Microbiome

RESEARCH

Open Access

DRAMMA: a multifaceted machine learning approach for novel antimicrobial resistance gene detection in metagenomic data

Ella Rannon[1], Sagi Shaashua[1] and David Burstein[1*]

Abstract

Background Antibiotics are essential for medical procedures, food security, and public health. However, ill-advised usage leads to increased pathogen resistance to antimicrobial substances, posing a threat of fatal infections and limiting the benefits of antibiotics. Therefore, early detection of antimicrobial resistance genes (ARGs), especially in pathogens, is crucial for human health. Most computational methods for ARG detection rely on homology to a predefined gene database and therefore are limited in their ability to discover novel genes.

Results We introduce DRAMMA, a machine learning method for predicting new ARGs with no sequence similarity to known ARGs or any annotated gene. DRAMMA utilizes various features, including protein properties, genomic context, and evolutionary patterns. The model demonstrated robust predictive performance both in cross-validation and an external validation set annotated by an empirical ARG database. Analyses of the high-ranking model-generated candidates revealed a significant enrichment of candidates within the *Bacteroidetes/Chlorobi* and *Betaproteobacteria* taxonomic groups.

Conclusions DRAMMA enables rapid ARG identification for global-scale genomic and metagenomic samples, thus holding promise for the discovery of novel ARGs that lack sequence similarity to any known resistance genes. Further, our model has the potential to facilitate early detection of specific ARGs, potentially influencing the selection of antibiotics administered to patients.

Introduction

Antibiotic substances, drugs targeting bacterial species, have drastically reduced the threat of infections and have become essential for many medical procedures such as surgeries, organ transplants, and cancer treatment [1–3]. These drugs are invaluable thanks to their ability to kill or inhibit the bacteria causing the infection while not causing any harm to the host's cells [1]. Nevertheless, prolonged overuse of antibiotics has resulted in the emergence and worldwide spread of antibiotic-resistant pathogens, which threatens the continued benefits of antibiotics [4]. The pathogens' resistance, also known as antimicrobial resistance (AMR), allows these strains to grow and spread due to the strong selective pressure of antibiotics, becoming dominant in their environment [5].

It is estimated that globally in 2019, 4.95 million deaths were associated with drug-resistant infections, with approximately 1.27 million of these deaths directly attributable to antibiotic-resistant bacteria [6]. This number is projected to reach ten million by 2050 if no solutions are devised to slow down the emergence of antibiotic-resistant bacteria [5]. This value is probably an underestimation, as key medical procedures, such as surgeries and chemotherapy, may become too dangerous to perform

*Correspondence:
David Burstein
davidbur@tauex.tau.ac.il
[1] The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

if antibiotics lose their effectiveness. Additionally, drug resistance has already made an economic impact. Resistance to first-line antibiotic treatments costs the US health system 20 billion USD per year [7], and the global cost of antibiotic resistance is predicted to exceed 100 trillion USD throughout the next few decades [5].

Most of the antibiotics used today are either compounds discovered during the "golden era" of antibiotic discovery between the 1940s and the 1960s, or their derivatives [8]. Since the 1980s, the rate of antibiotic discovery has fallen drastically, and only a few antibiotics have reached the market in the last two decades [5]. Further, since the 1960s, no new class of broad-spectrum compounds has been discovered [8]. Hence, we are currently lacking new drugs to battle antibiotic-resistant bacteria [5]. Therefore, other approaches are required in order to combat antimicrobial resistance, one of which is reliable surveillance of antimicrobial resistance. Surveillance data can help improve patient treatment and health, inform health policies, shape responses to health emergencies, provide early warnings of emerging threats, and help identify long-term trends [5]. The foresight of drug-resistant bacteria emergence should enable a proactive development of next-generation treatment strategies before the dissemination of resistance threats [1].

Metagenomic data is essential for the characterization of the global antibiotic resistome, as many relevant AMR genes evolved in environmental microorganisms [1]. Metagenomic research focusing on environmental samples from soil, sewage, and other sources in urban and rural areas worldwide has highlighted differences in the diversity, abundance, and distribution of ARGs, their class, and their resistance mechanisms across various regions [9–12]. These studies also reveal correlations between ARG abundance and socio-economic, health, and environmental factors, as well as associations with mobile genetic elements. Compared to human-derived samples, environmental samples offer significant advantages: they are readily accessible, enable real-time analysis, and are cost-effective, with no ethical constraints [9]. It was revealed that resistance genes in human pathogens share, in some cases, more than 99% nucleotide identity with resistance genes from soil bacteria [13]. Moreover, the synteny of resistance genes with mobility elements suggests these genes have likely undergone horizontal gene transfer [1]. Therefore, inferring the mobility of a gene across a wide variety of taxonomic groups can help detect antibiotic resistance genes and assess the threat they might impose.

Numerous bioinformatic tools have been developed for ARG annotation in metagenomic datasets, most of which are based on sequence similarity to a predefined gene database [14–25]. The gene repertoire that these methods can discover is thus limited to the current, incomplete, ARG knowledge base, and they lack the ability to generalize and identify novel ARGs. Recently, a few machine learning models were developed for ARG detection without the need for a predefined database. For example, HMD-ARG [26], a hierarchical deep-learning framework, utilizes the protein's raw amino acid sequence to predict multiple ARG properties. These properties include gene classification to ARG or non-ARG, the antibiotic family to which it confers resistance, the gene's resistance mechanism, whether the ARG is intrinsic or acquired, and the specific subclass of the beta-lactamase. Another recently developed algorithm is PLM-ARG [27], which utilizes the publicly available pre-trained protein language model ESM-1b [28] with two consecutive XGBoost [29] models for ARG identification task and resistance category prediction. An additional deep-learning model, ARGNet [30], processes either short reads or complete genes as input and applies an autoencoder model to identify ARGs and a convolutional neural network (CNN) multiclass classifier to predict ARG categories. However, since these models only utilize the sequence of the ARG or its product, they cannot take into account biological knowledge beyond the sequence, which can be crucial for ARG detection.

Here, we trained DRAMMA, a Random Forest model on global-scale metagenomic data. The model was trained on a wide variety of tailored features based on biological knowledge and understanding of ARG characteristics. These features take into account protein biochemical, physical, and structural properties, as well as genomic and evolutionary context. This approach allows us to predict new ARGs that are genuinely unknown and thus present no detectable sequence similarity to any known resistance gene. The model demonstrated strong performance on both the training set and an independent validation set annotated using an external ARG database based on functional metagenomics experiments. This led to the identification of novel ARG candidates, which were subsequently subjected to rigorous analysis. We anticipate that further investigation of the top candidates identified by DRAMMA, along with the application of our model on newly generated metagenomic datasets, will facilitate the early detection of previously unknown antimicrobial resistance genes. This, in turn, has the potential to significantly advance our understanding of antibiotic resistance and inform strategies to combat this growing global health threat.

Results

Dataset compilation

An extensive dataset was compiled of genes from genomic and metagenomic sources [31]. This data was acquired from various ecosystems, including human and animal microbiomes, groundwater, sewage, marine, and soil (Table 2). Only protein-coding genes in large contigs (≥ 10 kbp) were used. Overall, 492.1 million proteins were retrieved from assemblies of 22,241 metagenomes (Table 2).

The known ARGs in the dataset were annotated using DRAMMA-HMM-DB, a database of profile Hidden Markov Models (HMMs, Supplementary Datasets 1–3) that we compiled based on several AMR databases (Resfams [22], CARD [23], and HMD-arg-DB [26]). Positive examples (ARGs) for our classification scheme were genes with high similarity to known ARG families. Negative examples (non-AMR proteins) were randomly sampled from the gene pool to establish a ratio of 1:10 resistance to non-resistance genes, and duplicate (highly similar) proteins were then removed.

Feature extraction

We extracted 512 features for each protein. The features can be divided into four main categories: (1) amino acid properties, such as gene and contig length, physical and chemical attributes of the protein, the proportion of each amino acid in the protein, the proportion of groups of amino acids sharing similar attributes, and averages of amino acid indices that represent different physicochemical and biochemical characteristics for each amino acid [32]; (2) amino acid patterns, including 8-mers of hydrophilic/hydrophobic residues, Helix Turn Helix (HTH) domain, DNA binding domains, and transmembrane domains; (3) HGT signals, e.g., GC content differences between the gene and the contig it is coded on, the distance between DNA *k*-mer distribution vectors of the gene and its contig, and the distribution of each gene across diverse taxonomic groups; (4) genomic context, including the presence of known ARGs and genes of mobile genetic elements in the genomic region of the analyzed gene (Fig. 7B, Supplementary Table 1).

Model and feature selection

Following a comparison of several machine learning models (see Hyperparameter optimization in Methods), Random Forest was chosen due to its favorable trade-off between predictive accuracy and computational efficiency. To choose the optimal subset of features for the classification model, we utilized Random Forest's feature importance, known as impurity-based importance or Gini importance, and selected the best features according to these scores. These scores reflect the reduction in node impurity, weighted by the proportion of samples reaching the node, and averaged across all trees in the ensemble [33]. To choose the optimal number of selected features, we examined how varying the number of features impacted the mean performance (measured as the area under the precision-recall curve, PR-AUC) of the model across a five-fold cross-validation on the development set. The optimal number of features was approximately 30 (Fig. 8).

The model's features were selected based on the feature importance values of the Random Forest model trained on the entire training set. Prominent features included information about the presence of the proteins in different taxonomic groups, amino acid composition and patterns (percentage of different amino acids or groups of amino acids, presence of HTH domains, and frequency of hydrophilic-hydrophobic signatures), features regarding the proteins' physical and biochemical properties (gene product size, grand average of hydropathy (GRAVY) value, where negative values indicate hydrophilicity and positive values indicate hydrophobicity, and molar extinction coefficient), and features regarding the presence of ARGs within the gene's genomic region (see Supplementary Fig. 1A, Supplementary Table 1). We observed a distinct difference between the distribution of these feature values within the AMR and non-AMR populations (Supplementary Fig. 2). Specifically, this distinction becomes evident when examining the most significant feature, gene product size. ARGs were typically comprised of 500–1500 amino acids, making them, on average, larger than non-ARGs. This observation is supported by the SHAP values associated with this feature (Supplementary Fig. 1B), which tended to be negative for smaller values and thus contributed to negative classification. A similar pattern was observed for the taxonomic distribution features.

Model performance evaluation

The trained model, which we named DRAMMA for Detection of Resistance to AntiMicrobials using Machine-learning Approaches, was evaluated using the mean ROC-AUC and mean PR-AUC in a five-fold cross-validation process. The results indicate highly accurate classification (Fig. 1A), with a mean ROC-AUC of 0.98, and a mean PR-AUC of 0.857.

In order to ensure that the high performance is not the result of data leakage from genes of the same taxonomic groups that share multiple genomic properties, we decided to re-evaluate DRAMMA's performance using an NCBI WGS genomes dataset, which we divided into five folds according to major taxonomic groups: *Actinobacteria*, *Gammaproteobacteria*, *Firmicutes*, *Alphaproteobacteria*, and *Bacteroidetes*. The model's performances were

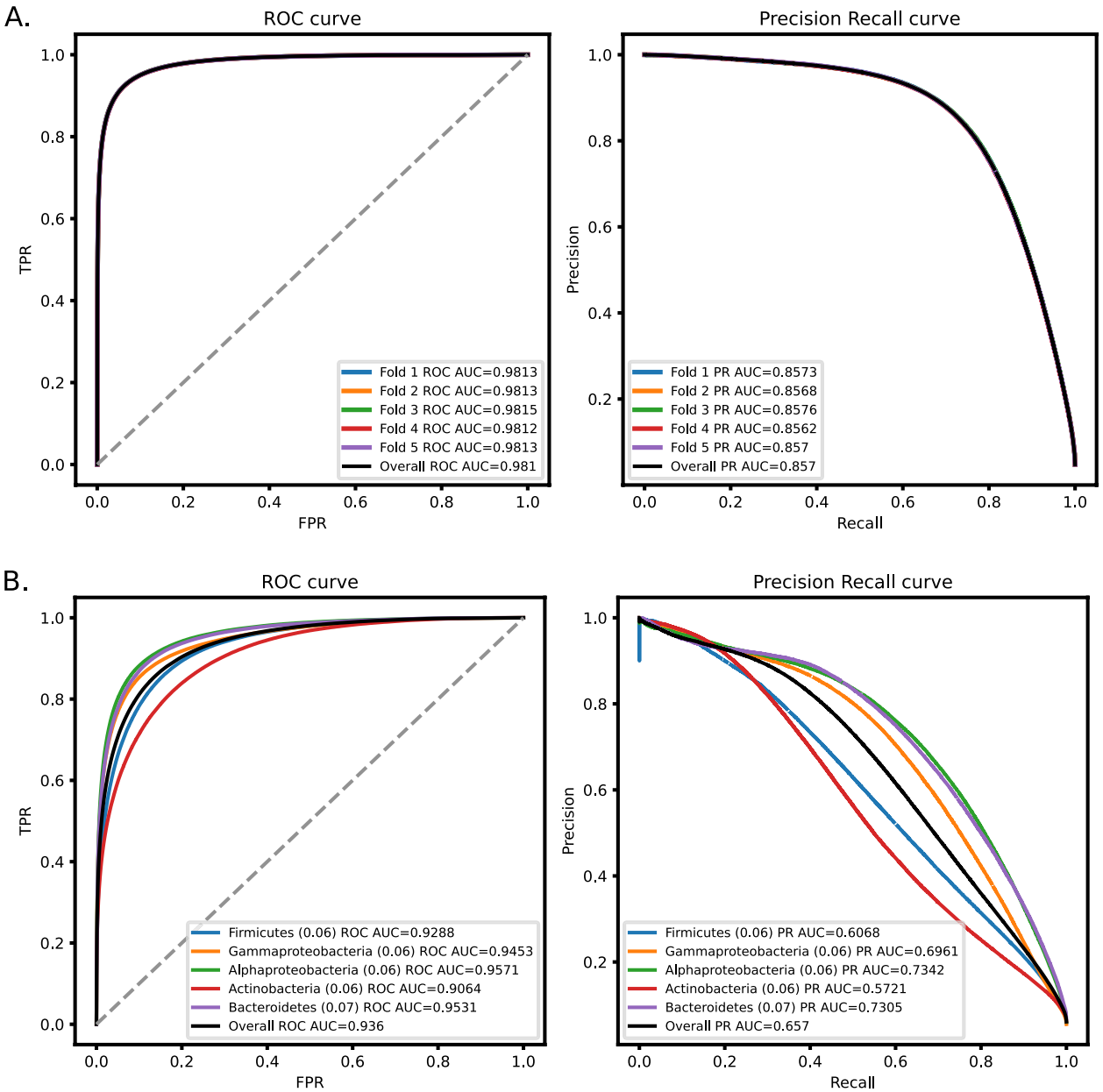

Fig. 1 Classification performances of DRAMMA over five-fold cross-validation measured as ROC-AUC and PR-AUC. A Performance on a genomic and metagenomic dataset split to folds by contigs. The dataset is comprised of 30.9M proteins, containing 5% AMR genes annotated using HMMs from the DRAMMA-HMM-DB database. B Performance on a genomic dataset split into folds by taxonomic groups. The frequency of positive proteins in each fold is noted in brackets

expected to be lower since in this evaluation the model was tested on genomes from taxa that were evolutionarily distant from any species in the training set. Indeed the model had lower performances but was still accurate with a mean ROC-AUC of 0.938, and a mean PR-AUC of 0.668 (Fig. 1B).

We further tested DRAMMA's performance on an external validation set taken from the Global Sewage Surveillance Project [9], which is comprised of sewage metagenomic samples collected worldwide (see Dataset compilation in the Methods section). The microbial communities in these samples are expected to include both environmental microbes as well as microorganisms prevalent in human microbiomes. This dataset was assembled from read data and annotated using two ARG databases, our DRAMMA-HMM-DB database and ResfinderFG

v2.0. The latter is an external experimental database of ARGs obtained by functional metagenomics. The model was evaluated on each annotation scheme. The model's performance on this dataset was high as well, with a mean ROC-AUC of 0.99 and a mean PR-AUC of 0.91 when tested on the annotation according to DRAMMA-HMM-DB. It achieved a mean ROC-AUC of 0.91 and a mean PR-AUC of 0.59 when tested on the experimental ResfinderFG annotation (Fig. 2). It was observed that, reassuringly, DRAMMA's score tends to be low for proteins with negative annotation (non-ARGs) according to ResfinderFG (Supplementary Fig. 3).

In addition, we assessed the runtime for feature extraction and model prediction on a 64-CPU machine, using the *E. coli* K-12 MG1655 genome and a random selection of approximately 100,000 metagenomic proteins from the Global Sewage Surveillance Project [9]. This process took 11.1 min for the 4329 proteins in the *E. coli* genome and 21.23 min for the 100,532 proteins in the sewage dataset.

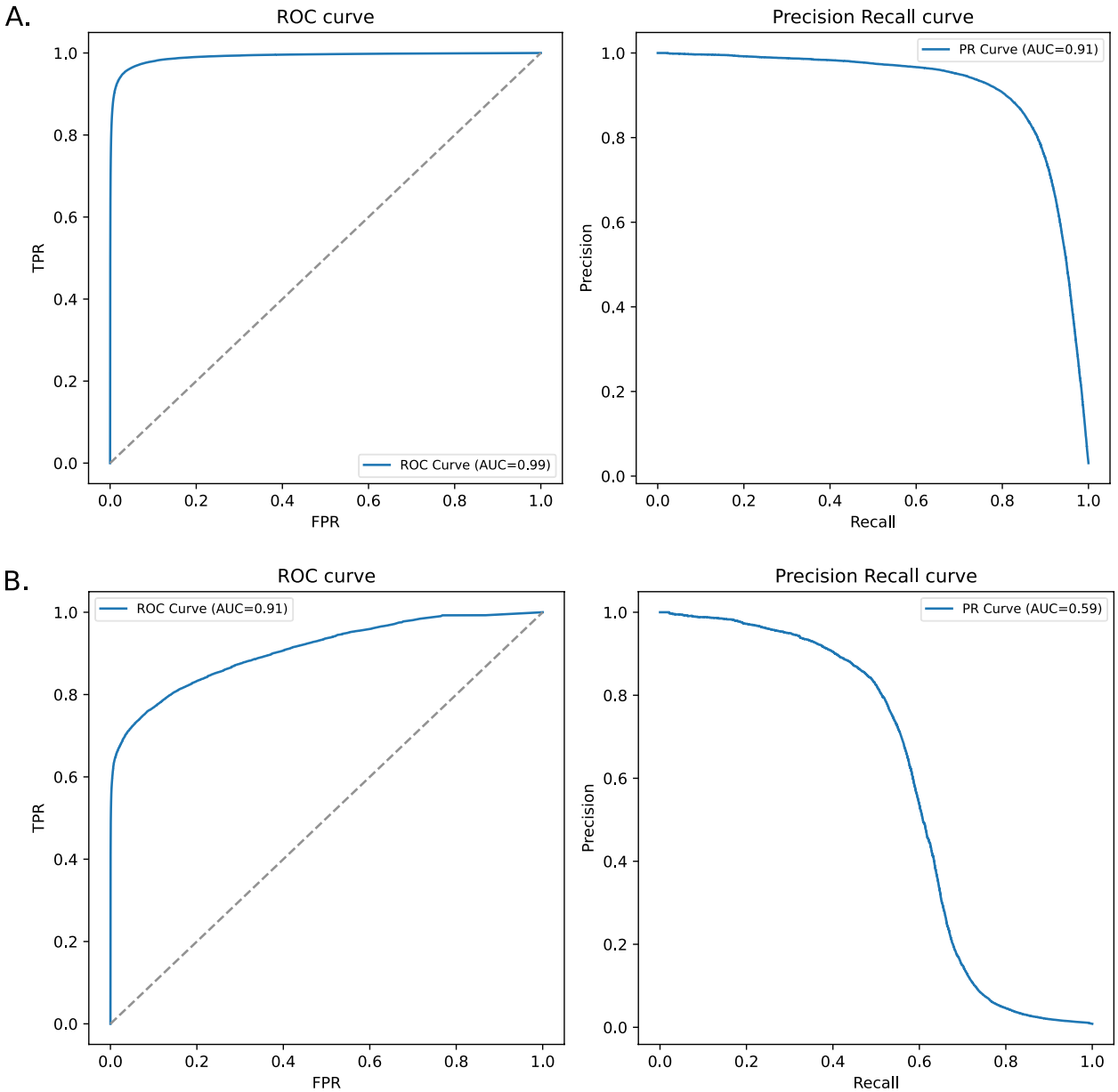

Fig. 2 Classification performances of DRAMMA over a validation set of sewage samples measured as ROC-AUC and PR-AUC. **A** The model's performance on the dataset annotated by our ARG HMM database, DRAMMA-HMM-DB. **B** The model's performance on the dataset annotated by the experimental ResfinderFG v.2.0 database

Impact of sequence disruption on DRAMMA predictions

Given that DRAMMA utilizes a range of biological features rather than relying on sequence homology to predefined ARG sequences, we sought to evaluate the impact of the protein sequence on model predictions. To investigate this, we scrambled the protein sequences from the *E. coli* genome and the ~ 100,000 proteins selected from the Global Sewage Surveillance Project [9], used for the runtime evaluation. As expected, following scrambling, none of the proteins in these datasets were labeled as ARGs by our DRAMMA-HMM-DB database (in contrast to 149 and 2780 of the original proteins in the *E. coli* genome and sewage samples, respectively). Despite DRAMMA's strong reliance on contextual and content features, its misclassification rate was low: only 0.02% (one of 4329) of the *E. coli* proteins and less than 0.1% (96 of 100,532) of the scrambled sewage proteins were classified as ARGs. An analysis of the SHAP values of the scrambled sequences' features has shown that the features contributing to positive classifications are indeed those unaffected by the sequence order. These included features capturing amino acid composition irrespective of their order and context-dependent features (see Supplementary Fig. 4).

Benchmarking

The performance of the DRAMMA model was compared with that of previous algorithms for ARG prediction: (1) Resfams [22], (2) DeepARG [21], (3) ARGNet [30], (4) PLM-ARG [27], (5) CARD October 2020 release [23], and (6) CARD October 2023 release [34]. The two CARD releases were selected since our DRAMMA-HMM-DB database was comprised of ARGs in the 2020 release, while the October 2023 release was the latest available at the time of the benchmarking. The performance evaluation was conducted on the sewage test set annotated based on the ResFinderFG database, an external ARG database collecting information from functional metagenomic experiments. In our pipeline, regulatory genes, efflux pumps, and resistance conferred via point mutations are not considered positive cases of ARGs. Some of the approaches to which we compared DRAMMA do consider these genes as positive. This might bias the results toward DRAMMA and unjustifyingly increase the false positive rates of the other algorithms. To ensure a fair and unbiased comparison, proteins that our approach labeled as non-ARGs but were considered ARGs by other approaches were excluded from the test set. The performance of each algorithm was assessed by MCC, true positive rate (TPR), false positive rate (FPR), macro precision, recall, and F1, with comparisons to our classification model at two different score thresholds (0.75 and 0.95), corresponding to expected precision scores. Results revealed that DRAMMA achieved the best recall (75.1%) and CARD strict reached the highest precision (94.4%), which can be expected as it is based on strict sequence comparisons and thus is not expected to yield numerous false positives. Notably, CARD strict also received a low recall rate (50.6%). Our approach, on the other hand, achieved the best balance between precision and recall, as indicated by F1 and MCC scores (0.78 and 0.567, accordingly, see Table 1). Although CARD loose achieved the highest TPR scores (0.975 and 0.95 for the 2020 and 2023 releases, respectively), it also had a high FPR (0.903 and 0.807 for 2020 and 2023, respectively). This indicates that while it correctly classifies many ARGs, it also frequently misclassifies non-ARGs as ARGs, as reflected in its relatively low

Table 1 Benchmarking on a test set comprising sewage metagenomic samples, with ARGs annotated using ResFinderFG v2.0. The number in brackets corresponds to the expected precision determined by the chosen DRAMMA model score. The high score threshold for DeepARG was selected based on their recommended setting. "Strict" and "Loose" refer to CARD's rgi search parameters. *TPR* true positive rate, *FPR* false positive rate

Algorithm	TPR	FPR	Precision	Recall	F1	MCC
DRAMMA (0.75)	0.527	0.024	0.82	**0.751**	**0.78**	**0.567**
DRAMMA (0.95)	0.408	0.013	0.853	0.698	0.748	0.529
Resfams	0.418	0.021	0.804	0.698	0.737	0.491
DeepARG High Score (> = 0.8)	0.027	0.0002	0.913	0.513	0.502	0.147
DeepARG All Scores	0.039	0.0003	0.923	0.519	0.514	0.18
ARGNet	0.056	0.077	0.488	0.49	0.488	-0.023
PLM-ARG	0.335	0.012	0.837	0.662	0.711	0.467
CARD 2020 Strict	0.013	**0.00003**	**0.944**	0.506	0.489	0.106
CARD 2020 Loose	**0.975**	0.903	0.537	0.536	0.177	0.073
CARD 2023 Strict	0.028	0.0002	0.915	0.514	0.503	0.151
CARD 2023 Loose	0.95	0.807	0.54	0.571	0.256	0.107

In bold are the best score for each performance measure

precision scores (0.537 and 0.54 for the 2020 and 2023 releases, respectively). Conversely, CARD strict achieved the lowest FPR (0.00003) but exhibited low TPR (0.013) and recall (0.506) scores.

Candidate analysis

To identify genuinely novel ARGs, we first used the DRAMMA model trained on the entire metagenomic training set to classify approximately 650 million proteins from genomic and metagenomic data. We focused on the top-ranking 18.1 million proteins that received a score equivalent to precision of > 95% on the training set. Subsequently, we categorized the candidates based on their source, differentiating between those originating from genomic and metagenomic samples. We then assessed the distribution of novel candidates (high-scoring genes annotated as non-ARG) with regard to taxonomic groups and environments (Fig. 3). Our analysis revealed that the most prevalent taxonomic groups among predicted ARGs were *Gammaproteobacteria* (20.66%), *Firmicutes* (18.46%), *Bacteroidetes/Chlorobi* group (14.45%), *Alphaproteobacteria* (13.35%), and *Actinobacteria* (12.14%). Among the metagenomes, predicted ARGs were detected primarily in samples originating from the human microbiome, accounting for a significant portion of the metagenomic ARG candidates (73.02%).

The genomic databases are highly biased toward specific bacteria (mostly pathogens and model organisms) and ecosystems (human-associated bacteria). Therefore, to properly assess enrichment, we normalized the number of novel candidates in the different groups according to the total number of non-ARG proteins within the same taxonomic groups and environments: $\text{enrichment}(g) = \log_2 \frac{\text{percentage of novel candidates from group } g}{\text{percentage of non-ARGs from group } g}$ (Fig. 4 and Supplementary Fig. 5). Our observations revealed that only the *Bacteroidetes/Chlorobi* and *Betaproteobacteria* groups displayed a notable enrichment of ARG candidates (with enrichment scores of 0.947 and 0.535, respectively), whereas the *Firmicutes*, *Alphaproteobacteria*, and *Bacteria Candidate Phyla* groups exhibited a more modest enrichment (with enrichment scores of 0.221, 0.199, and 0.121, respectively). In contrast, *Stenosarchaea* and *Euryarchaeota* exhibited notable depletions (with enrichment score of −1.237 for both groups). In addition, within metagenomic samples, it was evident that ARG candidates were highly enriched in human and animal microbiomes, and, to a lesser extent, in plant-associated bacteria.

We also investigated the distribution of the drugs they confer resistance to and predicted mechanisms by which the candidates confer resistance across the various taxonomic groups and environments (Fig. 5, Supplementary Fig. 6). These findings revealed that beta-lactam antibiotics were the most common drugs the candidates provided resistance to and that the most frequent resistance mechanisms among our predictions were target alteration and antibiotic inactivation. Notably, a similar distribution of resistance mechanisms and antibiotic drugs

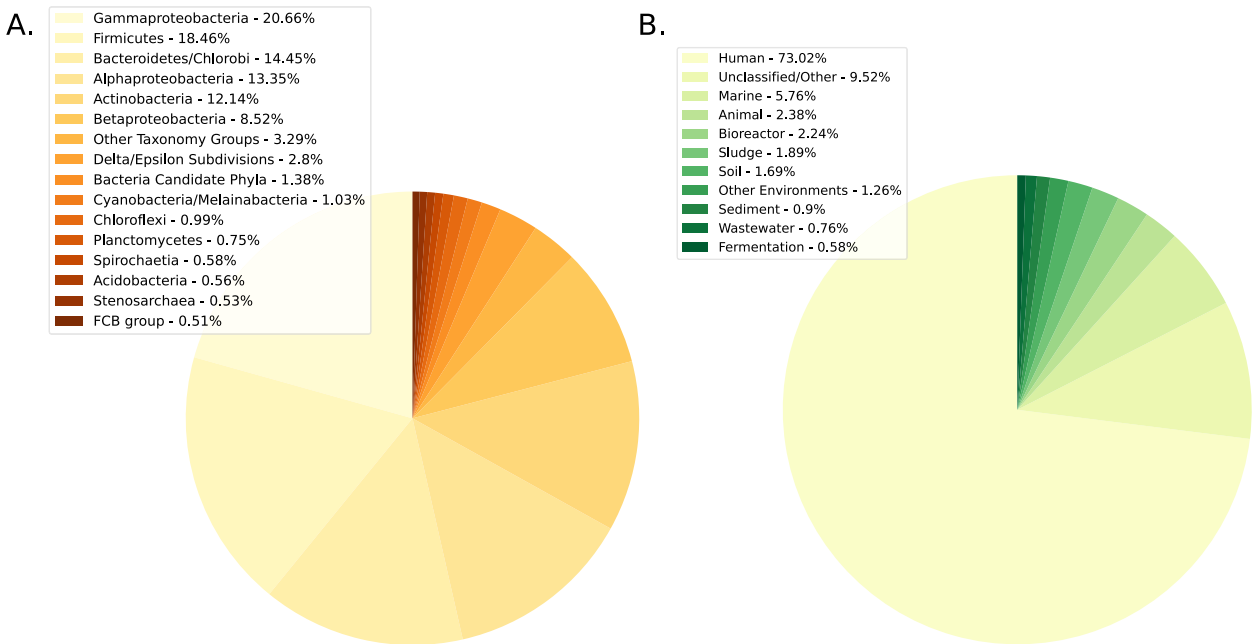

Fig. 3 Distribution of novel ARG candidates. **A** Distribution of novel candidates from genomic samples across different taxonomic groups. **B** Distribution of novel candidates from metagenomic samples across different ecosystems

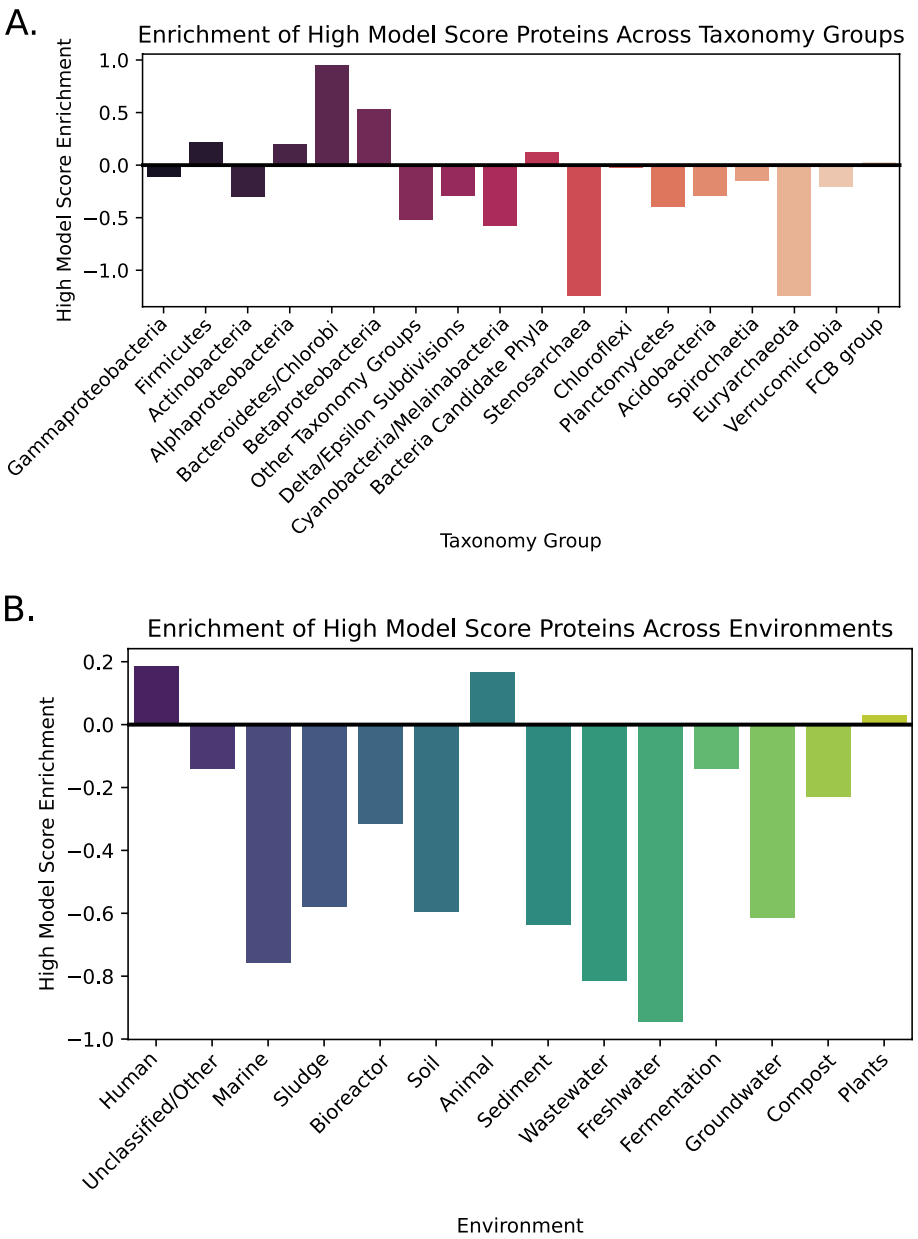

Fig. 4 Enrichment analysis of novel candidates across different groups. The enrichment of high-ranking candidates for each group g is calculated as $\text{enrichment}(g) = \log_2 \frac{\text{percentage of novel candidates from group } g}{\text{percentage of non-ARGs from group } g}$. **A** Enrichment of novel candidates from genomic samples in different taxonomic groups. **B** Enrichment of novel candidates from metagenomic samples from the ecosystem analyzed

was predicted for the ARGs across the different taxonomy groups and environments. *Betaproteobacteria* is the only taxonomy group to exhibit a significant percentage of the "reduced permeability to antibiotics" mechanism. Finally, we examined the novel candidates for the enrichment of known domains from Pfam [35]; however, the vast majority of these proteins (99.6%) did not contain a well-characterized domain. Furthermore, none of the few domains found exhibited a prevailing presence among the novel candidates (Supplementary Dataset 4).

Candidate selection

We aimed to highlight ARG candidates of most interest among the analyzed genes, focusing on genes that could not have been identified using traditional sequence-based approaches. We thus identified predictions with no annotation across several databases, as well as top-ranking

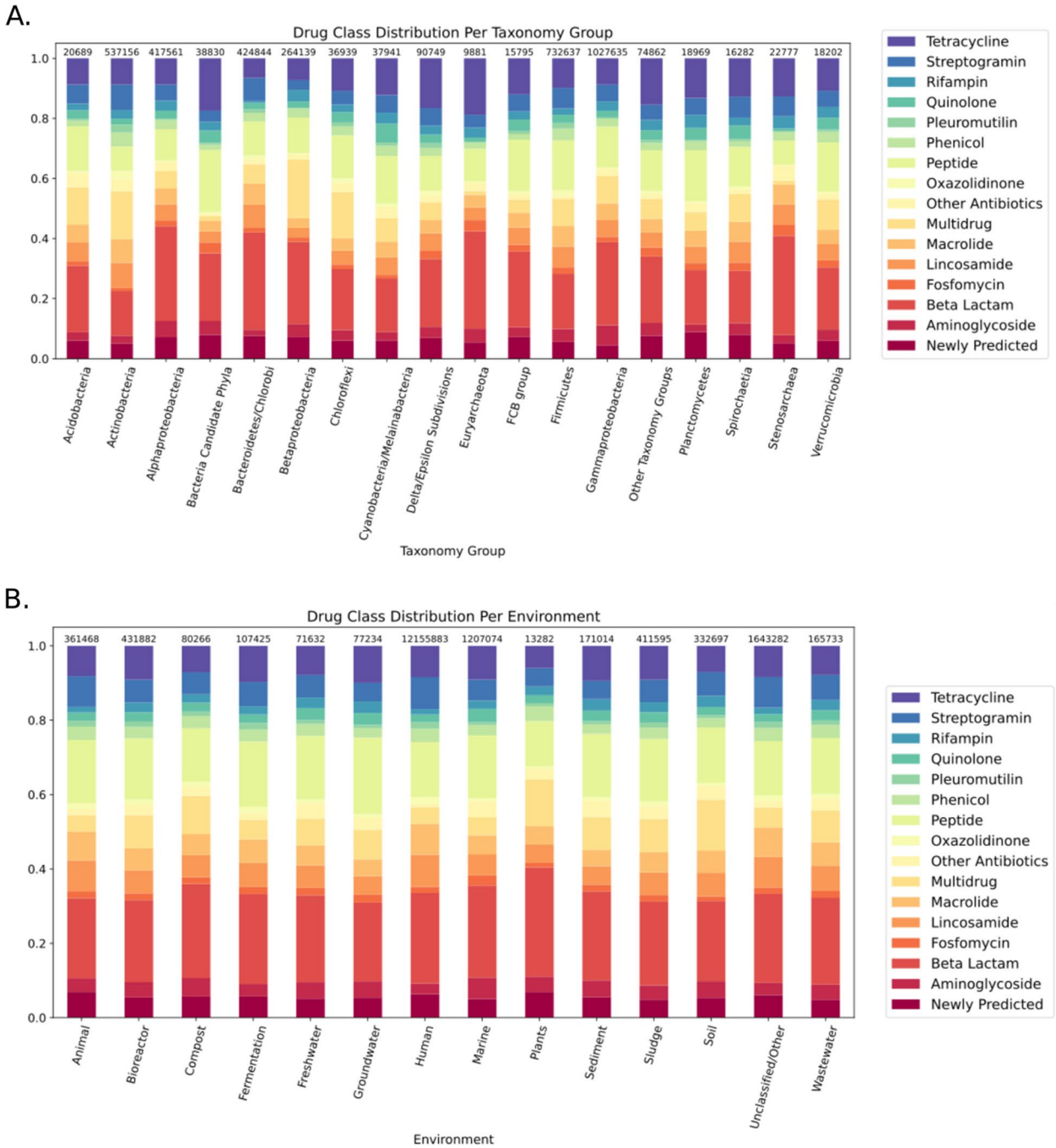

Fig. 5 The distributions of antibiotic drugs the high-ranking ARG candidates confer resistance to. The distribution of antibiotic drugs to which high-ranking ARG candidates confer resistance, across different taxonomic groups (**A**), and ecosystems (**B**). The number at the top of each bar indicates the total count of ARG candidates in the respective group

predictions with partial annotations or annotations not indicative of resistance. To that end, we first removed all the known ARGs from the list of model candidates, resulting in 1.28 million proteins. We then performed several other annotation steps, in which the ARG candidates were compared to different databases, namely the Kyoto Encyclopedia of Genes and Genomes (KEGG) [36], NCBI's non-redundant protein database [31], PDB [37], and Pfam [35] (Fig. 6).

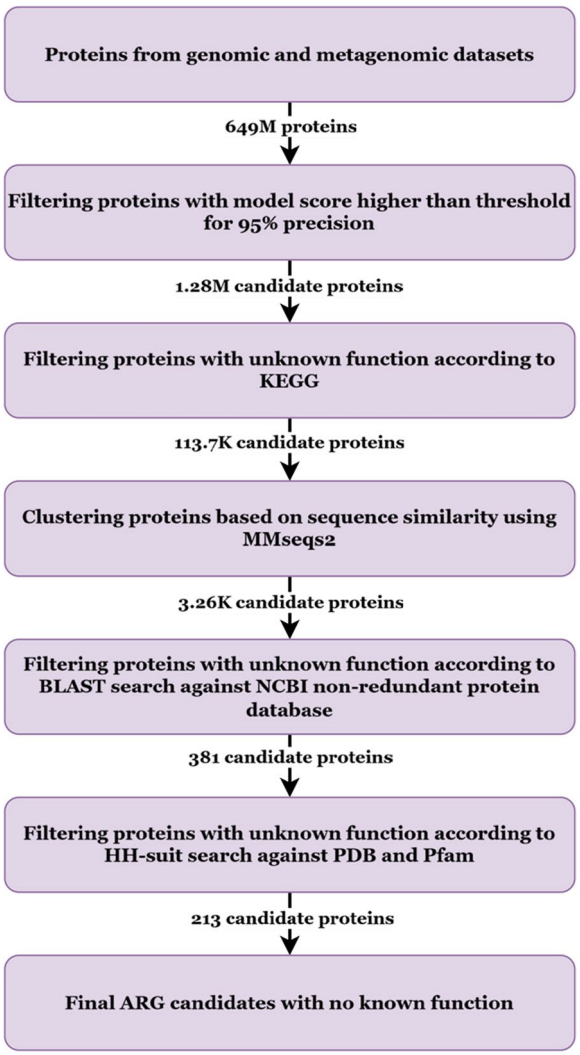

Fig. 6 Description of the annotation and filtration pipeline. The number of proteins that had no annotation in each step and were thus passed to downstream annotation is indicated below each step

In each step, the candidates were assigned to one of the following categories: (1) Known ARGs, (2) ARG-related, (3) A possible target of antibiotics, (4) Annotated, but not known to be associated with AMR, (5) Unknown function (Supplementary Fig. 7). Following each step, only the proteins of unknown function were used as input for the annotation next step to characterize them as much as possible. Since increasingly sensitive searches were applied, the number of proteins with unknown functions decreased with each step of the pipeline. However, the percentage of unknown proteins increased in each step of the pipeline, reaching 55.9% representing 213 unannotated proteins in the final step (HH-suite remote homology search [38]).

Table 2 Metagenomic assemblies from various ecosystems were downloaded from NCBI and EBI. The protein and base-pair count for informative contigs (≥ 10 kbp) were retrieved from assemblies of 22,241 metagenomes from different environments

Assemble sample type	Proteins in assemblies	Base-pairs in assemblies
Human microbiome	316,897,729	311.25 Gbp
Unclassified / Other	45,293,055	40.56 Gbp
Plants	420,898	0.39 Gbp
Sediment	6,811,745	6.25 Gbp
Fermentation	3,160,683	2.98 Gbp
Animal microbiome	10,420,960	10.12 Gbp
Marine	47,648,243	45.02 Gbp
Bioreactor	13,704,124	13.06 Gbp
Wastewater	6,527,929	5.74 Gbp
Sludge	13,847,044	13.64 Gbp
Compost	2,510,130	2.32 Gbp
Human/animal microbiome	6,193,314	6.17 Gbp
Soil	12,492,613	11.18 Gbp
Groundwater	2,950,435	2.73 Gbp
Freshwater	3,253,136	2.97 Gbp
Total	**492,132,038**	**474.38 Gbp**

Our candidates of interest thus included the 213 predicted ARGs that remained completely unannotated, without even remote homology to characterized proteins. In addition to these, we also included in our pool of potential ARGs of interest candidates with some annotations: (1) proteins that have annotation only based on HH-suite remote homology search, but with no annotation in a BLAST search against NCBI NR and HMM search of KEGG; (2) the 100 top-scoring predictions that had an annotation according to BLAST and no KEGG annotation; and (3) the top 200 candidates that had a KEGG annotation. This resulted in a total of 681 novel ARG candidates.

Finally, we wished to pinpoint top candidates that would be the most straightforward to test experimentally. First, we removed candidates with Helix-Turn-Helix (HTH) domains, which are indicative of regulatory function. We then assigned taxonomy to the remaining candidates by comparing all the genes on the relevant contigs to organisms with known taxonomy using MMseqs2 taxonomy search against UniRef100 [39] or by extracting the taxonomy of the DIAMOND hit with the lowest e-value against UniRef100 [39]. To focus on genes that are likely to be relevant for screening in *E. coli*, we filtered out genes originating from Gram-positive bacteria. We wished to pinpoint "standalone" ARGs, i.e., genes that confer resistance by themselves to facilitate experimental

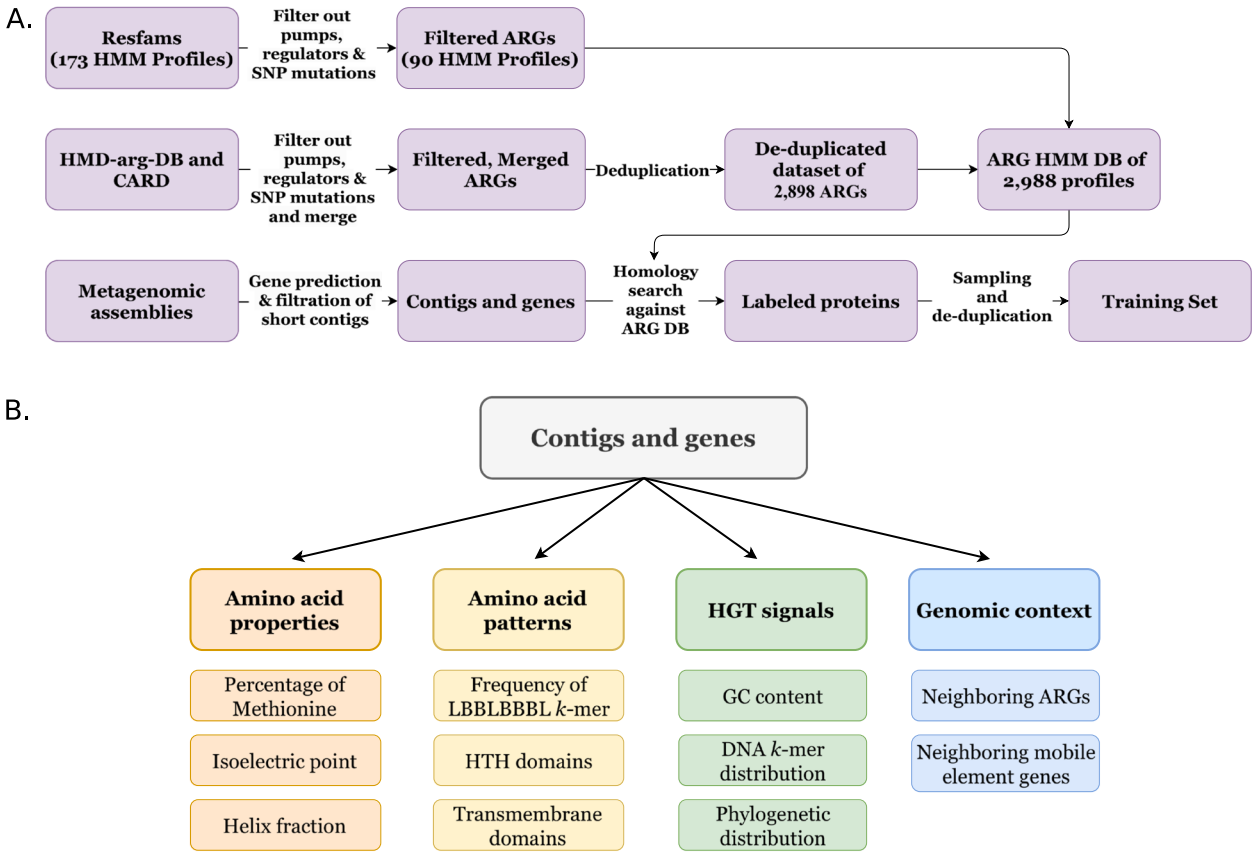

Fig. 7 Dataset compilation pipeline. **A** Gene annotation and selection. ARGs are collected from different public AMR databases and are then filtered and merged for the creation of DRAMMA-HMM-Db, an HMM ARG database that is used for ARG annotation of the metagenomics data. **B** Feature extraction. Illustration of the four main feature categories used by our model and examples of the features from each category

testing of top-ranking candidates. The neighboring genes of each candidate were thus examined to filter out candidates that consistently appeared with the same neighboring genes, assuming each of them may not function properly without the others. Using an additional machine learning classifier, we developed (see Supplementary Fig. 8), we also predicted the most likely resistance mechanisms of the potential ARGs. Subsequently, we utilized AlphaFold3 [40] to predict the structure of the candidate proteins and conducted searches against AlphaFold's protein structure database [41, 42] to obtain potential functional annotations. As a last step, we tested for similarity within the prediction to avoid testing two relatively similar proteins. The top candidates after filtering and adding the structural, syntenic, and taxonomic information are detailed in Supplementary Table 2. We anticipate that this rich source of information on potential ARGs will contribute to a better understanding of antimicrobial resistance mechanisms and their dissemination.

Discussion

The emergence and worldwide spread of antibiotic-resistant pathogens is a rising threat to public health [4]. The detection of ARGs has a pivotal role in enhancing patient well-being, issuing early alerts about emerging threats and informing administration policies [5]. This study introduced a novel machine learning approach designed to identify previously unknown antibiotic resistance genes (ARGs) within genomic and metagenomic data using a wide array of biological features. Most of the existing bioinformatical tools for ARG prediction rely solely on sequence similarity to a predefined ARG database [14–25, 43–48], which severely limits the detection of resistance genes that can be detected to genes that are similar enough to those in the database. By utilizing prior knowledge and biological properties characterizing ARGs, our method demonstrates the potential of discovering genuinely novel ARGs exhibiting no sequence similarity to any known resistance gene.

DRAMMA has been trained using an extensive dataset encompassing a wide array of genes originating from

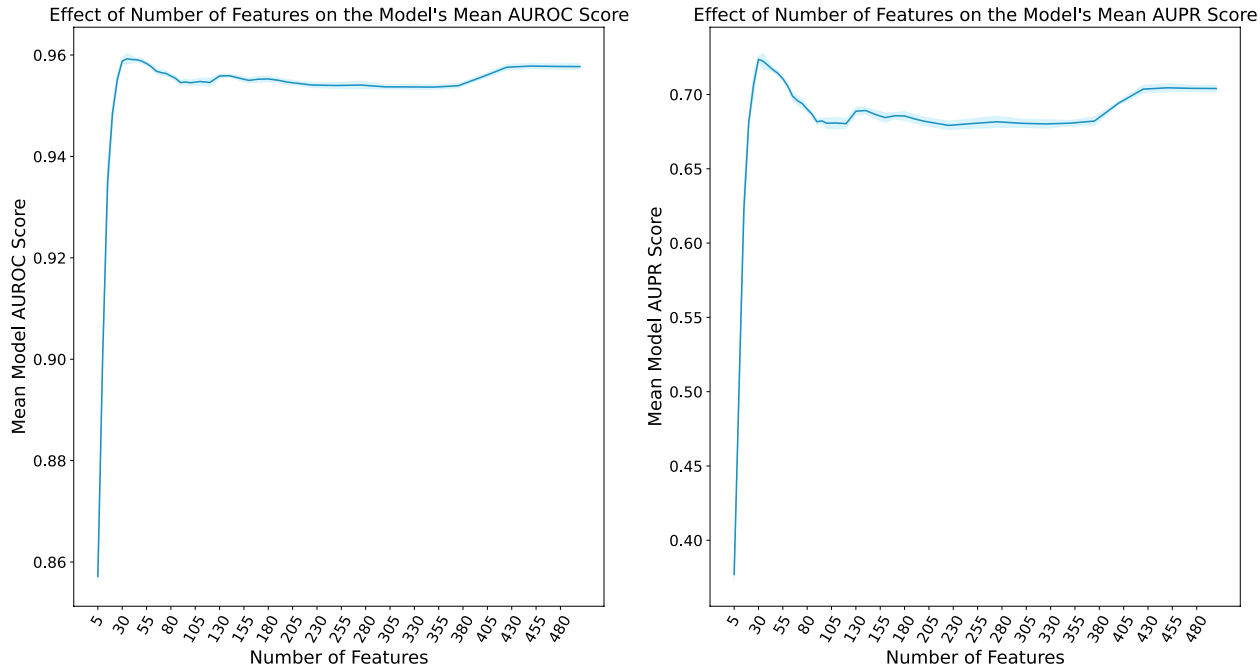

Fig. 8 Seeking the optimal number of features to select. Measurement of five-fold cross-validation classification performance (measured as ROC-AUC and PR-AUC) of a Random Forest algorithm trained on different numbers of features. Dark blue is the mean score across folds, light blue is the standard deviation. There is a decrease in the classifier's performance with models utilizing more than ~ 30 features

diverse environments and organisms. The ARG database compiled in this research encompasses a diverse spectrum of ARG gene families. Combined, these allow DRAMMA to exhibit robust generalization capabilities, effectively detecting ARGs in different environments, including genes conferring resistance to a variety of antibiotics through different resistance mechanisms. Our analysis of the ARG candidates identified by the model underscores this capability, revealing a consistent distribution of resistance mechanisms and antibiotics to which these genes confer resistance across diverse taxonomic groups and ecosystems (Fig. 3), implying that the model identifies ARGs in well-studied as well as in less-explored environments and taxonomic groups. The analysis also highlighted beta-lactam antibiotics as the most common antibiotic to which the candidates demonstrated resistance, aligning with the fact that they are indeed one of the most prescribed antibiotic classes [49].

The machine learning algorithm presented in this work demonstrated promising performance both in cross-validation and on an independent dataset. On the training set, the DRAMMA model received mean ROC-AUC scores of 0.98 and 0.938, along with mean PR-AUC scores of 0.857 and 0.668 for the metagenomic five-fold cross-validation and taxonomic five-fold cross-validation, respectively. The consistency in performance across the different folds of the metagenomic dataset can be attributed to the large size of the training data. The decreased performance quality on the taxonomic folds can be attributed to the inherent difficulty of this task. Unlike metagenomic datasets where organisms are expected to be similar to those encountered during training, taxonomic folds involve testing the model on organisms that are evolutionarily distant from those modeled in the training process. Despite this challenge, it is noteworthy that the model's performance on this task remains relatively high, especially when comparing PR-AUC results to those anticipated from a random classifier (approximately 0.06, the fraction of positive samples). Additionally, variations in the model's performance across different folds are noteworthy, with the *Actinobacteria* and *Firmicutes* folds exhibiting the lowest performance. This is probably because these two taxonomic groups predominantly comprise Gram-positive bacteria, making them highly dissimilar from the organisms in the other folds. DRAMMA demonstrated high performance on the external validation sewage samples as well. When assessed with the labels retrieved from DRAMMA-HMM-DB, which was the database used to label the training set, the model achieved a ROC-AUC score of 0.99 and a PR-AUC score of 0.91. Labeling the same data using ResFinderFG, a dataset of ARGs identified by functional metagenomic experiment, achieved a ROC-AUC score of 0.91 and PR-AUC score of 0.59. The

latter PR-AUC score, although lower than the former, remains significantly higher than what would be expected from a random classifier, as the frequency of positive genes in this case is only 0.0083. Furthermore, non-ARG proteins received considerably lower scores compared to ARGs (Supplementary Fig. 3). Among the metagenomic samples, only the human, animal, and plant microbiomes were enriched in ARG candidates (Fig. 4B). The enrichment within the plant microbiome environment was modest and could be attributed to the fact that plants may be exposed to antibiotics through waste disposal in soil and groundwater, thereby creating selective pressure for the emergence of ARGs. The considerable enrichment observed in the human and animal microbiomes can be explained by their continuous exposure to antimicrobial substances, which exerts significant selective pressure, driving the emergence of resistance genes that remain undiscovered, despite those environments being well studied. Within the groups, the *Bacteroidetes\Chlorobi*, *Betaproteobacteria*, *Firmicutes*, *Alphaproteobacteria,* and *Bacteria Candidate Phyla* groups exhibited an enrichment of ARG candidates (Fig. 4A). This enrichment can be attributed to the prevalence of *Bacteroidetes/Chlorobi*, *Proteobacteria*, and *Firmicutes* in human and animal intestinal microbiomes [50, 51]. These groups are thus exposed to antimicrobial substances, which could result in selective pressure for the emergence of ARGs. The enrichment observed in *Bacteria Candidate Phyla*, which is comprised of uncultured bacteria [52], suggests its potential role as a reservoir for undiscovered ARGs. In contrast, Stenosarchaea and Euryarchaeota exhibited notably negative enrichment values. This observation may suggest that the model is less adept at predicting ARGs originating from Archaea. Archaea have a natural resistance to particular classes of antibiotics, including those that target the synthesis or cross-linkage of the peptide subunit of murein, while exhibiting sensitivity to other types of antibiotics [53–55]. The depletion could be attributed to the limited investigation of ARGs in Archaea, to variation in their biophysical properties, or to their evolutionary distance from most well-characterized ARGs.

DRAMMAs' prediction was primarily driven by patterns within the taxonomic distribution of genes. In general, ARGs tended to have highly similar homologs across a wide range of taxonomic groups and even demonstrated more significant hits within specific taxonomic groups such as the *Firmicutes* and *Bacteroidetes\Chlorobi* groups (Supplementary Fig. 2). These findings align with the notion that ARGs confer an adaptive advantage and thus tend to disseminate to other organisms through HGT.

This study is subject to several limitations. First, inherent biases within the training data can affect the model's ability to generalize to new examples of resistance gene families, especially if resistance genes in less-represented taxa or ecosystems have unique patterns in terms of the features we measured. Second, the ARGs were labeled using HMM profiles of resistance genes, i.e., statistical models, as opposed to a manual curation process. Despite the stringent e-value we used to label ARGs, these profiles have the potential to generate false positives, which, when integrated into the model training, may impact its overall accuracy. Notably, the application of the HMM profiles used in this study for labeling ARGs led to the discovery of a novel ARG in *Nocardia*, which was experimentally validated [56]. Consequently, this approach not only facilitates model training on established ARGs but also provides an avenue for the identification and training based on potentially new ARGs.

By design DRAMMA is not trained to identify ARG-related genes: efflux pumps and genes that confer resistance through point mutation are not detected by our approach, as our focus is on discovering novel resistance genes. Given the critical role of these mechanisms in resistance, future work could benefit from integrating DRAMMA with models that specifically address these types of resistance genes. Additionally, this study only included contigs longer than 10 kbp, due to the model's reliance on genomic context features derived from neighboring genes. Shorter contigs would result in incomplete or inaccurate values for the encoded proteins and obscure potential signals. Consequently, while shorter contigs can also encode ARGs, our methodology is less suited for their analysis. Current features designed to detect HGT signals rely on profile HMM searches of neighboring genes against mobility-related genes, such as plasmid and phage genes, from the Pfam database. However, alternative databases or genome-based MGE detection tools, such as VIBRANT [57] and geNomad [58], could be employed in future studies to improve the detection of HGT signals. Finally, a range of bioinformatics tools was employed in this study, each with inherent limitations that could influence downstream analyses. Sequence similarity tools such as DIAMOND and MMseqs2 rely on local alignment, which might lead to local matches based on irrelevant regions of the proteins. While these tools are significantly faster than BLAST, they are slightly less sensitive, which could impact the accuracy of the phylogenetic distribution features and potentially result in missed homology or erroneous signals. This limitation may also affect the candidate filtration process, which depends on sequence similarity to previously annotated proteins. Additionally, CD-HIT, a clustering tool used for de-duplication of the training data, employs a greedy

incremental clustering algorithm that may generate suboptimal clusters. The clustering results can also be influenced by the order of input sequences, potentially leading to a partially redundant training set. Such issues could affect both model training and evaluation.

As part of DRAMMA, we developed a pipeline to compute novel biological-based features. Beyond their contribution to predicting novel antimicrobial resistance genes, they can be computed across diverse metagenomic samples and utilized for training machine learning models for a variety of biological questions, thereby providing new insights and characterizations of various genes.

In conclusion, DRAMMA offers the capability to supply rapid identification of both known and novel ARGs in large-scale genomic and metagenomic samples. By detecting sequences with low or no similarity to known ARGs, DRAMMA extends beyond the limitations of traditional sequence-based approaches. The model has the potential to expand the current ARG knowledge with genes with lower sequence similarity to those in the existing databases. Additionally, DRAMMA can potentially enable early detection of novel ARGs or genes associated with distinct resistance mechanisms before their widespread emergence in clinical settings. This could provide a crucial window for preventive intervention, allowing healthcare systems to implement modified treatment protocols before these resistance genes become widely distributed in bacterial populations. In the long term, our approach could contribute to the better use and effectiveness of existing and newly developed antimicrobial treatments.

Methods

Dataset compilation

The datasets used in this study included all assembled metagenomic contigs from various environments from both NCBI WGS [31] and EBI's Mgnify [59] (downloaded on March 14, 2020, Table 2). These were supplemented with all assembled genomes from NCBI GenBank Whole Genome Sequencing (WGS) database [31], downloaded on March 14, 2020, after filtering out *Fungi*, *Metazoan*, and *Viridiplantae*. In addition, as an external database, we assembled de novo all 235 metagenomic sewage samples from BioProject PRJEB27054. These samples were taken from 79 sites covering 60 different countries in seven different geographical regions as part of the Global Sewage Surveillance Project [9]. The assembly of the reads from the Global Sewage Surveillance Project was performed with MEGAHIT v1.2.2 [60], after using BBduk [61] for clipping adaptors and removing potential sequence contaminants (Illumina PhiX spike-ins, and other sequencing artifacts provided as part of BBTools [61]. Gene calling and initial gene annotation were performed using prodigal v.3.0[62]. (without restricting partial genes) and prokka v.1.12 [63], under the assumption that the vast majority of the contigs are from prokaryotic origin. Only coding sequences from contigs of length greater than 10 kbp were considered, since the genomic context (i.e., gene neighborhood) of ARGs is required for the model's features.

The training dataset was comprised of 34,311,250 proteins collected from NCBI's WGS metagenomes dataset [31] and EMBL-EBI Mgnify dataset [59]. It was divided into two: (a) a development set, comprised of ~ 10% of the training set (3,431,126 genes, $n_{pos} = 164{,}168, n_{neg} = 3{,}266{,}958$), which were used for the hyperparameter optimization and feature selection steps, (b) training set, comprised of the other ~ 90% of the proteins from the same dataset (30,880, 124 genes, $n_{pos} = 1{,}477{,}507, n_{neg} = 29{,}402{,}617$), which were used to train the final DRAMMA model.

For the classification task, ARGs were considered positive examples (see Data annotation below), and all other genes as negative examples. However, during the model development process, we noted that ARGs encoding efflux pumps were easily identified by the model. Therefore, two separate training sets were assembled: one for classifying efflux pumps that confer antibiotic resistance, and one for classifying AMRs that confer resistance through other mechanisms. We focused on the latter in order to detect genuinely novel ARGs.

To assess the performance of our methods on genes from species that were not part of the model's training, we compiled a genomic training dataset comprising 21,430,186 proteins ($n_{pos} = 1{,}308{,}921, n_{neg} = 20{,}121{,}265$) from NCBI's WGS genome dataset [31]. This dataset was divided into five major taxonomic groups: *Actinobacteria* (5,712,659 genes, $n_{pos} = 360{,}731, n_{neg} = 5{,}351{,}928$), *Gammaproteobacteria* (4,398,829 genes, $n_{pos} = 243{,}028, n_{neg} = 4{,}155{,}801$), *Firmicutes* (4,527,163 genes, $n_{pos} = 268{,}725, n_{neg} = 4{,}258{,}438$), *Alphaproteobacteria* (3,768,419 genes, $n_{pos} = 231{,}028, n_{neg} = 3{,}537{,}391$), and *Bacteroidetes* (3,023,116 genes, $n_{pos} = 205{,}409, n_{neg} = 2{,}817{,}707$).

The model was further evaluated on an external dataset of 6,118,656 genes from our assemblies of the Global Sewage Surveillance Project samples [9] (see above). This dataset was annotated for ARGs using two databases: (1) DRAMMA-HMM-DB, an in-house ARG HMM database containing genes acquired from Resfams [22], CARD October 2020 release [23], and HMD-ARG-DB [26] (see Data annotation section and Supplementary Datasets 1–3), which resulted in $n_{pos} = 187{,}700, n_{neg} = 5{,}930{,}956$ and (2) ResFinderFG v2.0 [47], a dataset of ARGs experimentally detected

using functional metagenomics, which resulted in $n_{pos} = 49{,}796, n_{neg} = 6{,}068{,}860$. To avoid high rates of "false positives" stemming from ARGs that are not represented in ResfinerFG, all proteins annotated as ARG by DRAMMA-HMM-DB database but not by ResfinderFG were excluded from the negative set. This process decreased the second dataset to $n_{neg} = 5{,}913{,}312$.

Data annotation

Known resistance genes were labeled using DRAMMA-HMM-DB, an ARG database of profile HMMs. HMMs are statistical models often used for homology search, as they allow rapid and sensitive detection of genes in large databases [64]. The database was comprised of profile HMMs from Resfams [22] as well as proteins from the CARD October 2020 release [23] and HMD-ARG-DB [26]. These databases were first curated to filter out genes that do not directly and specifically confer resistance. Hence, core genes that are targets of antibiotics were removed from the database, as were regulator genes and systems that are relevant to different types of substrates (e.g., ABC transporters that can transport antibiotics, as well as other compounds). Gene families that confer resistance by point mutations were removed as well, as we wished to focus on ARGs that are truly novel. This curation process reduced the number of Resfams HMM profiles from 173 to 90 (see Supplementary Dataset 1), the number of CARD proteins from 2734 to 2329, and the number of HMD-ARG-DB proteins from 17,282 to 10,483. We first merged CARD and HMD-ARG-DB and removed duplicate sequences using CD-HIT [65] version 4.6 (-g 1 -s 0.8 -c 0.9) which resulted in 2898 ARGs (see Supplementary Dataset 2). We then created a profile HMM for each sequence and united them with the profile HMMs from Resfams, resulting in our ARG HMM database of 2988 profiles (Fig. 7A).

For the benchmarking and external validation of the model, an additional ARG HMM database was compiled. This database was comprised of proteins from ResFinderFG v2.0, a recently published database based on experimentally validated ARGs through functional metagenomics. ResFinderFG was selected for external validation due to the several reasons: (1) It exclusively contains experimentally validated ARGs, (2) the ARGs in this database were detected in metagenomic samples, and (3) its reliance on functional metagenomics, an unbiased experimental approach that does not depend on a pre-defined set of known ARGs. The proteins from this database were also curated and de-duplicated in the same process described above, which resulted in 1067 ARGs. An HMM profile was then created for each of these proteins.

The genes were annotated using the HMMer suite [66] version 3.2.1 hmmsearch against our profile HMM databases described above. Proteins that passed the e-value threshold of 10^{-10} were considered positive examples (ARGs), and the negative examples were randomly sampled from the training set, excluding those ARGs, to prevent biases against a specific group of proteins. For each ARG in the positive set, we randomly included ten proteins in the negative set to provide an adequate representation of various non-ARG protein families and reduce false-positive rates. Duplicated sequences were then removed using CD-HIT with the same parameters as detailed above, resulting in a total of 34,311,250 proteins.

Feature extraction

To capture potentially relevant biological properties, hundreds of features corresponding to four main categories were extracted for each gene in our data: (1) amino acid properties, (2) amino acid patterns, (3) HGT signals, and (4) genomic context (Fig. 7B).

Amino acid properties

We measured the length of each gene, the length of the contig on which it resides, and the proportion of each amino acid in the gene's protein product. Moreover, the amino acid residues were grouped based on three different characteristics:

(1) Polarity of the residues: polar (NQYST) and nonpolar (AGILVFWPCM).
(2) The charge of the residues: positive (KR), negative (DE), and neutral charge (HAGILVFWPCMNQYST).
(3) Classes of residues: aliphatic (AGILV), amide (NQ), aromatic (FWY), hydroxyl-containing (ST), and sulfur-containing (CM).

The proportion of each of these amino-acids subgroups was then calculated to provide data on signals of enrichment of certain groups among AMR proteins.

In addition, more than 560 amino acid properties were extracted using AA indices, a database of numerical indices representing various physicochemical and biochemical properties of each amino acid [32]. The average of their values was calculated for each gene product using the amino acid proportion calculated beforehand. We next filtered correlated features reducing the number of amino acid indices to 39.

The physical and chemical properties of each protein were collected using ProtParam, an analysis module

from BioPython version 1.74 [67]. The properties calculated include molecular weight, aromaticity, grand average of hydropathy (GRAVY) value, instability index (i.e., an estimate of protein's stability in a test tube), isoelectric point (i.e., the pH at which the net electrical charge of the protein is zero), helix fraction, turn fraction, sheet fraction, and the molar extinction coefficient (i.e., the amount of light absorbed by a protein at a particular wavelength).

Amino acid patterns

Hydrophilic-hydrophobic signatures were created by dividing the amino acids of each protein into overlapping 8-mers that are converted into binary vectors according to whether each amino acid is hydrophilic (STNKYEQHDRZB) or hydrophobic (AGILMVPFWC). The frequency of each possible binary 8-mer signature was measured.

Information on helix-turn-helix (HTH) and transmembrane domains were also used as features. HTH domains were detected using HMMer suite [66] version 3.2.1 hmmsearch against a profile HMM database of families of HTH domains obtained from Pfam [35] (Supplementary Dataset 5). We extracted only domains that passed a bit-score similarity threshold of above 50 and collected as features the score and number of repeats of the highest-scoring domain. Transmembrane domains were detected using the tmhmm module [68] version 2.0c, which predicts transmembrane helices in proteins using HMMs. The predicted number of times the gene crosses the membrane was used as a feature.

HGT signals

GC content was retrieved, using BioPython [67] version 1.74, by calculating the combined proportion of guanine and cytosine in each gene and the contig on which it was found. The difference between the gene GC content and the contig GC content was measured to detect potential signatures of recent HGT, which often display atypical GC content [69, 70]. This was achieved by calculating the difference between the GC content of the gene and that of the contig after excluding the gene sequence. If the gene length was > 20% of the contig length, the difference was considered unreliable, and a null value was recorded instead. We also calculated the GC content differences while accounting for the amino acid sequence encoded by the genes. To this end, we considered the frequency of G and C only in synonymous positions, i.e., in which substitution between G/C and A/T does not change the amino acid in the product. For contigs encoding ≥ 5 genes, we calculated this feature as the difference between the gene GC content ratio and the mean ratio of all the other genes in the same contig. For each gene and contig, we further calculated the distribution of overlapping *k*-mers and their reverse complement for *k*s between 2 and 4. We calculated different distance measures (Euclidean distance, cosine distance, and correlation) between the vectors representing the *k*-mer frequencies in the gene and the contig.

The phylogenetic distribution of each gene was calculated using MMseqs2 [71]. Proteins from 12,643 proteome files of organisms across the three domains of life—Archaea, Bacteria, and Eukaryota, as well as viruses—were downloaded from the Universal Protein Resource (UniProt) [72] on November 2019 and organized into 53 different groups based on their taxonomy (Supplementary Table 3). In order to decrease the number of small taxonomic groups, we united all the viruses into one group and united all groups that had five members or fewer with other groups with shared taxonomy (on a higher level) unless the large group already had 24 or more organisms. This resulted in larger groups containing related organisms, while avoiding the creation of very large groups, thus providing manageable taxonomic clusters. We further assigned each group to a higher taxonomic level ("super-group") and its domain of life.

The proteins of each taxonomic group were used to create an MMseqs2 database. Each gene was searched against all these databases using mmseqs search (--search-type 1 -e 1e-6 --alignment-mode 1). The percentage of hits and the e-value exponent of the best hit in each group were extracted, as well as features taking into account the percentage of hits across the three taxonomic levels: the MMseqs2 database, the database's taxonomic supergroup, and the domain of life). We also calculated the quantiles (0.5, 0.75, 0.9) of the e-value exponents across all the databases. In case a protein had no significant hits against a specific database, the relevant features were filled with zero. In order to decrease the running time, taxonomic databases less relevant to this study were filtered out (Metazoa, Streptophyta, small sub-taxes of Opisthokonta, small sub-taxes of Rhodophyta). Further, highly similar proteins in each database were clustered using CD-HIT (-c 0.9 -s 0.8), and a representative protein of each cluster was mapped to all the taxonomic groups in which members of the cluster appeared. Finally, all the proteins were united into one database, while retaining an identifier of their original database.

Genomic context

For each gene, we searched for mobile element genes and ARGs in the same genomic region, which is defined either as a gene window of g genes ($g \in \{5, 10\}$) or as a nucleotide window of n nucleotides ($n \in \{5000,\ 10{,}000\}$). For each gene in question, the flanking genes within both windows, spanning upstream and downstream regions, were annotated using HMMer version 3.2.1 hmmsearch

against five in-house HMM profile databases, with an e-value threshold of 10^{-8}. The first profile includes AMR gene families, based on the same ARG HMM database we created for data annotation (see Supplementary Datasets 1–3). Other profiles include anti-defense genes such as anti-CRISPR genes, and anti-restriction genes, type VI secretion system (T6SS) immunity genes (Supplementary Dataset 6), and profile HMMs of mobility genes such as plasmid and phage genes from Pfam database [35] (Supplementary Dataset 7). In addition, we also utilized existing HMM profiles of phage "hallmark" genes (i.e., genes of viral origin that are annotated as major capsid protein, portal, terminase large subunit, spike, tail, coat, or virion formation proteins) from VirSorter [73], and HMM profiles of transposon genes from TnpPred [74]. For each database, the distance to the closest annotated gene and the number of genes within the defined windows were measured. If no genes were found in the relevant window, the distance feature was imputed with the value of twice the window size.

Feature selection

First, we removed highly correlated features by creating a correlation matrix between all features and applying hierarchical clustering using a correlation threshold of 95%. From each cluster, a single representative was selected. The representative was the feature with the highest importance score according to scikit-learn's [75] Random Forest, as calculated using a model trained on all the features.

Random Forest's default feature importance was used for selecting the optimal subset of features. In order to choose the number of top features to select, we trained models on the development set using different numbers of features and compared their performance in terms of mean PR-AUC and ROC-AUC (Fig. 8). The final set of features was then selected based on the feature importance values obtained by the model trained on the entire development set.

Hyperparameter optimization

Several machine learning models were compared in order to choose the optimal algorithm: Random Forest [76], LightGBM [77], Xgboost [29], Logistic Regression [78], SVM [79], and Multi-layer Perceptron [80]. These models were evaluated on the development set across five-fold cross-validation using mean PR-AUC, training time, and prediction time (see Supplementary Table 4). Random Forest was ultimately chosen for its balanced predictive performance and speed.

We later utilized a random grid search of the scikit-learn module [75] version 1.2.0 to assess the impact of various model hyperparameters on the model's five-fold cross-validation results on a random subset of 10% of our development set. We defined the best-performing combination of parameters as those that were present in all five folds and produced the best PR-AUC results (Supplementary Table 5).

A prominent hyperparameter that was tuned in our models was class weights, which can tackle the issue of class imbalance. By adjusting the loss function of the model, class weights penalize the misclassification of the minority class more severely than those of the majority class, thus enhancing the model's learning capabilities on the minority class. Therefore, we evaluated different strategies for assigning a high weight to the positive class (Supplementary Table 6).

Rather than searching all possible parameter combinations, we divided the hyperparameter optimization process into steps to reduce runtime. In each step, only a subset of features were optimized, while the remaining features were fixed with pre-determined values or the values chosen in previous steps. The first step was searching for the optimal max_depth and min_weight_fraction_leaf values while setting other parameters to pre-determined values (n_estimators = 250, bootstrap = True, oob_score = True, max_samples = 0.8, max_features = 0.8). The second step was searching for the optimal max_samples and max_features values while setting the parameters to the same pre-determined values or the values retrieved in the previous step. In the third step, the optimal values for the class_weight, oob_score, criterion, min_samples_leaf, and min_impurity_decrease parameters were sought, and the final step was searching for the best n_estimators value while setting the other parameters to values retrieved in the previous steps.

Model training and evaluation

The models' performances were evaluated in terms of mean PR-AUC and ROC-AUC across five-fold cross-validation and on external validation sets comprised of sewage metagenomic samples from the Global Sewage Surveillance Project [9]. To eliminate train-test leakage, we split the data into five folds while ensuring that genes encoded on the same contig appear in the same fold. We further mitigated leakage by employing Mmseqs2 linclust [71] to cluster the proteins based on their sequences, ensuring that highly similar proteins are not assigned to the different folds (see Supplementary Fig. 9). We also tried to eliminate leakage derived from evolutionary relatedness by training the model on genomic data and splitting it into folds according to five main taxonomic groups: *Actinobacteria*, *Gammaproteobacteria*,

Firmicutes, *Alphaproteobacteria*, and *Bacteroidetes* (Fig. 1B).

Runtime evaluation

The runtime of the feature extraction and model prediction processes was evaluated on the *E. coli* K-12 MG1655 genome, comprising 4329 proteins, as well as metagenomic samples from the Global Sewage Surveillance Project [9], which were randomly selected to compile a dataset of about 100 thousand proteins. The sewage dataset included four samples from various countries, totaling 4766 contigs and 100,532 proteins. The code was executed on a CentOS Linux 7 machine with two Intel(R) Xeon(R) Gold 6130 CPUs, utilizing all 64 available virtual CPUs.

Scrambled sequence analysis

DRAMMA's performance was further assessed on a set of scrambled protein sequences from the dataset compiled for the runtime evaluation. Each protein sequence was randomly shuffled, and the model's features were extracted using the scrambled sequences. In addition, features that rely on the sequences of neighboring genes were calculated using their unscrambled sequences. The model was then applied to the calculated features, with predictions made using the model score threshold that yielded a median precision score of 0.75 across the five-fold cross-validation conducted during model evaluation.

Benchmarking

The benchmarking evaluation was conducted on the sewage test set, labeled using mmseqs search (--search-type 1 -e 1e-4 --alignment-mode 1) against the filtered protein set from ResFinderFG v2.0 (see Data annotation section). A protein was considered positive (i.e., ARG) only if it received a hit with an e-value $\leq 10^{-10}$. A blacklist HMM database was compiled for proteins excluded from the evaluation. This database comprised the 80 profile HMMs representing efflux pumps, regulatory elements, and antibiotic targets from Resfams. Additionally, an HMM profile was generated for each protein filtered out of CARD 2020 and 2023 releases, HMD-ARG-DB, and ResFinderFG v2.0, and ARGNet-DB. Since ARGNet-DB is comprised of proteins from CARD and HMD-ARG-DB, an mmseqs search (--search-type 1 -e 1e-6 --alignment-mode 1) was performed against the excluded proteins from CARD, HMD-ARG-DB, and ResFinderFG v2.0 to filter out similar sequences from ARGNet-DB. This process resulted in a blacklist database containing 8508 profile HMMs (Supplementary Dataset 8). HMMer suite [66] version 3.2.1 hmmsearch was utilized to detect and filter out all proteins from the test set that have received a match with an e-value $\leq 10^{-6}$. All proteins that obtained an mmseqs search result against ResFinderFG with an e-value of $10^{-10} < e-value \leq 10^{-4}$ were considered ambiguous and excluded as well. In addition, protein sequences with codons identified as stop codons within their coding regions were removed as well. Negative examples were randomly sampled to achieve a ratio of 10 negatives per positive instance.

The performance of other ARG prediction tools on the resulting test set was assessed by true positive rate, false positive rate, MCC and macro precision, recall, and F1. The default parameters were employed for each algorithm unless specified otherwise. The algorithms assessed were as follows: (1) Resfams [22] (HMMer suite [66] version 3.2.1 hmmsearch --cut_ga), (2) DeepARG [21] v2 (deeparg predict --model LS --type prot), (3) ARGNet [30] (argnet.py --type aa --model argnet-l), (4) PLM-ARG [27] (plm_arg.py predict), (5) CARD's Resistance Gene Identifier strict search against CARD October 2020 release [23] and October 2023 release [34] (rgi -t protein), and (6) CARD's Resistance Gene Identifier loose search against CARD October 2020 release and October 2023 release (rgi -t protein --include_loose). The performance of DeepARG was evaluated using both a 0.8 score threshold and considering all results with no threshold. Their performance was compared to the performance of DRAMMA with two score thresholds corresponding to expected precision rates of 0.95 (for high precision) and 0.75 (for improved sensitivity).

Prediction of candidate ARGs using the trained model

The trained model was used to classify 649 million proteins from genomic and metagenomic genes acquired from a variety of ecosystems (Table 2). In order to analyze the most relevant proteins, we calculated the model score that achieved a median precision of at least 95% in the five-fold cross-validation performed as part of the model evaluation. All proteins that attained a model score higher than this threshold were considered "ARG candidates" and were subject to further analysis.

Analysis and filtration of ARG candidates

The ARG candidates were divided into two distinct groups, those derived from genomic samples and those originating from metagenomic samples. Subsequently, we further subdivided these groups based on their respective taxonomic groups and environmental origins, while defining taxonomic groups with a frequency smaller than 0.5% as "Other taxonomic groups." We next extracted information regarding the resistance mechanism and the antibiotics to which each positive candidate conferred resistance, leveraging data from the ARG

product that yielded the most significant HMM hit for that candidate. Antibiotics that appeared in less than 1% of the data were united into a single "Other antibiotics" category. Originally negative candidates above the threshold were marked as "Newly predicted." Since no specific information was available regarding the mechanisms of the "Aminotransferase class I and II" (ResfamID RF0024) and "Aminotransferase class IV" (ResfamID RF0025) ARGs, positive candidates associated with these hits were excluded from the mechanism and antibiotic drug analyses. Subsequently, we examined the distribution of the resistance mechanisms and antibiotic drugs across the different taxonomic groups and environments.

We further assessed the distribution of candidate ARGs across various taxonomic groups and environments. Subsequently, we determined the enrichment of these candidates within the taxonomical and environmental groups by comparing their distribution with that of all negative proteins (non-ARGs). The enrichment for each taxonomic group or environment g was calculated as follows: $\log_2\left(\frac{P_c(g)}{P_a(g)}\right)$, where $P_c(g)$ is the percentage of the group g among the novel candidates, and $P_a(g)$ is the percentage of group g among all the non-ARG proteins in our dataset.

We also analyzed the domains found in all novel candidates using an HMM search against Pfam [81] (downloaded on February 9, 2019) using an e-value threshold of 10^{-6}. The domains were determined by sorting the HMM hits by bit score, and removing hits that their alignment overlap with higher scoring domains. Next, a multi-step pipeline was developed and applied to detect candidates that are genuinely uncharacterized ARGs (see "Candidate selection," Fig. 6, Supplementary Fig. 7). The first step in this procedure was the removal of proteins with ontology annotation in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [36]. The remaining proteins (unannotated by KEGG's ontology) were clustered based on sequence similarity using mmseqs cluster (-s 7.5, -c 0.5), and the representative of each cluster was queried against NCBI's non-redundant protein database (nr) using DIAMOND [82] version 2.0.11 sensitive search (blastp -e 1e-6 --sensitive). The proteins with no result or no informative functional description were then searched for distant homologs to the remaining proteins using HH-suite [38] version 3 (hhblits -n -E e-10 -Z 500 -B 500) against both PDB [37] version 70 and Pfam [35] version 34.0. In all three steps, each protein was assigned the description of the top-scoring hit. The pipeline resulted in high-scoring proteins that have no functional annotation in close or remote homologs.

Automated processes were applied to gather as much information as possible on the resulting candidates. First, HMM search was used to detect proteins with sequence similarity to HTH or cross-membrane domains. We also utilized DIAMOND version 2.0.11 (blastp -e 1e-6) to search the candidates against each other to cluster similar proteins together. In the next step, we used MMseqs2's taxonomy search (-s 4) and DIAMOND protein search (-e 1e-6 –sensitive) against Uniprot's UniRef100 [39] in order to assign taxonomic classification of each protein and filter out Gram-positive bacteria, as the experimental testing at this stage was performed on *E. coli*, a Gram-negative bacterium.

Furthermore, candidate proteins that appeared consistently with common neighbors were filtered out since this could imply that the neighboring protein might be required for the candidate gene's function. First, MMseqs2 search (--search-type 1 -e 1e-10 --alignment-mode 1) was used in order to seek all occurrences of these proteins in our datasets and extract their neighboring genes (three flanking genes upstream and downstream). All the protein's neighbors were clustered using cd-hit version 4.6 (-s 0.5 -c 0.7) and proteins that shared similar neighbors in > 50% of the cases were filtered out.

AlphaFold3 [40] server was used for structure predictions of the candidate ARGs. These structures were subsequently searched against AlphaFold's structure database [41, 42], containing over 214 Million protein structures, which was downloaded in December 2022, using Foldseek [83] (-e 1e-6). The description of the annotated hit with the highest bit score was retrieved. In addition, the resistance mechanism of each candidate protein was predicted using a multi-class classification model (see below).

Resistance mechanism prediction using a multi-class model

A multi-class model for resistance mechanism classification was trained on the positive proteins from the main model's metagenomic training dataset using the hyperparameters and features selected for the main model. The ARGs were labeled one of five resistance mechanisms: (1) antibiotic target alteration ($N = 637{,}707$), (2) antibiotic inactivation ($N = 473{,}951$), (3) antibiotic target replacement ($N = 59{,}884$), (4) antibiotic target protection ($N = 133{,}519$), and (5) reduced permeability to antibiotic ($N = 11{,}089$), while ARGs with no known resistance mechanism were removed from the dataset, resulting in 1,316,150 proteins. The mechanism labels were created by mapping the results of the HMM search against DRAMMA-HMM-DB to the relevant resistance mechanism (Supplementary Dataset 1–2). The mechanism model was evaluated in terms of one-versus-all ROC-AUC and PR-AUC scores using both five-fold cross-validation on the training set and on the external sewage dataset used to evaluate the main

model ($N_{\text{alteration}} = 78{,}150, N_{\text{inactivation}} = 52{,}964, N_{\text{replacement}} = 8{,}041, N_{\text{protection}} = 23{,}096, N_{\text{reduced\_permeability}} = 2{,}059, N_{\text{total}} = 164{,}310$). The training set was split to folds in a stratified manner that also ensured that proteins encoded on the same contig will appear in the same fold. This multi-class model was used to predict the resistance mechanism of each candidate protein, such that the class that received the highest score and its score were returned.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02055-4.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
We thank Dr. Karin Mittelman for her feedback during the research and critical reading of the manuscript.

Authors' contributions
ER, SS, DB conceived and designed the analysis; ER, SS executed the feature extraction and model development; ER performed the data analysis and model evaluation; ER, SS, DB contributed to the study design; ER, DB wrote the manuscript.

Funding
ER and SS were funded in part by the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

Availability of data and materials
The data underlying this article are available in its online supplementary material and in Zenodo at the following links: https://doi.org/https://doi.org/10.5281/zenodo.14524530, https://doi.org/https://doi.org/10.5281/zenodo.14513933, and https://doi.org/https://doi.org/10.5281/zenodo.14524613. The datasets were derived from sources in the public domain: EMBL-EBI MGnify: https://www.ebi.ac.uk/metagenomics and NCBI Whole Genome Shotgun (WGS): https://www.ncbi.nlm.nih.gov/genbank/wgs. The code underlying this article is publicly available through GitHub at https://github.com/burstein-lab/DRAMMA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 August 2024 Accepted: 1 February 2025
Published online: 07 March 2025

References

1. Crofts TS, Gasparrini AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. Nat Rev Microbiol. 2017;15:422–34.
2. Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 2010;74:417–33.
3. Laxminarayan R, et al. Antibiotic resistance—the need for global solutions. Lancet Infect Dis. 2013;13:1057–98.
4. Van Boeckel TP, et al. Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. Lancet Infect Dis. 2014;14:742–50.
5. O'neill, J. Tackling drug-resistant infections globally: final report and recommendations. https://apo.org.au/node/63983 (2016).
6. Murray CJL, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 2022;399:629–55.
7. Smith R, Coast J. The true cost of antimicrobial resistance. BMJ. 2013;346:f1493.
8. Lewis K. Platforms for antibiotic discovery. Nat Rev Drug Discov. 2013;12:371–87.
9. Hendriksen RS, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun. 2019;10:1124.
10. Delgado-Baquerizo M, et al. The global distribution and environmental drivers of the soil antibiotic resistome. Microbiome. 2022;10:219.
11. Ryon KA, et al. A history of the MetaSUB consortium: tracking urban microbes around the globe. iScience. 2022;25:104993.
12. Danko D, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021;184:3376-3393.e17.
13. Forsberg KJ, et al. The shared antibiotic resistome of soil bacteria and human pathogens. Science. 2012;337:1107–11.
14. Lakin SM, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 2017;45:D574–80.
15. Zankari E, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67:2640–4.
16. Tsafnat G, Copty J, Partridge SRRAC. Repository of Antibiotic resistance Cassettes. Database. 2011;2011:bar054.
17. Gupta SK, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58:212–20.
18. Wattam AR, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42:D581–91.
19. Saha SB, Uttam V, Verma V. u-CARE: user-friendly comprehensive antibiotic resistance repository of Escherichia coli. J Clin Pathol. 2015;68:648–51.
20. Srivastava A, Singhal N, Goel M, Virdi JS, Kumar M. CBMAR: a comprehensive β-lactamase molecular annotation resource. Database. 2014;2014:bau111.
21. Arango-Argoty G, et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6:23.
22. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015;9:207–16.
23. Alcock BP, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48:D517–25.
24. Liu B, Pop M. ARDB–antibiotic resistance genes database. Nucleic Acids Res. 2009;37:D443–7.
25. Naas T, et al. Beta-lactamase database (BLDB) – structure and function. J Enzyme Inhib Med Chem. 2017;32:917–9.
26. Li Y, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome. 2021;9:40.
27. Wu J, et al. PLM-ARG: antibiotic resistance gene identification using a pretrained protein language model. Bioinformatics. 2023;39:btad690.
28. Rives A, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci. 2021;118:e2016239118.
29. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785.
30. Pei Y, et al. ARGNet: using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences. Microbiome. 2024;12:84.
31. Benson DA, et al. GenBank. Nucleic Acids Res. 2013;41:D36–42.
32. Kawashima S, Kanehisa M. AAindex: amino Acid index database. Nucleic Acids Res. 2000;28:374.

33. Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. Classification and Regression Trees. (Chapman and Hall/CRC, New York, 2017). https://doi.org/10.1201/9781315139470.
34. Alcock BP, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 2023;51:D690–9.
35. Mistry J, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9.
36. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
37. Sussman J, et al. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr. 1998;54:1078–84.
38. Steinegger M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20:473.
39. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23:1282–8.
40. Abramson J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
41. Varadi M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50:D439–44.
42. Barrio-Hernandez I, et al. Clustering predicted structures at the scale of the known protein universe. Nature. 2023;622:637–45.
43. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genomics. 2022;8:000748.
44. Feldgarden M, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11:12728.
45. Hunt M, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics. 2017;3: e000131.
46. Yin X, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics. 2018;34:2263–70.
47. Gschwind R, et al. ResFinderFG v2.0: a database of antibiotic resistance genes obtained by functional metagenomics. Nucleic Acids Res. 2023;51:W493–500.
48. Wallace JC, Port JA, Smith MN, Faustman EM. FARME DB: a functional antibiotic resistance element database. Database. 2017;2017:baw165.
49. Thakuria B, Lahon K. The beta lactam antibiotics as an empirical therapy in a developing country: an update on their current status and recommendations to counter the resistance against them. J Clin Diagn Res JCDR. 2013;7:1207–14.
50. Rajilić-Stojanović M, de Vos WM. The first 1000 cultured species of the human gastrointestinal microbiota. Fems Microbiol Rev. 2014;38:996.
51. Swanson KS, et al. Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. ISME J. 2011;5:639–49.
52. Tsurumaki M, Saito M, Tomita M, Kanai A. Features of smaller ribosomes in candidate phyla radiation (CPR) bacteria revealed with a molecular evolutionary analysis. RNA. 2022;28:1041–57.
53. Hilpert R, Winter J, Hammes W, Kandler O. The sensitivity of archaebacteria to antibiotics. Zentralblatt Für Bakteriol Mikrobiol Hyg Abt Orig C Allg Angew Ökol Mikrobiol. 1981;2:11–20.
54. Dridi B, Fardeau M-L, Ollivier B, Raoult D, Drancourt M. The antimicrobial resistance pattern of cultured human methanogens reflects the unique phylogenetic position of archaea. J Antimicrob Chemother. 2011;66:2038–44.
55. Khelaifia S, Drancourt M. Susceptibility of archaea to antimicrobial agents: applications to clinical microbiology. Clin Microbiol Infect. 2012;18:841–8.
56. Hershko, Y., Rannon, E., Adler, A., Burstein, D. & Barkan, D. WarA, a remote homolog of NpmA and KamB from Nocardia wallacei, confers broad spectrum aminoglycoside resistance in Nocardia and Mycobacteria. Int. J. Antimicrob. Agents 107089 (2024) https://doi.org/10.1016/j.ijantimicag.2024.107089.
57. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90.
58. Camargo AP, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2024;42:1303–12.
59. Gurbich TA, et al. MGnify genomes: a resource for biome-specific microbial genome catalogues. J Mol Biol. 2023;435:168016.
60. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
61. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA, USA (2014).
62. Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
63. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
64. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.
65. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
66. HMMER. http://hmmer.org/.
67. Cock PJA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
68. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes11Edited by F. Cohen J Mol Biol. 2001;305:567–80.
69. Lawrence JG, Ochman H. Reconciling the many faces of lateral gene transfer. Trends Microbiol. 2002;10:1–4.
70. Zhang R, Ou H-Y, Gao F, Luo H. Identification of horizontally-transferred genomic islands and genome segmentation points by using the GC profile method. Curr Genomics. 2014;15:113–21.
71. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
72. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.
73. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
74. Riadi G, Medina-Moenne C, Holmes DS. TnpPred: A web service for the robust prediction of prokaryotic transposases. Int J Genomics. 2012;2012:e678761.
75. Pedregosa F, et al. Scikit-learn: Machine learning in Python. JMLR. 2011;12:2825–30.
76. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
77. Ke G, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:52.
78. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Methodol. 1958;20:215–42.
79. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
80. Haykin S. Neural networks: a comprehensive foundation. USA: Prentice Hall PTR; 1998.
81. Bateman A, et al. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138–41.
82. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
83. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. 1–4 (2023) https://doi.org/10.1038/s41587-023-01773-0.

Publisher's Note