Deep Learning for Antibiotic Resistance Detection

Vedant Mahangade, Lucia Sanchez,
Matthew Mollerus
PUBH 8885: Computational Biology

Milken Institute School
of Public Health
THE GEORGE WASHINGTON UNIVERSITY

Outline

1. Motivation
2. Our Objectives
3. Reproducing Existing SotA Model
4. DNA Language Model
5. Ablation Testing

AMR - Global Threat

Why is it a problem?

Significant PH Burden

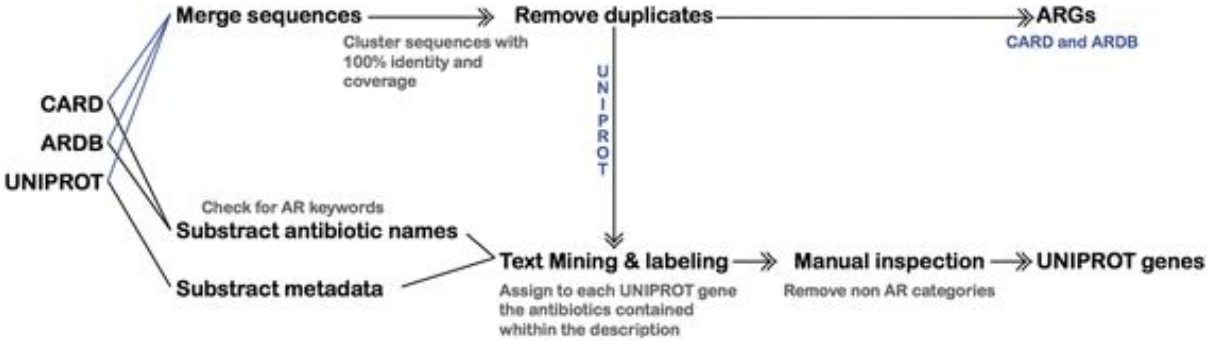
- Number of Deaths by 2050: 10 million annually
- Cost by 2050: 100 trillion USD globally

Importance of ARGs

- Antimicrobial Resistance Genes (ARGs) are pivotal in the development of antibiotic resistance
- ARG detection enables scientists and public health officials to identify where and how resistance genes are spreading in bacterial populations, both within healthcare settings and the environment
- Studying ARGs helps researchers uncover the molecular mechanisms by which bacteria resist antibiotics
  - this is important for developing new antibiotics or alternative treatments that can bypass or inhibit these resistance pathways.

DeepARG Overview

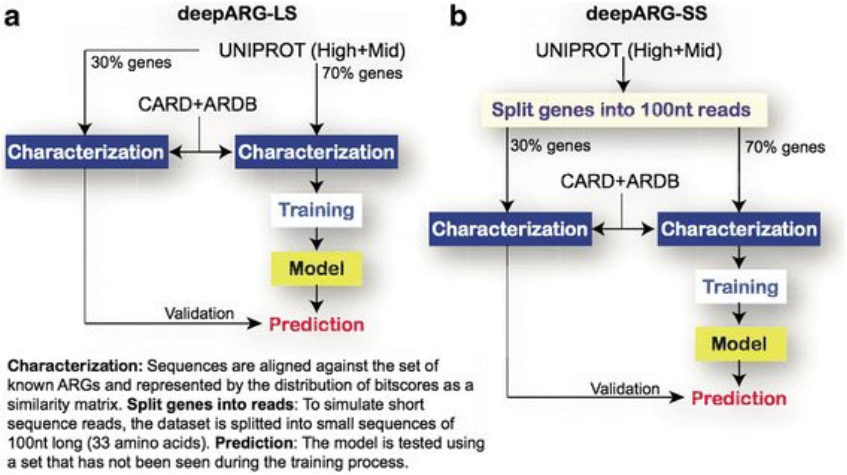
DeepARG Overview



Data

- Collected ARGs from ARG specific databases (CARD, ARDB), plus those annotated in general protein database (UNIPROT)
- QC on annotations and removed duplicates
- Used ARGs from ARG specific databases as reference

DeepARG Overview



Characterization: Sequences are aligned against the set of known ARGs and represented by the distribution of bitscores as a similarity matrix. **Split genes into reads**: To simulate short sequence reads, the dataset is splitted into small sequences of 100nt long (33 amino acids). **Prediction**: The model is tested using a set that has not been seen during the training process.

Models

- Aligned UNIPROT genes with reference genes via DIAMOND to create feature vector of normalized bitscore (characterization)
- Trained MLP with 4 hidden layers
- Two models, one for long reads (entire genes) and one for short (100bp)

DeepARG Problems

DeepARG Problems

- Amino acid sequences ≠ genes!

