Bioinformatics, 2024, pp. 1–7

doi:

ProtAlign-ARG: Antibiotic Resistance Gene Characterization Integrating Protein Language Models and Alignment-Based Scoring

Shafayat Ahmed,[1] Muhit Islam Emon,[1] Nazifa Ahmed Moumi,[1] Lifu Huang,[1] Dawei Zhou,[1] Peter Vikesland,[2] Amy Pruden[2] and Liqing Zhang[1,*]

[1]Department of Computer Science,Virginia Polytechnic Institute and State University and [2]Department of Civil & Environmental Engineering, Virginia Polytechnic Institute and State University
[*]Corresponding author. lqzhang@vt.edu

Abstract

Increasing antibiotic resistance poses a severe threat to human health. Detecting and categorizing antibiotic resistance genes (ARGs), genes conferring resistance to antibiotics in sequence data is vital for mitigating the spread of antibiotic resistance. Recently, large protein language models have been used to identify ARGs. Comparatively, these deep learning methods show superior performance in identifying distant related ARGs over traditional alignment-base methods, but poorer performance for ARG classes with limited training data. Here we introduce ProtAlign-ARG, a novel hybrid model combining a pre-trained protein language model and an alignment scoring-based model to identify/classify ARGs. ProtAlign-ARG learns from vast unannotated protein sequences, utilizing raw protein language model embeddings to classify ARGs. In instances where the model lacks confidence, ProtAlign-ARG employs an alignment-based scoring method, incorporating bit scores and e-values to classify ARG drug classes. ProtAlign-ARG demonstrates remarkable accuracy in identifying and classifying ARGs, particularly excelling in recall compared to existing ARG identification and classification tools. We also extend ProtAlign-ARG to predict the functionality and mobility of these genes, highlighting the model's robustness in various predictive tasks. A comprehensive comparison of ProtAlign-ARG with both the alignment-based scoring model and the pre-trained protein language model clearly shows the superior performance of ProtAlign-ARG.

Key words: ARG, Protein language model, Deep learning, Protein sequence

Introduction

Antibiotic resistance poses a grave threat to public health, with annual deaths projected to rise from 700,000 to 10 million by 2050 due to antibiotic resistance genes (ARGs) [1, 2]. These genes, which confer resistance to antibiotics, are widely transmitted among animals, humans, and environments [3, 4, 5, 6, 7].

Traditional DNA sequence alignment methods face challenges in detecting new ARGs, struggling with remote homology and large databases, thus failing to adequately capture the complexity of antibiotic resistance [8, 9]. Deep learning, particularly with protein language model embeddings, offers a more nuanced representation of biological data, excelling in contextualizing protein sequences and uncovering complex patterns missed by conventional methods [10]. This approach, leveraging transformer architecture, processes sequences in parallel, enabling better generalization from fewer sequences [11]. This approach can be utilized to enhance both the efficiency and accuracy of ARG characterization.

Deep learning-based tools have performed well in ARG identification and classification. For example, Deep-ARG [12] uses deep learning and considers a dissimilarity matrix, HMD-ARG [13] offers a hierarchical multi-task classification model using CNN, and ARG-SHINE [14] utilizes a machine learning approach to ensemble three component methods for predicting ARG classes. Recently, protein language models, trained on millions of protein sequences, have been utilized for developing ARG prediction models. Specifically, both our preliminary work, presented as a non-peer-reviewed poster [15], and a subsequent study [16], showed the efficacy of pre-trained protein language models in facilitating downstream ARG identification and classification tasks.

However, one limitation of deep learning models is the reduced performance in classifying ARGs with limited training data [12, 13, 14]. Comparatively, the alignment-based method shows more robust performance. To leverage the strengths of both approaches, we introduce ProtAlign-ARG, a novel model that integrates pre-trained protein language model (PPLM) based prediction and alignment-based scoring. This hybrid model enhances predictive accuracy, particularly in scenarios with limited training data. Additionally, it can classify antibiotic resistance mechanisms and analyze ARG mobility,

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

distinguishing between intrinsic ARGs and those acquired through horizontal gene transfer [17, 18]. The complete code, data, and results of our model are available on the GitHub repository `https://github.com/Shafayat115/ProtAlign-ARG`.

