Predicting *Escherichia coli* Drug Resistance through Different Deep Learning-Based Approaches using a Comprehensive Pan-genome Assembly

Abdoulfatah Abdillahi, Estefanos Kebebew, Gian Carlo L. Baldonado, Juvenal F Barajas, Myco Torres, Andrew Scott, MS., Anagha Kulkarni, PhD., Ilmi Yoon, PhD., Pleuni Pennings, PhD.

Introduction

Drug resistance, exemplified in *Escherichia coli (E. coli)*, is a global health threat. Traditional drug-resistance testing takes a long time, has low through-put, and is only possible with bacteria that can be cultivated in labs.

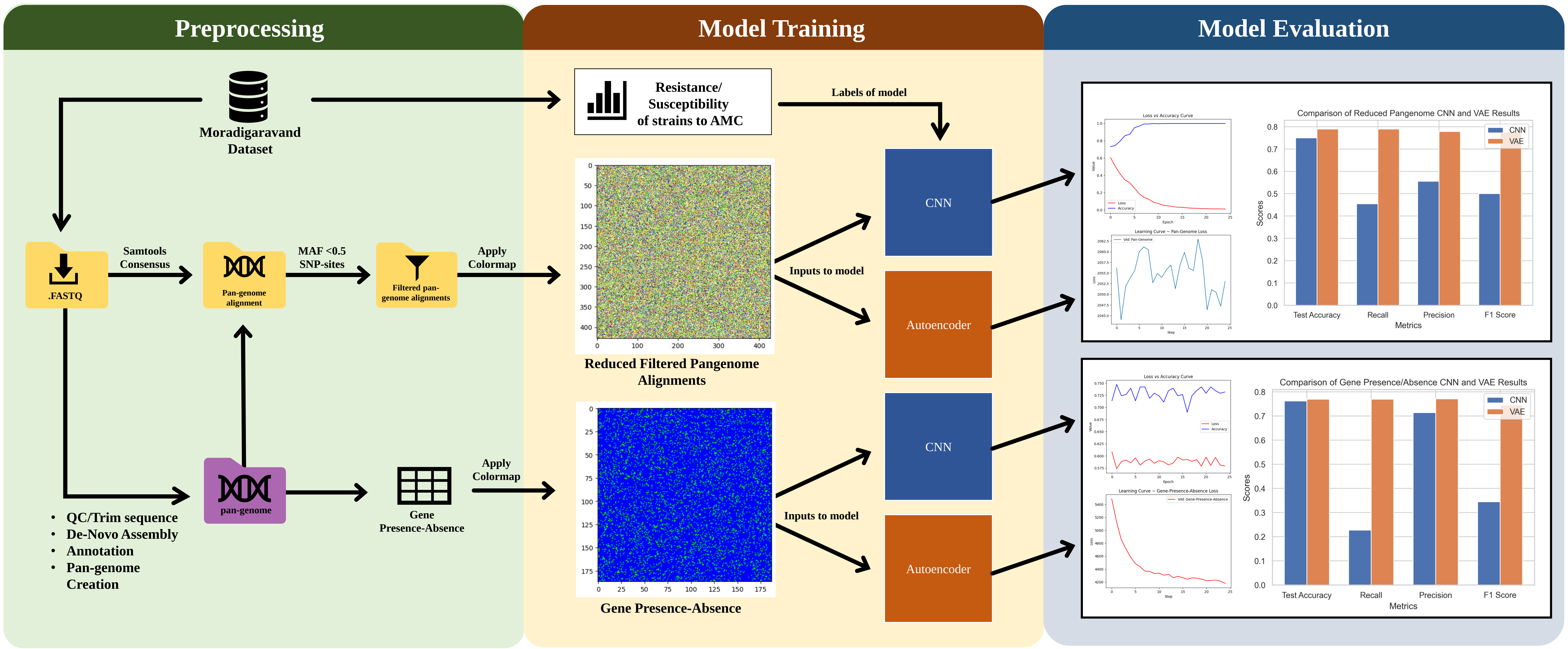
- Machine learning (ML) enables new possibilities in predicting drug resistance more efficiently.
- Previous ML-based studies have shown that single nucleotide polymorphisms (SNPs) and gene presence-absence tables are good predictors for drug resistance.
- Advancements in DNA sequencing enable us to create a comprehensive pan-genome assembly, also called pan-genome alignments, which contain both gene presence-absence and SNP information.

