`resLens`: genomic language models to enhance antibiotic resistance gene detection

Matthew Mollerus[1], Katharina Dittmar[1], Keith A. Crandall[1], Ali Rahnavard[1*]

[1]Computational Biology Institute, Department of Biostatistics and Bioinformatics, The George Washington University, Washington, 20052, DC, USA.

*Corresponding author(s). E-mail(s): rahnavard@gwu.edu;
Contributing authors: matthew.mollerus@gwu.edu; katharinad@gmail.com; kcrandall@gwu.edu;

Abstract

The pace of antibiotic resistance necessitates advanced tools to detect and analyze antibiotic resistance genes (ARGs). We present *resLens*, a family of genomic language models (gLM) leveraging latent genomic representations for ARG detection and analysis. Unlike alignment-based methods constrained by reference databases, *resLens* fine-tunes pre-trained gLMs on curated ARG datasets, achieving superior performance across several evaluation scenarios, including when ARGs exhibit dissimilar sequences and mechanisms to those in reference databases.

Keywords: Antibiotic Resistance, Language Models, Bioinformatics, Metagenomics, Genomics

Main

Rising levels of antibiotic resistance in microbial pathogens require new tools to study the presence and evolution of antibiotic resistance genes (ARGs). We propose *resLens*, a genomic language model (gLM) that leverages its latent understanding of genomic data to improve detection of ARGs from metagenomic data and to identify potential

novel ARGs. Existing tools are either alignment-based, like ResFinder[1] and ARG-ANNOT[2], use alignment-derived features for machine learning like DeepARG[3], or use protein sequence data like ARGNet[4]. These approaches perform poorly when variants do not closely match ARGs in their database. Reference databases represent a small fraction of the resistome and can struggle to keep up with the rapid pace of resistance evolution[5]. Relatedly, these approaches are incapable of identifying novel genes or mutations conferring resistance[6, 7]. Other efforts to predict resistance phenotype from genotype rely on combinations of whole genome data[8, 9], species-specific models[10], or causal information about the resistance evolution to predict novel genes[11], all of which are challenging in many real-world applications.

Applying large language model architectures, specifically transformer encoder models[12], to DNA sequence data yields a latent understanding of genomic elements and state-of-the-art performance on genomic classification tasks[13, 14]. gLMs divide a sequence into subsequences called tokens, embed them in a high-dimensional space, and update the embeddings based on the sequence context to generate an output embedding for unsupervised or supervised analysis. They encode information about sequence identity and biological function[15]. The models are pre-trained on large datasets of whole genomes(Figure 1a) via masked language modeling and then fine-tuned on task-specific labeled datasets (Figure 1b). *resLens* fine-tunes a *seqLens* 87M pre-trained gLM[16] on a broad ARG dataset, using the output embeddings to classify genes on their ability to confer resistance to different antibiotics. The fine-tuning

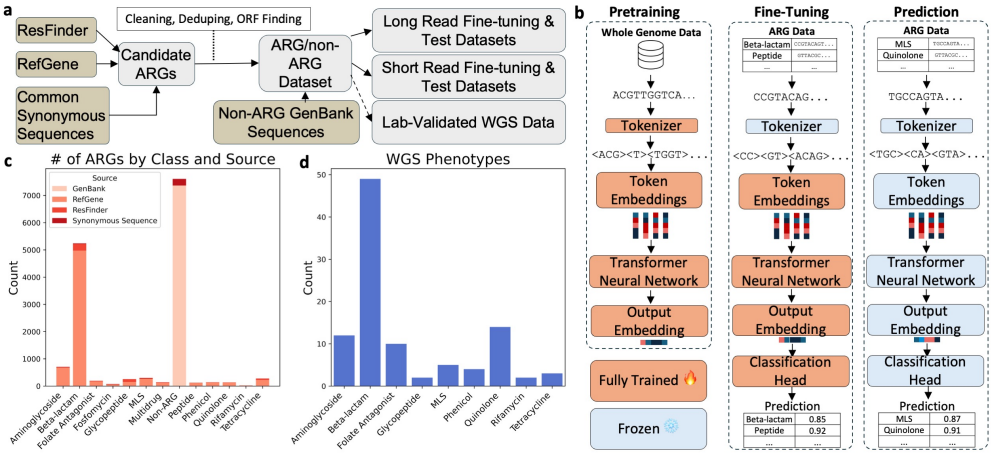

Fig. 1: Data and modeling overview. **a**, Flow chart of construction and application of datasets. **b**, Overview of gLM pretraining, fine-tuning, and inference, with frozen and trainable layers at each step displayed. **c**, Distribution of phenotypes presented in LR training and test data, colored by source. **d**, Distribution of phenotypes presented in lab-validated WGS data.