Model Development

Data Curation

We utilized HMD-ARG-DB[13] as it is one of the largest repositories of ARGs and boasts the most comprehensive annotations across various dimensions. HMD-ARG-DB was curated from seven well-recognized databases, namely AMRFinder[19], CARD[20], ResFinder[21], Resfams[22], DeepARG[12], MEGARes[23], and Antibiotic Resistance Gene-ANNOTation [24], and contain over 17,000 ARG sequences distributed among 33 antibiotic-resistant classes.

For the ARG identification task, our deep learning model requires a robust dataset not only comprising ARGs but also encompassing non-ARGs for effective differentiation. To curate the non-ARG dataset, we undertook the meticulous process of downloading the entire Uniprot dataset while excluding sequences labeled as ARGs.

In the following step, we performed diamond alignment with the HMD-ARG-DB. Sequences that had an e-value less than 1e-3 and a percentage identity below 40% were classified as non-ARG datasets for ARG identification. This enabled us to focus on non-ARG sequences demonstrating substantial similarity to ARG sequences. The goal was to enhance the model's capability to identify ARGs, even in scenarios where such similarities exist[13]. Furthermore, it facilitated comparisons with other state-of-the-art tools like HMD-ARG[13] and DeepARG[12], thus validating our model's effectiveness.

Graphpart Analysis

In biological sequence data segmentation, we partitioned the dataset into training and testing sets to avoid biases and accurately assess model performance. This partitioning was essential to ensure that the training and testing data were not overly similar, which could lead to artificially inflated accuracy metrics.

Our partitioning approach was designed to maintain a strict level of dissimilarity between the sets, defined by a critical threshold. We set this threshold at a specific percentage, representing the maximum allowed similarity between training and testing sequences, to balance dissimilarity with data representativeness, ultimately enhancing the model's predictive robustness on unseen data.

While widely used tools like CDHIT[25] and MMseq[26] offer various clustering modes for partitioning, these tools tend to prioritize maximizing similarity within clusters, often at the expense of minimizing similarity between partitions. In contrast, a novel partitioning tool named GraphPart[27] has proven to be highly effective. GraphPart ensures precise separation, regardless of sequence length, and retains most sequences until reaching the desired threshold.

For instance, in our utilization of CDHIT[25] clustering with a 40% threshold similarity on the HMD-ARG-DB dataset, we obtained 721 clusters. However, the analysis revealed that many sequences exhibited similarity exceeding 40% and, in some cases, even exceeding 90% when subjected to BLAST analysis (Supplementary Figure 2). In contrast, GraphPart provided exceptional partitioning precision.

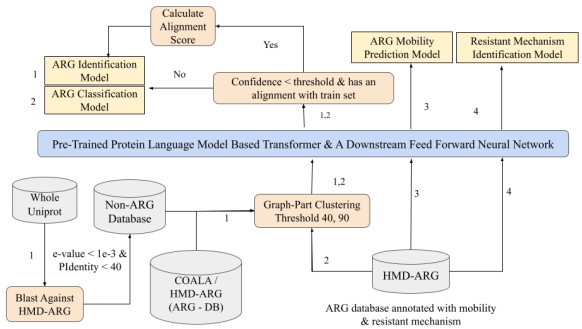

Fig. 1. The proposed pipeline for the ARG Identification & Classification. Each number (1-4) corresponds to the four different models.

For the preparation of our training and testing datasets, we employed GraphPart on the HMD-ARG-DB[13], allocating 80% of the sequences for training and 20% for the test set using a 40% similarity threshold. Subsequent BLAST analyses within the training and test sets consistently revealed an identity percentage below 40%, ensuring a clear distinction between the two sets and providing well-partitioned data for ARG class classification.

In the ARG identification task, the HMD-ARG-DB served as ARGs, while a customized non-ARG dataset was used. GraphPart was applied with both a stringent 90% similarity threshold and a 40% similarity threshold to obtain the train and test datasets for the ARG identification task. This approach allowed us to evaluate our model's performance across varying degrees of difficulty, ranging from datasets with high similarity to those with substantial dissimilarity.

