University of Hertfordshire UH

Comparison of Deep Learning Techniques, with a Focus in the Prediction of the Bacteriophage PhiC2 Mediating Horizontal Gene Transfer of Tn6215 within *Clostridium difficile*.

By

Sarah-Jayne Byrne

18001814

4th September 2022

A Project supervised by:

Dr Ashley Spindler

A Thesis submitted in partial fulfilment
of the regulations governing the award of

MSc DATA SCIENCE

and in agreement with the School of Physics, Astronomy and Mathematics of the University of Hertfordshires' Declaration of Academic Integrity.

WORD COUNT: 10,976

DECLARATION

I declare that:

(a) All the work described in this report has been carried out by me – and all the results (including any survey findings, etc.) given herein were first obtained by me – except where I may have given due acknowledgement to others;

(b) All the prose in this report has been written by me in my own words, except where I may have given due acknowledgement to others and used quotation marks, and except also for occasional brief phrases of no special significance which may be taken from other people's work without such acknowledgement and use of quotation marks;

(c) All the figures and diagrams in this report have been devised and produced by me, except where I may have given due acknowledgement to others.

I understand that if I have not complied with the above statements, I may be deemed to have failed the project assessment, and/or I may have some other penalty imposed upon me by the Board of Examiners.

Signed: **Date: 02/09/2022**

Name: Sarah-Jayne Byrne **Programme code: 7PAM2002-0209-2021**

ABSTRACT

Clostridium difficile (C. difficile), a gram-positive species of spore forming obligate anaerobic bacterial pathogen, is one of the most prevalent causes of healthcare-associated infections worldwide. The emergence of antibiotic resistant *C. difficile* strains contributes to the pathogenesis and virulence of *C. difficile* infection. The mechanism of acquisition of antimicrobial resistant (AMR) genes is horizontal gene transfer (HGT) which is classified into transduction, conjugation, and transformation. A particular erythromycin resistant gene, *ermB*, found on the transposon Tn6215 is of particular interest to this research. Goh et al. (2013) first demonstrated the bacteriophage PhiC2 mediated transduction of Tn6215 in *C. difficile* strains. Therefore, this relationship has clinical significance because Tn6215 confers erythromycin resistance. Hence, this research applied five *C. difficile* strains, obtained from Sir Charles Gardiner Hospital in Australia, to identify the transposon Tn6215, containing the *ermB* gene conferring erythromycin resistance.

The method comprised of a 'gold standard' method applying PHASTER to identify PhiC2 regions and BLAST to identify Tn6215 regions. The PhiC2-Tn6215 pairings were visualised in Artemis and both quantitatively and qualitatively analysed. Additionally, the study applied a Convolution Neural Network (CNN) to predict and classify genomic regions that contained PhiC2 and Tn6215 using a binary classifier. This is because, deep learning models can automatically extract features from the input. However, the data was supplied as strings and this is not compatible with the input to a CNN. Therefore DNA augmentation methods; one hot, label and *k*-mer encoding, were evaluated to alter the data into numerical forms.

Results from the 'gold standard' method confirmed successful identification of PhiC2 and Tn6215 in all 5 *C. difficile* genome isolates. However, PhiC2-Tn6215 pairings were only qualitatively identified in 2 out of the 5 genome isolates (E185B and S8); therefore, the project does not agree with the hypothesis; implying Tn6215 transference is not dependent on PhiC2. Additionally, there was an absence of a positive correlation found between the amount of PhiC2 genome sequences and Tn6215 sequences ($r=0.587$, $p>0.05$). Interestingly, 60% of the bacteriophage genome regions identified by PHASTER were categorised as incomplete, cryptic bacteriophage, contained phage nucleotide bases statistically similar to PhiC2.

The results from the employed CNN architecture using the identified most appropriate DNA augmentation method; One Hot encoding for DNA sequence classification of both PhiC2 and Tn6215. The models were evaluated on different classification metrics. From the experimental results, show that imbalanced datasets affected feature identification of PhiC2 in training sets. The CNN model had an accuracy of 92.5% however, further evaluation matrices highlight the CNN was predicting every class as 0. Hence, the accuracy was high because the majority of the training class was 0. Therefore, SMOTE was applied to both the Tn6215 and PhiC2 classification datasets to increase the number of

minority classes. The CNN model appeared to determine patterns in the E185B training set for both Tn6215 and PhiC2, with both training sets applied achieving high accuracies 93.3% and 96.4%, respectively. Additionally, both models had high AUC 0.9663 and 0.9985, respectively. However, the trained CNN model unsuccessfully classified any of Tn6215 or PhiC2 regions in the unseen test data, *C. difficile* S8 genome. The corresponding AUCs also highlighted the model is an inappropriate fit and is not suitable for further prediction of Tn6215 or PhiC2 on unseen test data. Despite the CNN results, application of a CNN will be useful for future directions exploring automation of biological sequence identification.

As a concluding statement, all the aspects investigated in this research have further reinforced the requirement to understand both the pathological relationship between PhiC2 and Tn6215 but also, the constituents involved and initiated in transduction. As well as adapting methods to apply machine learning and deep learning algorithms for the prediction and identification of potentially virulent genome regions.

KEYWORDS: ANTIMICROBIAL RESISTANCE, BLAST, *C. difficle,* CONVOLUTIONAL NEURAL NETWORK, PHASTER, PHIC2, TN6215,

ACKNOWLEDGEMENTS

Many people have helped me with this research. First and foremost, I would like to thank Dr Ashley Spindler for their guidance and knowledge that they have shared with me as my supervisor throughout my project at the University of Hertfordshire.

Additionally, to Dr Shan Goh and Dr Simon Baines for enabling the provision and procurement of the datasets used within this project. The research could not have taken place without their collaboration. I am also grateful to all the researchers at Sir Charles Gairdner Hospital in Australia through the generation of the datasets used herein.

Finally, I would like to thank the unending love and support of my mother and father, Karen and John, and my partner, James, who have always believed in me and supported me throughout my university journey, I would not have got through this without all of you.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	adenine
AMR	Anti-Microbial Resistance
AUC	Area Under the Curve
BLAST	Basic Local Alignment Search Tool
C	cytosine
C. difficile	*Clostridium difficile*
CDI	*Clostridium difficile* Infection
CNN	Convolutional Neural Network
DL	Deep Learning
DNA	deoxyribonucleic acid
E. coli	*Escherichia coli*
FN	False Negative
FP	False Positive
G	guanine
HGT	Horizontal Gene Transfer
MGE	Mobile Genetic Element
ML	Machine Learning
NW	Needleman-Wunsch
ORF	Open Reading Frame
PHASTER	PHAge Search Tool - Enhanced Release
PCR	Polymerase Chain Reaction
ROC	Receiver Operating Characteristic
SEM	Standard Error of the Mean
SMOTE	Synthetic Minority Oversampling Technique
SW	Smith-Waterman
T	Thymine
TN	True Negative
TP	True Positive

CHAPTER I: INTRODUCTION

This thesis has been carried out in the field of bioinformatics, applied to the potential ability for bacteriophages to enable horizontal gene transfer (HGT) of genetic material through transduction between members of the same bacterial species. However, there are conflicting views and an overall lack of research regarding the bacteriophage PhiC2 and transposon Tn6215, both species capable of residing and integrating into *Clostridium difficile (C. difficile)* genomes. This research aims to identify PhiC2 and Tn6215 regions in *C. difficile* genome isolates through the application of a 'gold standard' method and the machine learning technique, Convolutional Neural Network. The research will look at the algorithm itself and the computational and space complexity. From these, the research aims to compare the efficiencies of the above algorithm and observe the sacrifices the algorithm makes in exchange for speed.

This chapter introduces some general aspects the background and context, followed by the implications of the research, the hypothesis, aims and objectives, the focus and locus of the research, and finally, the limitations. Subsequent chapters contain additional background information relevant to the studies presented in each chapter.

I.I BACKGROUND

Antimicrobial resistance (AMR) has become one of the most serious global healthcare problems. The global mortality rate from drug-resistant bacteria is estimated to be 700,000, with prevalence projected to increase to 10 million by 2050. The common gram-positive anaerobic spore-forming bacterium *Clostridium difficile (C. difficile)* is considered one of the most important drug-resistant pathogens. Despite the Centers for Disease Control and Prevention listing *C. difficile* infection (CDI) as an urgent threat, the slow discovery to elucidate CDI through the production of new antibiotics and over-prescription has supplemented the emergence of a large number of antibiotic-resistant *C. difficile* strains.