DeepARG Problems

- Amino acid sequences ≠ genes!
- Incomplete/inaccurate public data

DeepARG Problems

- Amino acid sequences ≠ genes!
- Incomplete/inaccurate public data
- Lack of detail on method (even in code)

DeepARG Problems

- Amino acid sequences ≠ genes!
- Incomplete/inaccurate public data
- Lack of detail on method (even in code)
- Minimal ablation testing

Our Reproduction

- Use dataset closest to that discussed in paper, but track down as many associated DNA sequences as possible

Our Reproduction

- Use dataset closest to that discussed in paper, but track down as many associated DNA sequences as possible
- Train four models: AA long read and short read (their method), plus DNA long read and short read (still using AA sequences for reference ARGs)

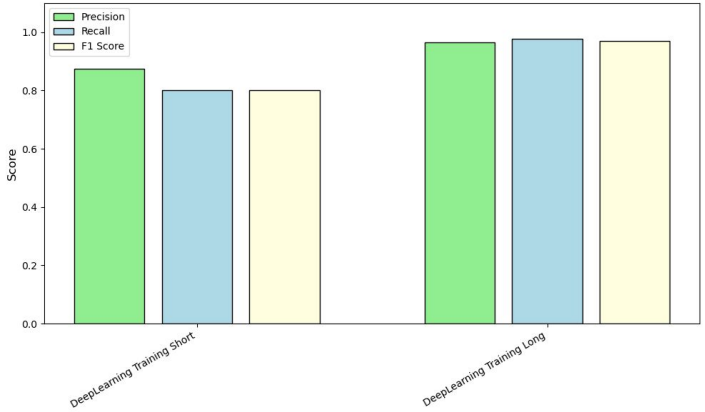
Our Reproduction

- Use dataset closest to that discussed in paper, but track down as many associated DNA sequences as possible
- Train four models: AA long read and short read (their method, plus DNA long read and short read (still using AA sequences for reference ARGs)
- More extensive ablation testing