To classify ARG classes, we applied a 40% similarity threshold to stratify the dataset, focusing on the 14 most prevalent ARG classes out of a total of 33. As depicted in (Supplementary Fig 1), the HMD-ARG-DB encompasses 17,282 ARG sequences distributed across these classes. However, 19 of these classes are represented by very few samples. The attempt to evenly split the dataset into training and testing sets while ensuring the representation of all classes at a 40% similarity threshold using GraphPart proved challenging due to substantial disparities in sample counts.

Additionally, we utilized the COALA (COllection of ALl Antibiotic resistance gene databases) dataset, categorizing it into three groups: "No alignment," "Less than 50% similarity between train and test sets," and "Greater than 50% threshold." This dataset served as a basis for comparing our model's accuracy with existing state-of-the-art models like DeepARG[12], ARG-SHINE[14], and HMD-ARG[13].

Furthermore, we performed a comparative analysis of the individual components of our models - the PPLM, the alignment-based scoring model, and the ProtAlign-ARG. This comparison was conducted on both the COALA dataset with a 90% similarity threshold and using GraphPart on the HMD-ARG-DB dataset with a 40% threshold.

For the prediction of resistant mechanisms and ARG mobility, we exclusively employed the HMD-ARG-DB[13]. For these, our model was trained and tested using a rigorous 5-fold cross-validation approach.

Model Components

Our system architecture comprises four distinct models, each dedicated to a specific task:

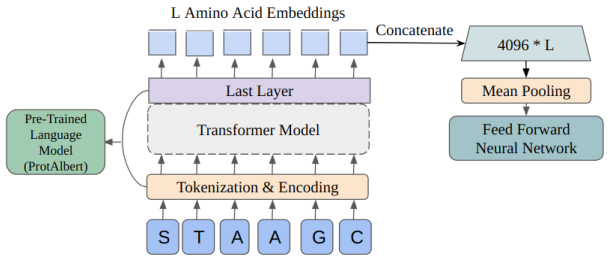

Fig. 2. Using pre-trained protein language model for ARG Identification & Classification

1. ARG Identification
2. ARG Class Classification
3. ARG Mobility Identification
4. ARG Resistance Mechanism

In Figure 1, we present an overview of the architecture of how these four models are developed, highlighting their respective designs and interrelationships. Detailed descriptions of the components are provided in the following sections.

Workflow for the Language Model

Generating Sequence Embeddings

To train protein language models (pLMs), annotations for protein sequences are not necessary. Instead, we predict the concealed amino acids within these sequences. These models are trained using millions of protein sequences, and we subsequently derive embeddings from them. These embeddings serve as inputs for predicting antibacterial resistance.

The performance of a language model is substantially influenced by the quality and diversity of its pre-training data. ProtAlbert, our chosen model, undergoes pre-training on an extensive and diverse collection of protein sequences. This approach equips ProtAlbert with the ability to generalize effectively across diverse biological contexts. Despite various options available at the time of this work, such as ProtXL, ProtElectra, and ProtT5, we deliberately selected the ProtAlbert model due to its superior performance. Notably, ProtAlbert outperformed its counterparts while maintaining a relatively modest size, comprising 224 million parameters. This choice was strategic, considering the balance between optimal performance and computational efficiency, especially when compared to larger models that are more time-consuming.

ProtAlbert, with its 12-layer architecture, undergoes training on UniRef100. For our downstream prediction task, we harnessed the pre-trained ProtAlbert model following the official GitHub repository provided by ProtTrans[28]. The training unfolds in two stages: an initial phase on sequences of up to 512 characters for 150,000 steps, succeeded by another 150,000 steps on longer sequences, spanning up to 2000 characters.

Transfer Learning with Supervised Models

The ProtAlbert model, pre-trained on UniRef100, enhances ARG classification through its embeddings, supporting transfer learning. This model adopts two approaches: per-residue, analyzing each amino acid individually, and per-protein, treating sequences as single entities. We selected the per-protein method for its holistic view of proteins, aligning with their structural and functional integrity. This choice offers computational simplicity and robustness against sequence variability, unlike the per-residue method, which, though detailed, demands more computation and is more sensitive to sequence alterations.

