Investigating the Relationship Between Bacteriophage Presence and Antimicrobial Resistance Phenotype in Different Environments

Corey Woods – 40296019

Supervisor: Prof. Chris Creevey

Word Count:

Abstract: 330

Lay Summary: 550

Article: 4400

I certify that the submission is my own work, all sources are correctly attributed, and the contribution of any assistive technologies is fully acknowledged and conforming the use of AI agreed with my supervisor.

Abstract:

Antimicrobial resistance (AMR) is a significant global health threat. Defined as the ability of a microorganism to become able to adapt and grow in the presence of medications that once inhibited their growth, it leads to ineffective treatment and persistent infections, and is driven by antibiotic overuse and misuse. This project considered the relationship AMR has with bacteriophages, hypothesising that they directly influence AMR phenotypes in *Escherichia coli* against Amoxicillin.

Bioinformatic and machine learning techniques were utilised to analyse 272 bacterial genomes to identify the presence of resistance genes, and prophage communities, that may influence antimicrobial resistance in bacteria. The J48 Decision Tree algorithm employed by Weka predicted AMR to a high degree in three datasets – bacterial genomes, prophage communities, and a dataset which combined the two, with true positive rates of 0.934, 0.886, and 0.926 respectively. 10 antimicrobial resistance genes (AMGs) and 17 prophage communities were highlighted by the models as significant, such as the beta-lactamase gene '*bla-TEM1B',* which produces enzymes against beta-lactam antibiotics including Amoxicillin. Other AMGs were highlighted as conferring resistance against Amoxicillin, such as the aminoglycoside '*aac(3)-ila',* illustrating the complexity of AMR mechanisms and the potential for cross-resistance among different antibiotic classes. The combined data dataset highlighted one route of resistance involving both an AMG and a prophage community - the gene 'drfA5' and Community 58, demonstrating the topical relationship between AMGs and prophages in conferring resistance phenotypes.

The prophage communities were labelled in a way that did not provide specific identifiers that linked them to known prophage sequences, limiting the ability to determine how they directly contribute to AMR. However, StarAMR analysis of these sequences found the presence of several AMGs, including the beta-lactamases '*blaTEM-1B' and 'blaCTX-M-15'*, which not only

supported the finding that bacteriophage presence influences AMR phenotypes, but they themselves harbour AMGs.

Future research initiatives must involve mitigating phage enabled AMR, such as novel therapeutics targeting phage infection, while the importance of antibiotic stewardship, to minimise further AMR emergence, should be stressed globally.

Lay Summary:

The creation and manufacturing of antibiotics in the 20th Century has enabled us to combat and control countless diseases, such as smallpox and tuberculosis, which were once rampant throughout society, infecting and killing many. However, what now poses a threat to this is antimicrobial resistance (AMR), which occurs when the microbes that cause these diseases evolve to resist the drugs designed to kill them, leading to treatments failing and infections persisting, increasing the amount of people that become sick with and die from diseases which were once controlled. AMR can arise by antibiotic overuse and misuse, such as patients not completing their course or using them to treat viral infections.

This project considered the relationship AMR has with bacteriophages, which are viruses which infect and replicate only in bacterial cells. It was suggested that bacteriophage presence within the genetic makeup of the bacteria they infect has a role in making that bacteria resistant to a particular antibiotic. We studied how they may influence the resistance of the bacteria *Escherichia coli* to the antibiotic Amoxicillin. Using advanced computational methods, we examined the DNA of 272 *E. coli* samples to unravel how their genes relate to resistance against the antibiotic, using algorithms to identify patterns to predict which bacteria would be resistant. Our algorithms were incredibly accurate in identifying the genes and bacteriophage communities linked to AMR, showing high accuracy across three groups of data, measured as 0.934, 0.886, and 0.926.

We discovered 10 genes, and 17 communities of bacteriophages, that played a role in making bacteria either resistant or susceptible to the antibiotic. We found genes that were directly involved in making bacteria resistant against Amoxicillin, such as '*bla-TEM1B'*, and other genes, such as '*aac(3)-ila*', which interestingly, combat different drugs. The fact that our algorithms highlighted this gene as important however suggests that bacteria may have a more complex defence strategy than we thought, where many different factors are involved in shielding them from the drugs designed to kill them.

Our study faced a challenge; the bacteriophage communities were labelled in such a way that did not tell us anything about the genetic makeup of them. Although our algorithms showed us that these communities contribute to antimicrobial resistance in bacteria, the lack of information about their genetic makeups made it difficult to pinpoint just how they do this. This challenge led to the decision to examine the DNA of these communities, to reveal any present genes which may cause antimicrobial resistance. We found several genes in these communities, such as '*blaTEM-1B' and 'blaCTX-M-15'* genes which have been shown to directly cause resistance against antibiotics like Amoxicillin. Not only did this support the theory that bacteriophages play a role in causing AMR, it showed us that these viruses hold the building blocks of resistance themselves.

Our research demonstrated that there is a strong relationship between bacteriophage presence in the genetic makeup of bacteria, and the ability of bacteria to become resistant to the antibiotics designed to kill them. Going forward, we hope that this work can be built upon, looking at a wide range of bacteria and antibiotics, in hopes that this relationship is consistent across the spectrum. Research should be dedicated to understanding how to reduce the spread of bacteriophage enabled AMR, and the responsible use of antibiotics should be encouraged.

Research Article:

1. Introduction:

The 1928 discovery of penicillin as the first true antibiotic by Alexander Fleming (Fleming, 1929) marked arguably the start of the modern age of medicine. The subsequent isolation and mass production of the antibacterial by the Allied militaries over a decade later paved the way for unprecedented advancements in healthcare (Rao, 1944). Antibacterials contributed to the control of many infectious diseases that were once the leading cause of human morbidity and mortality for most of human existence (CDC, 2023). The average life expectancy at birth before the 20th Century was 47 years, while infectious diseases such as plague, smallpox and tuberculosis were rampant (CDC, 2023). However, the golden age of antibiotics witnessed the discovery of new antibiotics that would treat and cure diseases that were once deemed incurable, increasing the average life expectancy in the United States to 78.8 years (CDC, 2023).

Later in the 20th century saw the rise of antimicrobial resistance (AMR), defined as the ability of a microorganism to become able to adapt and grow in the presence of medications that once inhibited their growth (Founou et al., 2017). The phenomenon has existed for as long as antimicrobials themselves, however the accelerated development and spread of AMR during this time was attributed to a range of factors. The excitement around the chemotherapeutic power of antibiotics caused for their extensive use, across not just human medical landscapes, but agriculture and animal husbandry also (Manyi-Loh et al., 2018). This was paired with a lack of awareness and, in turn, inappropriate use of the drugs. Misuses included the treatment of viral infections and patients not completing treatment courses – practises considered by the World Health Organisation as the main drivers of resistance (WHO, 2024). This was parallel to a decline in new antibiotic development, as easier targets of intervention had already been exploited, and processes for antibiotic approval became increasingly complex and costly

(Ardal et al., 2019), discouraging new development efforts (Ardal et al., 2019). Today, the general population remains unaware of the effects of antibiotic overuse, and there remains a lack of new antibiotic discoveries (Ardal et al., 2019).