datasets consisted of known ARGs from ResFinder[1] and the NCBI Pathogen Detection RefGene database[17], as well as synonymous variants of common ARGs. We deduplicated the dataset and kept the longest gene in the case that one gene was a perfect subsequence of another(Figure 1a). We considered only genes that conferred resistance to antibiotic classes with at least 20 instances. We then passed the sequences through Prodigal[18] to ensure they contained only open reading frames (ORF). This preprocessing yielded 7,606 ARGs over 12 classes of antibiotics(Figure 1c). This is fewer classes than many tools classify, as merging databases requires consolidating more specific classes into parent classes. We combined these ARGs with an equal number of non-resistance genes of similar length, processed with the same ORF finding process. We created a short-read dataset by splitting up the sequences into 150 bp reads. We finetuned four models, two each for both the whole gene and short read data. For each dataset, one model performed a binary classification of ARG or non-ARG, then a second classified the predicted ARGs into specific ARG classes. We split each dataset into 80% train and 20% test data and performed 10-fold cross-validation with the train dataset, then evaluated the best model checkpoint from cross-validation via weighted F1 on the test set. We split sequences longer than the model's context into two sequences that inherited the same label. We evaluated the *resLens* models, DeepARG, and ResFinder on both test datasets. To test resLen in a realistic setting, we evaluated the long read model against a sample of WGS data of organisms with laboratory-confirmed resistances compiled by VanOeffelen et al.[19]. Filtering and mapping classes to those predicted by *resLens* yielded 83 genomes with resistance to one to three antibiotic classes, resulting in 101 resistance phenotypes total(Figure 1d). Many had genes that confer resistance to antibiotics other than their annotations or were selected in our sampling, likely due to the laboratory testing being non-exhaustive. Specific genes conferring resistance were also not annotated. Given these constraints, we considered a model correct for a given genome and resistance phenotype if it identified at least one gene associated with the resistance phenotype. We evaluated the long-read *resLens* model, re-fine-tuned with our whole dataset, and ResFinder against the WGS data. We used ORFs identified by Prodigal as inputs for *resLens* and considered *resLens* predictions that were above the 90th percentile prediction probability for their ARG class on the whole dataset. *resLens* outperformed ResFinder and DeepARG on both test datasets (Figure 2a-d). It performed notably better on short reads, potentially indicating that the contextual understanding of gLMs improve performance on partial sequences. On the WGS dataset, *resLens* outperformed ResFinder in the number of genomes with a correctly identified phenotype (92.8% vs 83.1%) and in phenotypes with a corresponding identified ARG (94% vs 77.2%) (Figure 2e). *resLens* performed slightly worse in identifying enzymes highly similar to those in reference databases, like well-known beta-lactamases, while it substantially outperformed ResFinder on multidrug efflux pumps and genes in epistatic or multicomponent mechanisms of resistance (Figure 2f). This could be due to alignment-based algorithms' ability to handle degeneracy, while gLMs rely on a small fraction of possible synonymous sequences in training. We intend to validate this behavior of gLMs on degenerate sequences in future work.

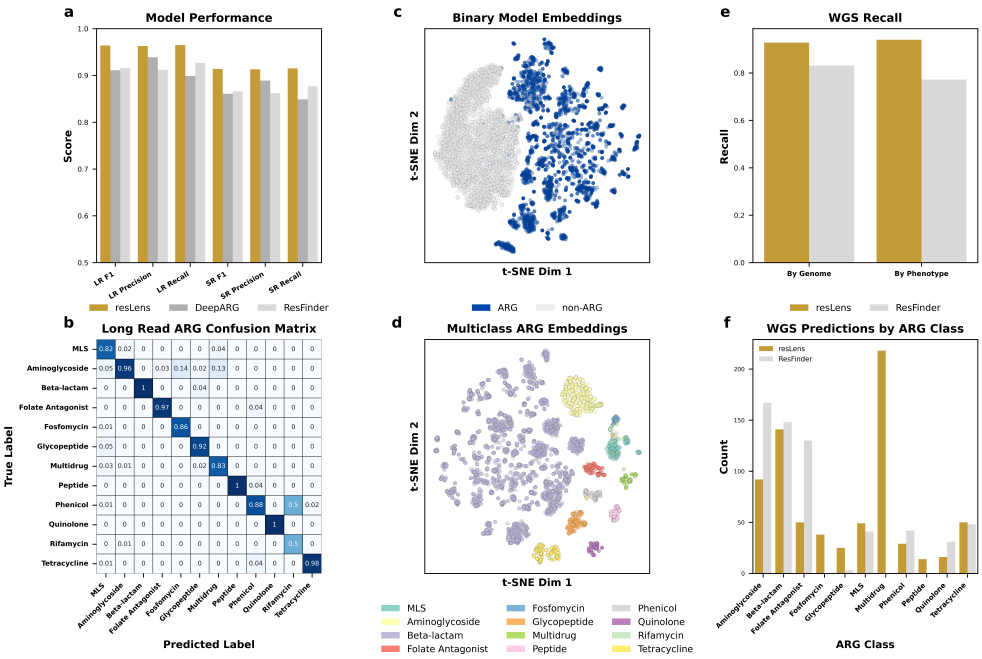

Fig. 2: Comparison of classification performance across different methods and datasets. **a**, Bar plot showing weighted F1 scores for resLens, DeepARG, and Best-Hit across Long Read, Short Read, and Simulated Data, with *resLens* highlighted in green. **b**, Confusion matrix for *resLens* using Long Read data, illustrating the distribution of true positives, true negatives, false positives, and false negatives, with darker shades representing higher counts. **c**, t-SNE plot of CLS token embeddings for the binary ARG/non-ARG model, colored by class. **d**, t-SNE plot of CLS token embeddings for the multiclass ARG model, colored by class. **e**, *resLens* and ResFinder performance on WGS recall by genome and phenotype. **f**, Distribution of ARGs identified by *resLens* and ResFinder in WGS data.

To evaluate the performance of the model on novel ARGs, we identified two gene families, one conferring resistance to beta-lactams and one conferring resistance to aminoglycosides and had low sequence similarity to other gene families conferring resistance to the same antibiotics. These were the blaADC family of beta-lactamases and the aminoglycoside nucleotidyltransferase (ANT) gene family; the blaADC sequences had a maximum of 61.6% identity with the closest other beta-lactam ARG sequence, while the ANT sequences had a maximum of 49.8% identity with the closest other aminoglycoside ARG sequence. The pre-trained model was fine-tuned with the whole dataset, excluding the held-out subgroups. The model then made a forward pass with all the data, and we extracted the CLS token from the final layer. It classified the heldout genes at a slightly lower rate than in the initial testing (100% for the blaADC genes in both cases, 84.7% vs 99.7% for ANT genes). This indicates that *resLens* can accurately predict the resistance phenotype of genotypes fundamentally different from

those seen in training. The model is not just 'memorizing' the dataset, but potentially developing representations of mechanisms by which genes provide resistance to an antibiotic.

This conclusion is further supported by examining *resLens*'s false positives from the WGS dataset. For instance, it identified an ATP-binding cassette domain-containing protein from *Streptococcus pneumoniae* as an ARG to macrolide-lincosamide-streptogramin (MLS) antibiotics. While no literature suggests that this protein confers resistance to MLS antibiotics, many of its structures share extensive similarity with those in a protein controlling resistance to tylosin, a macrolide antibiotic, in *Streptomyces fradiae*, suggesting potential common origin and function[20]. Using tblastn[21], the protein sequence had only 28.4% identity with the closest nucleotide sequence in our ARG dataset. This sequence dissimilarity validates that *resLens* relies on other features, potentially involving latent understanding of gene structure and function.

Correctly classifying novel ARGs represents an improvement over alignment-based methods, as accuracy is less tied to the quality and recency of a reference database. It enables researchers to identify and analyze potential new mechanisms of resistance more rapidly and comprehensively, speeding up the loop between in silico and in vitro experimentation. *resLens*'s performance indicates that gLMs could improve analysis of novel genotype-to-known-phenotype relationships in other domains.

References

[1] Florensa, A. F. et al. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb. Genomics* **8**, 000748 (2022).

[2] Gupta, S. K. et al. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).

[3] Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).

[4] Pei, Y. et al. ARGNet: using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences. *Microbiome* **12**, 84 (2024).

[5] Allen, H. K. et al. Call of the wild: antibiotic resistance genes in natural environments. *Nat. Rev. Microbiol.* **8**, 251–259 (2010).

[6] Galhano, B. S. P. et al. Antimicrobial Resistance Gene Detection Methods for Bacteria in Animal-Based Foods: A Brief Review of Highlights and Advantages. *Microorganisms* **9**, 923 (2021).

[7] Berglund, F. et al. *Microbiome* **7**, 52 (2019).

[8] Drouin, A. et al. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *BMC Genomics* **17**, 754 (2016).

[9] Davis, J. J. et al. Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.* **6**, 27930 (2016).

[10] Yang, Y. et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* **34**, 1666–1671 (2018).

[11] Drouin, A. et al. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **9**, 4071 (2019).

[12] Vaswani, A. et al. Attention is all you need. *NeurIPS* 6000–6010 (2017).

[13] Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* (2024).

[14] Ji, Y. et al. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[15] Sanabria, M. et al. Distinguishing word identity and sequence context in DNA language models. *BMC Bioinformatics* **25**, 301 (2024).

[16] Baghbanzadeh, M. et al. seqLens: optimizing language models for genomic predictions. *bioRxiv* https://doi.org/10.1101/2025.03.12.642848 (2025).

[17] The NCBI Pathogen Detection Project (2016).

[18] Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

[19] VanOeffelen, M. et al. A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes. *Brief. Bioinform.* **22**, bbab313 (2021).

[20] Rosteck, P. R. et al. Homology between proteins controlling Streptomyces fradiae tylosin resistance and ATP-binding transport. *Gene* **102**, 27–32 (1991).

[21] Altschul, S. F. et al. Basic local alignment search tool. *J. Mol. Biol.* **25**, 215:403–410 (1990).

Methods

Data

We initially collected ARGs from two sources: version 2.4.0 of the ResFinder database and RefGene sequences labeled as conferring antimicrobial resistance. Full datasets are available for download from our HuggingFace repository (see *Data Availability* section).

Given that the datasets varied in their level of granularity in assigning ARG classes to genes, we used the lowest level classifications available across all three datasets; however this turned out to be a relatively high level (e.g., 'cephalexin', 'cephalosporins', and 'beta-lactam' were all coalesced into the class beta-lactam, as cephalexin is a type of cephalosporin, which is a subclass of beta-lactam antibiotics). A full class mapping is available in Table S1.

We then combined these datasets and removed genes that were either perfect duplicates or perfect subsequences of other genes and that conferred resistance to the same class of antibiotic, prioritizing retaining ResFinder then RefGene. We then retained only sequences that had at least 20 instances in the dataset and fed the data through Prodigal to ensure they all constituted only ORFs, yielding 7,606 ARGs over 12 classes.

To acquire negative examples, we queried GenBank for 100,000 random bacterial sequences between 100 and 6,000bp, approximately the same length distribution as our ARGs. We then randomly selected an equal number of these sequences and performed the same ORF finding process with Prodigal to construct the non-ARG class in our dataset.