Amino Acid Model Results

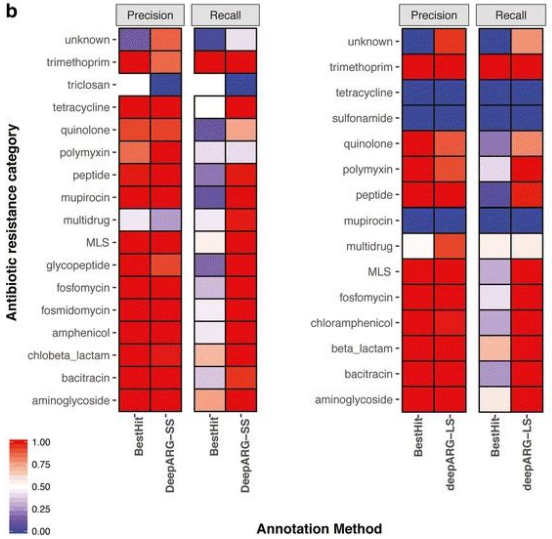
DeepARG

Reproduction

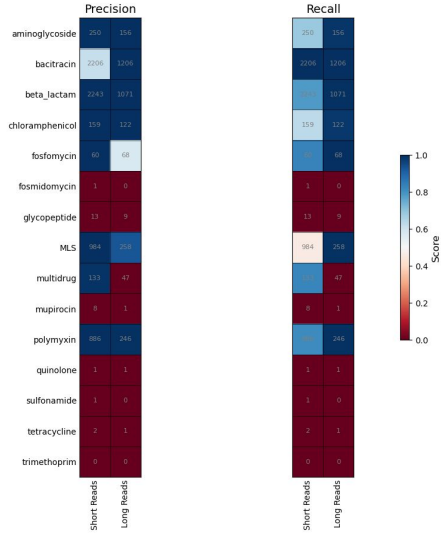


Amino Acid Model Results

DeepARG



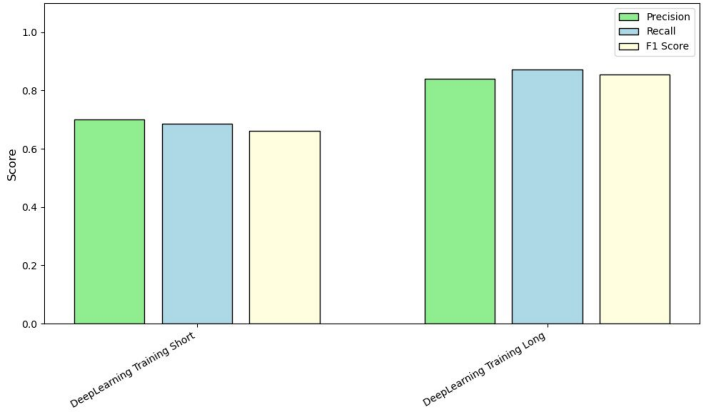
Reproduction



DNA Model Results

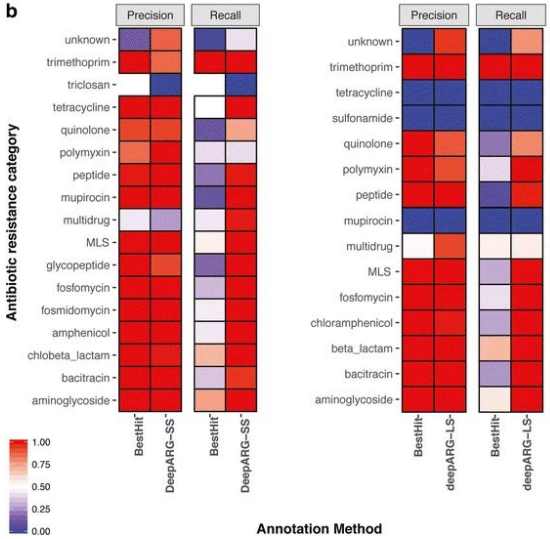
DeepARG

Reproduction

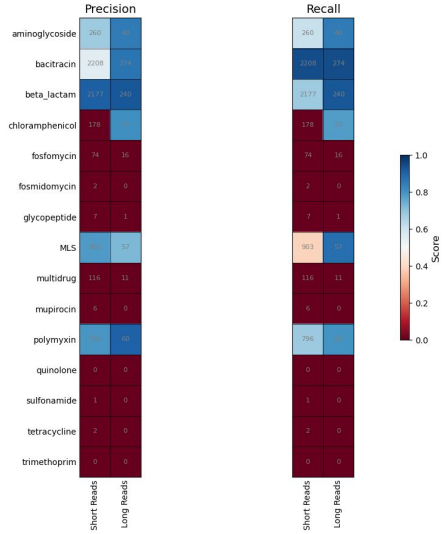


DNA Model Results

DeepARG



Reproduction

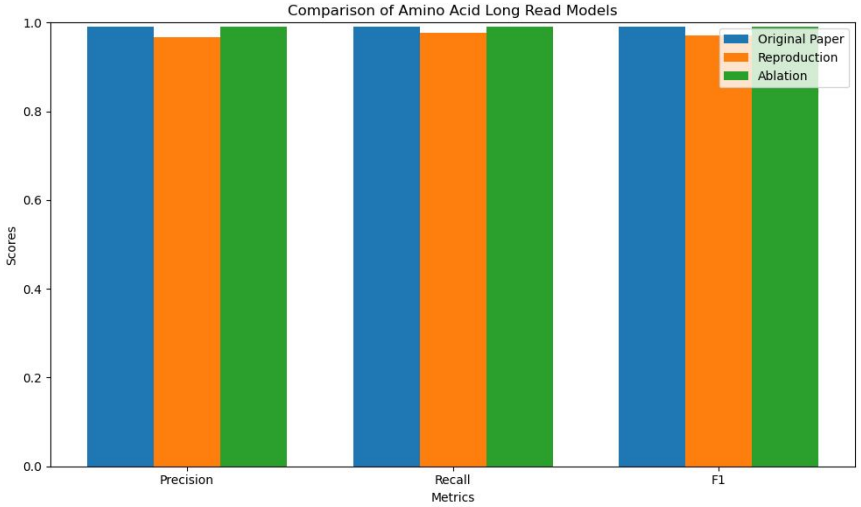


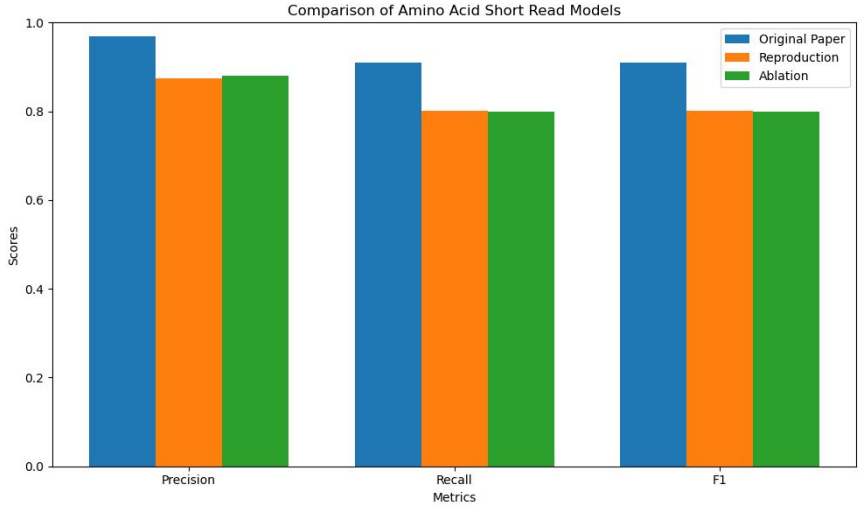
Ablation Testing

- Removes complexity to model to determine if complexity is justified and which portions of the model contribute to performance
- Last layer of MLP is just logistic regression, so here we trained a logistic regression model on the input feature vector
- MLP is ~12M parameters and uninterpretable; logistic regression is ~150k and highly interpretable