For our analysis, we utilized the model's last layer embeddings, summarized into a consistent-size vector via mean pooling. This vector feeds into a feed-forward neural network layer with 32 neurons for ARG tasks, effectively leveraging ProtAlbert's deep learning capabilities for precise ARG identification and classification.

Workflow for the Alignment-Based Scoring

Deep learning models, like neural networks, are effective at learning from large data sets, but they may not perform well with small amounts of data. They require data to learn effectively and can struggle with new, unseen data if trained on limited data. On the other hand, alignment-based models, which include some traditional machine learning algorithms, can be better when there's less data. These models, such as k-nearest neighbors (KNN) [29] and support vector machines (SVM) [30], focus on data similarities or specific rules, not on large-scale learning. So, while deep learning performs well on large amounts of data, alignment-based models are often better for smaller data sets. This is important to consider when choosing the right model for a task, especially when large data collection is not feasible.

For our alignment-based scoring, we utilized DIAMOND [31] for matching, following a modified version of the ARG-KNN model proposed by the paper ARG-SHINE[14]. Initially, we align the query sequence with the training data using DIAMOND, setting an e-value threshold of $< 1e-3$ to identify similar sequences (homologs). If a query sequence fails to align with any sequence in the training dataset, we cannot employ the alignment-based scoring to label it. We applied this scoring method to both ARG identification and classification.

For each query sequence, we computed alignment scores for each label. The label with the highest score is assigned to the query sequence. For a given query sequence, denoted as p_q, where q is the number of sequences from the same label. The score for the label C_i, $S(C_i, p_q)$ is defined by the following equation:

$$S(C_i, p_q) = \frac{\Sigma_{p \in T_q} I(C_i, p) B(p_q, p)}{q} \tag{1}$$

Where, T_q represents the set of proteins associated with label C_i and their bit scores for p_q,

p signifies any protein in T_q,

$I(C_i, p)$ is a binary indicator indicating whether p belongs to the label C_i, and

$B(p_q, p)$ signifies the bit score of the alignment between protein p_q and p.

The label with the highest score is considered the result of the similarity model.

ProtAlign-ARG:

ProtAlign-ARG synergizes PPLM-based scoring with alignment-based scoring to enhance the accuracy of antimicrobial resistance gene (ARG) classification.

In our methodology, we introduce confidence thresholds as a metric for evaluating the reliability of PPLM-based predictions. These thresholds—95%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, and 20%—represent the model's certainty in its classification. We assessed the accuracy of predictions falling below each threshold and observed a marked decline in reliability (Supplementary Figure 3). Specifically, predictions

Table 1. Dataset Split for HMD-ARG-DB using Graphpart with 40% and 90% threshold

Similarity Threshold	Data	ARG (Sequence Count)	Non-ARG (Sequence Count)
40%	Train	14882	9754
	Test	2400	3480
90%	Train	13826	10587
	Test	3456	2647
	Overall	17282	13234

with less than 90% confidence yielded a 45% accuracy rate, indicating that over half of these predictions were incorrect. Consequently, we established 90% as the minimum confidence threshold for utilizing PPLM-based classification for both ARG identification and class classification.

For predictions falling below this 90% confidence threshold, we defaulted to alignment-based scoring. This scoring system is more robust in scenarios with limited data, as it does not rely on the voluminous training datasets that PPLM requires. It improves prediction reliability by comparing the query sequence against a database of known sequences.

However, when the query sequence does not find a match within our training data—a situation that may arise with novel or rare ARGs—we head toward the PPLM for the final classification. The PPLM, leveraging its comprehensive understanding of protein sequence language, is a decision-making tool in such cases.

For ARG identification task we generated training and testing datasets using GraphPart, dividing the data approximately into an 80-20 split. We applied a strict similarity threshold of 40% and 90%. Table 1, represents the data distribution for ARG identification.

In the case of ARG class classification, we conducted experiments under multiple settings. We used GraphPart to partition the HMD-ARG data samples into training and test classes, ensuring that the between-cluster similarity was below 40%. We employed all three approaches and reported the accuracy.

