Enhancing Antimicrobial Drug Resistance Classification by Integrating Sequence-Based and Text-Based Representations

Hyunwoo Yoo, Bahrad Sokhansanj, James R. Brown
Drexel University
{hty23, bas44, jb4633}@drexel.edu

Abstract

Antibiotic resistance identification is essential for public health, medical treatment, and drug development. Traditional sequence-based models struggle with accurate resistance prediction due to the lack of biological context. To address this, we propose an NLP-based model that integrates genetic sequences with structured textual annotations, including gene family classifications and resistance mechanisms. Our approach leverages pretrained language models for both genetic sequences and biomedical text, aligning biological metadata with sequence-based embeddings. We construct a novel dataset based on the Antibiotic Resistance Ontology (ARO), consolidating gene sequences with resistance-related textual information. Experiments show that incorporating domain knowledge significantly improves classification accuracy over sequence-only models, reducing reliance on exhaustive laboratory testing. By integrating genetic sequence processing with biomedical text understanding, our approach provides a scalable and interpretable solution for antibiotic resistance prediction.

1 Introduction

The prevalence of antibiotic resistance genes (ARGs) has risen rapidly over the past decade, posing a severe threat to public health and medical treatment strategies (Zhang et al., 2022). The emergence of multidrug-resistant pathogens has further complicated treatment options, increasing the urgency of developing accurate methods for identifying and classifying ARGs. While traditional antibiotic resistance screening relies on phenotypic testing, these methods are time-consuming and require extensive laboratory resources. In contrast, bioinformatics-based approaches enable in silico prediction of resistance from genetic sequences, offering a scalable and efficient alternative. The primary computational approach for identifying antibiotic resistance genes (ARGs) has been sequence alignment, which compares nucleotide sequences to known ARG databases (Bonin et al., 2023). While effective, alignment-based methods struggle with novel mutations and require substantial computational resources. Alternative machine learning-based strategies have been explored to address these challenges but remain limited in capturing broader sequence dependencies (Wood and Salzberg, 2014; Eddy, 1998; McIntyre et al., 2017). To overcome these limitations, recent studies have applied natural language processing (NLP) models to genomic or protein sequences, leveraging contextual embeddings for improved classification and interpretability (Brandes et al., 2022; Ji et al., 2021; Zhou et al., 2024).

Despite their advancements, existing classification models predominantly focus on predicting a single resistance label per gene sequence (Kang et al., 2022). However, antibiotic resistance databases such as CARD (Alcock et al., 2023; Jia et al., 2017) and MEGARes (Bonin et al., 2023; Doster et al., 2020) provide richer annotations beyond a single resistance label. In particular, two critical attributes—Gene Family and Resistance Mechanism—offer valuable insights into how resistance manifests at a molecular level. These attributes provide a higher-level understanding of resistance beyond individual nucleotide variations, but current sequence-based models do not leverage this structured information. By incorporating Gene Family and Resistance Mechanism into predictive models, we can enhance interpretability and classification accuracy. In this work, we propose a novel NLP-based model that integrates genetic sequence data with structured textual annotations, specifically Gene Family and Resistance Mechanism, to improve antibiotic resistance classification. Our key contributions are as follows:

- We integrate biological knowledge with

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 263–273

sequence-based models for more accurate resistance prediction.

- We unify resistance classification by aligning CARD and MEGARes annotation systems.
- We generate synthetic samples to improve classification in rare resistance categories.

2 Related Work

Traditional methods for predicting antibiotic resistance rely on sequence alignment techniques, where unknown DNA sequences are compared to reference databases (Bonin et al., 2023). While effective for known resistance genes, alignment-based methods struggle with novel mutations and require high computational resources for large-scale datasets. Alternative computational approaches, such as Hidden Markov Models (HMMs) (Eddy, 1998) and k-mer-based classification (Wood and Salzberg, 2014), have been explored to recognize sequence patterns beyond direct alignment. However, these methods still face limitations in capturing broader contextual dependencies within genomic sequences. To address these limitations, sequence-based machine learning approaches, such as nucleotide transformers and DNABERT, have been introduced (Ji et al., 2021; Zhou et al., 2024). These models capture contextual representations of DNA sequences and offer improved classification performance over traditional alignment methods. However, existing sequence-based models primarily predict antibiotic resistance based on nucleotide sequence patterns alone, without incorporating additional biological knowledge. Antibiotic resistance is not solely determined by genetic sequence variations, but also by gene function, regulatory mechanisms, and evolutionary relationships (Kang et al., 2022). As a result, sequence-only models may fail to generalize across diverse resistance mechanisms and gene families.