DeepARG Model Ablation Testing

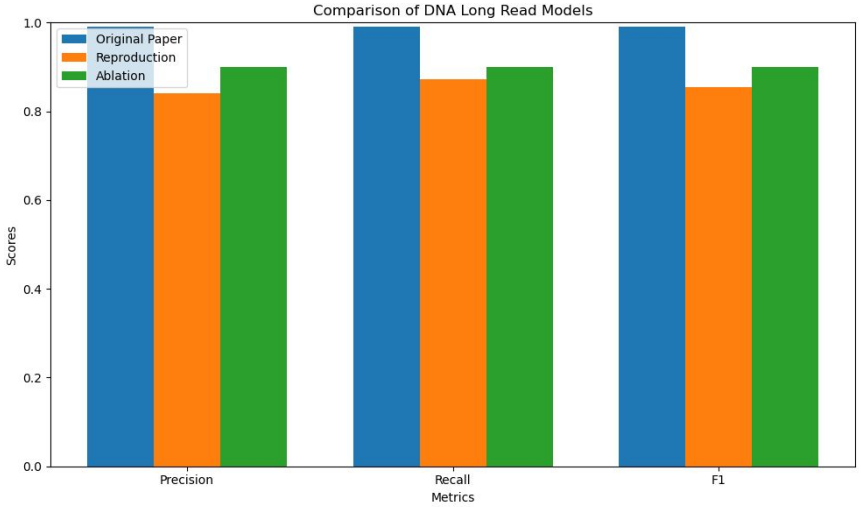
- Trained logistic regression models on just the input feature vector of alignment bit scores to reference sequences
- With amino acids, the logistic model performed as well or better than our reproduction of the MLP model

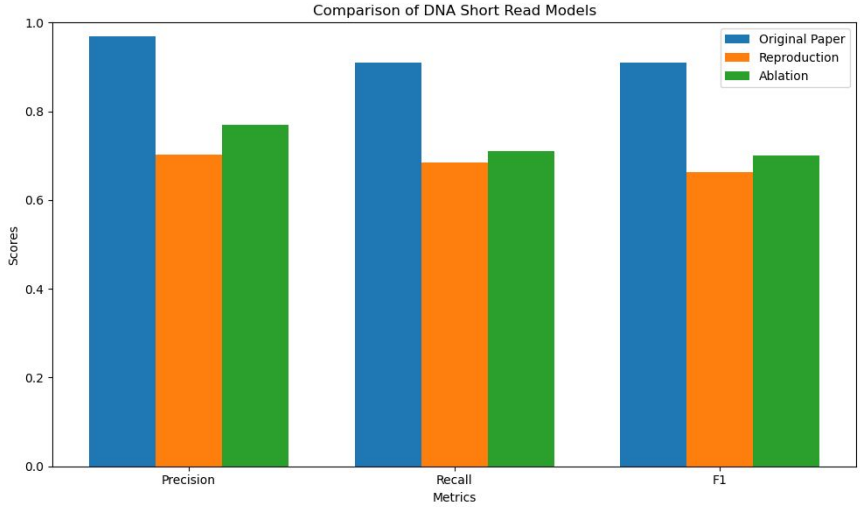




DeepARG Model Ablation Testing

- Logistic models performed even better compared to the DNA-based MLP models
- Takeaway: deep learning may add unnecessary complexity to alignment-based methods of detecting ARGs





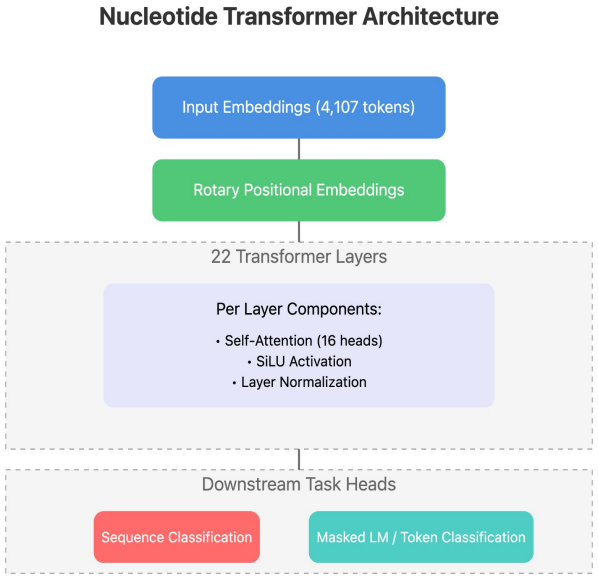
Nucleotide Transformer

Why use a Transformer based approach?

- Long range dependencies in DNA sequences are modeled effectively
- Pre-training provides deep understanding of patterns across multiple species

Pretrained Model

- The pretrained model used here was [InstaDeepAI/nucleotide-transformer-v2-100m-multi-species]
- Pretrained on collection of 850 genomes, representing a total of 174B nucleotides



Training Considerations

Limitations:

- The maximum token length the pretrained model could handle was 2048.
- The maximum sequence length in dataset for long reads was 3594.
- The fine-tuning was performed on A100 GPU with 40 GB memory.
- Although the memory is enough, the system could only handle token length of 1024.