As part of our evaluation, we also conducted independent test set validation. Following the validation approach employed by the DeepARG[12] framework, we introduced a set of 76 metallo beta-lactamase genes sourced from a separate study conducted by Berglund et al. [32]. These newly discovered genes had undergone rigorous experimental validation through functional metagenomics techniques, which confirmed their ability to bestow resistance to carbapenem in *E. coli*. The study encompassed a comprehensive analysis of thousands of metagenomes and bacterial genomes, focusing on a meticulously curated selection of beta-lactamases. Consequently, it is reasonable to assume that these 76 beta-lactamase genes primarily represent authentic ARGs. This presented a unique opportunity to subject our model to further testing and validation.

Experiments & Results

ARG Identification

We conducted training and testing using two different setups, employing both a 40% threshold and a 90% threshold to assess our model's performance. Table 2, provides an overview of the accuracy achieved in both setups, distinguishing between individual and ProtAlign-ARG models.

Table 2. Accuracy for component models on ARG Identification on HMD-ARG-DB using 40 and 90 percent similarity threshold using GraphPart.

Train-Test Similarity	Model	Prec.	Rec.	F1-Score
40%	PPLM	0.85	0.83	**0.84**
	Alignment	0.55	0.55	0.55
	ProtAlignARG	**0.86**	**0.84**	**0.84**
90%	PPLM	0.94	**0.98**	**0.97**
	Alignment	0.71	0.71	0.71
	ProtAlignARG	**0.95**	0.96	0.96

The model consistently outperforms alignment-based models, even when the train and test samples have a similarity of under 40%. It achieves an impressive overall accuracy of 84%. However, it is worth noting that for ARG identification, ProtAlign-ARG does not substantially improve overall accuracy. This result is expected since both the labels, ARG and non-ARG, have an ample amount of data in both the training and test sets.

ARG Class Classification

Table 3. Accuracy for the different component models for ARG class classification on HMD-ARG-DB on 40% threshold

Model	Metric	Precision	Recall	F1-Score
PPLM	Macro	0.58	0.62	**0.56**
	Weighted	0.88	0.83	0.83
Alignment-Scoring	Macro	**0.69**	0.52	0.58
	Weighted	**0.95**	0.76	0.84
ProtAlignARG	Macro	0.63	**0.67**	**0.64**
	Weighted	0.90	**0.89**	**0.89**

From Table 3, it is evident that the PPLM excels in achieving better recall compared to the alignment-based scoring model. However, the PPLM lags in terms of precision when compared to the alignment-based model. The ProtAlign-ARG, on the other hand, strikes a balance by achieving significant precision and recall, leading to an overall improvement in accuracy over both of these models.

From Figure 3, we see that the PPLM performs well for aminoglycoside, beta-lactam, fosfomycin, glycopeptide, tetracycline, and trimethoprim, with high scores across all metrics. The Alignment model exhibits high precision and recall for certain classes such as aminoglycoside and trimethoprim, but performs poorly for polymyxin and rifampin. The Hybrid model performed well for classes like aminoglycoside and beta_lactam. It also improved the performance of the PPLM model for quinolone, sulfonamide, and MLS class for the F1-score.

Additionally, we conducted tests using a 90% similarity threshold. This was particularly important because 19 of the ARG class samples had a limited sample size (Supplementary Fig 1). Table 4 presents the performance of the three models in this scenario.

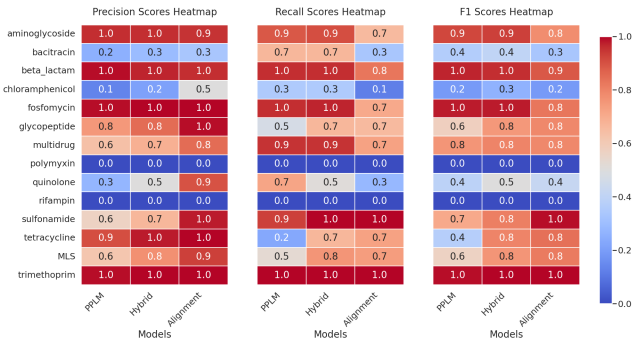

Fig. 3. ARG Classification Accuracy for different models based on precision, recall, and F1 score.