Recent advancements in biomedical NLP and knowledge-driven machine learning have demonstrated the potential of integrating structured domain knowledge into predictive models. In fields such as protein function prediction and clinical text mining, hybrid approaches combining structured knowledge with sequence-based embeddings have shown promising results (Brandes et al., 2022). This motivates the need for similar methods in antimicrobial resistance (AMR) classification. Antibiotic resistance databases such as CARD (Alcock et al., 2023) and MEGARes (Bonin et al., 2023) provide valuable metadata beyond sequence-based labels, including Gene Family classifications and Resistance Mechanisms. These attributes capture biologically meaningful relationships between genes and their resistance properties. However, existing AMR classification models do not fully leverage these structured annotations, treating resistance prediction as a single-label classification problem from raw sequences. While sequence-based language models have improved antibiotic resistance prediction, they still lack biological interpretability and fail to incorporate structured knowledge from domain-specific databases. The integration of sequence embeddings with domain knowledge has the potential to enhance classification performance and interpretability. This motivates further exploration of hybrid models that combine genetic sequence processing with structured textual annotations, enabling more comprehensive and generalizable resistance prediction.

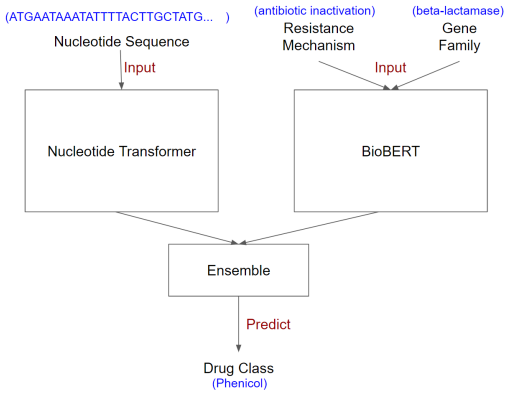

Figure 1: Overview of hybrid model for antibiotic resistance drug class classification. The model takes as input a nucleotide sequence, gene family, and resistance mechanism, and predicts the corresponding drug class by combining outputs from Nucleotide Transformer and BioBERT.

3 Methods

Our model integrates sequence-based and text-based representations to improve antibiotic resistance drug class classification. Given a nucleotide sequence (e.g., ATGC...), its associated gene family (e.g., "beta-lactamase"), and resistance mechanism (e.g., "antibiotic inactivation"), the model predicts the corresponding drug class (e.g., "Phenicol"). As illustrated in Figure 1, we utilize two pretrained models such as a Nucleotide Transformer for processing sequence input and BioBERT for encoding

structured biological metadata. Their outputs are combined using a weighted soft-voting ensemble (Dietterich, 2000). The overall model architecture is illustrated in the Appendix A.

3.1 Nucleotide Sequence Based Antibiotic Resistance Drug Class Classification

To classify antibiotic resistance genes, we fine-tune a nucleotide transformer (NT) model (Dalla-Torre et al., 2023). We consider the NT model as a strong sequence-only baseline that represents current methods that rely solely on nucleotide features without structured annotations. Unlike conventional models primarily trained on human genomes (Sanabria et al., 2024), NT is pre-trained on a diverse collection of genomic sequences from bacteria, fungi, and protozoa, allowing for a more comprehensive representation of microbial resistance patterns. For input processing, nucleotide sequences are tokenized using a 6-mer tokenizer, a widely used k-mer tokenization technique in genomic analysis (Mejía-Guerra and Buckler, 2019). The input length is restricted to 1000 nucleotides, corresponding to the model's pretraining constraints. The classification task is fine-tuned using Low-Rank Adaptation (LoRA), which inserts low-rank decomposed matrices into transformer layers while keeping the original model weights fixed (Hu et al., 2022). This significantly reduces trainable parameters while maintaining model efficiency and accuracy.

3.2 Text Information-Based Antibiotic Resistance Classification

To complement sequence-based models, we fine-tune BioBERT (Lee et al., 2020), a biomedical language model pre-trained on PubMed and PMC articles, to extract Gene Family and Resistance Mechanism attributes from textual descriptions of resistance genes. The input text is formatted using structured markers to enhance contextual understanding, improving attribute recognition and classification accuracy. Fine-tuning is conducted with a single classification layer, linking biological domain knowledge with sequence-based predictions. A comparison of different entity representation techniques is provided in Appendix D. Although resistance mechanism and gene family annotations may correlate with drug class labels, they are curated independently from the target labels in standardized resources such as CARD and MEGARes. These structured attributes often co-occur but not always perfectly aligned, providing complementary biological context that enhances classification robustness and interpretability.