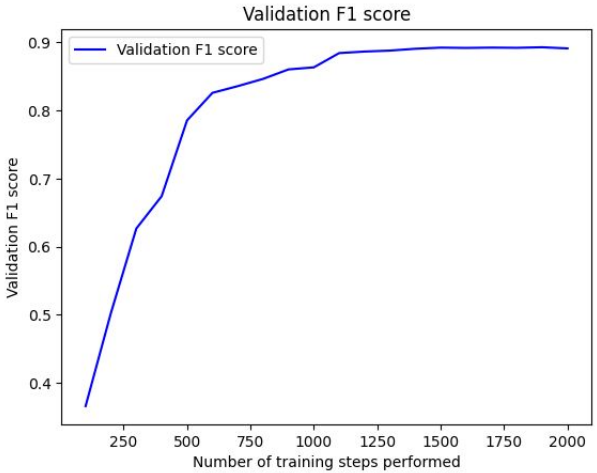
To overcome these limitations 3 steps were taken:

- Mixed precision training - fp16
- Effective Batch Size: 64
  - Using batch size of 8 and
  - Gradient accumulation steps of 8 for each iteration
- Sequence splitting and batch processing
  - max_length of 1024 with
  - overlap of 512

Fine-tuning

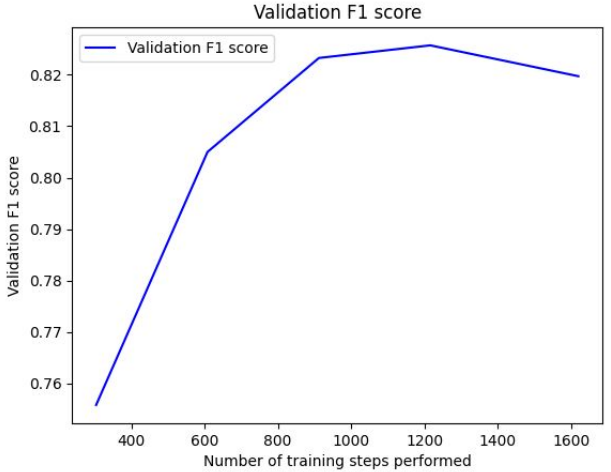
Long Reads

F1	Recall	Precision
0.892495	0.891986	0.894948



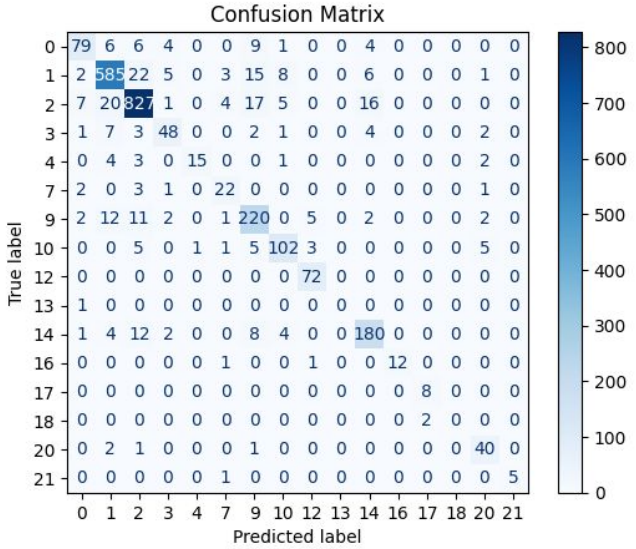
Short Reads

F1	Recall	Precision
0.82574	0.82755	0.8256



Testing on Long Reads

- # of Samples: 2018
- Consistent scores across metrics indicate that model generalizes well
- Class 12 (non resistant) also gets high accuracy
- Highly Scalable with more data
- Confusions could also be linked to biologically similar sequences