Table 4. Accuracy for the different component models for ARG class classification on HMD-ARG-DB on 90% threshold

Model	Metric	Precision	Recall	F1-Score
PPLM	Macro	0.41	0.45	0.42
	Weighted	0.96	0.97	0.97
Alignment-Scoring	Macro	**0.80**	**0.80**	0.78
	Weighted	**0.98**	**0.98**	**0.98**
ProtAlign-ARG	Macro	0.80	0.79	0.78
	Weighted	**0.98**	**0.98**	**0.98**

In this case, the alignment-scoring model outperformed both the PPLM and the ProtAlign-ARG in precision. This is because the PPLM tends to perform poorly when the sample size is limited for some of the ARG classes. When we experimented with all 33 ARG classes, including the 19 ARG classes with just only 219 sequences in total, the accuracy dropped significantly for the PPLM model. Whereas the alignment-scoring model performed much better. However, when we considered the 90% similarity threshold just for the rest of the 14 frequent ARG classes, the PPLM outperformed the alignment-scoring-based model. Our hybrid ProtAlign-ARG strikes a 78% F1-score matching with the alignment-scoring-based model.

Moreover, we conducted experiments with our models on the COALA dataset (Table 5). We used the COALA90 dataset. As mentioned, the COALA90 dataset was prepared using the CD-HIT tool, which could not maintain a strict threshold between cluster similarity. In this dataset, the performance of the PPLM deteriorated, especially for the ARG classes STREPTOGRAMIN, RIFAMYCIN, and BACITRACIN, which had very low amounts of data and resulted in poor performance. While ProtAlign-ARG outperformed both of them.

Table 5. Accuracy for the different component models for ARG class classification on COALA90 dataset

Models	Accuracy	Precision	Recall	F1-score
PPLM	Macro	0.68	0.65	0.67
	Weighted	0.82	0.81	0.81
Alignment-Scoring	Macro	0.79	0.7	0.71
	Weighted	0.89	0.74	0.80
ProtAlign-ARG	Macro	**0.86**	0.81	**0.83**
	Wighted	0.85	**0.84**	**0.84**

Furthermore, we divided the test dataset into three subsections based on the alignment with the training sequences: those with no alignment, those with less than 50% alignment, and those with greater than 50% alignment using the CDHIT tool. In Table 6 are the accuracies achieved by the PPLM and the ProtAlign-ARG for these test sets.

We observed that for cases with no alignment and very low alignment, the PPLM performed better than the alignment-scoring model. However, as expected, for those with greater than 50% alignment, the ProtAlign-ARG outperformed the PPLM.

Comparison with Other Methods

ARG Identification

In our comparison with other state-of-the-art ARG identification tools, ProtAlign-ARG outperforms them. We conducted a 5-fold cross-validation on the HMD-ARG-DB and repeated this process 10 times to report our accuracy. Our model achieved an accuracy of 97% reported in Table 7.

ARG Classification

When we compared our ARG class classification model on the COALA dataset, it outperformed most alignment-based classification tools and delivered performance comparable to deep learning-based tools. Our model was evaluated on the COALA90 dataset, which was generated using CDHIT clustering. It is worth noting that while our model was trained on the same set of training data, we tested it on a dataset three times larger than that used by other models. Table 8, summarizes the performance of our model relative to other models.

We furthermore compared ProtAlign-ARG using the three splits of the COALA90 dataset. In this case, we reported the overall accuracy of our model, emphasizing that our models were tested on a test set three times larger than other tools in Table 9.

We also show the accuracy of our model using a 5-fold cross-validation approach on the HMD-ARG-DB in (Supplementary Table 1). ProtAlign-ARG consistently outperformed the other models in this evaluation.

Additionally, we expanded the scope of our prediction tasks to include ARG resistance mechanism prediction and mobility identification. In both cases, PPLM achieved better performance than other SOTA tools when using 5-fold cross-validation (Supplementary Table 2 & 3).

Independent Test Set Validation

ProtAlign-ARG, trained on GraphPart-based clustering with a 40% similarity threshold on the HMD-ARG-DB dataset, successfully identified all 76 beta-lactamase samples from Berglund et al. [32] as ARGs. Furthermore, the ARG class classification model categorized these genes as beta-lactamases. To ensure that the training set did not contain these beta-lactamase genes, we conducted DIAMOND alignments, excluding any sequences with similarity above 40% during the model's training. This performance underscores the effectiveness of our model in identifying and annotating novel ARGs compared to existing tools.

Table 6. Accuracy for the different component models for ARG class classification on COALA dataset for varying percentage similarity

Models	Similarity	Accuracy	Precision(%)	Recall(%)	F1-score(%)
PPLM	Greater 50	Macro	74	71	72
		Weighted	90	91	90
	Less 50	Macro	58	53	55
		Weighted	72	73	72
	No Alignment	Macro	42	47	**42**
		Weighted	41	45	**41**
Alignment-Scoring	Greater 50	Macro	85	81	82
		Weighted	93	92	92
	Less 50	Macro	46	41	42
		Weighted	72	65	68
	No Alignment	Macro	0	0	0
		Weighted	0	0	0
ProtAlign-ARG	Greater 50	Macro	94	89	**91**
		Weighted	95	95	**95**
	Less 50	Macro	56	56	**56**
		Weighted	73	73	**73**
	No Alignment	Macro	42	47	**42**
		Weighted	41	45	**41**

Table 7. Accuracy for the different component models for ARG Identification on HMD-ARG-DB

Models	Precision	Recall	F1-Score
HMD-ARG	0.939	0.971	0.948
CARD	0.999	0.421	0.592
DeepARG	0.998	0.93	0.963
AMRPlusPlusb	0.867	0.449	0.592
Meta-MARC	0.847	0.85	0.848
ProtAlign-ARG	0.97	0.98	0.97

Table 8. Accuracy for the different models for ARG Classification on COALA90 dataset

Models	Macro Avg. Score	Weighted Avg. Score
BLAST best hit	0.8258	0.8423
DIAMOND best hit	0.8103	0.8423
DeepARG	0.7303	0.8419
HMMER	0.4499	0.4916
TRAC	0.7399	0.8097
ARG-SHINE	0.8555	0.8591
PPLM Model	0.67	0.81
Alignment-Score	0.71	0.80
ProtAlign-ARG	0.83	0.84

Conclusion

Our study presents a robust pipeline and benchmarking standard for the identification and class classification of ARGs. ProtAlign-ARG outperformed existing methods in both ARG identification and class classification. Additionally, its impressive performance in resistance mechanism prediction and mobility classification further establishes its superiority in prediction tasks. One limitation of the current pipeline is that the clustering methods for ensuring the quality of the training and testing data split can be time-consuming. Moreover, the performance of the model is low for sequences with low identities to the training data. As protein structure is more conserved than sequences, we plan to incorporate the protein 3D structure information into the models to further improve the performance.

Table 9. Accuracy for the different component models for ARG Classification on three splits of COALA90 dataset

Models	No-Alignment	Alignment <50	Alignment >50
BLAST	0	0.6243	0.9542
DIAMOND	0	0.5740	0.9534
DeepARG	0	0.5266	0.9419
HMMER	0.0563	0.2751	0.6051
TRAC	0.3521	0.6124	0.9199
ARG-SHINE	0.4648	0.6864	0.9558
PPLM Model	0.45	0.73	0.91
Alignment	0	0.65	0.92
ProtAlignARG	0.45	0.73	0.95

Acknowledgments

This work was partly funded by the National Science Foundation (NSF grants #2319522, #2125798, and #2004751).

References

1. Viktória Lázár and Roy Kishony. Transient antibiotic resistance calls for attention. *Nature Microbiology*, 4(10):1606–1607, 2019.
2. Jim O'Neill. Tackling drug-resistant infections globally: final report and recommendations. 2016.
3. José L Martínez. Antibiotics and antibiotic resistance genes in natural environments. *Science*, 321(5887):365–367, 2008.
4. Luria Leslie Founou, Raspail Carrel Founou, and Sabiha Yusuf Essack. Antibiotic resistance in the food chain: a developing country-perspective. *Frontiers in microbiology*, 7:1881, 2016.
5. Antti Karkman, Thi Thuy Do, Fiona Walsh, and Marko PJ Virta. Antibiotic-resistance genes in waste water. *Trends in microbiology*, 26(3):220–228, 2018.
6. Qiang Wang, Panliang Wang, and Qingxiang Yang. Occurrence and diversity of antibiotic resistance in untreated hospital wastewater. *Science of the Total Environment*, 621:990–999, 2018.

7. W-Y Xie, Q Shen, and FJ Zhao. Antibiotics and antibiotic resistance from animal manures to soil: a review. *European journal of soil science*, 69(1):181–195, 2018.
8. Tymor Hamamsy, James T Morton, Robert Blackwell, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Charlie EM Strauss, Julia Koehler Leman, Kyunghyun Cho, and Richard Bonneau. Protein remote homology detection and structural alignment using deep learning. *Nature biotechnology*, pages 1–11, 2023.
9. Kevin Liu, C Randal Linder, and Tandy Warnow. Raxml and fasttree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PloS one*, 6(11):e27731, 2011.
10. Mesih Kilinc, Kejue Jia, and Robert L Jernigan. Protein language model performs efficient homology detection. *bioRxiv*, pages 2022–03, 2022.
11. Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.
12. Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6:1–15, 2018.
13. Yu Li, Zeling Xu, Wenkai Han, Huiluo Cao, Ramzan Umarov, Aixin Yan, Ming Fan, Huan Chen, Carlos M Duarte, Lihua Li, et al. Hmd-arg: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome*, 9:1–12, 2021.
14. Ziye Wang, Shuo Li, Ronghui You, Shanfeng Zhu, Xianghong Jasmine Zhou, and Fengzhu Sun. Arg-shine: improve antibiotic resistance class prediction by integrating sequence homology, functional information and deep convolutional neural network. *NAR Genomics and Bioinformatics*, 3(3):lqab066, 2021.
15. Shafayat Ahmed, Muhit Islam Emon, Nazifa Ahmed Moumi, and Liqing Zhang. Lm-arg: Identification & classification of antibiotic resistance genes leveraging pre-trained protein language models. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3782–3784. IEEE, 2022.
16. Jun Wu, Jian Ouyang, Haipeng Qin, Jiajia Zhou, Ruth Roberts, Rania Siam, Lan Wang, Weida Tong, Zhichao Liu, and Tieliu Shi. Plm-arg: antibiotic resistance gene identification using a pretrained protein language model. *Bioinformatics*, 39(11):btad690, 2023.
17. Georgina Cox and Gerard D Wright. Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology*, 303(6-7):287–292, 2013.
18. Howard Ochman, Jeffrey G Lawrence, and Eduardo A Groisman. Lateral gene transfer and the nature of bacterial innovation. *nature*, 405(6784):299–304, 2000.
19. Michael Feldgarden, Vyacheslav Brover, Daniel H Haft, Arjun B Prasad, Douglas J Slotta, Igor Tolstoy, Gregory H Tyson, Shaohua Zhao, Chih-Hao Hsu, Patrick F McDermott, et al. Using the ncbi amrfinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of narms isolates. *BioRxiv*, page 550707, 2019.
20. Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, et al. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, 48(D1):D517–D525, 2020.
21. Valeria Bortolaia, Rolf S Kaas, Etienne Ruppe, Marilyn C Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, et al. Resfinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12):3491–3500, 2020.
22. Molly K Gibson, Kevin J Forsberg, and Gautam Dantas. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, 9(1):207–216, 2015.
23. 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. 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, 2020.
24. Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. Arg-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy*, 58(1):212–220, 2014.
25. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.
26. K-K Ma and Sarah A Rajala. A comparison of absolute moment block truncation coding and the minimum mean square error quantizer. In *1991., IEEE International Sympoisum on Circuits and Systems*, pages 296–299. IEEE, 1991.
27. Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. Graphpart: homology partitioning for biological sequence analysis. *NAR genomics and bioinformatics*, 5(4):lqad088, 2023.
28. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
29. Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.
30. Shan Suthaharan and Shan Suthaharan. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235, 2016.
31. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.
32. Fanny Berglund, Nachiket P Marathe, Tobias Österlund, Johan Bengtsson-Palme, Stathis Kotsakis, Carl-Fredrik Flach, DG Joakim Larsson, and Erik Kristiansson. Identification of 76 novel b1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, 5:1–13, 2017.