3.3 Weighted Soft-voting Ensemble

To integrate predictions from the nucleotide sequence-based model and the text-based model, we implement a soft-voting ensemble strategy. The ensemble model is designed to leverage the complementary strengths of both approaches (Kuncheva and Whitaker, 2003), combining genetic sequence representations with structured textual knowledge for improved classification accuracy. The ensemble takes two types of inputs: (1) the nucleotide sequence, processed through the sequence-based language model, and (2) textual annotations, including Gene Family and Resistance Mechanism attributes, extracted from the text-based model. To optimize classification performance, we determine the weight ratio of each model's contribution using a validation dataset. This validation set is separate from the training and test datasets and is used to fine-tune the weight distribution for optimal ensemble decision-making. Final prediction probabilities are computed using a weighted soft-voting scheme:

$$P(y \mid x) = \lambda \cdot P_{\text{NT}}(y \mid x_{\text{seq}}) + (1-\lambda) \cdot P_{\text{BB}}(y \mid x_{\text{text}})$$

where λ is a weight parameter determined from validation performance. In our experiments, λ ranged between 0.35 and 0.55 depending on the dataset, reflecting the relative contributions of sequence-based and text-based predictions.

3.4 Integrating Classes Based on Antibiotic Resistance Ontology

Antibiotic resistance classification varies across databases, with CARD and MEGARes using different resistance labels and hierarchical structures. To address these inconsistencies, we employ the EBI Antibiotic Resistance Ontology (ARO) (Cook et al., 2016) to standardize resistance annotations across datasets. Each database entry is mapped to the ARO ontology by querying the EBI API and retrieving hierarchical Gene Family relationships. Instead of using fine-grained subcategories, we adopt the third-level hierarchy in ARO, ensuring that class representations remain general enough for robust classification across different datasets. This hierarchical integration harmonizes classification schemes, reducing discrepancies in resistance

annotations between databases. This mapping process ensures consistency across heterogeneous labels by aligning them to a shared third-level ARO hierarchy, as detailed in Appendix B.

3.5 Data Augmentation Using a Large Language Model

To mitigate data imbalance in antibiotic resistance gene classification, we employ BioGPT (Luo et al., 2022) for generating synthetic samples in underrepresented categories. Augmenting resistance descriptions improves classification performance, particularly in Macro F1 score. The effectiveness of this approach is detailed in Appendix E.

4 Experiments

We evaluate the performance of sequence-based and text-based models for antibiotic resistance drug class classification using three datasets: CARD, MEGARes, and an integrated dataset combining both sources. We compare Nucleotide Transformer (NT), BioBERT (BB), and an ensemble of both models, analyzing their effectiveness in different dataset settings.

4.1 Experimental Setup

We finetune NT on genetic sequences and BioBERT on structured text annotations describing resistance genes. The ensemble model uses a weighted soft-voting approach, integrating both modalities. All models are trained on CARD, MEGARes, and Integrated datasets, following the standard pre-processing pipeline described in Methods. In addition, experiments using read-level data generated based on the Integrated dataset is conducted. Further details can be found in the Appendix C

4.2 Datasets

We use the CARD and MEGARes v3 datasets, integrating Drug Class, Gene Family, and Resistance Mechanism labels using the EBI ARO ontology. Following standard preprocessing, classes with fewer than 15 samples are removed. Dataset details are provided in the Appendix B.

4.3 Classification Results

Table 1 presents the classification results, demonstrating the impact of integrating structured biological knowledge into sequence-based models. Compared to sequence-only models, incorporating Gene Family and Resistance Mechanism attributes led to significant performance improvements. Specifically, our method improved accuracy by 9.53 points and Macro F1 by 30.34 points on CARD, while on MEGARes, the improvement was 10.38 points and 50.57 points, respectively. These findings indicate that sequence-based models alone struggle to capture higher-level biological relationships necessary for robust resistance classification. By integrating structured textual annotations, our model achieves superior interpretability and generalization, particularly for low-resource resistance categories. Furthermore, using integrated data from multiple annotation systems enhances classification performance, demonstrating the advantage of leveraging domain-specific knowledge for a unified prediction model.

4.4 Ablation Analysis

To assess the contribution of each component in our hybrid model, we conduct an ablation analysis comparing individual models (NT and BB) versus their ensemble, and dataset configurations (individual vs. integrated). As shown in Table 1, the ensemble consistently outperforms NT and BB alone across all datasets, confirming the complementary nature of sequence-based and text-based representations.

The integrated dataset includes more diverse and heterogeneous resistance profiles from both CARD and MEGARes, offering a broader and more realistic evaluation setting. Despite this increased complexity, our ensemble model maintains strong and consistent performance, demonstrating its robustness and generalizability across databases.

5 Discussion

Our results demonstrate that incorporating structured biological knowledge significantly enhances antibiotic resistance classification. Sequence-based models alone struggle to capture higher-order biological relationships that influence resistance mechanisms. By integrating Gene Family and Resistance Mechanism annotations, our model improves interpretability and generalization, particularly for low-resource resistance categories. Furthermore, class integration using the EBI ARO ontology standardizes resistance classification across datasets, increasing training data availability and improving consistency. This standardization not only enhances model performance but also facilitates broader applicability across different resistance gene databases. Notably, the near-perfect performance observed on the MEGARes dataset may par-

Dataset	Method	Accuracy	Macro F1	Precision	Recall
CARD	NT	87.92	63.08	66.46	61.51
CARD	BB	97.22	89.68	92.09	90.54
CARD	Ensemble	97.55	93.44	95.72	92.86
MEGARes	NT	89.61	46.42	54.92	43.94
MEGARes	BB	99.64	99.47	99.96	99.03
MEGARes	Ensemble	99.99	99.99	99.99	99.99
Integrated	NT	82.89	65.79	81.84	58.67
Integrated	BB	90.26	79.34	84.05	77.14
Integrated	Ensemble	92.11	80.95	83.52	78.94
Integrated with reads	NT	83.11	62.82	74.81	57.32
Integrated with reads	BB	90.24	79.34	84.05	77.14
Integrated with reads	Ensemble	93.40	81.85	84.34	80.25

Table 1: Result of using the CARD, MEGARes, and Integrated databases for antibiotic resistance drug class prediction using Nucleotide Transformer(NT), BioBERT(BB), and a weighted ensemble of both. The weighted ensemble with Nucleotide Transformer(NT) and BioBERT(BB) shows better performance in every datasets.

tially reflect the benefits of ontology-based class harmonization and the high consistency of resistance annotations in MEGARes. While these results highlight the model's capacity to leverage structured knowledge, they also suggest that annotation quality and class structure play a key role in enabling robust classification. Additionally, our ensemble model maintains strong performance even when using sequencing reads instead of full-length genes, demonstrating its robustness in practical applications. Beyond classification performance, incorporating structured biological knowledge also provides practical advantages in reducing experimental complexity and time (see Appendix F). By bridging the gap between sequence-based and knowledge-driven classification, our approach offers a scalable and interpretable solution for antimicrobial resistance prediction. However, our approach still relies on the quality of existing resistance gene annotations, which may not always reflect emerging resistance mechanisms. Additionally, maintaining up-to-date structured knowledge requires continuous curation, posing a scalability challenge.

6 Conclusion

We present a hybrid model that integrates sequence-based and text-based representations to improve antibiotic resistance classification. By incorporating structured biological knowledge, including Gene Family and Resistance Mechanism annotations, our approach enhances interpretability and outperforms sequence-only models. Additionally, we standardize resistance classification using the EBI ontology and utilize large language models for data augmentation, improving performance in low-resource settings. These results demonstrate the effectiveness of combining genetic and textual information for more accurate and scalable resistance prediction.

7 Limitation

While our approach improves antibiotic resistance classification by integrating sequence-based and text-based models, certain limitations remain. First, our reliance on curated databases, such as CARD and MEGARes, means that model performance may be affected by biases in annotation quality and completeness. Additionally, while integrating Gene Family and Resistance Mechanism improves interpretability, the hierarchical structure of these annotations may introduce inconsistencies across datasets. Another limitation is the challenge of handling rare or novel resistance genes, where even with data augmentation, model generalization remains an open problem. Computational efficiency remains a concern, as training large-scale sequence and text models requires significant resources, which may limit accessibility for some research applications. Finally, beyond domain-specific models, evaluating the potential of recent general-purpose LLMs such as ChatGPT-4o or Claude 4 Sonnet for antibiotic resistance prediction remains an open direction for future research.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grant Number 2107108.