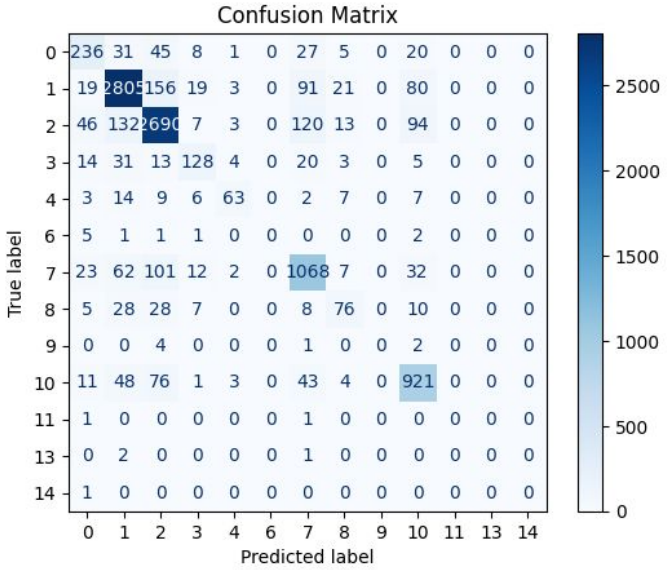


F1	Recall	Precision	Accuracy
0.88099	0.88177	0.88225	0.88177

Testing on Short Reads

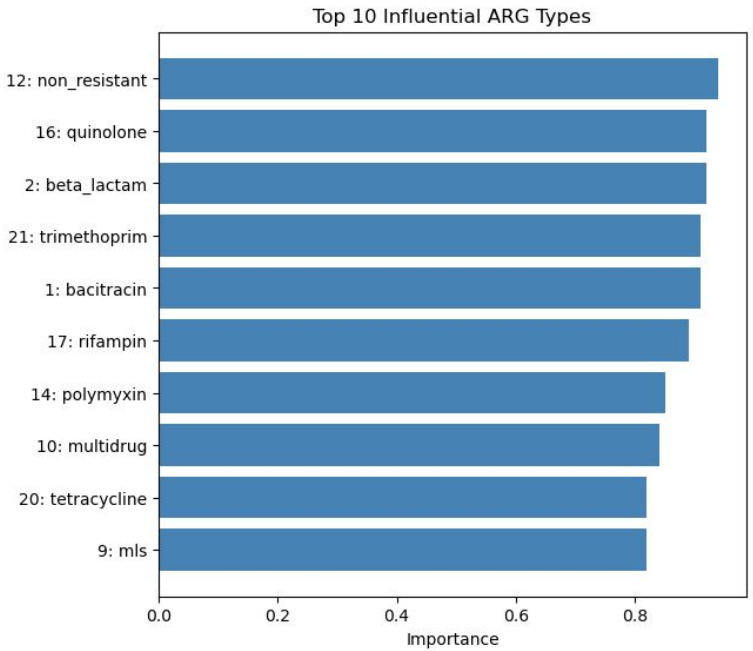
- # of Samples: 9615
- Performance of model for dominant classes is consistent with that of long reads
- Misclassification here is higher than observed in the long read model
- Short reads have max length of 100 which might not be enough for model to learn biological patterns.