The World Health Organisation (WHO, 2023) stresses AMR as one of the greatest threats to global health. Current estimates state that 1.27 million global deaths occur annually that are attributed to AMR (Murray et al., 2022), with projections rising to 10 million by 2050 if current trends persist (WHO, 2023). It has a significant economic impact, potentially costing the global economy up to $100 trillion by 2050, increasing hospital stays and treatment costs (Cueni, 2024). The phenomenon fundamentally undermines the decades of medical advancements that have shaped the world today as we know it. These advancements, including transplant medicine and cancer treatments, have significantly increased life expectancies and reduced mortality rates of the global population (Singh, 2010). AMR poses a significant threat to these marvels of modern medicine. If current attitudes towards antibiotic use persist, it may be the case in the future that simple surgeries and therapies become compromised by the presence of once curable infections (Jasovský et al., 2016).

Reygaert et al. (2018) describes mechanisms of resistance as intrinsic or acquired. Bacteria with intrinsic resistance have natural immunity, possessing genes that make resistance an inherent part of their genetic makeup. Whereas those which acquire resistance do so through the spontaneous mutagenesis of their DNA, potentially leading to changes in the target site of the antimicrobial, reducing its effectiveness. Another mechanism of acquired resistance is horizontal gene transfer (HGT), in that microbes acquire resistance genes from neighbouring organisms through transformation – the uptake of DNA from the environment, transduction – DNA transfer by bacteriophages, and conjugation – direct DNA transfer between bacteria through a pilus (Reygaert, 2018). Molecularly, the main mechanisms of resistance include drug uptake limitation, modification, inactivation, and active efflux (Reygaert, 2018).

AMR can be spread among bacterial populations through several mechanisms, such as HGT, as previously discussed. Mobile genetic elements, such as plasmids and transposons, which antimicrobial resistance genes (AMGs) often reside on, can move throughout the chromosome of the bacteria or be transferred to other bacteria through HGT (da Silva et al., 2022). AMR can also be clonally spread. Once resistance has been acquired, a bacterium can reproduce to create a clonal population of resistant bacteria, which in the presence of an antibiotic is more dominant and naturally selected to survive (Silva et al., 2023).

This research project will consider the relationship antimicrobial resistance has with bacteriophages - viruses which infect and replicate only in bacterial cells (Simmonds, Aiewsakun, 2018). Recognised as the most abundant biological agent on earth, they exhibit extreme diversity in size, morphology, and genomic organisation (Hatful, Hendrix, 2011) and are extremely species specific (Kassman, Porter, 2022).

Bacteriophages replicate within their bacterial hosts via either the lytic or lysogenic cycles, both involving the introduction of phage genomes to the cytoplasm of the bacterium (Kassman, Porter, 2022). During the lytic cycle, the phage utilises host ribosomes to manufacture its proteins, creating multiple copies of itself. The host bacterium then dies, after which new bacteriophages are released which infect other susceptible bacteria (Kassman, Porter, 2022). During the lysogenic cycle, the phage genome is integrated into the chromosome of its host or remains as an episomal element. The phage is then replicated and passed to daughter cells through vertical gene transfer. These phages are known as temperate and remain latent to ensure survival and propagation of phage genetic material through successive bacterial generations (Kassman, Porter, 2022). They may convert to the lytic replication cycle, often in response to environmental conditions, such as ultraviolet light exposure and nutrient limitation (Kassman, Porter, 2022).

Chiang et al., (2019) explains that bacteriophages are directly involved in HGT through transduction, which is categorised as general or specialised. Lytic phages may undergo generalised transduction, accidentally packaging random fragments of bacterial DNA during the assembly of new phage particles. Temperate phages undergo specialised transduction, carrying specific bacterial genes near their integration site during excision from the bacterial chromosome. In both cases, the bacterial genetic material may contain AMGs (Chiang et al., 2019). Lytic phages may contribute to the spread of AMR in bacterial populations through selective pressure as they lyse susceptible bacteria, creating a niche where resistant strains can thrive and proliferate (Chiang et al., 2019). Lysogenic phages may carry AMGs themselves, which become integrated into the bacterial chromosome, or their integration into the bacterial chromosome may activate AMGs that would have otherwise lay dormant. In both cases, resistance is conferred down generations (Chiang et al., 2019).

AMR gene identification tools such as StarAMR have revolutionized the identification of AMGs within bacterial genomes, pinpointing potential markers of antimicrobial resistance (Bharat et al., 2022). However, such tools cannot directly correlate these genes with AMR phenotypes. It is crucial to recognise that the mere presence of AMGs within a bacterium does not guarantee an AMR phenotype. Instead, the expression of resistance often results from complex interactions between multiple genes, such as those between AMGs and non-AMG accessory genes (Dillon et al., 2024). The use of AMR gene identification tools to predict AMR phenotypes oversimplifies the mechanisms underpinning AMR, indicating that a single gene is solely responsible for a resistance phenotype (Dillon et al., 2024).

Bridging the identification and functional analysis gap, machine learning algorithms such as those employed by ‘Weka’ delve deeper, aiding researchers in their efforts to understand the mechanisms that underly the emergence of AMR phenotypes. These algorithms specialise in identifying patterns and correlate them directly with resistance phenotypes (Noman et al., 2023). Among these algorithms is the J48 decision tree, which offers a user friendly,

hierarchical structure that mimics the human decision-making process (Yasir et al., 2022). Several studies highlight the increased accuracy of machine learning tools, such as the J48 decision tree, in predicting AMR phenotypes compared to AMR gene identification tools. Dillon et al. (2024) found the average prediction accuracy of AMR gene identification tools in inferring AMR phenotypes as 57.6%, while the J48 decision tree model had an average prediction accuracy of 91.1% (Dillon et al., 2024). This accuracy increase highlights how machine learning applications are instrumental in pinpointing pathways of resistance. Leveraging their capabilities has proven crucial in advancing our understanding of AMR dynamics. Such understandings can aid in facilitating the design of novel diagnostic and theraupticc tools that contribute to combatting the increasingly alarming AMR crisis (Khaledi et al., 2020).