References

Brian P. Alcock, William Huynh, Romeo Chalil, Keaton W. Smith, Amogelang R. Raphenya, Mateusz A. Wlodarski, Arman Edalatmand, Aaron

Petkau, Sohaib A. Syed, Kara K. Tsang, Sheridan J. C. Baker, Mugdha Dave, Madeline C. McCarthy, Karyn M. Mukiri, Jalees A. Nasir, Bahar Golbon, Hamna Imtiaz, Xingjian Jiang, Komal Kaur, Megan Kwong, Zi Cheng Liang, Keyu C. Niu, Prabakar Shan, Jasmine Y. J. Yang, Kristen L. Gray, Gemma R. Hoad, Baofeng Jia, Timsy Bhando, Lindsey A. Carfrae, Maya A. Farha, Shawn French, Rodion Gordzevich, Kenneth Rachwalski, Megan M. Tu, Emily Bordeleau, Damion Dooley, Emma Griffiths, Haley L. Zubyk, Eric D. Brown, Finlay Maguire, Robert G. Beiko, William W. L. Hsiao, Fiona S. L. Brinkman, Gary Van Domselaar, and Andrew G. McArthur. 2023. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, 51(D1):D690–D699.

J. M. Andrews. 2001a. Determination of minimum inhibitory concentrations. *Journal of Antimicrobial Chemotherapy*, 48(Suppl 1):5–16.

Jennifer M. Andrews. 2001b. Determination of minimum inhibitory concentrations. *Journal of Antimicrobial Chemotherapy*, 48(suppl_1):5–16.

Frederik Otzen Bagger, Line Borgwardt, Andreas Sand Jespersen, Anna Reimer Hansen, Birgitte Bertelsen, Miyako Kodama, and Finn Cilius Nielsen. 2024. Whole genome sequencing in clinical practice. *BMC Medical Genomics*, 17:Article 39.

Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S Morley, and Christina Boucher. 2023. MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Research*, 51(D1):D744–D752.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Karen Bush and George A. Jacoby. 2010. Updated functional classification of -lactamases. *Antimicrobial Agents and Chemotherapy*, 54(3):969–976.

Carolina Cason, Maria D'Accolti, Irene Soffritti, Sante Mazzacane, Manola Comar, and Elisabetta Caselli. 2022. Next-generation sequencing and pcr technologies in monitoring the hospital microbiome and its drug resistance. *Frontiers in Microbiology*, 13:969863.

Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. 2016. The european bioinformatics institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Genomics*.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS 2000)*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer.

Enrique Doster, Steven M Lakin, Christopher J Dean, Cory Wolfe, Jared G Young, Christina Boucher, Keith E Belk, Noelle R Noyes, and Paul S Morley. 2020. Megares 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research*, 48(D1):D561–D569.

S. R. Eddy. 1998. Profile hidden markov models. *Bioinformatics*, 14(9):755–763.

Hadrien Gourlé, Oskar Karlsson-Lindsjö, Juliette Hayer, and Erik Bongcam-Rudloff. 2019. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, 35(3):521–522.

Manuel Holtgrewe. 2010. Mason – A Read Simulator for Second Generation Sequencing Data. Technical report, FU Berlin.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. 2021. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811.

Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. Art: A next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E. Williams, Jonathan G. Frye, Tariq Elsayegh, Daim Sardar, Erin L. Westman, Andrew C. Pawlowski, Timothy A. Johnson, Fiona S.L. Brinkman, Gerard D. Wright, and Andrew G. McArthur. 2017. Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573.

Hyeunseok Kang, Sungwoo Goo, Hyunjung Lee, Jung-Woo Chae, Hwi-Yeol Yun, and Sangkeun Jung. 2022. Fine-tuning of bert model to accurately predict drug-target interactions. *Pharmaceutics*, 14(8):1710.

Beata Kowalska-Krochmal and Ruth Dudek-Wicher. 2021. The Minimum Inhibitory Concentration of Antibiotics: Methods, Interpretation, Clinical Relevance. *Pathogens*, 10(2):165.

Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Alexa B. R. McIntyre, Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, David Danko, Jonathan Foox, Sofia Ahsanuddin, Scott Tighe, Nur A. Hasan, Poorani Subramanian, Kelly Moffat, Shawn Levy, Stefano Lonardi, Nick Greenfield, Rita R. Colwell, Gail L. Rosen, and Christopher E. Mason. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1):182.

María Katherine Mejía-Guerra and Edward S. Buckler. 2019. A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biology*, 19(1):103.

Pauline C. Ng and Ewen F. Kirkness. 2010. Whole genome sequencing. *Methods in Molecular Biology*, 628:215–226.

Melissa Sanabria, Jonas Hirsch, Pierre M. Joubert, and Anna R. Poetsch. 2024. DNA language model GROVER learns sequence context in the human genome. *Nature Machine Intelligence*, 6:911–923.

Patrick Schorderet. 2016. NEAT: a framework for building fully automated NGS pipelines and analyses. *BMC Bioinformatics*, 17:53.

Derrick E. Wood and Steven L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15.

Koshi Yamada, Makoto Miwa, and Yutaka Sasaki. 2023. Biomedical Relation Extraction with Entity Type Markers and Relation-specific Question Answering. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 377–384, Toronto, Canada. Association for Computational Linguistics.

Zhenyan Zhang, Qi Zhang, Tingzhang Wang, Nuohan Xu, Tao Lu, Wenjie Hong, Josep Penuelas, Michael Gillings, Meixia Wang, Wenwen Gao, and Haifeng Qian. 2022. Assessment of global health risk of antibiotic resistance genes. *Nature Communications*, 13.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2024. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Model Overview

Figure 1 illustrates the overall architecture of our proposed model, which integrates sequence-based and text-based representations for antibiotic resistance classification. We fine-tuned two pre-trained language models—Nucleotide Transformer and BioBERT—for DNA sequence classification tasks involving the prediction of antimicrobial drug classes. The Nucleotide Transformer model was fine-tuned using parameter-efficient LoRA-based adaptation. DNA sequences were truncated to a maximum length of 1000 nucleotides and tokenized using a domain-specific tokenizer. Training data was structured with input DNA sequences and corresponding drug class labels. The model was fine-tuned using a sequence classification objective on a multi-class dataset. Performance was evaluated on a separate test set using macro-average F1 score, accuracy, precision, recall, and balanced accuracy. For BioBERT, the input consisted of textual descriptions including gene family and resistance mechanism information, formatted into natural language prompts. These were tokenized using a BERT tokenizer with a fixed input length. A classification head was added to predict the drug class labels. The model was trained for multiple epochs and evaluated using the same metrics as for the Nucleotide Transformer. Both models showed effective performance in multi-class classification tasks, demonstrating the potential of sequence- and text-based pretraining approaches in genomic classification problems.

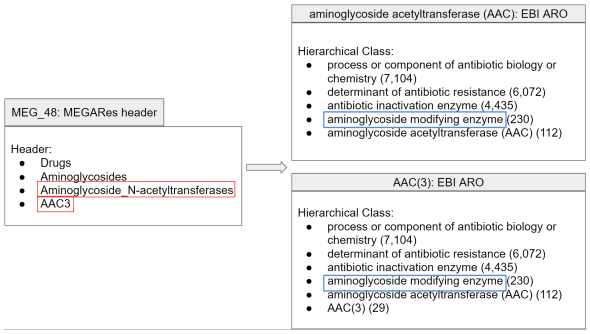

Figure 2: EBI ARO Gene Family mapping: search to find mapping information with header and ontology by using API.

B Dataset Details

The CARD and MEGARes v3 datasets are used for training and evaluation. Classes with fewer than 15 samples are removed because obtaining meaningful results from the data split is difficult. The remaining data is split into 75% for training data, 20% for test data, and 5% for validation data. EBI ARO ontology search is used to integrate the data, which is then split similarly to the above. Classes with difficult-to-obtain meaningful results are also removed. The MEGARes dataset consists of 9733 Reference Sequences, 1088 SNPs, 4 antibiotic types, 59 resistance classes, and 233 mechanisms. The CARD dataset consists of 5194 Reference Sequences and 2005 SNPs, 142 Drug Classes, 331 Gene Families, and 10 Resistance Mechanisms. The EBI ARO ontology provides hierarchical group information for genes. Using the EBI ARO Ontology, Gene Family class information can be integrated into a higher-level hierarchy. The number of Gene Family text information classes in the case of MEGARes is 589, while for CARD, it is 331. There are 300 and 166 datasets with only one sample in their respective classes for Gene Family in the case of MEGARes and CARD, respectively. Resistance Mechanism is integrated based on the 6 categories of CARD. The original 8 categories were reduced to 6, excluding cases of various class combinations and those with very few samples. Drug Class is integrated using 9 common Drug Classes found in competing models. Integration is done based on names and theories and has been verified. Macro f1 score, accuracy, balanced accuracy, and precision are used as performance metrics, and the results are listed in the Table 1.

Figure 3, Figure 4 and Figure 5 represent the distribution of training dataset which is integrated

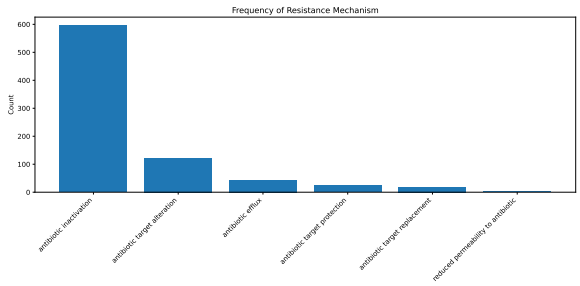

Figure 3: Counts of the frequent Resistance Mechanism in training dataset.

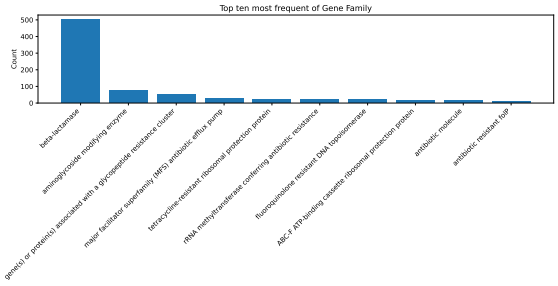

Figure 4: Counts of the frequent Gene Family in training dataset.

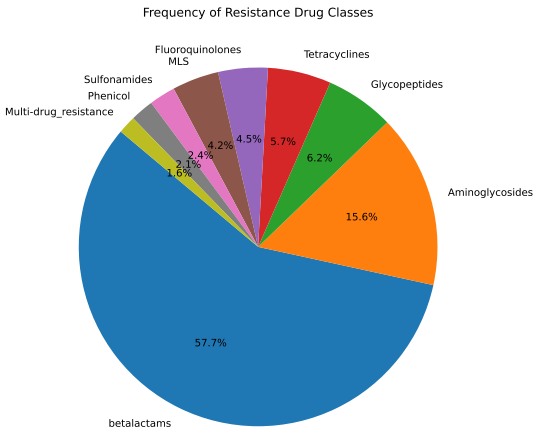

Figure 5: Counts of the frequent Antimicrobial Resistance Drug Classes in training dataset.

with CARD and MEGARes. We observe a long-tail distribution for Resistance Mechanism, Gene Family, and Drug Class classes.

The distribution indicates that certain resistance mechanisms, gene families, and drug classes are significantly overrepresented in the dataset, while many others occur with low frequency. Specifically, antibiotic inactivation is the most common resistance mechanism, while beta-lactamase genes

dominate the gene family distribution. Similarly, beta-lactams appear as the most frequently associated drug class.

This imbalance in distribution suggests that models trained on this dataset may exhibit biased performance, favoring well-represented categories while struggling with rare classes. Furthermore, the presence of diverse resistance mechanisms and gene families emphasizes the complexity of antimicrobial resistance (AMR) prediction.

The dataset used in this study is publicly available at https://zenodo.org/records/15213479.

C Read Generation

Read generation is a computational process used to simulate short DNA or RNA sequences, commonly referred to as "reads", from reference genomes or annotated genetic sequences. This technique is designed to mimic the output of next-generation sequencing (NGS) technologies (Hu et al., 2021), providing a way to generate data for various applications such as machine learning model training, benchmarking, or evaluating bioinformatics pipelines. In the context of antibiotic resistance prediction, read generation is often performed using curated databases like CARD, MEGARes, or the Integrated database, which contain known resistance genes and associated metadata.

To simulate realistic reads, researchers commonly use specialized tools such as ART (Huang et al., 2012), InSilicoSeq (Gourlé et al., 2019), DWGSIM, NEAT (Schorderet, 2016), or Mason (Holtgrewe, 2010). These simulators can generate Illumina-style short reads with configurable read lengths, sequencing errors, mutation rates, and coverage depth. In this study, we used ART to generate synthetic reads based on the Integrated database. ART supports detailed customization of error profiles and is widely used for simulating realistic Illumina sequencing data.

The generated reads can serve as a substitute when real-world sequencing data is limited or unavailable. By generating reads from known reference sequences, researchers can perform controlled experiments with clearly defined ground truth, assess model robustness under noisy or imperfect conditions, and evaluate how well different models generalize to simulated real-world data. Overall, read generation combined with realistic simulators plays a crucial role in creating labeled datasets that facilitate the development and validation of genomic analysis tools.

D Entity Representation Techniques

To improve antibiotic resistance classification, we experimented with different entity representation techniques for encoding Gene Family and Resistance Mechanism attributes in BioBERT-based models. Table 2 compares the impact of these techniques on classification performance.

These representations were designed to help the model better distinguish between biological attributes and general text (Yamada et al., 2023). The Base format uses plain-text input without additional markers, while the Entity Marker (punct) format introduces brackets around key attributes. The Typed Entity Marker (Zhou and Chen, 2022) explicitly labels entities, providing more structured input, and the Typed Entity Marker (punct) format further combines these strategies.

Results indicate that using entity markers improves classification performance. In particular, the Typed Entity Marker (punct) approach achieves the highest Macro F1 score, demonstrating that structured formatting helps the model capture contextual relationships between resistance mechanisms and gene families more effectively. Results indicate that explicit formatting, such as typed entity markers with punctuation, enhances BioBERT's contextual understanding about Gene Family and Resistance Mechanism attributes from general text. This suggests that structured annotations provide useful inductive bias, allowing the model to better capture domain-specific relationships.

E Impact of LLM-Based Data Augmentation

Despite ontology-based class standardization, certain resistance categories remain underrepresented due to natural imbalances in antibiotic resistance gene distributions. To address this, we employ BioGPT (Luo et al., 2022) for generating synthetic samples in low-resource categories. BioGPT is prompted to generate contextually similar resistance gene descriptions, maintaining the linguistic characteristics of real annotations to ensure realistic and informative augmentation.

By integrating BioGPT-based augmentation, we observe consistent improvements in classification performance, particularly in Macro F1 scores for rare classes. Table 3 presents the results of this

Output	Input Example	BioBERT
Base	Gene Family: Beta-lactamases, Resistance Mechanism: Antibiotic inactivation	78.20
Entity marker (punct)	[Gene Family]: Beta-lactamases, [Resistance Mechanism]: Antibiotic inactivation	77.41
Typed entity marker	*Beta-lactamases*, #Resistance Mechanism#	77.70
Typed entity marker (punct)	*[Gene Family]: Beta-lactamases*, #[Resistance Mechanism]: Antibiotic inactivation#	78.46

Table 2: Test Macro F1 score of different entity representation techniques in antibiotic resistance classification with BioBERT.

augmentation strategy, demonstrating its positive impact on model robustness.

F Practical Advantages of Using Gene Family and Resistance Mechanism

Incorporating Gene Family and Resistance Mechanism information in antibiotic resistance classification provides practical advantages, particularly in reducing experimental complexity and time. Traditional laboratory-based methods, such as Minimum Inhibitory Concentration (Kowalska-Krochmal and Dudek-Wicher, 2021; Andrews, 2001a) (MIC) assays and Disk Diffusion Tests, require separate testing for each antibiotic, which involves overnight incubation and may take longer for certain organisms (Andrews, 2001b). Testing multiple antibiotics increases time and resource consumption, and experimental conditions such as growth medium and gene expression variability can further complicate results.

Sequence-based approaches, such as Polymerase Chain Reaction (PCR) and Whole Genome Sequencing (WGS), enable the identification of resistance-related genes directly from genomic data (Bagger et al., 2024; Ng and Kirkness, 2010). PCR/qPCR can provide results relatively quickly, typically within hours, whereas WGS requires a longer processing time, often taking multiple days to complete (Cason et al., 2022).

Leveraging Gene Family and Resistance Mechanism attributes allows for a more efficient computational approach to resistance prediction, minimizing reliance on exhaustive in vitro testing. Many resistance mechanisms are well-characterized and strongly associated with specific gene families. For instance, betalactamase genes are well-known indicators of resistance to betalactam antibiotics, such as penicillins and cephalosporins (Bush and Jacoby, 2010). By integrating structured biological knowledge with sequence-based models, resistance predictions can be made with greater confidence and interpretability, supporting a scalable and practical framework for antimicrobial resistance classification.

Method	Accuracy	Macro F1	Precision	Recall
NT	84.15	64.04	72.78	59.28
NT with data augmentation	83.42	64.85	80.15	58.65
NT with reads	82.85	61.02	68.32	57.06
NT with reads and data augmentation	83.11	62.82	74.81	57.32

Table 3: Effect of BioGPT-based data augmentation on resistance classification performance. Augmentation improves Macro F1, particularly for low-resource categories.