F1	Recall	Precision	Accuracy
0.830177	0.831979	0.829460	0.831979

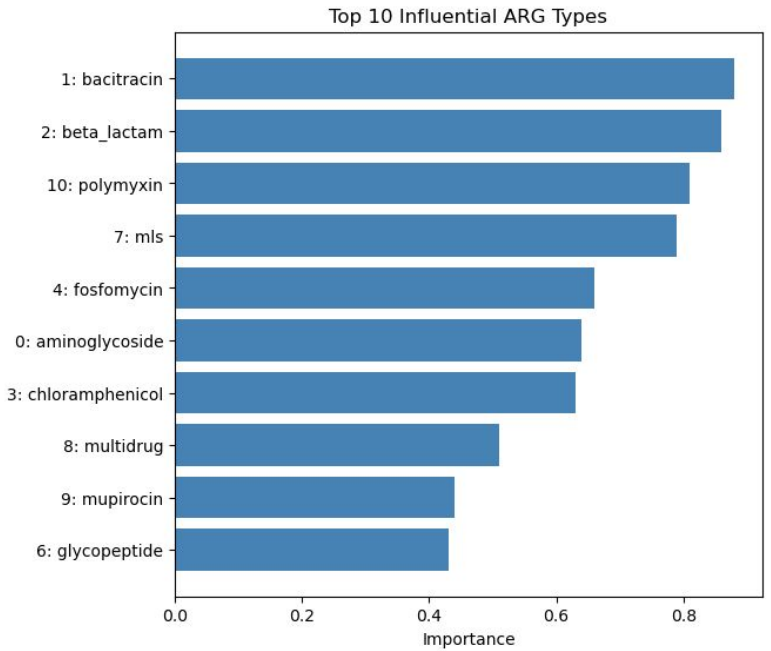


Feature Importance

Long Reads



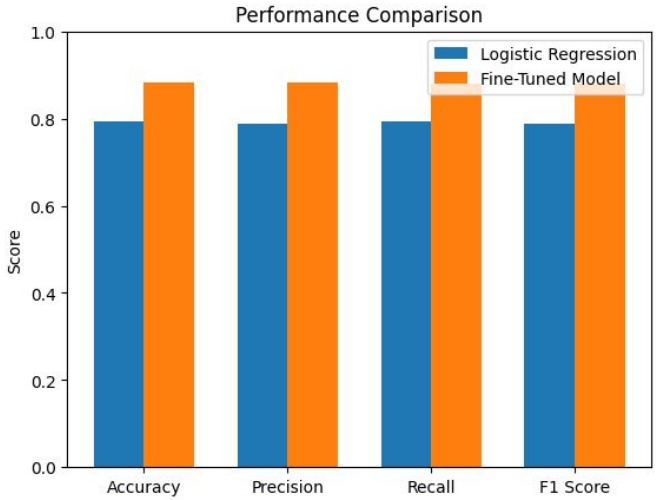
Short Reads



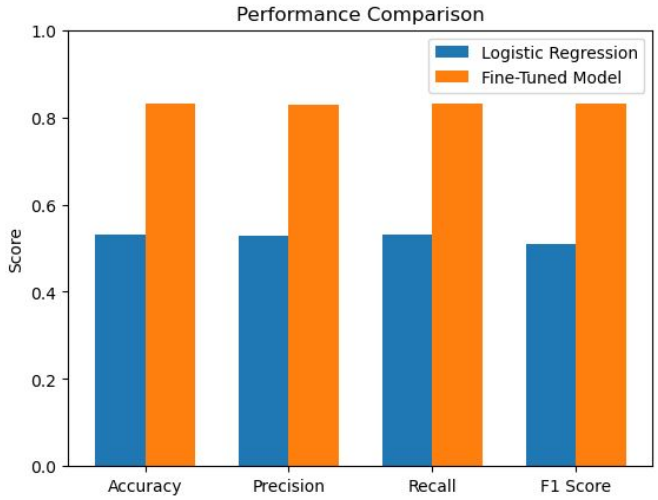
Language Model Ablation Using Logistic Regression Head

- We extract embedding from the pretrained model using tokenized data
- These CLS embedding serve as an input to Logistic Regression Model
- The logistic regression model is trained for 1000 iterations to classify sequences

For Long Reads

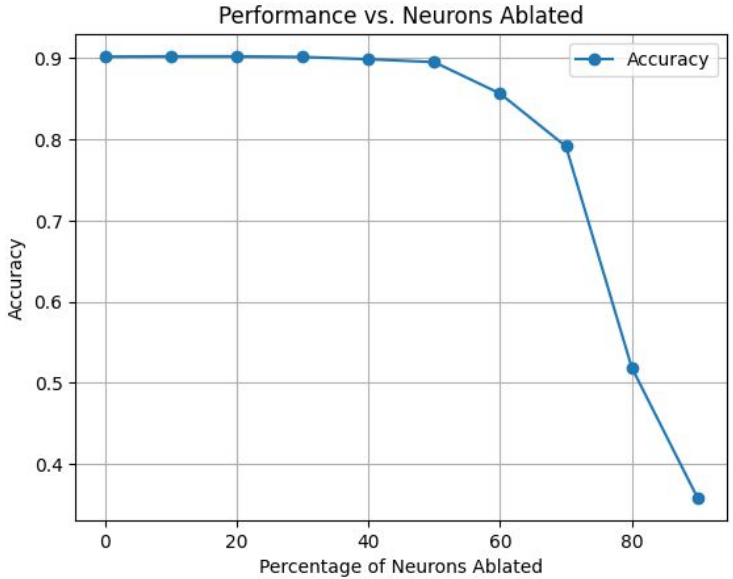


For Short Reads

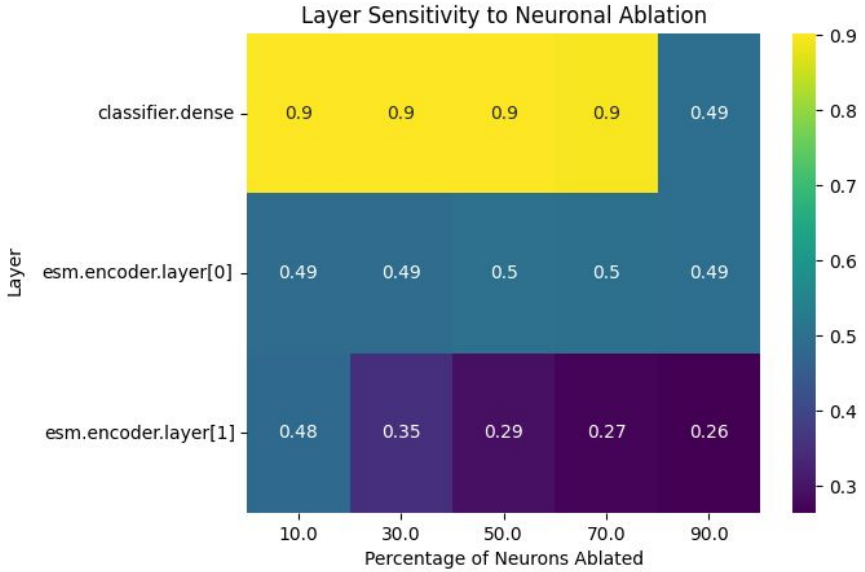


Neuron and Layer Ablation

Neuron Ablation



Layer-Wise Sensitivity



Overall Comparison

Metric	DeepARG-SR	DeepARG-LR	Repro-AA-SR	Repro-AA-LR	Repro-DNA-SR	Repro-DNA-LR	NT-SR	NT-LR
Precision	0.97	0.99	0.87	0.97	0.70	0.84	0.83	0.89
Recall	0.91	0.99	0.80	0.98	0.69	0.87	0.83	0.89
F1	0.91	0.99	0.80	0.97	0.66	0.86	0.83	0.89