Horizontal gene transfer (HGT) mechanisms are responsible for the increased spread of antibiotic-resistant *C. difficile* strains. Conjugation, transformation, and transduction are the primary mechanisms dissemination of antibiotic resistance genes occurs (Daubin et al., 2016). The notion that bacteriophages mediate transduction is a major driver of HGT is increasingly becoming recognised (Goh et al., 2003; Goh et al., 2007; Goh et al., 2013). AMR genes are often found on various mobile genetic element (MGEs) such as plasmids, genomic islands, and transposons, and, as such, can be horizontally transferred by bacteriophage transduction (Chiang et al., 2019). PhiC2, a bacteriophage with high specificity to *C. difficile*, has been demonstrated to mediate the transduction of the transposon Tn6215, encoding erythromycin resistance, between *C. difficile* strains (Goh et al., 2013). Therefore, this research applies bioinformatics, a science that combines biology, statistical methods

and computer science together with a specific focus on the role that the bacteriophage PhiC2 plays in horizontal transfer of the AMR gene Tn6215.

Genomic samples were provided by Sir Charles Gairdner Hospital, Australia from when a patient was infected with CDI and the genomes were sequenced. The DNA nucleotide can be defined as a word over the four-letter alphabet of nucleotides; {Adenine: A, cytosine: C, guanine: G, and thymine: T} (Portin, 2014). The DNA of each species, bacteriophage, mobile genetic element is unique, and the pattern of nucleotide base arrangement determines the unique characteristics of a virus. Therefore, DNA sequence identification plays a vital role in the classification of genomic regions.

Sequence alignment is a very important and successful method to analyse molecular sequences and thus, is one of the most important areas in bioinformatics. It is the process of searching similarity in two or more DNA or protein sequences (Mullan, 2006). At present, modern high-throughput genomic methods are an attractive way to unravel the molecular AMR pathomechanisms behind complex diseases, allowing, e.g., studies of whole genomes of patients (Goh et al., 2013). These methods are increasingly applied to understand AMR. However, huge amounts of data are generated, and the greatest challenge in performing meaningful bioinformatic sequence alignments has been a trade-off between accuracy and efficiency. The 'gold standard' methodology using Phage search tool enhanced release (PHASTER) (Arndt et al., 2016), and Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) tend to be used as a ground truth for sequence identification. Older algorithms like Needleman-Wunsch and the Smith-Waterman algorithm tend to have very high computational complexities and slow speed (Murata et al., 1990; Olsen et al., 1990). Thus, optimisation is not achieved owing to the requirement of high throughput computer processing. Newer algorithms, utilising the characteristics of deep learning (DL), such as, Convolutional Neural Networks (CNNs) sacrifice some of the accuracy to make the alignments faster (Nguyen et al., 2016). Comparisons between the 'gold standard' method and the newer CNN DL algorithm are required because as the amount and scale of DNA sequence identification grows larger, faster algorithms become more important to be able to quickly compare a given sequence to the entire database.

I.II IMPORTANCE

C. difficile is a prevalent human pathogen causing diarrhoea that can be persistent, especially in hospitals, and difficult to resolve using antibiotics. The acquisition of AMR genes is changing rapidly within *C. difficile*; potential zoonosis, horizontal and vertical gene transfer, and mutations are heralding a new approach to AMR (Lim et al., 2020). A large proportion of *C. difficile* genomes contain horizontally acquired AMR genes; elements acquired through conjugation are studied well however a mechanism of transduction with PhiC2 has only been demonstrated once (Goh et al., 2013). It was evident therefore that further research regarding the transduction mechanism of PhiC2 needed to be explored because transduction may be a useful tool for the genetic manipulation of *C. difficile* and applied in drug discovery. The completion of this research has provided an opportunity to use up-to-date tools and technologies in an attempt to implement and explore something new within the theory of PhiC2 enabling transduction of Tn6215.

I.III HYPOTHESIS, AIMS AND OBJECTIVES

Prior to the emergence of machine learning, bioinformatic algorithms had to be programmed manually; for problems including AMR gene transfer through transduction assessment in *C. Difficile,* this proved difficult. Therefore, genomic analysis will be beneficial to determine the prevalence of phage transduction and mobilisation of Tn6215 which may aid in the early detection of drug resistant *C. difficile* stains, and prevention of dissemination worldwide.

The primary aim of this study is to provide a comprehensive review of the relationship between the Australian identified prophage, phiC2, and transposon, Tn6215. The study hypothesises phiC2 is necessary for the transfer of erythromycin resistance, hence, *C. difficile* isolates containing Tn6215 will also always be accompanied by phiC2 to enable phage mobilisation of the transposon.

Additionally, the study aims to compare various machine learning and deep learning algorithms to compare the speed and accuracy of identifying PhiC2 and Tn6215 regions between DNA-sequencing algorithms and against the proposed ‘gold standard’ method. The application of machine learning algorithms will serve to learn and identify phiC2 features and further learn how to combine the occurrence of Tn6215, a low-level feature, into abstract features. Thus, enabling the system to make sophisticated predictions, once trained properly, and enable the data to be interpreted in unanticipated ways.

Furthermore, the study aims to utilise the pathophysiological properties of phiC2-Tn6215 pairings in *C. difficile* stains to explain the alternative pathogenesis pathway causing symptomatic disease expression in humans. The validation of this aim in the current study can be applied *in vivo* to observe the susceptibility of pharmaceutical drugs.

I.IV FOCUS AND LOCUS OF THE RESEARCH

The focus and locus of the research has been both general and specific. The hypothesis, aims and objectives listed above seek to explore the identification of PhiC2 and Tn6215 regions within *C. difficile* genome isolates. This encompasses the assessment of the corresponding relationship between these two elements and interpretations of their pathophysiological properties. This broad exploration has relied upon a 'gold standard' method of genome sequence identification and has been focussed on the use of PHASTER and BLAST for identification of PhiC2 and Tn6215, respectively.

The decision to focus a significant part of the research on the hypothesis phiC2 mediates transduction of Tn6215 aligns with the use of bioinformatics in qualitative research. Bioinformatics is a broad term and does not relate to a particular methodological choice but rather to a choice of what is to be studied, and according to Bayat (2002) is "defined as the application of tools of computation and analysis to the capture and interpretation of biological data". Owing to the research highlighting bacteriophages can disseminate AMR genes in *C. difficile* being in infancy this research has solely focused on the relationship between PhiC2 and Tn6215. Investigation into this narrow focus has enabled the research to explore a greater breadth and depth and formed conclusions that would not have been feasibly possible if exploring all bacteriophage species in *C. difficile*. It is important to highlight the generalisation of the necessity of PhiC2 in the transduction of Tn6215 in other bacterial species will not be explored. This is owing to the time constraints of the research and an absence of literature showing Tn6215 in other species.

However, the research also seeks to compare machine learning and deep learning methods as an alternative and to challenge the 'gold standard' method. A detailed comparison of all the machine learning algorithms and tools for gene identification are beyond the scope of this project. The research emphasises only that each of these methods has its own advantages and limitations, with none of these being perfect. As a result, the research has also had the specific focus of the DNA sequence similarity algorithm, a Convolutional Neural Network (CNN).

I.V LAYOUT OF THE THESIS

The research described in the following chapters is set out in the following way: In chapter two, the chapter uses the history of bacteriophages facilitating AMR gene transfer through transduction to explore the novel theory Tn6215 is dependent on phiC2 to enable HGT intra-specially. The chapter also explores some of the key problems that arise during genome sequence identification and presents a new process to challenge the established 'gold standard' method. Leading into chapter three exploring the research question, aims and exploratory hypothesis.

In chapter three, the research methodology is described, and a research paradigm is established. Multiple methods of nucleotide sequence augmentation and analysis are outlined.

Chapter four is a significant chapter in the research and highlights the results achieved from the research paradigm.

Chapter five looks to genome sequence regions of Tn6215 and PhiC2 identification to explain the relationship between these two elements. The chapter evaluates the results from chapter four. The concepts of the pathophysiological relationship between Tn6215 and PhiC2 are also explored, and it is argued that these are entwined and interdependent processes that can be manipulated as targets during drug development.

Chapter six, the final chapter, weaves together the discussions that have been had in earlier chapters and conclusions are drawn.

CHAPTER II: LITERATURE REVIEW

II.I INTRODUCTION

This chapter provides a review of the literature and secondary data that have been collected, collated, and incorporated into discussion throughout this chapter. In keeping with the focus given to the design within the research, this process has been both an evolving process and ongoing throughout the duration of the research. Accordingly, this chapter will initially discuss the components of antimicrobial resistance in *C. difficile*; it will then move on to describe and analyse the evolution of horizontal gene transfer in response to emerging theories of the bacteriophage PhiC2 and the transposon Tn6215. The concept of PhiC2 being closely linked to Tn6215 and thus there will also be an analysis of how these two variables function. This chapter will also provide an analysis of the current pairwise sequence identification methods to provide a context for analysis and future usage. Finally, there will be a critical analysis of the benefits and challenges of implementing pairwise sequence identification methods and a consideration of the failures of implementation and the implications of this.

II.II CLOSTRIDIUM DIFFICLE

Clostridium difficile (C. difficile), a gram-positive species of spore forming obligate anaerobic bacterial pathogen, is one of the most prevalent causes of healthcare-associated infections worldwide (Fortier, L., 2018). A review of epidemiological studies from the EU/ EEA, report an estimated 125,000 cases annually (Robertson et al., 2020) with *C. difficile* infection (CDI) responsible for a quarter of all cases of infectious diarrhoea and associated complications including sepsis, pseudomembranous colitis, and kidney failure. In turn, it is associated with high healthcare demands and related costs, with approximately 30% of patients relapsing owing to perturbation of gut microbiota (Robertson et al., 2020). CDI is challenging to control and resolve because treatment guideline advice antibiotics, despite paradoxically being a pre-disposing factor for symptomatic disease expression (Pohl et al., 2011).

Antimicrobial resistance (AMR) contributes to the pathogenesis and spread of CDI through the acquisition of virulence genes through plasmids and transposons, (Fortier, L., 2018; Goh et al., 2005) described to significantly contribute to genome plasticity (Goh et al., 2013). Currently, approximately 11% of *C. difficile* genomes contain horizontally acquired genetic elements (Knight et al., 2017). Despite a tenth of the entire genome containing potentially virulent material, research regarding AMR genes acquired through horizontal gene transfer remains in its infancy. Therefore, this research applied five *C. difficile* strains to identify the transposon Tn6215, containing the *ermB* gene conferring erythromycin resistance. Additionally, the research aimed to identify the bacteriophage, PhiC2 demonstrated to mediate HGT of Tn6215 (Goh et al., 2013).

II.III THE ANTIMICROBIAL RESISTANT GENE *ermB*

One of the most prominent antibiotics administered to combat CDI is erythromycin, a macrolide (Brittain, 1987). erythromycin induces *C. difficile* cell death by inhibiting protein synthesis. The macrolide binds to the 23S ribosomal RNA molecule in the 50S subunit of ribosomes, as shown in Figure II.I, preventing the assembly of the 50S ribosomal subunit and subsequently preventing protein synthesis, causing susceptible bacterial organism to die (Patel and Hashmi, 2022; Wang et al., 2021).

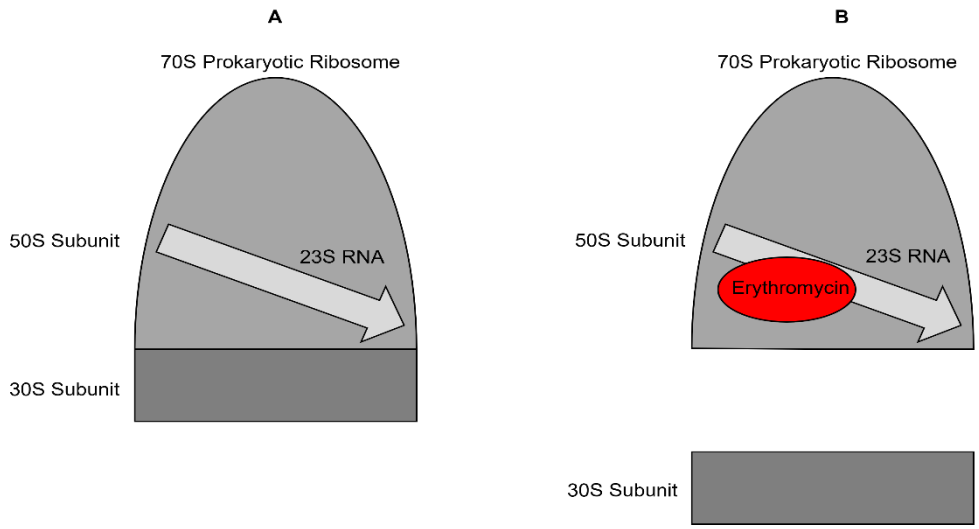


Figure II.I. *Visualisation figure of the 70S prokaryotic ribosome formation to carry out protein synthesis* ***A*** *without erythromycin and* ***B*** *with erythromycin.*

However, the emergence and expression of the AMR gene *ermB* has led to erythromycin resistance in several gram-positive bacteria including *C. difficile* but is increasingly found in gram-negative bacteria (Schroeder and Stephens, 2016; Vazquez-Laslop et al., 2018). Several studies have found prolonged exposure to erythromycin induced the expression of four *erm* genes (A, B, C, and D) in *C. difficile* (Sothiselvam, Liu, Han & Mankin, 2014). The *ermB* gene encodes an enzymatically active ribosomal methylase, shown in Figure II.II, that demethylates a single A in the 23S ribosomal RNA molecule (Wang et al., 2021). This single action prevents compatible binding of erythromycin leading to resistance and cell survival.

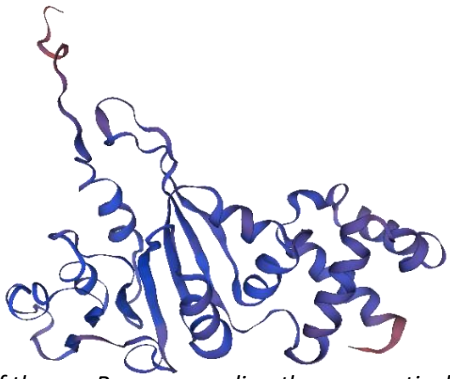


Figure II.II. *Protein 3D structure of the ermB gene encoding the enzymatically active ribosomal methylase. Homology modelling was performed by Swiss-model server (https://swissmodel.expasy.org/).*

II.IV TRANSPOSON: TN6215

In *C. difficile ermB* genes, granting erythromycin resistance, are location on conjugative transposons such as Tn6215. Conjugative transposons, also referred to as "jumping genes", are modular MGEs containing core areas required for regulation and recombination (Goh et al., 2013; Wasels et al., 2015). They are characterised by their capability to move from one location on the genome to another. However, the induction of *ermB* expression within Tn6215 has not been extensively studied, and its transfer mechanism is still not fully understood (Goh et al., 2013).

Previous literature has described some of the transposons initially found in *C. difficile* were horizontally transferred by conjugation-like mechanisms to other species. The transposons; Tn5398 and Tn5397, have been shown to be transferable from *C. difficile* to *Bacillus subtilis* and *Enterococcus faecalis* (Farrow, Lyras & Rood, 2001; Jasni et al., 2010; Mullany et al., 1990). However recently, Goh et al. (2013) has identified the transfer of Tn6215 to be mediated by the bacteriophage PhiC2. This report is the first of its kind showing PhiC2 can contribute to AMR of *C. difficile* through the sparsely studied HGT mechanism of generalised transduction; interspecies transfer of MGEs through mispackaging of lytic bacteriophages. This novel finding is the basis of this research to explore the phenomenon in the genomes of five *C. difficile* strains.

II.V BACTERIOPHAGE: PHIC2

PhiC2 is a *C. difficile* infecting bacteriophage capable of lysogeny and integration into the host genome and form infectious lytic particles spontaneously or following mitomycin C induction, shown in Figure III.III.

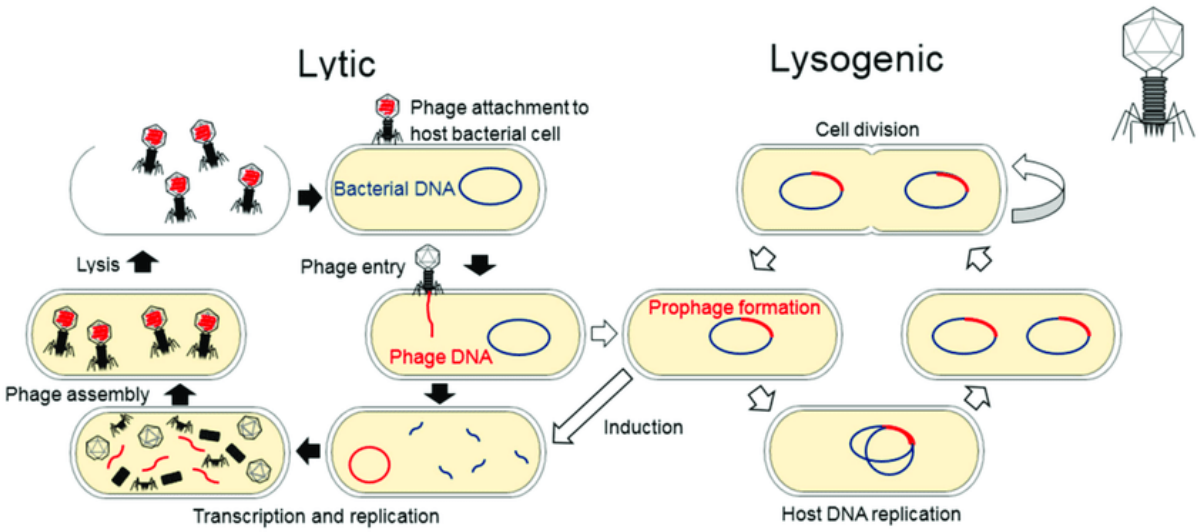


Figure II.III. *The lifecycle of a bacteriophage displaying the lytic cycle (left) and lysogenic cycle (right).* [Batinovic et al., (2019). Bacteriophages in Natural and Artificial Environments. *Pathogens (Basel, Switzerland)*, *8*(3), 100.

Copyright to use this figure was obtained from https://marketplace.copyright.com/rs-ui-web/mp on 25/08/2022.

Goh et al., 2007 successfully sequenced the complete genome of PhiC2 and subsequently first reported PhiC2 mediates transduction of Tn6215, encoding erythromycin resistance (Goh et al., 2013). This method of HGT occurs during the packaging of bacteriophage DNA, within the lytic cycle, has low fidelity and small pieces of host bacterial DNA can become packaged into the bacteriophage genome. This process leads to the transfer and recombination of potentially virulent genetic material into another bacterium upon bacteriophage cell infection (Abbas & Vinberg, 2021; Chiang, Penades & Chen, 2019; Ludwig et al., 2019).

Despite transduction being able to transfer AMR genes, in comparison to bacteriophages of E*scherichia coli* (Cieplak et al., 2018; Pacificio et al., 2019), *Staphylococcus aureus* (Abatangelo et al., 2017; Feng et al., 2021) and *Lactobacillus* species (Kyrkou et al., 2020; Sechaud, Rousseau & Accolas, 1988), the study of clostridial bacteriophages is in its infancy. Therefore, it is important to investigate PhiC2 in this research not only because the bacteriophage can contribute to AMR, but because there is potential to use the properties of transduction to be a useful tool in the genetic manipulation of *C. difficile.*

II.VI PAIRWISE SEQUENCE ALIGNMENT

This current research applies two methods; 'gold standard' pipeline and machine learning (ML) algorithms, to identify PhiC2 and Tn6215 regions in *C. difficile* genomes. Therefore, because the research focuses on DNA sequence identification and classification, sequence analysis algorithms were adopted. There are two main branches of sequence alignment algorithms: pairwise alignment algorithms and multiple alignments methods (Li et al., 2018). This thesis applied pairwise alignment algorithms because multiple alignment methods try to align all the sequences in a given query set (PhiC2 or Tn6215) whereas the aim of this research was to identify alignments between only two biological DNA sequences. Thus, pairwise alignment algorithms were the most appropriate to use.

II.VI.I EXISTING PAIRWISE ALIGNMENT ALGORITHMS

II.VI.I.I NEEDLEMAN-WUNSCH ALGORITHM

The Needleman-Wunsch (NW) algorithm first developed 1970, is a global alignment algorithm using a dynamic programming technique to compute an optimal alignment for two sequences (Aspland et al., 2017). Global alignment is the process of comparing two or more sequences as a whole. The way this algorithm works is that DNA sequences are aligned by matching and shifting to obtain the maximum global similarity of the two DNA sequences with complexity O(*nm*) (Isa et al., 2019). Several previous studies concluded the NW algorithm had superior performance for finding an optimal solution. Notably, Naidu & Narayanan (2016) applied NW to align two polymorphic malware viruses. However, this algorithm was of limited value for this research because the global alignment method tries to make matches along the entirety of the two sequences. Thus, the method could not be applied to a whole genome to identify a short query region with a different size. To accommodate this problem, the research proposed splitting the data into chunks of the same size as the query region however, the algorithm is very time-consuming for aligning a whole genome sequence because the algorithm runs in O(*nm*) time. Thus, this running time is not feasible for real biological applications.

II.VI.I.II SMITH-WATERMAN ALGORITHM

The Smith-Waterman (SW) algorithm was developed in 1981 to compute the optimal local alignment of two sequencing of length *n* and *m* (Rucci et al., 2018). The algorithm works similarly to the NW algorithm, but the algorithm starts at 0. However, the SW algorithm requires a very large extra memory space for saving the matrix. This is impractical for this thesis because the alignment aims to be carried out over whole *C. difficile* genome and therefore, the algorithm requires O(*nm*) memory (Naidu & Narayanan, 2016; Xia et al., 2022). Additionally, the high time complexity of the algorithm to allow deletions or insertions of any length with a penalty along the query sequence, caused the cost of making the SW method time-consuming with speed O(m^2n) (Rucci et al., 2018). Thus, this method calculates optimal alignment in a longer time than the NW algorithm.

II.VI.II GOLD STANDARD PIPELINE

In accordance with the literature documenting existing pairwise alignment algorithms, the research concluded it was inappropriate to apply either the NW or SW. Therefore, the most appropriate qualitative methods implemented in the research aimed to identify PhiC2 and Tn6215 regions within *C. difficile* genome isolates was through the production of a 'gold standard' DNA-sequence identification pipeline through the optimisation of APIs and databases. These methods selected within the identification pipeline were the most appropriate because they provided optimal characteristics of speed, accuracy, and ease-of-use.

II.VI.II.I PHASTER

Detecting bacteriophages in bacterial DNA is a challenging computational problem which lead to the development of two new tools for bacteriophage annotation: PHAST (PHAge Search Tool), published in 2011 (Arndt et al., 2011) and its successor PHASTER (PHAge Search Tool – Enhanced Release), released in 2016 (Arndt et al., 2016). Several software improvements and hardware enhancements have made PHASTER faster, more efficient, and much more user-friendly. PHASTER is 4.3x faster than PHAST when analysing a typical bacterial genome and software optimizations have made the backend of PHASTER 2.7x faster (Arndt et al., 2016).

Despite PHASTER having improvements in the areas of speed and usability application of the PHASTER API was applied to the research to identify the bacteriophage PhiC2 as opposed to using the website enabled automation of bacteriophage searches. This was because less human effort was required, and the workflow was easily updated to become faster. This method was more productive, despite the requirement of internet accessibility, because code was reused in complex but repetitive processes, whereby a function could be called preventing starting from scratch and using greater amounts of computer memory storage of variables.

However, APIs have been shown to be vulnerable to security breaches owing to poor integration. This factor was taken into consideration because the genome isolates are unpublished however the data did not contain any information that required ethical approval and therefore, this was deemed a low risk.

II.VI.II.II BLAST

Basic Local Alignment Search Tool (BLAST) is a sequence similarity search program that can be applied via a web interface or as a stand-alone tool (). The BLAST server at the National Center of Biotechnology Information has a varieties of BLAST searches by type; Nucleotide, Protein, Translated and Genomes. This research used data containing only DNA nucleotide bases therefore BLAST Nucleotide 'blastn' search was employed. BLAST has been applied in several studies for genome identification of specific genes.

The use of BLAST for genome identification of nucleotide sequences has been in employed in several studies. Luhmann, Holley & Achtman (2021) emphasised BLAST found the identification of the presence and absence of individual genes using *Salonella* genomes within minutes. Pratama et al. (2021) demonstrated how dataset composition and assembly affect viral taxonomic classification and identification of metabolic genes, contributing to virus virulence. Che et al. (2021) used BLAST to link conjugative plasmid and phylogenetically distant pathogens to suggest an evolutionary mechanism for the HGT of AMR genes. Hence, BLAST was the most appropriate sequence similarity search program to identify Tn6215 in *C. difficile* genomes.

However, the web interface of BLAST can only hold query and subject nucleotide lengths of 100,000 bases whereas the stand-alone tool does not have a nucleotide length limit (Smith, 2022). Thus, the stand-alone tool was implemented because each genome used in this research had approximately 4.5 million bases.

II.VI.III CONVOLUTIONAL NEURAL NETWORK

Machine learning (ML) has been proposed in metagenomics for rapid taxonomic assignments of bacteria (Chaudhary et al., 2016; Wang et al., 2007) and fungi (Liu et al., 2012). Previously, a Bayesian classifer (Wang et al., 2007) was applied to metagenomic studies however, the emergence of deep learning (DL) has been demonstrated as a successful paradigm for big data classification and clustering (LeCun, Bengio & Hinton, 2015). The class of DL algorithms most commonly applied for pattern analysis are convolutional neural networks (CNNs). In contrast to shallow networks, CNNs are characterised by a significantly increased number of successively connected neural layers. This increased number of layers can reveal higher-level features and more abstract concepts uncovering more complex and hierarchical relationships.

The effectiveness of CNNs has revolutionised recognising genes and characteristics in the DNA sequence. Reference is made here to previous research, including Zeng, Edwards & Gifford (2016) on the general principles and potentials of CNN for predicting DNA protein bindings using transcription factor datasets. Wang et al. (2018) summarising how CNNs can accurately predict the binding intensities of type-specific binding of transcription factors to regulatory elements in a given DNA sequence. Dasari & Bhukya (2021) highlight CNN-based methods automatically extract features for viral genome prediction. Gunasekaran et al. (2021) providing perspectives on how CNNs can classify mutated eukaryotic DNA sequences to identify virus origin. Therefore, a CNN is generated and applied in this research to predict the genome regions of Tn6215 and PhiC2 in *C. difficile.*

Previous studies demonstrated inappropriate input data types may result in numerous errors in genome annotation (Sarigul et al., 2019; Wang et al., 2021). The *C. difficile* biological sequences contain DNA nucleotide bases, represented as a sequence or string of characters, an inappropriate input data type. Therefore, DNA data augmentation methods were employed to convert the categorical data into numerical data. This process was a fundamental step in the methodology because machine learning techniques require numerical data to be able to form statistical relationships between the inputs (Sarigul et al., 2019). Several research papers identified one-hot, *k*-mer, sequential and ordinal encoding methods (Gunasekaran et al., 2021; Liimatainen et al., 2021). However, this thesis did not implement ordinal encoding owing to the similarity to sequential encoding.

CHAPTER III: METHODOLOGY

III.I INTRODUCTION

This chapter outlines the research design and methodology. First a research paradigm is established. The second section details the genome isolates. The third section discusses methodology within the 'gold standard' identification of genome sequences method. The fourth section outlines the DNA augmentation methods to convert categorical data for input in machine learning algorithms. The final section describes the methodology of each of the selected machine learning algorithms used to identify the desired genome sequences.

Note that the *C. difficile* genome isolates were all input into the PC-based computer analysis software; Python (Van Rossum, 1995) where all methodological processes were completed.

III.II GENOME ISOLATES

The experimental work was carried out on secondary data; 5 *C. difficile* isolate datasets (D133, E185B, E011, S4 1, S8), provided by Dr Shan Goh, Senior Lecturer in Microbiology at the University of Hertfordshire. The datasets were derived from experiments performed by researchers at Sir Charles Gairdner Hospital, Perth, Western Australia (Imwattana et al., 2014). The *C. difficile* genome isolates (n=5) with phiC2 susceptibility and relevant properties are described in Table III.I. The format of the DNA sequence data was FASTA file, and the metadata is provided in Table III.II.

Table III.I. *Clostridium Difficile genome isolates description and metadata applied to this study.*

Strain	Erythromycin gene argument	Erythromycin transposon	Type
S4-1	ermB	Tn5398	String; DNA sequence
S8	ermB	Tn6218	String; DNA sequence
E011	ermB	Tn6215	String; DNA sequence
D133	ermB	Tn6215	String; DNA sequence
E185B	ermB	Tn6215	String; DNA sequence

Detection of *ermB,* resistance to the erythromycin, was confirmed by disc diffusion where erythromycin-impregnated filter paper discs were placed on agar plates inoculated with each strain of *C. difficile* (Goh et al., 2013). The absence of a zone of inhibition confirmed the bacterial strains were not sensitive to erythromycin conferring resistance. Additionally, Polymerase Chain Reaction (PCR) detection of *ermB* was carried out under optimal PCR conditions using genomic DNA and primers which are extensively described in the original papers of Imwattana et al., 2021. The presence of *ermB* mobile element integration site was confirmed by Southern Hybridisation.

The GC content distribution of each isolate with the length of sequence is shown in Table III.II. The GC content in Table III.II of each dataset was interesting to observe because a low GC content, observed in this research, has low thermostability. The standard range of GC content in *C. difficile* is 3-75% (Wang et al., 2021). This highlights DNA with low GC content is not stable owing to an absence of more hydrogen bonds and is more capable of transference of AMR genes (Wang et al., 2021).

Table III.II. *Distribution of each C. difficile genome isolate and the length of each nucleotide sequence.*

	D133	**E185B**	**S4 1**	**S8**	**E011**
Length of sequence	4,617,490	4,215,595	4,293,381	4,326,616	4,683,519
GC content (%)	29.3	29.3	28.9	29.5	28.9

The sample DNA sequences from the dataset with the complete genomic sequence of the *C. difficile* isolate are shown in Figure III.I.

	Cluster	**Code**
0	>cluster_001_consensus	ATGGATATAGTTTCTTTATGGGACAAAACCCTACAATTAATAAAAG...
1	>cluster_003_consensus	GGGAAACCTCCAGAGTCAAGGATAAAAAACCTTGAGCTAATAACAA...
2	>cluster_004_consensus	ATACAACAGTTGATGGAAGCGTTTAATGTTGCAGGTGGTACATTTA...
3	>cluster_005_consensus	ACGCTTCTCAGCTAAAAGACTTTGTACCCGATACAATTACTTTTAA...
4	>cluster_007_consensus	TGGCTACAGAGGAGCAGAAAACAAGAGTTGAGAAGTTGAAATGTGA...
5	>cluster_009_consensus	CTACTCGTGGCAAGAGTAAAAGTGGCTGGGAACTTGGTTTTAAAAT...

Figure III.I. *Sample dataset D133 C. difficile genome isolate with genomic sequences.*

III.III SEQUENCE IDENTIFICATION PIPELINE

Owing to the infancy of bacteriophage exploration there is a lack of directly linked literature to potential methods. A 'gold standard' workflow, shown in Figure III.II, for DNA sequence identification was implemented to mine the *C. difficile* isolate genomes.

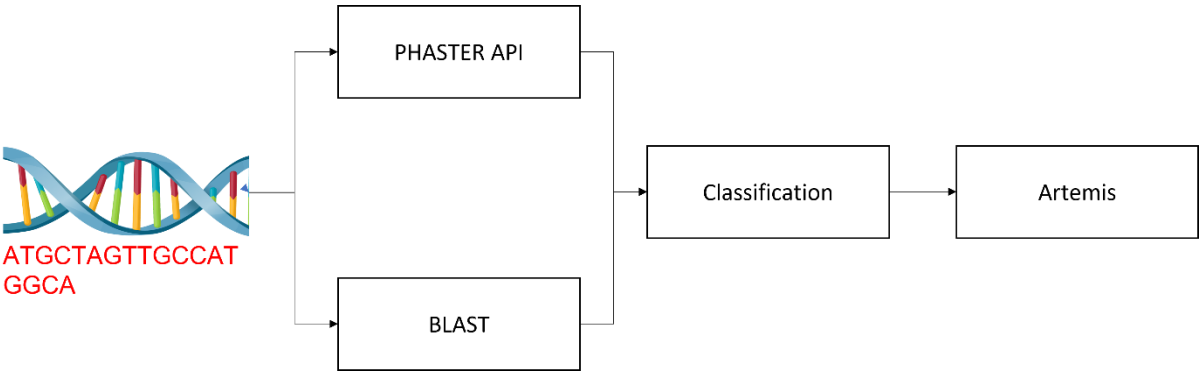


Figure III.II. *'Gold standard' DNA sequence identification pipeline applied in this research.*

The data was supplied in the format of FASTA files with 2 *C. difficile* genome isolates; D133 and E011 containing multiple circular genome consensus'. The data was subject to several data pre-processing stages and to convert the raw data into a more understandable, useful, and efficient format. The data was parsed, split, and converted into a pandas dataframe.

III.III.I PHASTER API

The identification of complete and incomplete bacteriophage prophages in *C. difficile* genomes was achieved through the integration of the software PHASTER (Ardnt et al., 2016) using the URL API (http://phaster.ca/phaster_api) in Python. The PHASTER API was utilised as opposed to the default web-version of PHASTER was its ability to run several queries in conjunction, reducing the speed of PhiC2 and other bacteriophage identification.

Successful connection to the API was confirmed through the status code '200: healthy connection with the API on web'. The corresponding accession number generated corresponded to three output files; 'Summary.txt', 'Detail.txt' and 'phage_regions.fna'. The summary file contained the most useful and extractable information regarding the bacteriophage species in each genome file and was applied for analysis of the nucleotide base regions PhiC2 resided.

Table III.III shows an example of the alignment of the PhiC2 target obtained from the implementation of the PHASTER API as a summary file. The cells highlighted green highlight PhiC2 as the most common bacteriophage region found. Interestingly, the presence of the bacteriophage attachment site designates whether the bacteriophage is complete, capable of successful excision from the bacterial cell, or not.

Table III.III. *An example of the summary file generated text output between the C. difficile genome D133 and the search for the bacteriophage PhiC2 using the PHASTER API.*

REGION	REGION LENGTH	COMPLETENESS (score)	SPECIFIC KEYWORD	REGION POSITION	ATTATCHMENT SITE SHOWUP	MOST COMMON PHAGE NAME (hit_genes_count)
1	56.1Kb	intact(150)	integrase,terminase,portal,head,capsid,tail	cluster_001_consensus: 263328-319431	yes	PHAGE_Clostr_phic2
2	18.1Kb	incomplete(40)	transposase,tail	cluster_001_consensus: 337175-355300	no	PHAGE_Escher_ESCO5
3	66.4Kb	intact(150)	integrase,terminase,portal,head,capsid,tail	cluster_001_consensus: 1286427-1352855	yes	PHAGE_Clostr_phic2
4	27.3Kb	incomplete(30)	lysin,tail	cluster_001_consensus: 1597494-1624862	no	PHAGE_Clostr_phiCDHM19
5	14Kb	incomplete(30)	transposase,portal	cluster_001_consensus: 2629263-2643262	no	PHAGE_Clostr_phic2
6	4.5Kb	incomplete(50)	virion,integrase	cluster_001_consensus: 2875549-2880123	yes	PHAGE_Clostr_phic2
7	74.6Kb	intact(150)	head,tail,portal,terminase,capsid,integrase,re...	cluster_001_consensus: 3333672-3408315	yes	PHAGE_Clostr_phic2
8	126.4Kb	intact(150)	tail,lysin,recombinase,integrase,transposase,c...	cluster_003_consensus: 10614-137056	yes	PHAGE_Clostr_phiCD211
9	45.8Kb	intact(150)	tail,plate,lysin,integrase,portal,head,capsid	cluster_004_consensus: 1-45865	yes	PHAGE_Clostr_phiCDHM19
10	36.3Kb	intact(98)	fiber,lysin,capsid	cluster_005_consensus: 3-36308	yes	PHAGE_Clostr_phiCD111

III.III.II BLAST

Subsequently, *C. difficile* genomes positive for PhiC2 were subject to identification of the complete sequence of Tn6215, deposited in GenBank (Benson et al., 2013), with accession number KC166248. This was performed through the application of Local BLAST (McGinnis et al., 2004); BLAST over the Internet was not suitable for this study because it limits nucleotide queries to 100,000 bases. A database only containing KC166248 was created to be used as the subject nucleotide base sequence, the corresponding *C. difficile* isolates were input as the query nucleotide base sequence, and the output was a text file that contained the nucleotide base regions corresponding to Tn6215.

Figure III.III below shows an example alignment (BLASTn) between Tn6215 and a sample, D133, *C. difficile* genome. As seen in the figure, often a lot of alignment metrics are shown, such as a score and the associated E-value, number of identical letters, the number of gaps etc. Percent identity or E-value is often used as a measurement of sequence similarity and contributes to the homology between two sequences.

```
 Score = 569 bits (630),  Expect = 2e-161
 Identities = 324/330 (98%), Gaps = 0/330 (0%)
 Strand=Plus/Minus

Query  8093     TTGTAGGAAAATGTCCTAAGTGTGGCAACAATATTGTATTaaaaaaaTCGTTTTATGGTT  8152
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4040863  TTGTAGGAAAATGTCCTAAGTGTGGCAACAATATTGTATTAAAAAAATCGTTTTATGGTT  4040804

Query  8153     GTTCAAATTATCCTGAATGTAAGTTTACTTTAGCTGAACATTTTagaaagaaaaaactca  8212
                |||||||||||||||||||  || ||||||||||||||||||||||||||||||||||||
Sbjct  4040803  GTTCAAATTATCCTGAATGTACCTTCACTTTAGCTGAACATTTTAGAAAGAAAAAACTCA  4040744

Query  8213     ccaaaacaaatgtaaaagaattactagagggaaaagaaaccctggtaaaaGGAATCAAAA  8272
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4040743  CCAAAACAAATGTAAAAGAATTACTAGAGGGAAAAGAAACCCTGGTAAAAGGAATCAAAA  4040684

Query  8273     CGAAAGATAGAAAGTCCTACAATGCCGTTGTAAAAATCGGAGAAAAGGGATATATTGATT  8332
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4040683  CGAAAGATAGAAAGTCCTACAATGCCGTTGTAAAAATCGGAGAAAAGGGATATATTGATT  4040624

Query  8333     TTATATCTTTCTCAAAATAAACATAAAAGCCCTTTAAAGAGGGCTTTTATATATTAATCA  8392
                |||| ||||| ||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  4040623  TTATCTCTTTTTCAAAATAAGCATAAAAGCCCTTTAAAGAGGGCTTTTATATATTAATCA  4040564

Query  8393     CAAATCACTTATCACAAATCACAAGTGATT  8422
                ||||||||||||||||||||||||||||||
Sbjct  4040563  CAAATCACTTATCACAAATCACAAGTGATT  4040534
```

Figure III.III. *An example of alignment from the generated text output file between the C. difficile genome D133 and the transposon Tn6215 using the nucleotide library using the software BLAST.*

As mentioned in chapter II, it is difficult to ascertain whether an alignment is biologically relevant or not. Addressing this issue, BLAST provided an E-value for each alignment to help assess whether it is a statistically significant alignment or not. The E-value declares how frequent a particular score appears by chance, given the alignment model, query and database length. Despite this, even if an alignment is statistically significant it does not prove that two sequences are homologs or that the alignment is biologically relevant. Therefore, the E-value of the alignment output was assessed individually to determine importance.

III.III.III ARTEMIS

The genomic sequences were visualised qualitatively in 2D using Artemis (Carver et al., 2008) to observe the integration of all the found bacteriophage species using PHASTER and the transposon Tn6215 to analyse and annotate the genetic organisation.

III.III.IV STATISTICAL ANALYSES

The identification of PhiC2-Tn6215 pairings in *C. difficile* isolates were visually analysed through the production of 2D circular genome plots. This was the most appropriate owing to a lack of defined number of nucleotide bases bacteriophages, specifically PhiC2, are capable of mis-packaging by. Therefore, visual identification of PhiC2-Tn6215 pairings were performed within all five *C. difficile* strains.

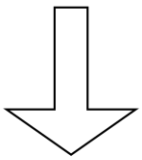
A correlation analysis, followed by a regression analysis was conducted to observe the relationship between phiC2 and Tn6215. A correlation analysis was the most appropriate to apply in this experiment because the research sought to explore the degree association between these two variables. Data with statistical significance was considered at a value of $P < 0.05$, because it indicates strong evidence against the null hypothesis and there is less than a 5% probability the null hypothesis is accepted. All quantitative data are shown as mean values $\pm$ standard error of the mean (SEM).

III.IV NUCLEOTIDE AUGMENTATION METHODS

Pre-processing data is the most critical step in most machine learning and deep learning algorithms that involve numerical rather than categorical data. The genomic sequences applied in this research were 1D categorical. There are several encoding techniques currently available to convert categorical data of nucleotide into numerical form. In this research, three methods for sequence encoding were used; one-hot encoding, sequential encoding, and k-mer encoding. The effect of the encoding technique on the classification accuracy was analysed and compared.

In One-Hot encoding, each nucleotide of the DNA sequence is assigned a 2D numerical matrix ([A, T, G, C] = [[1, 0, 0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]], with any other character recorded as [[0, 0, 0, 0]]). The entire DNA sequence was converted into an array of matrix using OneHotEncoder() from sklearn.

A	T	G	C	T	A	G

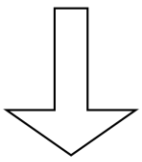


1	0	0	0	0	1	0
0	1	0	0	1	0	0
0	0	1	0	0	0	1
0	0	0	1	0	0	0

Figure III.IV. *Sequence data encoding using OneHotEncoder().*

In Sequential encoding, each nucleotide of in the DNA sequence is assigned an index value ([A, T, G, C] = [1, 2, 3, 4], with any other character recorded as 0) as shown in Figure X. The entire DNA sequence was converted into an array of number using LabelEncoder() function from sklearn.

A	T	G	C	T	A	G



1	2	3	4	2	1	4

Figure III.V. *Sequence data encoding using LabelEncoder().*

In *k*-mer encoding, the DNA sequence is converted into a text-like format by generating *k*-mers for the DNA sequence (k=3, [ATGCAATGC] = [ATG, CAA, TGC]). Each DNA sequence was decomposed into a *k*-mer of size *m,* as shown in Figure Y. This research employed a Convolutional Neural Network (CNN) with an input of (x, 100, 4) therefore, each sequence was decomposed into 100 nucleotide bases. The entire DNA sequence was converted into *k*-mers using the user-defined function kmer(), shown in Figure X.

```
def kmer(sequence, ksize):
    kmers = []
    n_kmers = len(sequence) - ksize + 1

    for i in range(n_kmers):
        kmer = sequence[i:i + ksize]
        kmers.append(kmer)

    return kmers
```

Figure III.VI. *The user-defined kmer(sequence, ksize).*

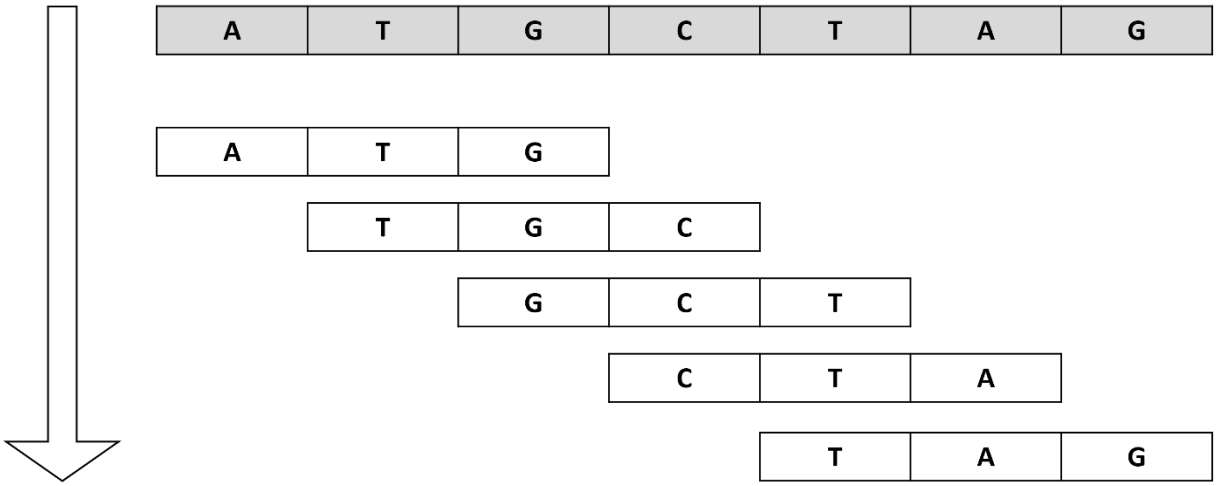


Figure III.VII. *Sequence data encoding (k=3) using the user-defined k-mer(sequence, ksize).*

III.IV.I DNA AUGMENTATION METHOD ANALYSIS

The One-Hot, Sequential and *k*-mer encoding techniques from the above section were used to encrypt the DNA sequence. The most appropriate nucleotide augmentation method that was used as the input for the deep learning algorithm was determined through both the individual and mean speed of converting categorical nucleotide bases.

$\frac{\sum speed}{N} \times 100$, where N = 5, the total number of *C. difficile* genome isolates.

III.V DNA SEQUENCE SIMILARITY ALGORITHM PRE-PROCESSING

III.V.I BINARY CLASSIFICATION

The corresponding most appropriate DNA augmentation method was applied to E185B, with significant PhiC2-Tn6215 pairings, because of the properties that will contribute to successful training with a deep learning algorithm. Owing to the provided datasets being raw and uncategorised two binary classification datasets were produced corresponding to PhiC2 and Tn6215 as 1 if the nucleotide bases in the split dataset into 100 chunks contained these regions and 0 if the nucleotide bases did not.

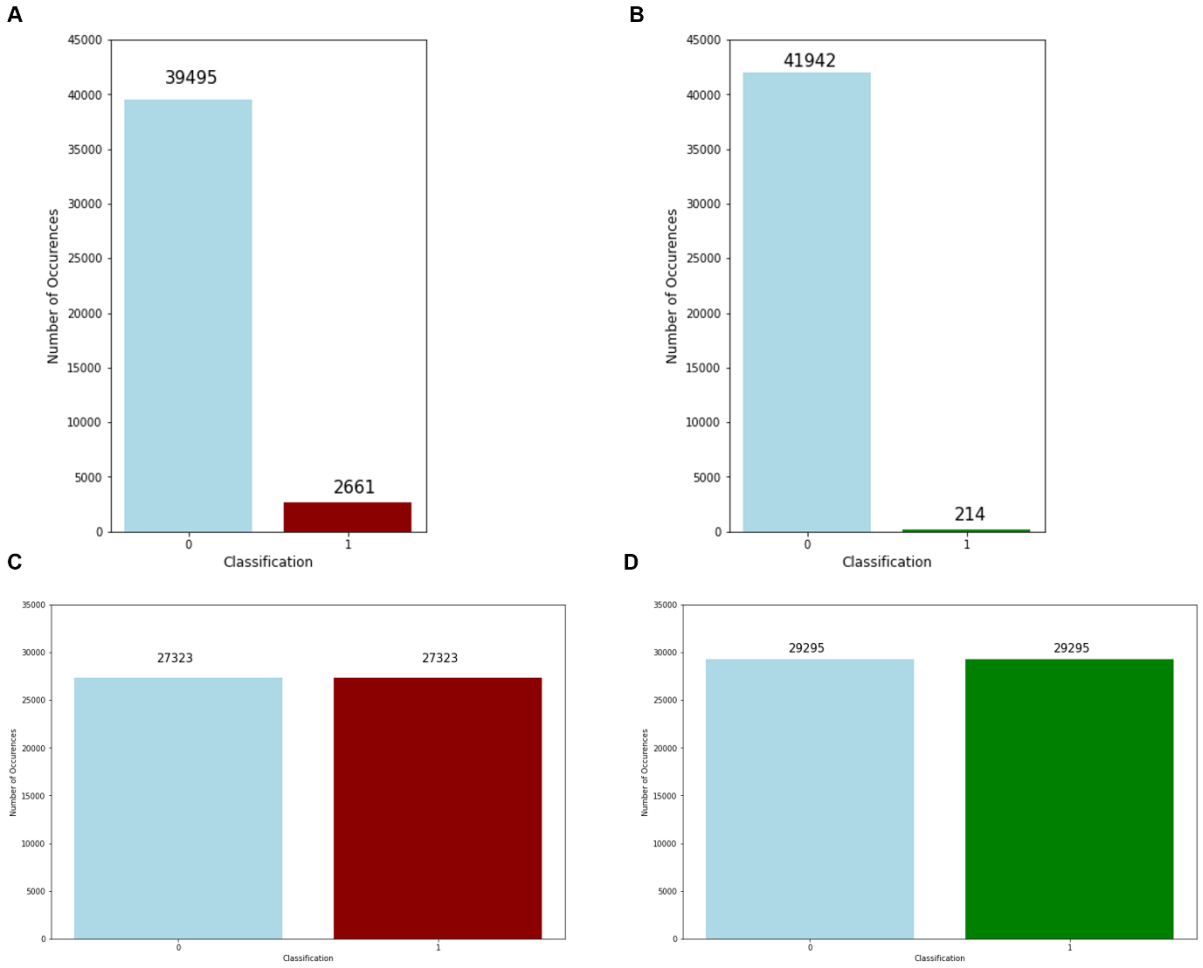


Figure III.VIII. *Distribution of each class of **A** unbalanced PhiC2, **B** unbalanced Tn6215, **C** balanced PhiC2 using SMOTE and **D** balanced Tn6215 using SMOTE in the sample E185B.*

From the binary classifier datasets in Figure III.VIII, it is clearly showing that there is an imbalanced dataset problem. Therefore, Synthetic Minority Oversampling Technique (SMOTE) was employed to handle this problem (Gunasekaran et al., 2021). The E185B dataset, the number of 100 nucleotide bases categorised as PhiC2 and Tn6215 were deficient in numbers. Synthetic samples for these minority classes were generated using the SMOTE algorithm to match the majority class closely. SMOTE picked the minority class instance randomly and searches for its closest minority class neighbours to contrast synthetic instances based on the nearest neighbours. This procedure was used to make an artificial instance for the minority classes.