The bacterium *Escherichia coli* is the most widely studied organism to date and was instrumental in developing many fundamental concepts in biology (Ruiz, Silhavy, 2022). Naturally present in the intestines of warm-blooded mammals, it is ubiquitous and easily accessible for research (Smati et al., 2015). Its extensive research history provides abundant, readily accessible data (Basavaraju, Gunashree, 2023), and so was chosen as the focal microorganism of this project.

Amoxicillin is a broad spectrum, penicillin class β-lactam antibiotic. Its method of action involves inhibiting bacterial cell walls synthesis, causing cell lysis and death (Bhattacharjee, 2016). A bacterium may become resistant to the antibiotic by harbouring β-lactam genes, which through transcription and translation produce β-lactamase enzymes. These genes hydrolyse the antibiotic's β-lactam ring, preventing the disruption of cell wall synthesis, promoting the survival of the bacterium in the presence of the antibiotic (Mora-Ochomogo et al., 2021). The drug was chosen as the focal antimicrobial of the project due to reasons similar to *E. coli*.

2. Aims and Hypothesis:

Project Aim:

- To determine if bacteriophage presence influences the antimicrobial resistance phenotype of their host *E. coli* bacterium against Amoxicillin.

Hypothesis:

- There was a relationship between the bacteriophage presence in *Escherichia coli* and its phenotype of resistance against Amoxicillin.

It was predicted if a bacteriophage was present, its host *E. coli* bacterium phenotype was resistant to Amoxicillin, and if a bacteriophage was not present, its host bacterium phenotype was susceptible to Amoxicillin. It was found that upon analysis of some 272 genomes that bacteriophages were present within the genomes of bacteria susceptible to Amoxicillin, while bacteriophages were not present within the genomes resistant to Amoxicillin. However, it was expected that there would be a statistically significant increase in resistance against Amoxicillin where bacteriophages were present, as there would be an increase in susceptibility with bacteriophage absence.

Once the aim of establishing the relationship between bacteriophage presence and the resistance phenotype of *E. coli* against Amoxicillin had been achieved, the project delved deeper, aiming to identify specific AMGs present in prophage sequences that may have been responsible for such resistance. This moved beyond the simplistic association of resistance with phage presence, instead aiming to pinpoint specific markers of resistance within phages themselves. This allowed for a greater understanding of the mechanisms that underly phage enabled AMR emergence, holding the potential to refine our understanding of AMR mechanisms, and may guide the discovery of novel interventions that disrupt these mechanisms at their source (Kondo et al., 2020).

3. Methods:

GitHub Repository: https://github.com/cwoods2001/AMR-Scripts-Data.

This repository contains the necessary scripts, reference datasets, and other relevant information required to review, replicate, or build upon the works of this research project.

Some 272 *E. coli* genomes were selected against Amoxicillin and downloaded as FASTA files from the Bacterial and Viral Bioinformatics Research Centre (BVBRC). The genome database was filtered to select genomes with a broth dilution laboratory typing method (BVBRC, 2024). Each genome corresponded with a Minimum Inhibitory Concentration value, defined as the lowest concentration of an antimicrobial agent that inhibits the growth of a given strain of bacteria (Andrews, 2001), which was then matched against Clinical Breakpoints set by the European Committee on Antimicrobial Susceptibility Testing. In this case, the Clinical Breakpoint was 8 micrograms per litre (EUCAST, 2024).

The genome FASTA files were then uploaded to StarAMR, a tool which scans genomes against ResFinder, PlasmidFinder, and PointFinder databases searching for AMGs (Galaxy, 2024). The tool was run with the 'Type' of AMG set to auto-detect and the 'Reference' information selected as unspecified (Galaxy, 2024). Upon completion, a csv file was produced containing information regarding the presence of AMGs within each bacterial genome, and the antibiotics the tool had predicted each genome was resistant to.

The genomes were then uploaded to a geNomad workflow that identifies virus and plasmid sequences in assembled scaffolds and estimates the quality of viral genomes (NERSC, 2024). The output file of provirus DNA sequences was then processed using a series of Unix shell commands and Python scripts. The main steps included splitting FASTA files based on provirus IDs, creating Sourmash sketches for DNA sequences, executing external Python scrips to generate edge tables and identify elbow points, and using R for additional analysis.

A step-by-step guide on processing the output file can be found in the GitHub repository dedicated to this project.

The two files, one containing information regarding AMGs present within the bacterial isolates, and one containing provirus DNA sequences, were then processed using Python scripts. Steps were taken involving data loading, cleaning, aggregation, and transformation to generate two files - the frequency of AMGs in each bacterial isolate, and the frequency of phage communities in each bacterial isolate. A comprehensive understanding of these procedures, including the scripts used, is available in the GitHub repository.

The files were then combined to create one unified dataset, with features such as AMGs and prophage communities aligned with their corresponding Isolate IDs. This dataset and the steps taken for its creation are available in the GitHub repository, for reference and reproducibility needs. This dataset was then used to make three ARFF files, one for the AMG data, the prophage data, and the combined data. The AMGs and prophage communities were noted as attributes, and a class attribute indicating AMR status was designated (Gonzalez-Lopez et al., 2019) based on the comparison to EUCAST. The exact ARFF files are available in the GitHub repository for reference.

The files were then loaded into Weka (version 3.8.6). Each ARFF file was uploaded in ‘Explorer’ and pre-processed before the ‘Classify’ tool was selected to build the models. The J48 Decision Tree was selected with a Cross Validation of 10 Folds (Witten, Frank, 2005). Once the models were built, key figures were noted, such as the True Positive Rate, measuring how well a model identifies positive outcomes from datasets (Monaghan et al., 2021), Cohen’s Kappa coefficient, a measure of reliability for two raters that are rating the same thing, correcting for how often the raters may agree by chance (McHugh, 2012), and the F measure, which evaluates a model’s accuracy (Hand et al., 2021).

The prophage communities FASTA file, obtained from the geNomad workflow, was then uploaded to StarAMR, with the ‘Type’ and ‘Reference’ options selected as before. The output csv file was then analysed to identify AMGs present within the viral genomes. Found genes were researched to draw conclusions on how they may confer the resistance phenotypes against Amoxicillin in *E. coli* isolates.

4. Results:

Displayed below are three J48 Decision Trees labelled as Bacterial Genomes, Prophage Communities, and Combined Data. The trees are available as higher resolution images in the GitHub repository. Additionally, there is a bar chart displaying the performance metrics across the datasets.

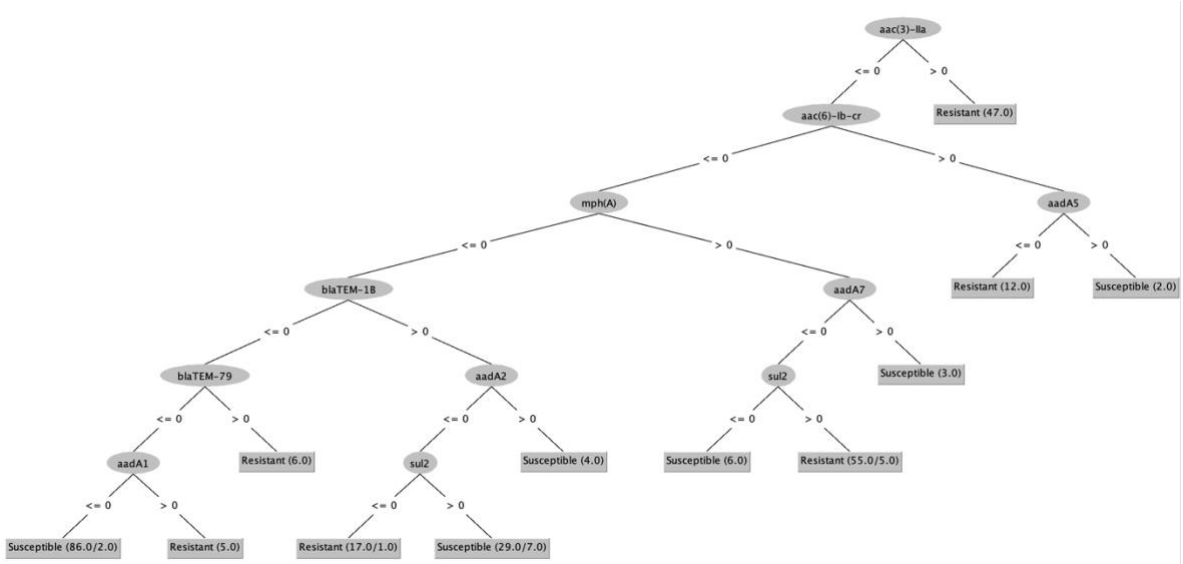

Figure 1 (Above): The Bacterial Genomes Decision Tree

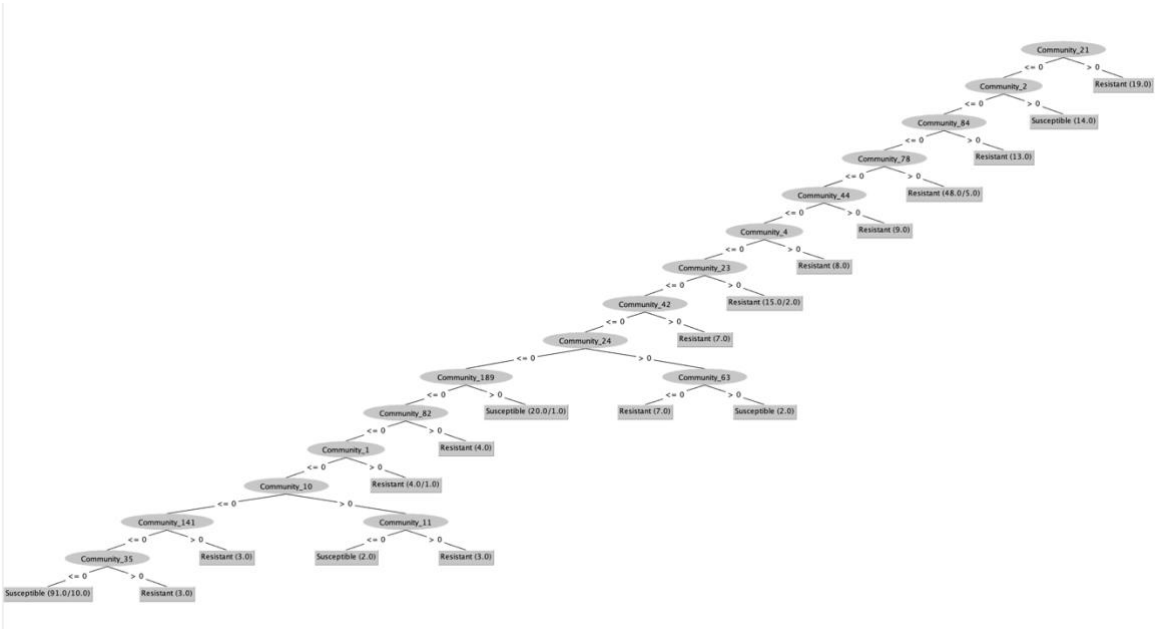

Figure 2 (Above): The Prophage Communities Decision Tree

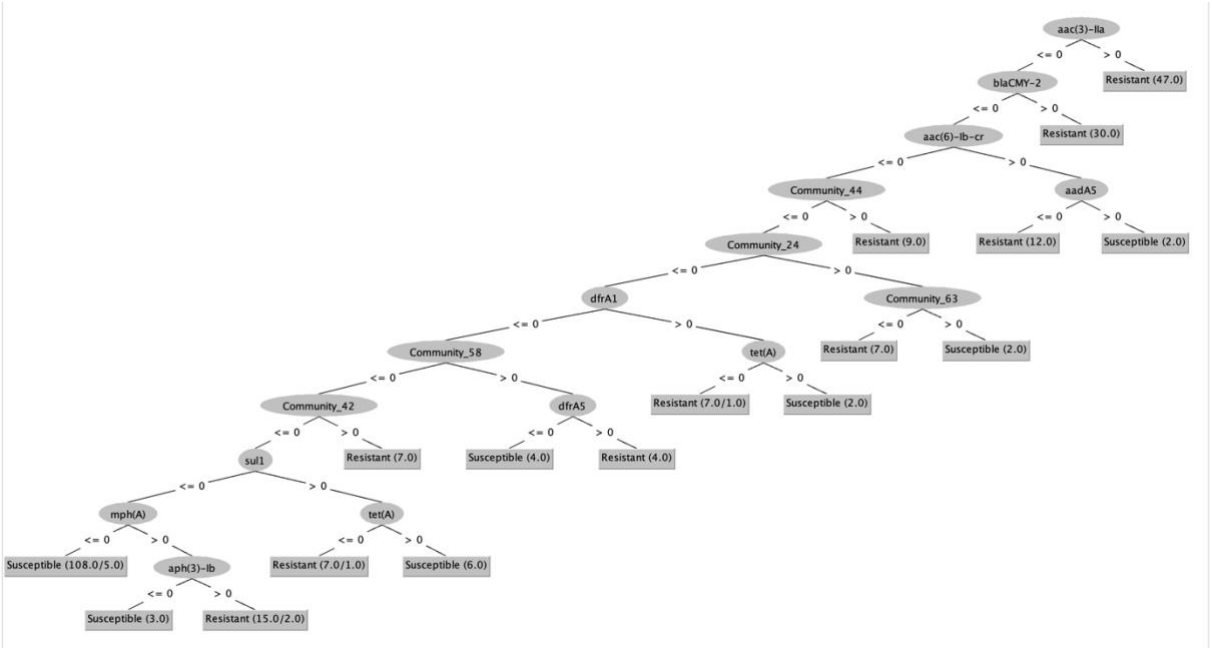

Figure 3 (Above): The Combined Data Decision Tree

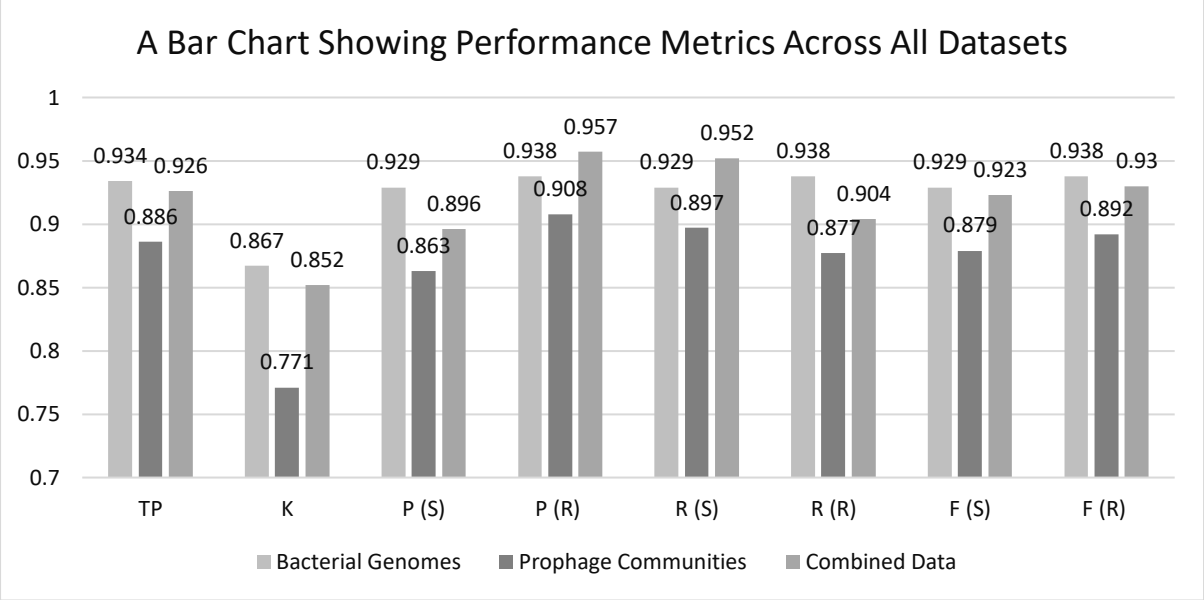

Figure 4 (Above): A Bar Chart Showing Performance Metrics Across All Datasets, Where TP: True Positive; K: Cohen's Kappa Statistic; P(S): Precision Susceptible; P(R): Precision Resistant; R(S): Recall Susceptible; R(R): Recall Resistant; F(S): F Measure Susceptible; F(R): F Measure Resistant

4.1 Decision Trees Analysis:

Analysis of the decision trees allowed for an understanding of the factors influencing AMR in *E. coli* to Amoxicillin. The trees revealed various pathways leading to both susceptible and resistance phenotypes.

In the Bacterial Genomes decision tree, 10 genes were highlighted. The presence or absence of these genes lead to 13 routes of resistance and 5 of susceptibility. Major routes that the genomes conformed to include the presence of '*aac(3)-ila*', wherein resistance was conferred in 47 genomes. Other routes include the absence of the genes '*mph(a)*' and '*aad7*' with the presence of '*sul2*', conferring resistance in 55 genomes correctly and 5 genomes incorrectly. Where '*blaTEM-18*' was present and '*aadA2'* nor '*sul2'* were not, 17 genomes were correctly labelled resistant and 1 genome incorrectly. Where no AMGs were present, 86 genomes were correctly resistant and 2 incorrectly.

In the Prophage Communities decision tree, 17 communities were significant. The presence or absence of these communities lead to 13 routes of resistance and 5 of susceptibility. Where no communities were present, 91 genomes were correctly labelled susceptible and 10 genomes incorrectly. Elsewhere, the model correctly conferred resistance in 48 genomes and incorrectly in 5 where Community 78 was present. The model also highlighted how communities can interact to confer phenotypes, seen in Community 24 presence without Community 63, where 7 genomes were resistant. Where no prophage communities were present, 91 genomes were identified as susceptible correctly and 10 incorrectly.

In the Combined Data decision tree, 11 AMGs and 5 Communities were significant. The presence or absence of these lead to 10 routes of resistance and 7 of susceptibility. Where no AMGs or communities were present, 108 genomes were correctly susceptible and 5

incorrectly. 1 route led to resistance involving the presence of both AMGs and prophage communities – the gene ‘drfA5’ and Community 58, conferring resistance in 4 genomes.

4.2 Performance Metrics Analysis:

Performance metric analysis allowed for an evaluation the model’s efficacy across the datasets, revealing distinct differences and notable implications for each model’s ability to predict AMR.

The True Positive Rate (TP) demonstrated an outperformance in the Bacterial Genomes and Combined Data datasets, 0.934 and 0.926 respectively against the Prophage Communities dataset, 0.886. The latter however exhibited a measure that remained commendably high.

For Cohen’s Kappa Statistic (K), Bacterial Genomes lead with a Kappa of 0.867, followed by the Combined Data figure of 0.852. The Kappa for Prophage Communities was notably lower at 0.771, signalling that the model’s predictions for this dataset were less reliable, although still commendable.

Precision (P) is a measure of the quality of positive predictions made by the model (Huilgol, 2023). The precision for predicting resistant strains was higher than susceptible strains across all three datasets, with P(R) highest within the Combined Dataset at 0.957, followed by 0.938 and 0.908 for Bacterial Genomes and Prophage Communities respectively.

Recall is the ratio of true positives to the sum of true positives and false negatives, measuring the model’s ability to detect all actual positives (Huilgol, 2023). Recall was lowest for resistant strains of Prophage Communities, 0.877. It was highest for the susceptible strains of the Combined Dataset, 0.952.

The F Measure (F) balances precision and recall (Hand et al., 2021). The Prophage Communities dataset scored the lowest F measures for both susceptible (F (S) 0.879) and resistant strains (F (R) 0.892). The Bacterial Genomes dataset scored the highest for both susceptible (F (S) 0.929) and resistant strains (F (R) 0.938).

5. Discussion:

5.1 Performance Metrics Discussion:

The performance metrics illustrated that all three models were able to predict AMR to a high degree. However, those applied to the Bacterial Genomes and Combined Data datasets were more balanced and effective in their predictions, higher across all metrics. The Prophage Communities metrics however remained notably high, with the disparity in performance being attributed to several possibilities. The dynamics surrounding prophage-bacterial interactions, prophage sequence variability, and their influence on bacterial phenotypes contribute to the complexity of accurately predicting AMR within this dataset (Li et al., 2021). This led to the decision to upload the data to StarAMR to uncover AMGs present within the viral genomes.

5.2 Decision Trees Discussion

As previously highlighted, the presence of the gene '*aac(3)-ila'* indicated a strong association with Amoxicillin resistance. The gene encodes enzymes known as aminoglycoside acetyltransferases, encoded by plasmids, transposons, and integrons in a range of bacteria, including *E. coli* (McArthur, 2024). These enzymes confer resistance to aminoglycoside antibiotics by acetylating them, causing inactivation and the inability to bind to bacterial ribosomes (McArthur, 2024). However, it must be noted that Amoxicillin is a β-lactam antibiotic (Bhattacharjee, 2016), and not an aminoglycoside. The presence of this gene regarding its association with Amoxicillin resistance is indicative of complex interactions between genes and multiple resistance mechanisms in the genome. Its presence may act as a marker for genomes that also possess β-lactamase genes or other resistance mechanisms effective against Amoxicillin. This finding aligns with the research of Altayb et al. (2022), which investigates genomic co-occurrence of β-lactam and aminoglycoside resistance determinants,

stressing resistance as a polygenic phenomenon, suggesting that a gene conferring resistance to one antibiotic class may be indicative of a multi-drug resistance phenotype.

The coexistence of '*blaTEM-1B'* and '*sul2'*, without '*aad2'*, was associated with susceptibility, while the presence of '*blaTEM-1B'* alone indicated resistance. '*blaTEM-1B'* is a member of the *blaTEM* gene family that encodes β-lactamase enzymes, providing resistance to β-lactam antibiotics, including Amoxicillin, by hydrolysing the antibiotic's β-lactam ring (Mora-Ochomogo et al., 2021). '*aadA2'* encodes aminoglycoside adenylyl transferases (Cameron et al., 2018), while '*sul2'* provides resistance to sulphonamides (Venkatesan et al., 2023). The co-existence of '*blaTEM-1B'* and '*sul2'* predominantly leading to susceptibility makes it initially intuitive to argue that the presence of '*sul2'* suppresses the ability of '*blaTEM-1B'* to confer resistance. However, both genes have distinct mechanisms of actions, making the case of a causative suppression interaction unlikely. Instead, the pattern could be indicative of complex genomics within these isolates, wherein the genetic backgrounds of isolates containing both genes includes other factors that influence susceptibility to Amoxicillin, such as regulatory elements affecting gene expression (López-Siles et al., 2022). The presence of '*sul2'* may indicate the presence of specific characteristics that mitigate the effect of '*blaTEM-1B'*, such as the inherent propensity for certain resistance profiles, a phenomenon supported by the findings of Ramamurthy et al. (2022) and Arabi et al. (2015). These patterns highlight the need for further experimental investigation to illuminate the underlying mechanisms driving phenotype manifestation.

The decision tree models for the Prophage Communities and Combined Data datasets posed a challenge for in depth analysis. The prophage communities were labelled in a way which did not provide specific identifiers that linked them to known prophage sequences. The models successfully demonstrated that there is a level of interaction at play both within prophage communities and between prophage communities and bacterial genomes to confer

phenotypes, however the lack of knowledge regarding the specific genetic compositions and functions of these communities limited the ability to precisely determine how they contribute to AMR. This challenge, along with the performance metrics of the Prophage Communities dataset, supported the reasoning for uploading the data to StarAMR.

5.3 StarAMR Analysis of Prophage Sequences Discussion:

StarAMR analysis of the Prophage Communities dataset highlighted several AMGs present within the communities – *dfrA1*, *aadA1*, *blaTEM-1B* and *blaCTX-M-15*. The genes '*dfrA1'* and '*aadA1'* do not directly confer resistance to β-lactam antibiotics such as Amoxicillin, but instead confer resistance to trimethoprim and aminoglycoside antibiotics respectively (McArthur, 2024). However, their presence alongside β-lactam genes is indicative of broad-spectrum resistance profiles throughout the prophage communities present within the isolates, supporting the theory that phages harbour AMGs which confer resistance to a wide range of antibiotics.

The identification of the β-lactamase genes '*blaTEM-1B'* and '*blaCTX-M-15'* (Negrete-González et al., 2022) in the prophage sequences is a pivotal finding of this project. It not only supports the fact that bacteriophage presence in *E. coli* isolates is a driving force of resistance against Amoxicillin, it pinpoints markers of resistance found within the phages themselves. This expands upon the existing body of research that has increasingly recognised bacteriophages as significant contributors to the horizontal gene transfer mechanisms that facilitate the spread of AMR across bacterial populations. Studies by Modi et al. (2013) and Colavecchio et al. (2017) discuss the role of phages in mediating AMG transfer, but the specific identification of these AMGs within prophage sequences emphasises phages as a direct vector for β-lactam resistance – phages are not simply passive enablers of resistance, but active agents in the distribution of resistance determinants.

Understanding that prophages are direct vectors for the spread of AMGs highlights the needs for surveillance and therapeutic strategies that target phage invasion and mediated gene transfer. Surveillance methods may include the implementation of widespread programmes that monitor the presence and spread of prophages carrying AMR genes in clinical samples (Kondo et al., 2020). Future research could involve the development of bacterial strains that are resistant to phage infection and integration (Zou et al., 2022). Furthermore, phage therapy, treatment that utilises the parasitic nature of bacteriophages to eliminate pathogenic bacteria and combat bacterial infections (Kakkar et al., 2023), could be designed with bacteriophages engineered to lack the ability to carry or transfer AMGs.

5.4 Project Strengths and Weaknesses:

A key strength of this study is the utilization of bioinformatics and machine learning to highlight the relationship between AMR phenotypes and bacteriophage presence. This approach provided novel insights into phage-mediated AMR mechanisms. However, a limitation is the challenge of linking specific prophage communities to known sequences. This limits the precision of conclusions that can be drawn regarding their role in AMR. Overcoming this limitation can significantly enhance our understanding of AMR. Future directions could include the utilization of enhanced bioinformatic tools capable of more accurately identifying prophage sequences within bacterial genomes. One such tool is ‘PROPHETESS’, recently designed for the prediction of prophage loci in bacterial genomes (Nair, Amudha, 2022). Elsewhere, focusing on a singular bacterial species and antimicrobial limits the applicability of findings to other bacteria and antibiotics (Wen et al., 2016). Future studies should address this by expanding the range of bacterial species and antimicrobial agents analysed.