In this project, we investigate the efficacies of deep learning architectures convolutional neural networks (CNN) and variational auto-encoder (VAE) in drug resistance prediction in *E. coli* for amoxicillin (AMC).

Key Points

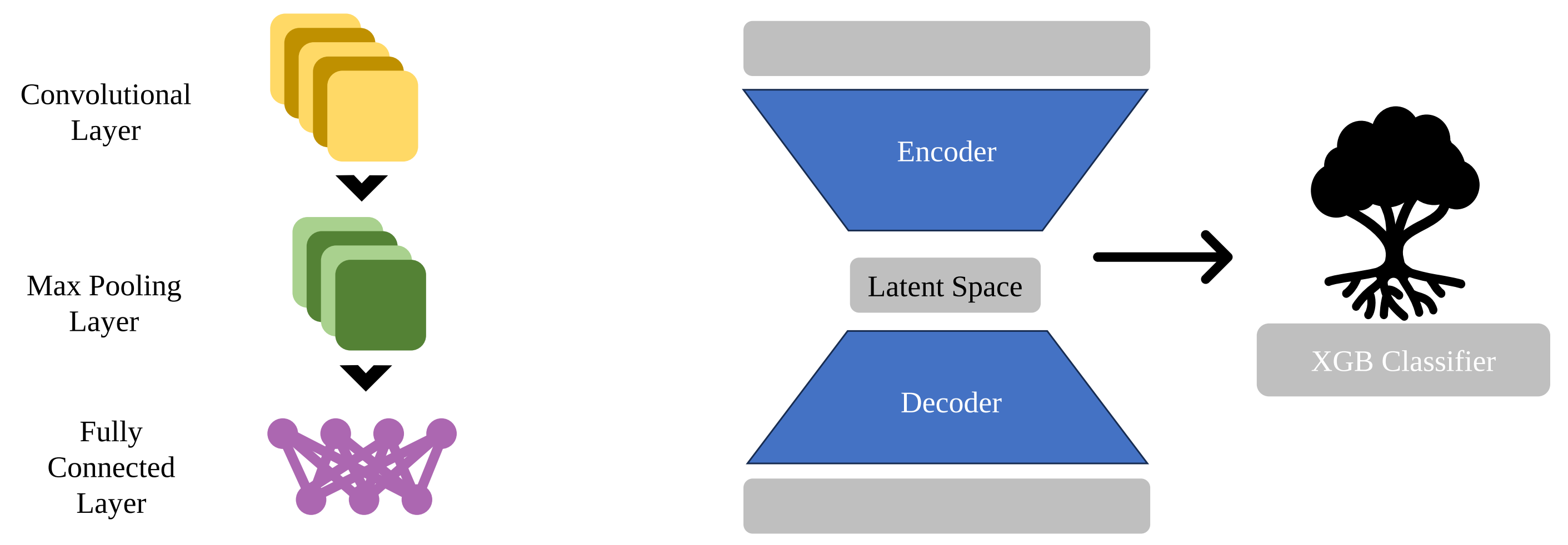
- Using convolutional neural networks (CNN) and variational auto-encoder (VAE) in the task of drug resistance.
- Aligning the pangenome allows us to use both single nucleotide polymorphisms (SNPs) and gene presence-absence in our training.
- Using visual colormaps to densely embed DNA sequence data as input for CNN and VAE.
- Reducing pangenome size with minor allele frequency.

Pipeline



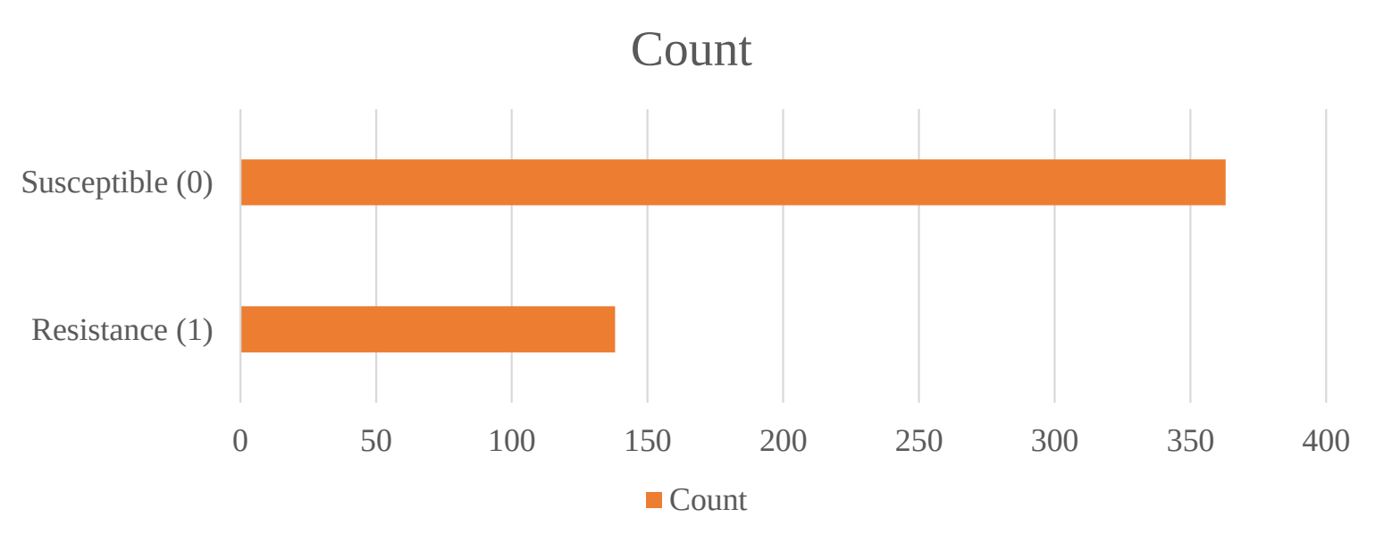
Architectures

CNNs are a type of neural network that excel at image recognition. They use filters that slide across the input, identifying patterns like edges and shapes.
Auto-encoders are a type of neural network that learn to compress data into a latent space, which is then used as inputs to an XGB Classifier, an ensemble-based classifier.



Dataset

Moradigaravand Dataset n=501 subset	Count, Percentage
Susceptible (0)	363, 72.46
Resistance (1)	138, 27.54
Total	501



Acknowledgement

Funding:

- Student Enrichment Opportunities @SFSU
- NIH
- Bristol Myers Squibb

Conclusion

- Model evaluation reveals that reduced pan-genome models have better performance than gene presence-absence models, which may indicate that the reduced pan-genome dataset is a better predictor for drug resistance.
- VAE consistently outperformed CNN across all evaluated metrics.
- Both CNN and VAE have better training and testing performances with the reduced pan-genome dataset than the gene presence-absence dataset.

Future Direction

- Investigate the efficacy of the VAE and CNN by increasing the dataset.
- Explore other architectures such as the multi-layer perceptron and Vision Transformers.
- Compare the performance of color maps and DNA sequences using techniques in natural language processing.
- Multi-label binary classification of multiple drugs.

References

- Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., & Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Computational Biology, 14(12), e1006258. https://doi.org/10.1371/journal.pcbi.1006258
- Muneeb, M., Feng, S. F., & Henschel, A. (2022). Can We Convert Genotype Sequences Into Images for Cases/Controls Classification? Frontiers in Bioinformatics, 2. https://www.frontiersin.org/articles/10.3389/fbinf.2022.914435
- Poirel, L., Madec, J.-Y., Lupo, A., Schink, A.-K., Kieffer, N., Nordmann, P., & Schwarz, S. (2018) Antimicrobial Resistance in Escherichia coli. Microbiology Spectrum, 6(4), 10.1128/microbiolspec.arba-0026–2017. https://doi.org/10.1128/microbiolspec.arba-0026-2017
- Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., Hauschild, A.-C., Schwengers, O., & Heider, D. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics, 38(2), 325–334. https://doi.org/10.1093/bioinformatics/btab681