After initial testing on an exploratory WGS dataset drawn from the same database as our test WGS dataset, we supplemented the dataset with synonymous ARG sequences from gene families that *resLens* was frequently predicting incorrectly, as well as with non-ARG sequences that *resLens* was frequently classifying as ARGs. A full table of genes by class is available in Supplementary Table 2, and distribution of gene lengths is in Figure S1.

We used this dataset as is to fine-tune the long-read model. For the short read model, we split up the genes into 150bp short reads, with each inheriting the class label of the original gene. We then used 80% of each of these datasets for training and validation, and the remaining 20% as a test set. We used the same training and test splits for the two stages of each model (see *Language Model Training* below), with non-ARG sequences excluded from the dataset for fine-tuning the second stage model to distinguish between different ARG classes.

The WGS dataset consisted of 100 randomly selected sequences from the database assembled by VanOeffelen et al., each of which comes from studies that validated one or several resistance phenotypes of an organism in a lab. As the lab testing was non-exhaustive for all ARG classes and not all were selected in our sampling, many genomes contained ARGs to classes other than the labeled ones. Additionally, the exact mechanisms of resistance or genes implicated were not annotated. These constraints are inherent to any large lab-validated database, so we chose to adapt our evaluation methods to them rather than construct synthetic datasets. Filtering out genomes with resistances to antibiotics that *resLens* did not predict left 83 genomes with 103 resistance phenotypes. The distribution of resistance phenotypes is in Figure 1c. Each file contained several hundred contigs ranging from 1 kbp to 100 kbp.

Language Model Training

Our foundation gLM was the SeqLens 89M model from Baghbanzadeh et al. [16]. It uses BERT architecture and was pretrained on large quantities of genomic data from across all kingdoms of life and demonstrated state-of-the-art performance on genomic understanding tasks involving prokaryotic genomes.

We finetuned two models each for the long read (LR) and short read (SR) versions of *resLens*, one binary classification model to distinguish between ARGs and non-ARGs, and one multiclass classification model to categorize those sequences that the first model had identified as ARGs. Each model was fully fine-tuned for 8 epochs with a learning rate of 1e-5, a batch size of 8, and 2 gradient accumulation steps. We conducted ten-fold cross-validation, recorded model performance on the validation set every 500 steps, and saved a checkpoint of the best model observed by validation weighted F1 during the training process; this best model was used to evaluate performance on the test set. All training was conducted on NVIDIA Tesla V100 GPUs at the George Washington University High Performance Computing cluster.

Long and Short Read Evaluation

We evaluated the models against the test LR and SR datasets in two different ways. First, we evaluated each *resLens* model on the appropriate version of the test dataset, meaning that we scored the binary model on its ability to distinguish ARGs and non-ARGs in the test dataset and the multiclass ARG model on its ability to distinguish between different the types of ARGs in the test dataset, with non-ARGs excluded. Second, we evaluated *resLens*, ResFinder, and DeepARG on a 'full pass' of the test dataset, meaning that we had the *resLens* binary model predict which genes were ARGs and then had the multiclass ARG model make predictions for only the genes the binary model had classified as ARGs; we ran the sequences through ResFinder and DeepARG with their default parameters, except for the short read data, we ran ResFinder with a minimum coverage of 10% to account for the shorter sequences.

Figure S2 and Table S3 show the ten-fold validation results of the four *resLens* models. Figure S2 and Table S4 outline the *resLens* test scores on the first method. Figure 2a outlines the performance of each tool via the second method. Figure S3-Figure S6 present further training and evaluation data from the *resLens* models.

WGS Evaluation

Due to the WGS data coming from non-exhaustively tested organisms and lacking annotation of specific genes or mechanisms conferring a given resistance phenotype, we focused our analysis on 1) the recall rate for the lab-validated phenotypes and 2) the amount of the resistome correctly identified.

We evaluated the *resLens* long read models, fully re-fine-tuned on our entire dataset to ensure that the models had the opportunity to learn all of the resistance gene families, and ResFinder on the 83 genomes we selected. We used Prodigal to identify ORFs in the genomes, then used the two-stage *resLens* models to make predictions on each ORF. We wanted to restrict the *resLens* output to only high-confidence predictions,

so we used only those predictions that were in or above the 90% percentile in prediction probability for each ARG class. We ran ResFinder with its default parameters of 60% overlap and 90% identity thresholds, provided the species name when available, and restricted our analysis to those antibiotic classes that *resLens* also predicts for fair comparison. This yielded 1,046 predictions from *resLens* and 713 from ResFinder.

We validated the *resLens* predictions by comparing the outputs to NCBI's non-redundant protein sequence database via blastp and examining results with 100% identity. If the entry's page or a literature review of that protein indicated that it conferred resistance to the predicted antibiotic class, we labeled it as correct. In some cases, proteins confer resistance to multiple antibiotic classes or have mixed evidence, which we outline in Supplementary Table 5. Only those genes with unambiguous evidence or resistance to the correct class were marked as correct predictions for our analysis. This left us with 702 (67.1%) unambiguous true positives, 246 (23.5%) unambiguous false positives, and 97 (9.3%) ambiguous cases or cases where the model identified an ARG but for the wrong class of antibiotic. We took ResFinder's results as entirely true positives due to the strict identity cut-offs.

Novel ARG Classification

To evaluate the ability of this methodology to predict the phenotype of novel or unseen ARGs, we re-fine-tuned the foundation model with certain gene families removed from the dataset, then evaluated the ability of the new model to classify and cluster genes from the heldout families.

First, we searched our long-read dataset for gene families that were 1) in ARG classes that were well-represented in our data and had a high diversity of gene families conferring resistance and 2) had low sequence similarity to other gene families that also conferred resistance to the same class of antibiotics. We relied on the gene family annotations from the ResFinder and RefGene databases for this search. This resulted in us selecting the blaADC gene family of class A beta-lactamases from the ARGs that conferred resistance to beta-lactam antibiotics and the aminoglycoside nucleotidyltransferase (ANT) gene family from the ARGs that conferred resistance to aminoglycoside antibiotics. The blaADC sequences had a maximum of 61.6% and a median of 60.4% identity with the closest other genes conferring beta-lactam resistance; the ANT sequences had a maximum of 64.8% and a median of 50.9% to the closest other genes conferring aminoglycoside resistance. We then removed the genes in these two families from the dataset and placed them in a separate holdout set. Additionally, we removed any genes in the beta-lactam and aminoglycoside ARG classes that lacked a gene family annotation from the remaining dataset in order to eliminate the possibility of data leakage. Table S7 details the breakdown of the remaining datasets and heldout genes.

Removing the blaADC sequences posed an easier version of the problem of classifying novel ARGs, as there were other class A beta-lactamase genes in the training dataset; despite their modest sequence similarity, all class A beta-lactamase genes confer resistance via similar mechanisms of serine hydrolysis. On the other hand, removing the ANT sequences posed a harder version of the problem as they both had lower

sequence similarity with and a fundamentally different mechanism of resistance to other aminoglycoside ARGs.

We fine-tuned the same foundation models for multiclass ARG classification with the same hyperparameters and the main models, but this time for only one fold, with 20% of the training data used as a validation set and the best model selected via weighted F1 score on this set. We then evaluated these models' performance on the two heldout gene families.

We additionally performed a forward pass with all of the data, both training and heldout data, and extracted the embeddings for the CLS token for each gene from the second stage model used to predict specific ARG classes. We then performed dimensionality reduction using t-SNE on this embedding data to visualize the results.

Data Availability

All of the data used for training and testing the models, as IDs to download the whole genome sequencing data from PATRIC, is available at https://huggingface.co/collections/omicseye/reslens-6834aae78d6a59c46d156744.

Code Availability

All of the code used to train the models, perform inference, and to generate figures is available at https://github.com/omicsEye/resLens.

Acknowledgments

This work was supported by the National Science Foundation grant 2109688 to AR and KAC. We are grateful to the GW High Performance Computing team for providing outstanding resources and support for this work.

Author contributions

A.R. and M.M. conceived the method; M.M. implemented the approach; M.M. tested and packaged the software and evaluated the performance; M.M. and A.R. provided online documents and software. M.M., A.R. analyzed application datasets. KAC and KD guided data interpretation and result structuring. M.M. and A.R. drafted the manuscript. A.R. and K.C. acquired funding for the project. All authors discussed the results and wrote the manuscript.

Competing interests

AR and KAC are inventors of a pending patent related to the work presented and co-founders of seqSight, a company associated with the technology. The other authors declare no competing interests.

Supplementary Materials

Table S1: Antibiotic classes and antibiotics assigned to them.

Antibiotic Class	Antibiotics
Aminoglycoside	Amikacin, aminoglycosides, capreomycin, gentamicin, kanamycin, neomycin, tobramycin
Beta-lactam	Amoxicillin, ampicillin, aztreonam, carbenicillin, cefazolin, cefepime, cefoperazone, cefotaxime, cefotetan, cefoxitin, cefpirome, cefpodoxime, ceftazidime, ceftriaxone, cefuroxime, cephalothin, clavulanic acid, doripenem, ertapenem, imipenem, meropenem, methicillin, penicillin, piperacillin, sulbactam, tazobactam, temocillin
Folate pathway antagonist	Sulfamethoxazole, sulfisoxazole, sulfonamide, trimethoprim, trimethoprim
Fosfomycin	Fosfomycin, phosphomycin, phosphonomycin
Glycopeptide	Vancomycin, glycopeptide
MLS	Azithromycin, clarithromycin, clindamycin, erythromycin, lincosamide, macrolide, streptogramin, streptogramin A, streptogramin B, virginiamycin
Multidrug	Any sample conferring resistance to two or more other classes
Peptide	Actinomycin, colistin, polymyxin
Phenicol	Amphenicol, chloramphenicol, florfenicol, thiamphenicol
Quinolone	Ciprofloxacin, fluoroquinolone, gemifloxacin, levofloxacin, moxifloxacin, nalidixic acid, norfloxacin, ofloxacin, quinolone
Rifamycin	Rifabutin, rifampicin, rifamycin
Tetracycline	Doxycycline, glycylcycline, tetracycline, tigecycline

Table S2: Number of samples by antibiotic class in long read (LR) and short read (SR) training and test sets.

Antibiotic Class	LR Train	LR Test	SR Train	SR Test
Aminoglycoside	575	131	2,281	549
Beta-lactam	4,211	1,028	23,889	5,798
Folate pathway antagonist	155	37	532	131
Fosfomycin	62	13	144	30
Glycopeptide	208	49	1,214	281
MLS	235	65	1,449	390
Multidrug	115	23	1,381	272
Non-ARG	6,047	1,559	22,459	5,716
Peptide	97	28	959	283
Phenicol	116	25	632	144
Quinolone	114	21	529	84
Rifamycin	20	2	110	20
Tetracycline	214	62	2,120	604

Table S3: 10-fold validation results from *resLens* models. Reported values are mean ± standard deviation across folds.

Model	Accuracy	F1	MCC
LR Binary	0.978 ± 0.005	0.978 ± 0.005	0.956 ± 0.010
LR Multiclass	0.982 ± 0.007	0.982 ± 0.007	0.965 ± 0.014
SR Binary	0.901 ± 0.002	0.878 ± 0.003	0.797 ± 0.040
SR Multiclass	0.915 ± 0.002	0.913 ± 0.002	0.834 ± 0.004

Table S4: *resLens* model performance on held-out test datasets.

Model	Accuracy	F1	MCC
LR Binary	0.977	0.978	0.954
LR Multiclass	0.977	0.977	0.955
SR Binary	0.897	0.876	0.790
SR Multiclass	0.904	0.902	0.816

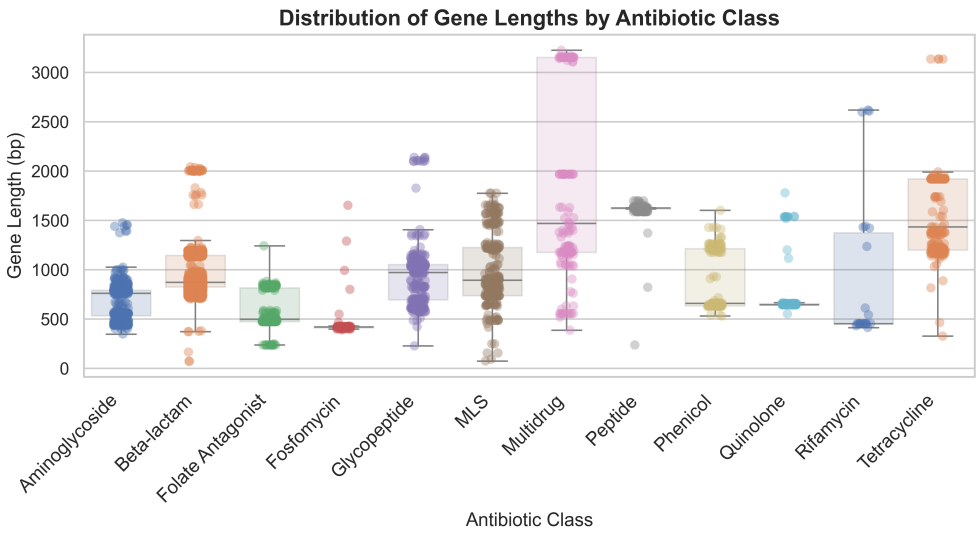

Fig. S1: Distribution of gene lengths by ARG class.

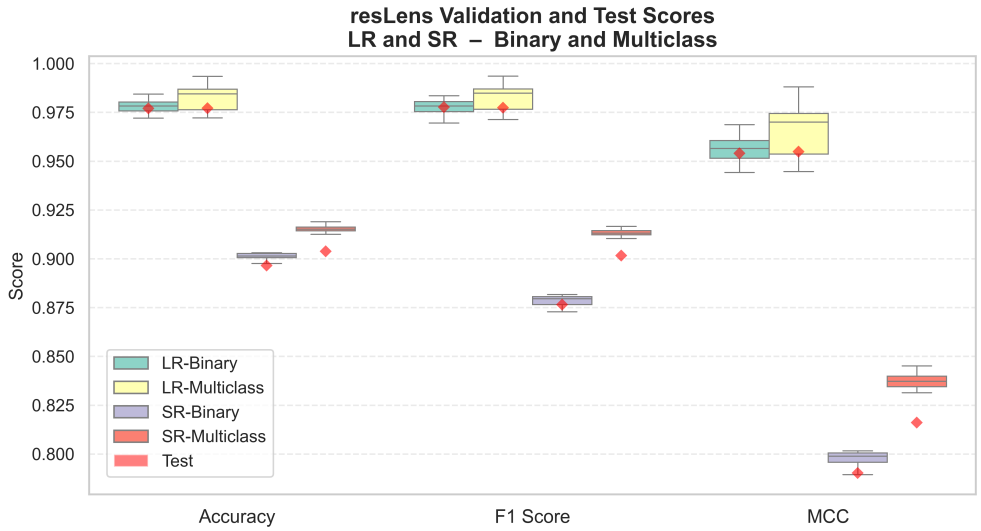

Fig. S2: *resLens* models' performance on test data. Boxplots indicate 10-fold cross validation scores, and diamonds indicate test scores.

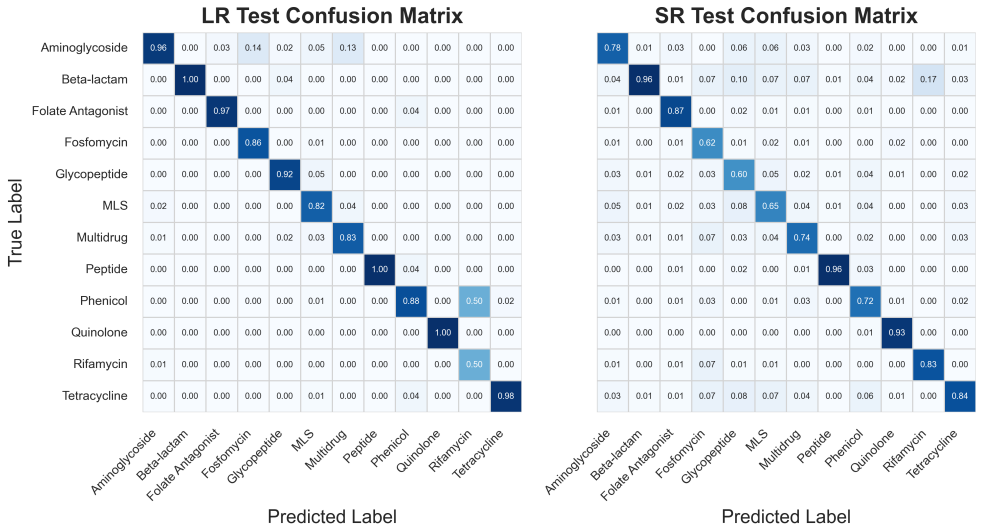

Fig. S3: Confusion matrices of the multiclass model on the multiclass long-read and short-read test datasets.

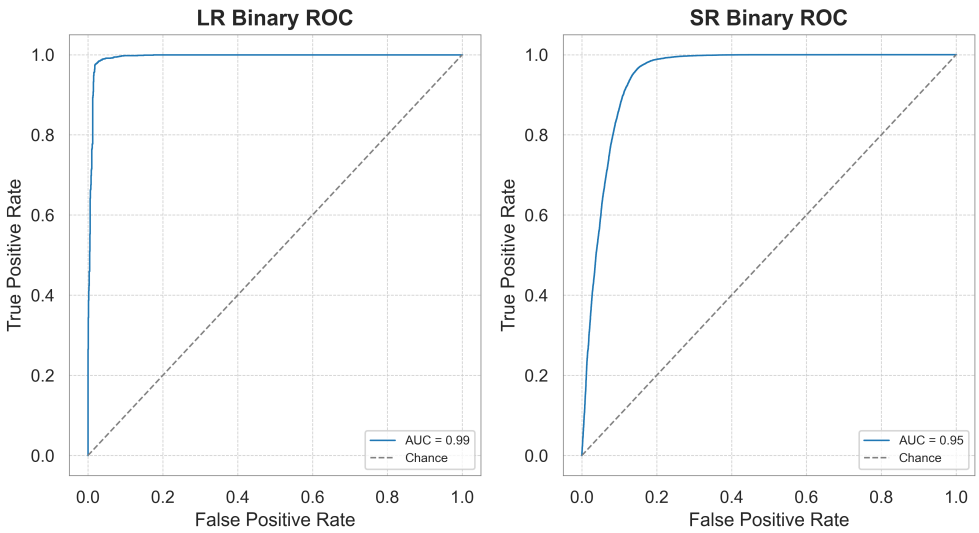

Fig. S4: ROC curves for LR and SR Binary models.

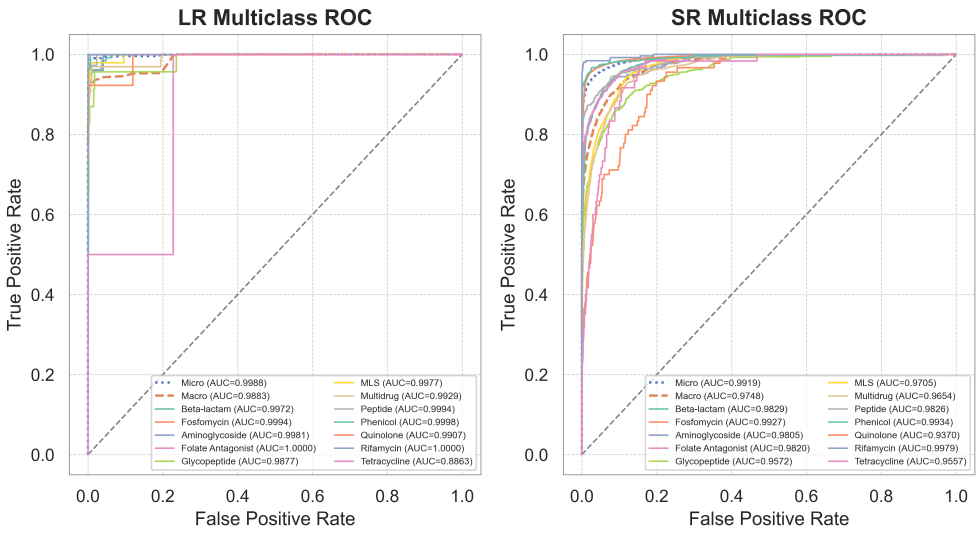

Fig. S5: ROC curves for LR and SR Multiclass models.

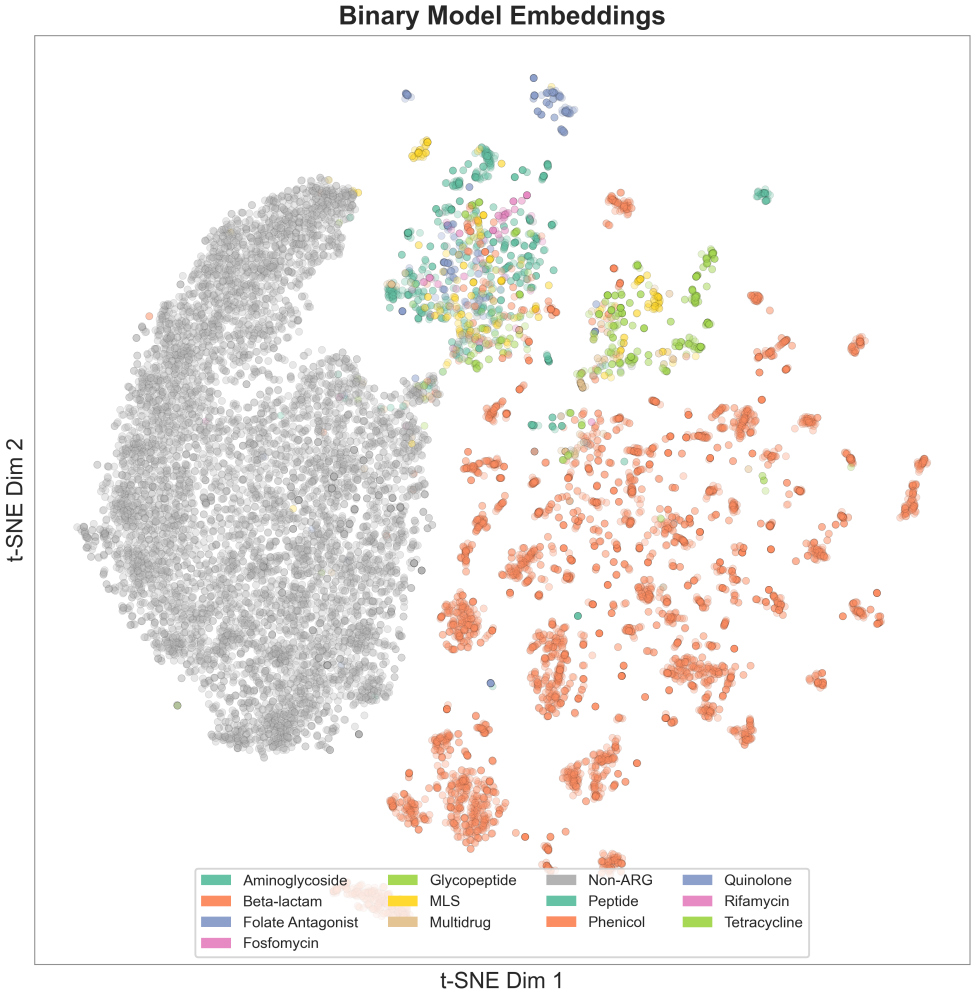

Fig. S6: Ordination plot from *resLens* binary long read model CLS token embeddings, colored by ARG class, including non-resistance genes. Despite not being trained to discriminate between ARG classes, the model shows some clustering separation of ARGs by class while grouping all non-ARGs together.

Table S5: Labels assigned to genes identified by the *resLens* model in the WGS dataset and examples of reference genes with 100% identity found via BLASTP. Sources used to determine label assignments are indicated where applicable. See supplementary data for the full dataset.

Label	**# of Occurrences**	**Example Target Class**	**Example Reference Gene & Supporting Evidence**
Unambiguous match	689	MLS	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(C) [*Staphylococcus pseudintermedius*]
Unambiguous miss	246	Glycopeptide	Alkaline phosphatase synthesis transcriptional regulatory protein PhoP [*Streptococcus pneumoniae PNI0008*]; source
Efflux pump	67	Multidrug	p-hydroxybenzoic acid efflux pump subunit AaeA, partial [*Escherichia coli*]
Confer resistance to other antibiotic class	15	Fosfomycin	GNAT family N-acetyltransferase [*Klebsiella*]
Multidrug resistance gene	13	MLS	ABC-F type ribosomal protection protein CplR [*Clostridioides difficile*]; source
Regulatory gene with some evidence of involvement in resistance	10	Fosfomycin	TetR/AcrR family transcriptional regulator [*Acinetobacter*]
Ambiguous; mixed evidence	5	MLS	beta-1,6-N-acetylglucosaminyltransferase [*Enterobacteriaceae*]; source

Table S6: Performance of *resLens* and ResFinder on WGS data, shown by the number of genomes with at least one phenotype correctly identified and the total number of phenotypes recalled.

	# of Genomes (% of Total)	**# of Phenotypes (% of Total)**
Recalled by *resLens*	77 (92.8%)	95 (94.0%)
Recalled by ResFinder	69 (83.1%)	78 (77.2%)
Total	83	101

Table S7: Distribution of ARGs in training and holdout datasets used for novel ARG evaluation. A dash (–) indicates that no samples for that class were included in the holdout dataset.

Antibiotic Class	**# Training Samples**	**# Holdout Samples**
Aminoglycoside	432	229
Beta-lactam	4,850	324
Folate pathway antagonist	187	–
Fosfomycin	64	–
Glycopeptide	150	–
MLS	280	–
Multidrug	138	–
Peptide	94	–
Phenicol	137	–
Quinolone	133	–
Rifamycin	22	–
Tetracycline	269	–