5.5 Conclusion:

The findings of this study highlight the critical role bacteriophages play in conferring antimicrobial resistance among bacterial populations. However, to fully understand the scope and implications of these mechanisms, further, research is crucial. Future efforts should encompass a much wider range of bacterial species, and antibiotics of differing classes. Broadening the scope of research will provide greater insights into the mechanisms of bacteriophage mediated AMR transmission and will aid in efforts to combat the increasing threat of antimicrobial resistance to ensure the protection of global health against the phenomenon.

Going forward, facing this escalating challenge requires a united effort across a multitude of disciplines. Educating healthcare professionals and the public on the importance of antibiotic stewardship is vital, something than can be facilitated by the development and enforcement of regulations that limit antibiotic overuse and misuse across all sectors. AMR is a global issue, and so such policies should be implemented on a global scale. Research initiatives should take a collaborative approach across molecular biology, genetics, bioinformatics and beyond, establishing international and interdisciplinary alliance to develop novel diagnostic and therapeutic strategies, such as those which target phage infection and integration, to mitigate the increasingly alarming crisis that is antimicrobial resistance.

Bibliography:

1. Altayb, H.N. *et al.* (2022) 'Co-occurrence of β-lactam and aminoglycoside resistance determinants among clinical and environmental isolates of Klebsiella pneumoniae and escherichia coli: A genomic approach', *Pharmaceuticals*, 15(8), p. 1011. doi:10.3390/ph15081011.
2. Andrew G. McArthur, G.D.W. (no date a) *AAC(3)-IA, The Comprehensive Antibiotic Resistance Database*. Available at: https://card.mcmaster.ca/ontology/38928 (Accessed: 11 March 2024).
3. Andrew G. McArthur, G.D.W. (no date b) *AADA, The Comprehensive Antibiotic Resistance Database*. Available at: https://card.mcmaster.ca/ontology/39001 (Accessed: 11 March 2024).
4. Andrew G. McArthur, G.D.W. (no date c) *DFRA1, The Comprehensive Antibiotic Resistance Database*. Available at: https://card.mcmaster.ca/ontology/39288 (Accessed: 11 March 2024).
5. Andrews, J.M. (2001) 'Determination of minimum inhibitory concentrations', *Journal of Antimicrobial Chemotherapy*, 48(suppl_1), pp. 5–16. doi:10.1093/jac/48.suppl_1.5.
6. Arabi, H. *et al.* (2015) 'Sulfonamide resistance genes (SUL) m in extended spectrum beta lactamase (ESBL) and non-esbl producing escherichia coli isolated from Iranian hospitals', *Jundishapur Journal of Microbiology*, 8(7). doi:10.5812/jjm.19961v2.
7. *Arff stable* (no date) *Arff stable - Weka Wiki*. Available at: https://waikato.github.io/weka-wiki/formats_and_processing/arff_stable/ (Accessed: 04 April 2024).
8. *Bacterial and Viral Bioinformatics Resource Center*. Available at: https://www.bv-brc.org/view/Taxonomy/562#view_tab=amr&filter=and(or(eq(antibiotic,%22amoxici

llin%22),eq(antibiotic,%22amoxicillin%2Fclavulanic%20acid%22)),eq(laboratory_typing_method,%22Broth%20dilution%22)) (Accessed: 04 April 2024).

9. Basavaraju, M. and Gunashree, B.S. (2023) '*escherichia coli*: An overview of main characteristics', *Escherichia coli - Old and New Insights* [Preprint]. doi:10.5772/intechopen.105508.
10. Bhattacharjee, M.K. (2016) 'Antibiotics that inhibit cell wall synthesis', *Chemistry of Antibiotics and Related Drugs*, pp. 49–94. doi:10.1007/978-3-319-40746-3_3.
11. C Reygaert, W. (2018) 'An overview of the antimicrobial resistance mechanisms of bacteria', *AIMS Microbiology*, 4(3), pp. 482–501. doi:10.3934/microbiol.2018.3.482.
12. Cameron, A. *et al.* (2018) 'A novel *aada* aminoglycoside resistance gene in bovine and porcine pathogens', *mSphere*, 3(1). doi:10.1128/msphere.00568-17.
13. Colavecchio, A. *et al.* (2017) 'Complete genome sequences of two phage-like plasmids carrying the CTX-M-15 extended-spectrum β-lactamase gene', *Genome Announcements*, 5(19). doi:10.1128/genomea.00102-17.
14. da Silva, G.C. *et al.* (2022) 'Mobile genetic elements drive antimicrobial resistance gene spread in Pasteurellaceae species', *Frontiers in Microbiology*, 12. doi:10.3389/fmicb.2021.773284.
15. (*Data Mining: Practical Machine Learning Tools ... - academia.dk*. Available at: http://academia.dk/BiologiskAntropologi/Epidemiologi/DataMining/Witten_and_Frank_DataMining_Weka_2nd_Ed_2005.pdf (Accessed: 04 April 2024).
16. Dillon, L., Dimonaco, N.J. and Creevey, C.J. (2024) 'Accessory genes define species-specific routes to antibiotic resistance', *Life Science Alliance*, 7(4). doi:10.26508/lsa.202302420.

17. ESCMID - European Society of Clinical Microbiology and Infectious Diseases 2008 (no date) *Clinical breakpoints - breakpoints and guidance, EUCAST*. Available at: https://www.eucast.org/clinical_breakpoints (Accessed: 07 March 2024).
18. *Galaxy: Tool shed* (no date) *Galaxy | Tool Shed*. Available at: https://toolshed.g2.bx.psu.edu/repository?repository_id=b9822587a6ad2cbf (Accessed: 07 March 2024).
19. General, T. Cueni Director, General info@ifpma.org, A.T.C.D. and General info@ifpma.org, T. Cueni Director (2018) *By 2050, superbugs may cost the economy $100 trillion*, *IFPMA*. Available at: https://www.ifpma.org/insights/by-2050-superbugs-may-cost-the-economy-100-trillion/ (Accessed: 28 February 2024).
20. *Genomad#* (no date) *geNomad*. Available at: https://portal.nersc.gov/genomad/ (Accessed: 07 March 2024).
21. Gonzalez-Lopez, J., Ventura, S. and Cano, A. (2019) 'ARFF Data Source Library for distributed single/multiple instance, single/multiple output learning on apache spark', *Lecture Notes in Computer Science*, pp. 173–179. doi:10.1007/978-3-030-22744-9_13.
22. Hand, D.J., Christen, P. and Kirielle, N. (2021) 'F*: An interpretable transformation of the F-measure', *Machine Learning*, 110(3), pp. 451–456. doi:10.1007/s10994-021-05964-1.
23. Huilgol, P. (2023) *Precision and recall: Essential metrics for machine learning (2024 update), Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/ (Accessed: 11 March 2024).

24. Jasovský, D. *et al.* (2016) ‘Antimicrobial resistance—a threat to the world’s sustainable development’, *Upsala Journal of Medical Sciences*, 121(3), pp. 159–164. doi:10.1080/03009734.2016.1195900.

25. Kakkar, A. *et al.* (2023) *Engineered bacteriophages: A panacea against pathogenic and drug resistant bacteria* [Preprint]. doi:10.20944/preprints202305.0771.v1.

26. Khaledi, A. *et al.* (2020) ‘Predicting antimicrobial resistance in *pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics’, *EMBO Molecular Medicine*, 12(3). doi:10.15252/emmm.201910264.

27. Kondo, K., Kawano, M. and Sugai, M. (2020a) *Prophage elements function as reservoir for antibiotic resistance and virulence genes in nosocomial pathogens* [Preprint]. doi:10.1101/2020.11.24.397166.

28. Kondo, K., Kawano, M. and Sugai, M. (2020b) *Prophage elements function as reservoir for antibiotic resistance and virulence genes in nosocomial pathogens* [Preprint]. doi:10.1101/2020.11.24.397166.

29. Li, M. *et al.* (2021) ‘A deep learning-based method for identification of bacteriophage-host interaction’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5), pp. 1801–1810. doi:10.1109/tcbb.2020.3017386.

30. López-Siles, M., McConnell, M.J. and Martín-Galiano, A.J. (2022) ‘Identification of promoter region markers associated with altered expression of resistance-nodulation-division antibiotic efflux pumps in Acinetobacter baumannii’, *Frontiers in Microbiology*, 13. doi:10.3389/fmicb.2022.869208.

31. Manyi-Loh, C. *et al.* (2018) ‘Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential Public Health Implications’, *Molecules*, 23(4), p. 795. doi:10.3390/molecules23040795.

32. McHugh, M.L. (2012) ‘Interrater Reliability: The kappa statistic’, *Biochemia Medica*, pp. 276–282. doi:10.11613/bm.2012.031.
33. Modi, S.R. *et al.* (2013) ‘Antibiotic treatment expands the resistance reservoir and ecological network of the Phage Metagenome’, *Nature*, 499(7457), pp. 219–222. doi:10.1038/nature12212.
34. Monaghan, T.F. *et al.* (2021) ‘Foundational statistical principles in medical research: Sensitivity, specificity, positive predictive value, and negative predictive value’, *Medicina*, 57(5), p. 503. doi:10.3390/medicina57050503.
35. Mora-Ochomogo, M. and Lohans, C.T. (2021) ‘B-lactam antibiotic targets and resistance mechanisms: From covalent inhibitors to substrates’, *RSC Medicinal Chemistry*, 12(10), pp. 1623–1639. doi:10.1039/d1md00200g.
36. Nair, M.R. and Amudha, T. (2022) ‘Prophetess: A tool for prediction of prophage loci in bacterial genomes’, *ICT Analysis and Applications*, pp. 681–689. doi:10.1007/978-981-19-5224-1_68.
37. Negrete-González, C. *et al.* (2022) ‘Plasmid carrying *bla*CTX-m-15, *bla*per-1, and *bla*tem-1 genes in *citrobacter spp.*from Regional Hospital in Mexico’, *Infectious Diseases: Research and Treatment*, 15, p. 117863372110657. doi:10.1177/11786337211065750.
38. Noman, S.M. *et al.* (2023) ‘Machine learning techniques for antimicrobial resistance prediction of pseudomonas aeruginosa from whole genome sequence data’, *Computational Intelligence and Neuroscience*, 2023, pp. 1–11. doi:10.1155/2023/5236168.
39. Ramamurthy, T. *et al.* (2022) ‘Deciphering the genetic network and programmed regulation of antimicrobial resistance in bacterial pathogens’, *Frontiers in Cellular and Infection Microbiology*, 12. doi:10.3389/fcimb.2022.952491.

40. RAO, S.S. (1944) ‘Production of penicillin’, *Nature*, 154(3898), pp. 83–83. doi:10.1038/154083a0.
41. Silva, A. *et al.* (2023) ‘Antimicrobial resistance and clonal lineages of escherichia coli from food-producing animals’, *Antibiotics*, 12(6), p. 1061. doi:10.3390/antibiotics12061061.
42. Singh, A. (2010) ‘Modern Medicine: Towards Prevention, cure, well-being and longevity’, *Mens Sana Monographs*, 8(1), p. 17. doi:10.4103/0973-1229.58817.
43. Smati, M. *et al.* (2015) ‘Quantitative analysis of commensal *escherichia coli* populations reveals host-specific enterotypes at the intra-species level’, *MicrobiologyOpen*, 4(4), pp. 604–615. doi:10.1002/mbo3.266.
44. Stohr, J.J. *et al.* (2020) ‘Development of amoxicillin resistance in escherichia coli after exposure to remnants of a non-related phagemid-containing E. coli: An exploratory study’, *Antimicrobial Resistance & Infection Control*, 9(1). doi:10.1186/s13756-020-00708-7.
45. *Uniprot website fallback message* (no date) *UniProt*. Available at: https://www.uniprot.org/uniprotkb/A1YKW2/entry (Accessed: 11 March 2024).
46. Venkatesan, M. *et al.* (2023) ‘Molecular mechanism of plasmid-borne resistance to sulfonamide antibiotics’, *Nature Communications*, 14(1). doi:10.1038/s41467-023-39778-7.
47. Wen, X. *et al.* (2016) ‘Limitations of MIC as sole metric of pharmacodynamic response across the range of antimicrobial susceptibilities within a single bacterial species’, *Scientific Reports*, 6(1). doi:10.1038/srep37907.
48. Yasir, M. *et al.* (2022) ‘Application of decision-tree-based machine learning algorithms for prediction of antimicrobial resistance’, *Antibiotics*, 11(11), p. 1593. doi:10.3390/antibiotics11111593.

49. Zou, X. *et al.* (2022) ‘Systematic strategies for developing phage resistant escherichia coli strains’, *Nature Communications*, 13(1). doi:10.1038/s41467-022-31934-9.

50. Årdal, C. *et al.* (2019) ‘Antibiotic development — economic, regulatory and societal challenges’, *Nature Reviews Microbiology*, 18(5), pp. 267–274. doi:10.1038/s41579-019-0293-3.