Overall Comparison

Metric	DeepARG-SR	DeepARG-LR	Repro-AA-SR	Repro-AA-LR	Repro-DNA-SR	Repro-DNA-LR	NT-SR	NT-LR
Precision	0.97	0.99	0.87	0.97	0.70	0.84	0.83	0.89
Recall	0.91	0.99	0.80	0.98	0.69	0.87	0.83	0.89
F1	0.91	0.99	0.80	0.97	0.66	0.86	0.83	0.89

- Takeaway #1: DNA-based models always perform worse than amino acid-based models, but may be more useful in real-world analysis

Overall Comparison

Metric	DeepARG-SR	DeepARG-LR	Repro-AA-SR	Repro-AA-LR	Repro-DNA-SR	Repro-DNA-LR	NT-SR	NT-LR
Precision	0.97	0.99	0.87	0.97	0.70	0.84	0.83	0.89
Recall	0.91	0.99	0.80	0.98	0.69	0.87	0.83	0.89
F1	0.91	0.99	0.80	0.97	0.66	0.86	0.83	0.89

- Takeaway #1: DNA-based models always perform worse than amino acid-based models, but may be more useful in real-world analysis
- Takeaway #2: when comparing apples-to-apples, DNA language models outperform DNA-based alignment models

Overall Comparison

Metric	DeepARG-SR	DeepARG-LR	Repro-AA-SR	Repro-AA-LR	Repro-DNA-SR	Repro-DNA-LR	NT-SR	NT-LR
Precision	0.97	0.99	0.87	0.97	0.70	0.84	0.83	0.89
Recall	0.91	0.99	0.80	0.98	0.69	0.87	0.83	0.89
F1	0.91	0.99	0.80	0.97	0.66	0.86	0.83	0.89

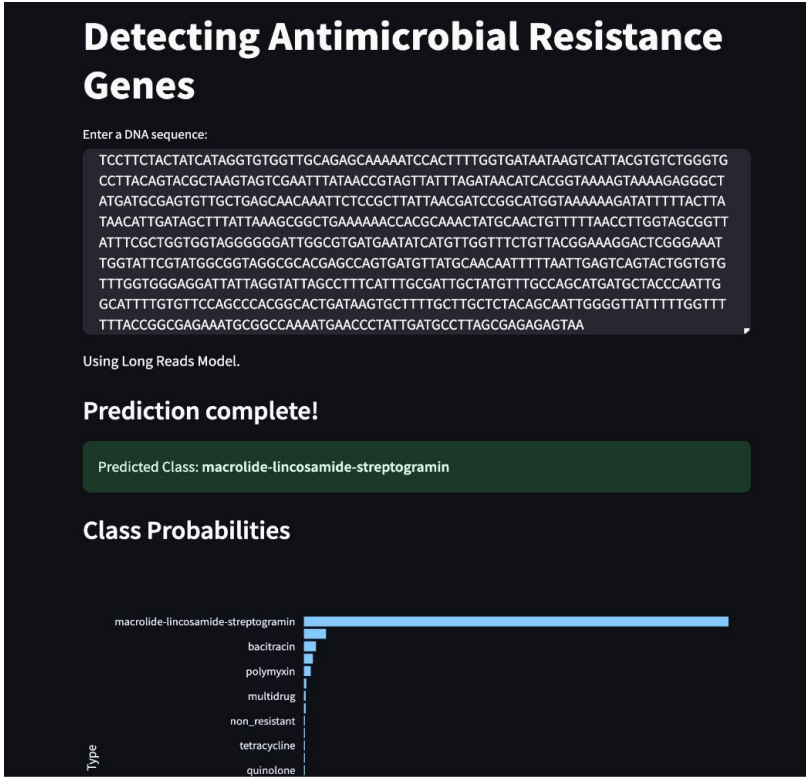
- Takeaway #1: DNA-based models always perform worse than amino acid-based models, but may be more useful in real-world analysis
- Takeaway #2: when comparing apples-to-apples, DNA language models outperform DNA-based alignment models
- Takeaway #3: deep learning may not be needed for alignment-based approaches

Conclusion

- Reproduced DeepARG Results
- Transitioning from alignment-based and deep learning methods to DNA language models improved ARG detection performance in the short read sequence data
- Leveraging DNA sequences, rather than amino acid sequences alone, can potentially show new patterns and insights into resistance evolution
- DeepARG: 654 citations, but still problems with reproducibility and clarity

Links

1. Short Reads Model
1. Long Reads Model
2. Demo App



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC