ARGs are shown. "+": P value < 0.005 (Welch's *t* test, one-tailed).

aware neural network model that predicts ARGs in a comprehensive manner (Figure 1). Here, we described the ARG ontology as a list of resistance types (e.g., drug classes), which were organized into a four-level tree structure (i.e., a graph containing no cycles),²⁴ in which higher-level resistance types cover lower-level resistance types (Methods, Figure S1).

ONN4ARG has innovative designs for unleashing its power in the ARG prediction task. First, ONN4ARG employs a novel ontology-aware layer that incorporates ancestor and descendent annotations to enhance annotation accuracy (Figures 1A-C, Figure S1). The ontology-aware layer is a partially connected layer that encourages annotation predictions satisfying the ontology rules (i.e., the hierarchical structure of the antibiotic resistance ontology). Second, ONN4ARG uses both sequence and profile Hidden Markov Model (pHMM) alignment similarities as features, which could help capture the evolutionary signals among remotely homologous genes and identify

unknown ARGs.

ONN4ARG takes alignment similarities between the query gene and ARG as inputs (Figure 1B). These sequence-alignment similarities and profile-alignment similarities are preprocessed by calling Diamond and HHblits.²⁵ ONN4ARG outputs hierarchical annotations that are compatible with the ARG ontology (Figure 1A & Figure 1C). To train and evaluate our ONN4ARG model and for rapid deployment of ARG discovery in multiple contexts, we also constructed an ARG database (Figure 1D), namely, ONN4ARG-DB (Methods).

Systematic evaluation and comparison

Systematic evaluation revealed that our model is highly efficient, accurate, and comprehensive for ARG identification. ONN4ARG is fast since it can complete ARG identification for all genes in the testing dataset within half an hour, which is equivalent to the identification of ten genes per second using a

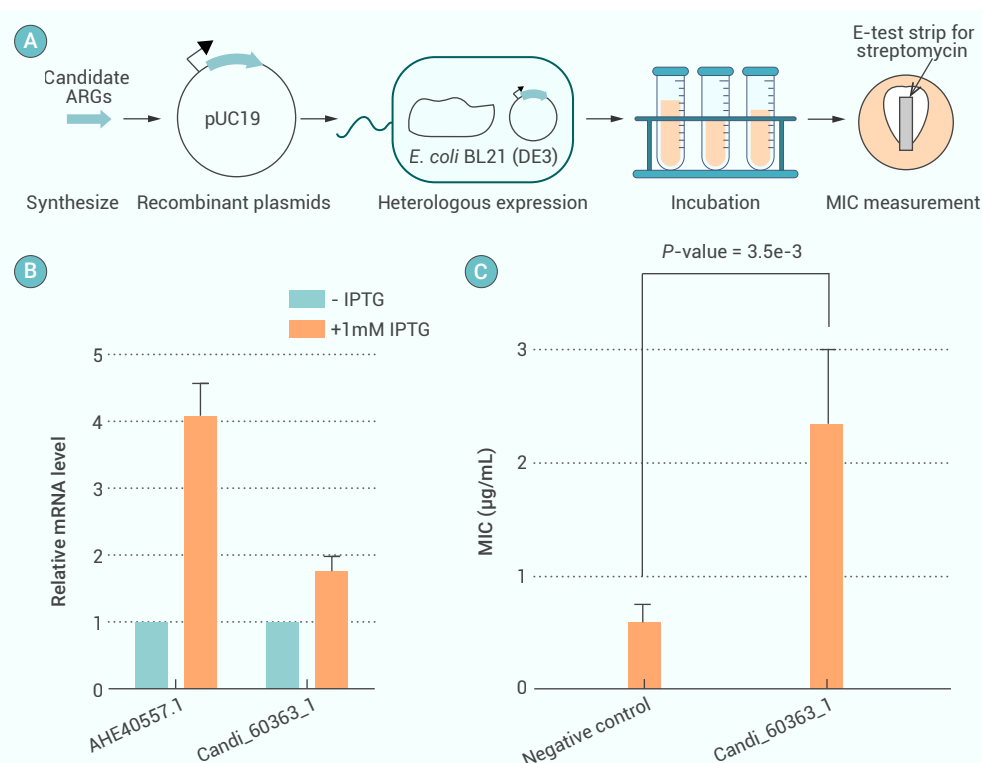


Figure 5. Functional validation of a candidate novel ARG (Candi_60363_1) revealed an increased minimal inhibitory concentration (MIC) of streptomycin (A) A diagram showing the procedure of heterologous expression and functional analysis of the predicted candidate ARG in the *E. coli* BL21 (DE3) host. (B) Gene expression validation of the predicted candidate ARG. The vertical axis indicates the relative mRNA level. (C) The MIC of the predicted candidate ARG and negative control. The vertical axis indicates the MIC value. The MIC of the predicted candidate novel ARG was significantly greater than that of the negative control (Welch's *t* test, one-tailed, *P* value = 3.5e-3).

cially those with low homology to the existing ARG database.

ARG mining of metagenomic data from diverse environments

We collected contig data from metagenomic samples from several published studies,^{28,29} and these samples were from different biomes (niches), including "marine", "soil", "human gut", "human oral", and "human skin". We considered each biome (niche) as an environment and investigated the distribution patterns of ARGs in these different types of environments. Details of these samples are provided in Table S5. Then, 240 million genes

single core. ONN4ARG achieved an overall precision of 75.59% and an overall recall of 89.93%, which were higher than those of DeepARG (68.30% and 77.84%), respectively (Figure 2). Furthermore, ONN4ARG was more accurate at identifying ARGs (overall accuracy of 97.70%; Table 1) than sequence alignment (overall accuracy of 69.11%), and ONN4ARG had only a slight advantage over DeepARG (overall accuracy of 96.39%). Moreover, ONN4ARG has advantages over other methods for identifying remotely homologous ARGs whose sequences are not similar to existing ARG sequences. For example, when tested with only remote homologs (i.e., the masking threshold of the testing set was equal to 0.4; see Methods), ONN4ARG achieved an accuracy of 94.26%, which was largely improved from the 89.85% accuracy of DeepARG (Table 1, Table S1). To test how much the ontology contributes to the performance of the neural network model, we evaluated the accuracy of ONN4ARG using random ontologies. The results showed that ONN4ARG had a large decrease in accuracy (16.50%) when using random ontologies (Table 1). These results validate ONN4ARG's better generalization ability than sequence-alignment and DeepARG, which makes ONN4ARG especially suitable for the identification of remotely homologous ARGs and indicates that ONN4ARG has the ability to identify novel ARGs.

ONN4ARG performed well in identifying known ARGs in existing ARG databases, including CARD and ResFinder. We first tested ONN4ARG in a validation dataset consisting of 681 newly added ARGs in the CARD database version 3.1.3. The results showed that our model outperformed the other methods in terms of accuracy and efficiency, given that the memory usage is acceptable for a regular laptop (Table S2). We also evaluated ONN4ARG on the ResFinder database version 4.1, which includes thousands of manually curated ARGs.⁸ The results showed that ONN4ARG achieved an accuracy higher than 90% for most types of resistance, while DeepARG was less accurate than ONN4ARG, except for fosfomycin (Table S3).

We further evaluated the ability of ONN4ARG to identify novel ARGs based on 13 recently experimentally validated ARGs.^{7,26,27} We searched for these 13 ARGs with both the DeepARG and ONN4ARG models. The results showed that ONN4ARG recognized 12 of these 13 ARGs and correctly classified their type of resistance (Table S4). For comparison, DeepARG identified only 8 of these 13 ARGs. Notably, among the four ARGs identified by ONN4ARG only, two had relatively low homology to known ARGs in the ONN4ARG-DB (NCBI GenBank accession MW234453 and OL806615, sequence identity <40%). Taken together, the ONN4ARG is suitable for identifying novel ARGs, espe-

were obtained from these samples by calling Prodigal³⁰ with default parameters. The ONN4ARG model was used to predict whether these unclassified genes were ARGs and their corresponding resistance types.

We investigated the broad-spectrum profile of these predicted ARGs among diverse environments. First, we investigated the proportion of predicted ARGs for different sequence lengths and found that more than half of the predicted ARGs had a length of 128–512 amino acid residues, which is close to the protein sequence length distribution of ARGs in the CARD database (Figure 3A & Figure S2). We also analyzed the protein domains of these predicted ARGs by searching the conserved domain database (CDD, last update Aug 2022) using the RPS-BLAST tool version 2.9.0. Most of these predicted ARGs (over 97%) had protein domains that were consistent with known catalytic activity and/or may bind to the antimicrobial agents they are predicted to elicit resistance against (Table S6). Second, we found that the ARG content of metagenomic samples from the human oral group was the highest (1.4%), followed by that of soil (1.1%) and marine (0.2%) samples (Figure 3B & Table S7). Previous studies have shown that the abundance of ARGs differs across environments. For example, a metagenomic study by Qian et al.³¹ revealed high ARG diversity (242 ARG subtypes) and abundance (0.184–0.242 ARG copies per 16S rRNA gene copy) in the soil ecosystem. The study by Ning et al.³² demonstrated that the ARG content of metagenomic samples from different niches varied, with the ARG content in sediment being the highest (0.35%), followed by that in water (0.2%). We should emphasize that although we have no exact knowledge of the ARG content for those contigs from real metagenomic samples, these relative abundances are meaningful and contribute to the knowledge of the results. Third, we tested the novelty of these predicted ARGs. We found that approximately one-third of the genes (42,848 out of all 120,726 ARGs) had a sequence identity of less than 40% with their homologs in the ONN4ARG-DB (Figure 3C). We defined these ARGs as candidate novel ARGs that had low sequence identities when aligned to their homologs in the reference database (i.e., ONN4ARG-DB). In total, 31 resistance types in the third level of ARG ontology were detected in these various environments, most of which were acquired ARGs (Figure 3D & Figure S3). The number of predicted ARGs for different resistance types varied greatly, from a few (i.e., nitrofurantoin) to thousands (i.e., fluoroquinolone). As expected, these abundant ARGs were usually associated with antibiotics used extensively in human activities, such as growth promoters in animals.^{33–35}

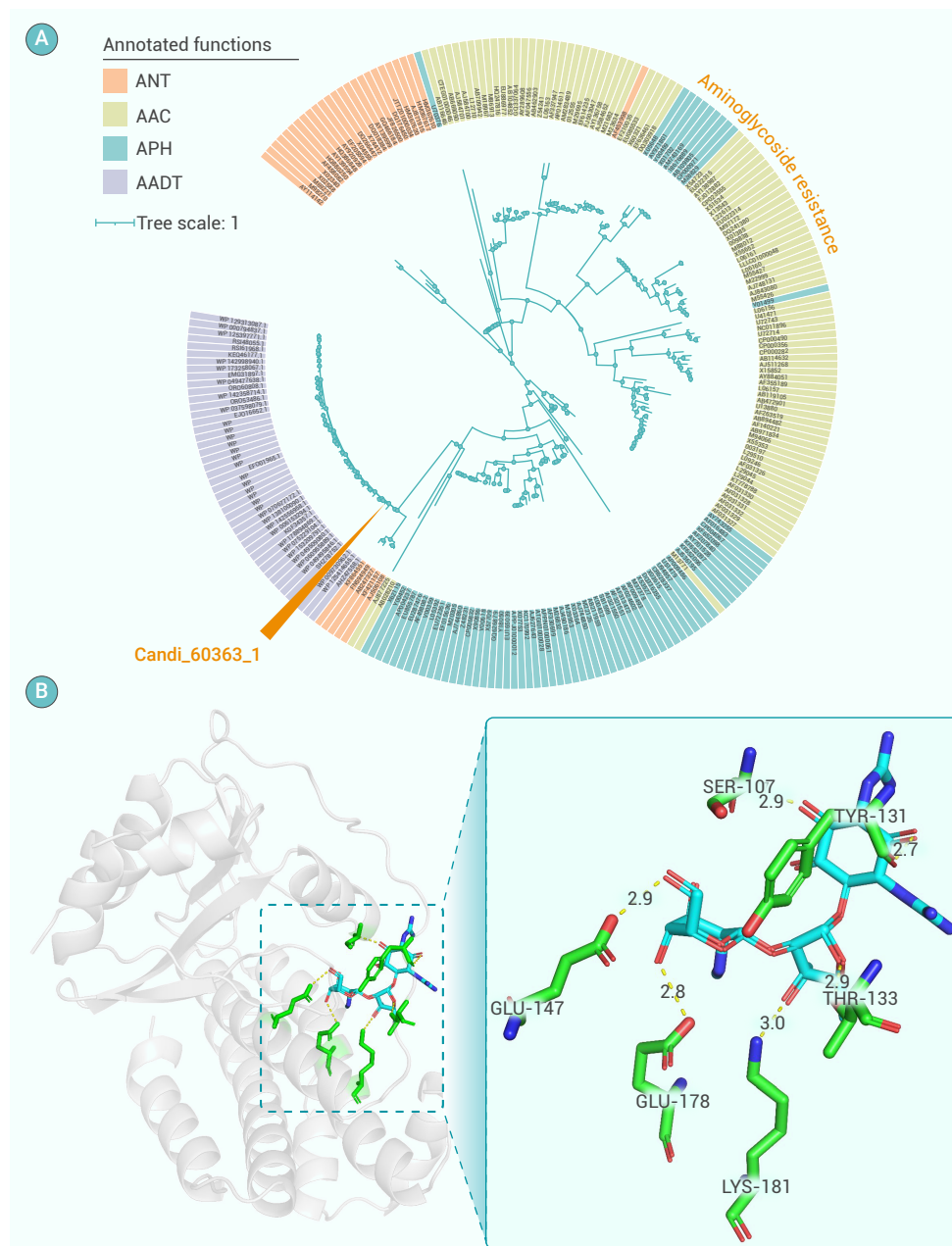


Figure 6. Phylogenetic analysis and structural investigation of Candi_60363_1 confirmed its close evolutionary relationship with known aminoglycoside adenyltransferases and showed great antibiotic resistance potential (A) Phylogenetic tree of the aminoglycoside resistance enzyme Candi_60363_1 and its homologs from the NCBI non-redundant protein database. ANT: O-nucleotidyltransferase, AAC: N-acetyltransferase, APH: O-phosphotransferase, AADT: aminoglycoside adenyltransferase. (B) The optimal Candi_60363_1-streptomycin complex structure (left) and the local interactions between the ligand and neighboring residues (right). The docking experiment revealed six neighboring residues whose distances were less than three angstroms.

proportion of 0.1%) and in Actinobacteria (with a proportion of 4.7%) (Figure 4 & Figure S6). Rifamycin is an important antibacterial agent that is active against gram-positive bacteria and has a wide range of applications.^{36,37} These enrichment results were not surprising because *Actinomycetes* is a representative genus widely distributed in various soil environments, and its resistance to rifamycin is compatible with its ability to produce rifamycin.³⁸⁻⁴¹

Functional verification of candidate novel resistance genes

To identify promising putative novel resistance genes, we used four criteria: had (i) remote homologs to reference ARGs, (ii) prediction with high confidence, (iii) prediction of single-type resistance, and (iv) known host. Despite the large number of candidate genes identified by the ONN4ARG model, only 4,365 ARGs underwent all the abovementioned criteria (Table S9). To showcase the actual function of the predicted ARGs, we analyzed six ARGs associated with streptomycin resistance (Table S10), and all of these ARGs were highly confidently predicted by the ONN4ARG model. The experimental results showed that Candi_60363_1 is one of the most promising ARGs and had a high minimal inhibitory concentration (MIC) of streptomycin compared to that of the negative control (Table S10).

Enrichment patterns of predicted ARGs among diverse environments and hosts

Rapid deciphering of potential antimicrobial-resistant pathogens is necessary for effective public health monitoring, and host tracking of ARGs allows accurate identification of pathogens. To track the hosts of these predicted ARGs, we performed a taxonomy analysis and discovered 949 genera harboring at least one type of ARG. (Table S8). The host distribution showed that these ARGs were primarily affiliated with Proteobacteria (38.2%), the most abundant of which were ARGs conferring drug resistance to fluoroquinolone, macrolide, peptide, penam, and tetracycline, accounting for approximately half of the total ARGs (Figure S4). Network inference based on strong (Spearman's $\rho > 0.8$) and significant (Welch's t test, P value < 0.01) correlations revealed co-occurrence patterns among ARGs and microbial taxa (Figure S5). For example, ARGs associated with beta-lactam resistance (e.g., cephamycin, penam, penem, and monobactam) were observed to occur together in Proteobacteria.

Enrichment analyses revealed that the ARGs were both environment-specific and host-specific (Figure 4). We found that the proportion of certain types of ARGs was greater in certain environments than in others. For example, rifamycin ARGs were found to be enriched in the soil environment (with a

However, the MICs of the other candidate ARGs were not significantly different from that of the negative control (Table S10). Thus, we selected Candi_60363_1 for further functional and structural validation (Table S11).

Candi_60363_1, detected in *Streptococcus* in the oral environment, was predicted to confer resistance to streptomycin (a subtype of aminoglycoside). One positive control (AHE40557.1, streptomycin resistant) from the CARD database was used for verification of the experimental system. All of these genes were heterologously expressed in the *E. coli* BL21 (DE3) host by the induction of Isopropyl β -D-1-thiogalactopyranoside (IPTG), after which the minimal inhibitory concentration (MIC) of streptomycin was tested (Figure 5A). The results showed that the mRNA level of the genes increased with the addition of 1 mM IPTG compared with that without IPTG (Figure 5B), which verified the expression of the genes induced by IPTG. Furthermore, the MIC for the strain containing the positive control gene AHE40557.1 was more than 1,024 $\mu\text{g/ml}$ (Figure S7), which is consistent with previous reports.^{42,43} This verified that our MIC measurement system works well. Our results showed that the MIC for the strain containing Candi_60363_1 was significantly greater than that for the negative control containing no insert (Welch's t test, one-tailed, P value $= 3.5\text{e-}3$), which demonstrated increased resistance to streptomycin by the novel candidate gene Candi_60363_1 (Figure 5C &

Figure S7).

There are remote similarities between Candi_60363_1 and all known ARGs in the reference database, including aminoglycoside resistance genes. The InterPro search results showed that the protein family matching Candi_60363_1 was IPR007530, which is also known as aminoglycoside 6-adenylyltransferase and confers resistance to aminoglycoside antibiotics. Then, we used BLAST to search for homologs of Candi_60363_1 from the NCBI non-redundant protein database. BLAST results revealed 44 homologs with sequence identities greater than 80%, and these homologs were from various organisms (Table S12), such as *Streptococcus oralis*, *Peptoniphilus lacrimalis* DNF00528, and *Mycobacteroides abscessus* subsp. Abscessus. Considering that Candi_60363_1 is harbored by distantly related species and that the contig containing Candi_60363_1 has hits in the ACLAME⁴⁴ plasmid database, Candi_60363_1 is likely to have mobility. Notably, the most similar protein, Candi_60363_1, from the NCBI non-redundant protein database (87.5% identity, SHZ78752.1) was also annotated as an aminoglycoside adenylyltransferase (Table S12). Taken together, these findings suggest that Candi_60363_1 is highly likely to be an ARG that confers resistance to aminoglycoside antibiotics.

Aminoglycoside-modifying enzymes are the most clinically important resistance mechanism against aminoglycosides.⁴⁵ These enzymes are divided into three enzymatic classes, namely, aminoglycoside N-acetyltransferase (AAC), O-nucleotidyltransferase (ANT), and O-phosphotransferase (APH). We investigated the phylogenetic relationship between Candi_60363_1 and known aminoglycoside-modifying enzymes. The phylogenetic tree of Candi_60363_1 and related proteins (Figure 6A) revealed that Candi_60363_1 is clearly separated from known aminoglycoside-modifying enzymes and is located among proteins mostly annotated as aminoglycoside adenylyltransferases. Phylogenetic analysis revealed the evolutionarily close relationships of this gene with known aminoglycoside adenylyltransferases.

The protein structure prediction results confirmed the functionality of Candi_60363_1. The optimal Candi_60363_1-streptomycin complex structure and the corresponding interaction details are described in Figure 6B. The optimal binding affinity between Candi_60363_1 and streptomycin was -7.7 kcal/mol (Table S13), which was 1.6 kcal/mol lower than that of the negative control. Wet-lab, phylogenetic, and protein structure docking analyses revealed that Candi_60363_1, predicted by ONN4ARG, is highly likely a real ARG.

DISCUSSION

In this study, we proposed an ontology-aware deep learning method, ONN4ARG, for the detection and understanding of ARGs. ONN4ARG employs a novel ontology-aware layer that incorporates ancestor and descendent annotations to increase annotation accuracy, enabling ONN4ARG to outperform the state-of-the-art DeepARG model, especially in the discovery of novel ARGs. To complement ONN4ARG for ARG mining applications, we also created a custom ARG database, ONN4ARG-DB, that contains 28,396 well-curated ARGs. ONN4ARG analysis revealed 120,726 ARGs in microbiome samples, 42,848 of which were novel, substantially expanding the existing ARG repositories.

The novelty of this work lies in three contexts. First, ONN4ARG has the potential to detect remotely homologous ARGs and generate a more comprehensive set of ARGs. The potential of ONN4ARG for detecting novel ARGs has been shown through the evaluation of 13 recently experimentally validated ARGs. ONN4ARG recognized 12 of these 13 ARGs and correctly classified their resistance types (Table S4), while DeepARG identified only eight. Notably, among the four ARGs identified by ONN4ARG only, two had relatively low homology to known ARGs in the ONN4ARG-DB (NCBI GenBank accession MW234453 and OL806615, sequence identity <40%). Furthermore, the antibiotic resistance ontology used in the ONN4ARG model consists of four levels and more than 100 resistance subtypes (i.e., terms at the most informative level on the ontology), which substantially expands the classification space of current tools.

Second, this approach enabled comprehensive enrichment analysis of ARGs among diverse environments and hosts. The environment-specific and host-specific enrichment of ARGs may be caused by specific bacteria evolving to possess specific types of ARGs in response to specific environments,

and horizontal gene transfer may be one of the mediating pathways of this process. For example, one published study reported that *Amycolatopsis* in the soil environment produces rifamycin and thus gains ecological advantages over other bacteria.³⁸

Third, our study demonstrated the importance and potential of complementing computational work with wet-lab experimental validation of gene function. Functional verification of a novel streptomycin resistance gene (i.e., Candi_60363_1) via wet-lab experiments demonstrated the ability of the ONN4ARG model to discover novel ARGs. Moreover, phylogenetic analysis and protein structure docking further confirmed that Candi_60363_1 is highly likely to be an ARG that confers resistance to aminoglycoside antibiotics.

We propose an ontology-aware deep learning approach, ONN4ARG, which is superior to existing methods such as DeepARG in terms of efficiency, accuracy, and comprehensiveness. It has detected novel ARGs that are remotely homologous to existing ARGs. Whereas Although ONN4ARG provides one of the most comprehensive profiles of ARGs, it could be further optimized. For more comprehensive ARG prediction, continuous improvement of the curating ARG nomenclature and annotation databases is needed. For novel ARG prediction, especially for those belonging to entirely new ARG families, deep learning models might need to consider more information other than sequence alone, such as protein structure. It has been shown that protein structure information can complement protein sequence information to predict protein function.⁴⁶ Computational methods such as AlphaFold generate high-quality protein structures but fail to predict protein families with few homologous sequences,^{47,48} which can be addressed by targeted recruitment of homologous sequences from metagenomic data.⁴⁹ We believe these efforts could lead to a holistic view of ARGs in diverse environments around the globe.

REFERENCES

- Zha, Y., Chong, H., Yang, P., et al. (2022). Microbial dark matter: from discovery to applications. *Genomics, Proteomics & Bioinformatics* **20**: 867–881. DOI: 10.1016/j.gpb.2022.02.007.
- Li, J., Jia, H., Cai, X., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**: 834–841. DOI: 10.1038/nbt.2942.
- Brogan, D.M., and Mossialos, E. (2016). A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *Global Health* **12**: 8–8. DOI: 10.1186/s12992-016-0147-y.
- Goossens, H., Ferech, M., Stichele, R.V., et al. (2005). Outpatient antibiotic use in Europe and association with resistance: A cross-national database study. *The Lancet* **365**: 579–587. DOI: 10.1016/S0140-6736(05)17907-0.
- Jia, B., Raphenya, A.R., Alcock, B., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**: D566–D573. DOI: 10.1093/nar/gkw1004.
- Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., et al. (2019). CARD 2020: Antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**: D517–D525. DOI: 10.1093/nar/gkz935.
- Böhm, M.-E., Razavi, M., Marathe, N.P., et al. (2020). Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities. *Microbiome* **8**: 1–11. DOI: 10.1186/s40168-020-00814-z.
- Bortolai, V., Kaas, R.S., Ruppe, E., et al. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**: 3491–3500. DOI: 10.1093/jac/dkaa345.
- Bateman, A., Martin, M.J., O'Donovan, C., et al. (2015). UniProt: A hub for protein information. *Nucleic Acids Res* **43**(D1): D204–D212. DOI: 10.1093/nar/gku989.
- Rowe, W., Baker, K.S., Verner-Jeffreys, D., et al. (2015). Search engine for antimicrobial resistance: a cloud compatible pipeline and web interface for rapidly detecting antimicrobial resistance genes directly from sequence data. *PLoS One* **10**: e0133492. DOI: 10.1371/journal.pone.0133492.
- Kleinheinz, K.A., Joensen, K.G., and Larsen, M.V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage* **4**: e27943. doi:10.4161/bact.27943.
- Arango-Argoty, G., Garner, E., Pruden, A., et al. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**: 23. DOI: 10.1186/s40168-018-0401-z.
- Davis, J.J., Boisvert, S., Brettn, T., et al. (2016). Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.* **6**: 27930. DOI: 10.1038/srep27930.
- Lakin, S.M., Kuhnle, A., Alipanahi, B., et al. (2019). Hierarchical Hidden Markov models enable accurate and diverse detection of antimicrobial resistance sequences. *Commun. Biol.* **2**: 294. DOI: 10.1038/s42003-019-0545-9.
- Doster, E., Lakin, S.M., Dean, C.J., et al. (2020). MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.* **48**: D561–D569. DOI:

- 10.1093/nar/gkz1010.
16. Li, Y., Xu, Z., Han, W., et al. (2021). HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome* **9**: 40. DOI: 10.1186/s40168-021-01002-3.
 17. Gupta, S.K., Padmanabhan, B.R., Diene, S.M., et al. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**: 212–220. DOI: 10.1128/AAC.01310-13.
 18. Feldgarden, M., Brover, V., Haft, D.H., et al. (2019). Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**: e00483–00419. DOI: 10.1128/AAC.00483-19.
 19. Inouye, M., Dashnow, H., Raven, L.A., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome. Med.* **6**: 90. DOI: 10.1186/s13073-014-0090-6.
 20. Rowe, W.P.M. and Winn, M.D. (2018). Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics* **34**: 3601–3608. DOI: 10.1093/bioinformatics/bty387.
 21. Altschul, S.F., Gish, W., Miller, W.C., et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410. DOI: 10.1016/S0022-2836(05)80360-2.
 22. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461. DOI: 10.1093/bioinformatics/btq461.
 23. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**: 59–60. DOI: 10.1038/nmeth.3176.
 24. Lakin, S.M., Dean, C., Noyes, N.R., et al. (2017). MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* **45**: D574–D580. DOI: 10.1093/nar/gkw1009.
 25. Steinegger, M., Meier, M., Mirdita, M., et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**: 1–15. DOI: 10.1186/s12859-019-3019-7.
 26. Wang, X.-R., Lian, X.-L., Su, T.-T., et al. (2021). Duck wastes as a potential reservoir of novel antibiotic resistance genes. *Sci. Total Environ.* **771**: 144828. DOI: 10.1016/j.scitotenv.2020.144828.
 27. Wang, S., Wei, L., Gao, Y., et al. (2022). Novel amikacin resistance genes identified from human gut microbiota by functional metagenomics. *J. Appl. Microbiol.* **133**: 898–907. DOI: 10.1111/jam.15615.
 28. Sunagawa, S., Coelho, L.P., Chaffron, S., et al. (2015). Structure and function of the global ocean microbiome. *Science* **348**: 1261359. DOI: 10.1126/science.1261359.
 29. Mitchell, A.L., Scheremetjew, M., Denise, H., et al. (2018). EBI metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**: D726–D735. DOI: 10.1093/nar/gkx967.
 30. Hyatt, D., Chen, G.L., LoCascio, P.F., et al. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**: 119. DOI: 10.1186/1471-2105-11-119.
 31. Qian, X., Gunturu, S., Guo, J., et al. (2021). Metagenomic analysis reveals the shared and distinct features of the soil resistome across tundra, temperate prairie, and tropical ecosystems. *Microbiome* **9**: 108. DOI: 10.1186/s40168-021-01047-4.
 32. Ning, K., Ji, L., Zhang, L., et al. (2022). Is rice-crayfish co-culture a better aquaculture model: from the perspective of antibiotic resistome profiles. *Environ. Pollut.* **292**: 118450. DOI: 10.1016/j.envpol.2021.118450.
 33. Li, B., Yang, Y., Ma, L., et al. (2015). Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J.* **9**: 2490–2502. DOI: 10.1038/ismej.2015.59.
 34. Plata, G., Baxter, N.T., Susanti, D., et al. (2022). Growth promotion and antibiotic induced metabolic shifts in the chicken gut microbiome. *Commun. Biol.* **5**: 293. DOI: 10.1038/s42003-022-03239-6.
 35. Wang, L., Zhang, Y., Xu, J., et al. (2023). Listening to enteric bacteria from the perspective of antibiotic alternatives in animal husbandry. *The Innovation Life* **1**: 100022. DOI: 10.59717/j.xinn-life.2023.100022.
 36. Qi, F., Lei, C., Li, F., et al. (2018). Deciphering the late steps of rifamycin biosynthesis. *Nat. Commun.* **9**: 2342. DOI: 10.1038/s41467-018-04772-x.
 37. Floss, H.G., and Yu, T.-W. (2005). Rifamycin-mode of action, resistance, and biosynthesis. *Chem. Rev.* **105**: 621–632. DOI: 10.1021/cr030112j.
 38. Yao, Y., Zhang, W., Jiao, R., et al. (2002). Efficient isolation of total RNA from antibiotic-producing bacterium *Amycolatopsis mediterranei*. *J. Microbiol. Methods* **51**: 191–195. DOI: 10.1016/S0167-7012(02)00078-7.
 39. Wilson, M.C., Gulder, T.A.M., Mahmud, T., et al. (2010). Shared biosynthesis of the saliniketals and rifamycins in *Salinispora arenicola* is controlled by the sare1259-encoded cytochrome P450. *J. Am. Chem. Soc.* **132**: 12757–12765. DOI: 10.1021/ja105891a.
 40. Saxena, A., Kumari, R., Mukherjee, U., et al. (2014). Draft genome sequence of the rifamycin producer *Amycolatopsis rifamycinica* DSM 46095. *Genome Announc.* **2**: e00662–00614. DOI: 10.1128/genomea.00662-14.
 41. Huang, H., Lv, J., Hu, Y., et al. (2008). *Micromonospora rifamycinica* sp. nov., a novel actinomycete from mangrove sediment. *Int. J. Syst. Evol. Microbiol.* **58**: 17–20. DOI: 10.1099/ijs.0.64484-0.
 42. Pinto-Alphandary, H., Mabilat, C., and Courvalin, P. (1990). Emergence of aminoglycoside resistance genes aadA and aadE in the genus *Campylobacter*. *Antimicrob. Agents Chemother.* **34**: 1294–1296. DOI: 10.1128/AAC.34.6.1294.
 43. Holden, M.T.G., Hauser, H., Sanders, M., et al. (2009). Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PLoS One* **4**: e6072. DOI: 10.1371/journal.pone.0006072.
 44. Leplae, R., Hebrant, A., Wodak, S.J., et al. (2004). ACLAME: A classification of mobile genetic elements. *Nucleic Acids Res.* **32**: D45–D49. DOI: 10.1093/nar/gkh084.
 45. Ramirez, M.S., Nikolaidis, N., and Tolmasky, M. (2013). Rise and dissemination of aminoglycoside resistance: the aac(6)-Ib paradigm. *Front. Microbiol.* **4**: 121. DOI: 10.3389/fmicb.2013.00121.
 46. Gligorijević, V., Renfrew, P.D., Kosciolk, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**: 3168. DOI: 10.1038/s41467-021-23303-9.
 47. Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589. DOI: 10.1038/s41586-021-03819-2.
 48. Liu, Y., Chen, Y., and Han, L. (2023). Bioinformatics: Advancing biomedical discovery and innovation in the era of big data and artificial intelligence. *The Innovation Medicine* **1**: 100012. DOI: 10.59717/j.xinn-med.2023.100012.
 49. Yang, P. and Ning, K. (2022). How much metagenome data is needed for protein structure prediction: The advantages of targeted approach from the ecological and evolutionary perspectives. *iMeta* **1**: e9. DOI: 10.1002/imt2.9.
 50. Wu, Y. and He, K. (2020). Group Normalization. *Int. J. Comput. Vis.* **128**: 742–755. DOI: 10.1007/s11263-019-01198-w.
 51. Hendrycks, D., and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv*. DOI: 10.48550/arXiv.1606.08415.
 52. He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
 53. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**: 1–13. DOI: 10.1186/s13059-019-1891-0.
 54. Kumar, S., Stecher, G., Li, M., et al. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**: 1547–1549. DOI: 10.1093/molbev/msy096.
 55. Letunic, I. and Bork, P. (2019). Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**: W256–W259. DOI: 10.1093/nar/gkz239.
 56. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., et al. (2004). Protein structure prediction using Rosetta. *Methods Enzymol* **383**: 66–93. DOI: 10.1016/S0076-6879(04)83004-0.
 57. Tian, W., Chen, C., Lei, X., et al. (2018). CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.* **46**: W363–W367. DOI: 10.1093/nar/gky473.
 58. Cole, J.C., Korb, O., McCabe, P., et al. (2018). Knowledge-based conformer generation using the cambridge structural database. *J. Chem. Inf. Model.* **58**: 615–629. DOI: 10.1021/acs.jcim.7b00697.
 59. Trott, O. and Olson, A.J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**: 455–461. DOI: 10.1002/jcc.21334.
 60. Benjamin, D.J., Berger, J.O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* **2**: 6–10. DOI: 10.1038/s41562-017-0189-z.

FUNDING AND ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 81774008, 81573702, 31871334, and 31671374) and the National Key R&D Program of China (Grant Nos. 2023YFA1800900, 2018YFC0910502 and 2018YFA0902400). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

K. N. and X.C. conceived and proposed the idea and designed the study. Y.Z., C.C., Q.J., X.Z., and X.C. performed the experiments and analyzed the data. Y.Z., C.C., X.Z., K.N. and X.C. contributed to editing and proofreading the manuscript. All the authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

DATA AND CODE AVAILABILITY

We collected metagenomic samples from several published studies^{28,29} and these samples were collected mainly from marine, soil and human-associated environments. The human-associated samples included two gut groups (one from Madagascar, i.e., GutM; the other from Denmark, i.e., GutD), one oral group and one skin group (both oral and skin groups are from the HMP project). The ONN4ARG-DB dataset can be accessed at <https://github.com/HUST-NingKang-Lab/ONN4ARG>.

SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-life.2024.100054>

LEAD CONTACT WEBSITE

<http://www.microbiinformatics.org/>