II.V.II SLICING THE CLASSIFICATION DATASET INTO TRAIN AND TESTING SAMPLES

Following balancing of the dataset, a random 70% of *n* number of nucleotide bases of the dataset was categorised as the training sample. The remaining 30% of *n* number of nucleotide bases of the dataset was categorised as the testing sample. Each sample had a *x:* the DNA nucleotide augmented code, and *y:* the binary label. Shuffle was not applied to the splitting of the data into train and test groups (n=4) because it was essential to conserve position information of each nucleotide. Table III.IV shows the number of samples from each dataset, number of sequences as training sets, and testing sets.

Table III.IV. *Table showing the datasets, number of samples in each dataset, number of training sets, number of testing sets and the total number of sequences in the CNN model to identify PhiC2 regions and the CNN model to identify Tn6215 regions.*

Group	**CNN-PhiC2**	**CNN-Tn6215**
Datasets	E185B and S8	E185B and S8
Number of samples in each dataset	39,970 and 43,267	39,970 and 43,267
Number of training sets	27,323	29,295
Number of testing sets	55,914	55,914
Total number of sequences	83,237	85,209

III.VI CONVOLUTIONAL NEURAL NETWORK

In this research, the DNA-sequence similarity algorithm of a Convolutional Neural Network (CNN) was used for DNA sequence classification. CNN is a commonly applied deep-learning technique that can perform well on not only image classification, but also on text data. The CNN can extract features from the input dataset automatically as opposed to user defined features. The most appropriate dimension of CNN for this research was 1D because 2D and 3D are used for image and video data, respectively.

Owing to the purpose of this work is not to develop the best model for the tasks, but to illustrate the potential of DL techniques for DNA sequence classification, minimal hyperparameter optimisation was performed and the same hyperparameters for the CNN model was used across tasks. Therefore, it is reasonable to expect additional performance improvement with optimised parameters. The CNN model was written in Python using the TensorFlow and Keras framework.

III.VI.I CONVOLUTIONAL NEURAL NETWORK ARCHITECHTURE

The CNN architecture in this study consisted of a 6-layer fully connected network with 4 hidden layer(s) including two convolution layers and two pooling layers as illustrated in Figure IX.

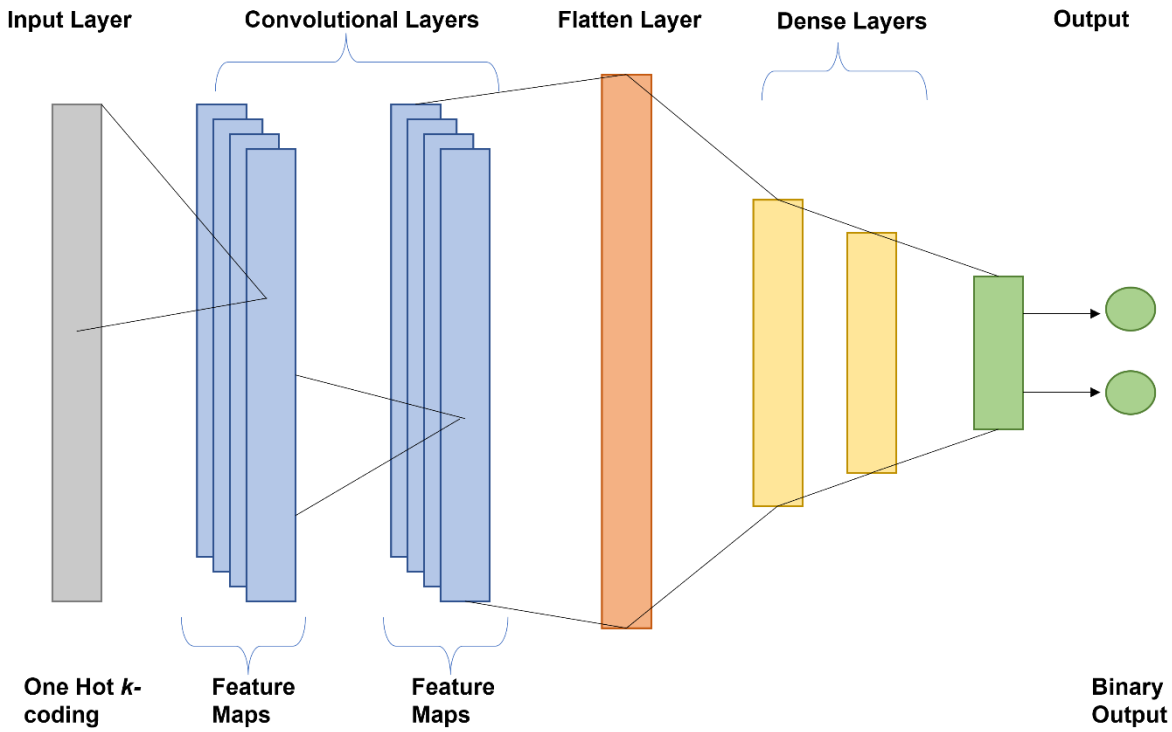


Figure III.IX. *1D CNN architecture for pairwise alignments of Tn6215 and PhiC2.*

A series of convolutional layers automatically extract features from the input dataset. Max pooling layers were applied after each convolutional layer and the dimensions of extracted features were reduced. Table III.IV shows the summary of the complete architecture of the proposed CNN model.

Table III.V. *Complete architecture specification of proposed CNN model.*

Layer (type)	Output shape	Parameter #
Conv 1	(None, 98, 32)	416
MaxPooling 1	(None, 42, 32)	0
Dropout (25%)	(None, 49, 32)	0
Conv 2	(None, 47, 16)	1552
MaxPooling 2	(None, 23, 16)	0
Dropout (25%)	(None, 23, 16)	0
Flatten	(None, 368)	0
Dense 1	(None, 16)	5904
Dropout (50%)	(None, 16)	0
Dense 2	(None, 1)	17

The first layer is the input layer. Two convolutional layers were added to the model with filters of 32 and 16, with the kernel of size (3 x 3) with ReLU as an activation function for feature extraction. ReLU activation was more appropriate than Sigmoid or Tanh because it was more computationally efficient to compute owing to ReLU computes max(0, x) whereas the other functions perform exponential operations that take longer. The feature map dimensions were reduced by adding a max pooling layer of size (2 x 2). A dropout rate of 25% was added to both convolutional layers to prevent overfitting. Finally, the feature maps were converted into single-column vectors using the flatten layer. The output was passed to a dense layer with neurons 16 and 1, respectively and a dropout layer of 50%. The ReLu function was used as the binary classification layer. The binary layer computes the probability of each sequence being containing the desired region (1) or not (0).

In the training phase, the binary crossentropy function was used as the loss function. This loss function calculates the error between the actual output and the target label, on which the training and update of the weights are done.

III.VI.II CLASSIFICATION ALGORITHM EVALUATION

III.VI.II.I CLASSIFICATION ACCURACY

CNN classification accuracy was one of the metrics applied for model evaluation. The accuracy of the model is the fraction of predictions that were correct using the parameters in Table V.

Table III.VI. *Definition of parameters used in the calculation of classification accuracy for the proposed CNN*

Parameter	Definition
True Positive (TP)	The true value is labelled as the positive class and the prediction is positive class.
True Negative (TN)	The true value labelled as the negative class and the prediction is negative class.
False Positive (FP)	The true value is labelled as the negative class, but the prediction is positive class.
False Negative (FN)	The true value is labelled as the positive class, but the prediction is negative class.

Based on the definitions of TP, FP, TN and FN, in Table X, the Accuracy was calculated as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

III.VI.II.II RECIEVER OPERATIVE CHARACTERISTIC CURVE

The application of Receiver Operative Characteristic (ROC) curve was used as an evaluation metric for the binary classification problems. The ROC curve is based on probability, plotting the TP against FP at various threshold values. Also, Area Under the Curve (AUC) was determined because it is a measure of the ability of the classifier to distinguish between classes and is used as a summary of the ROC curve. While the CNN was implemented by TensorFlow, the ROC and AUC values were computed a user-defined implementation.

CHAPTER IV: RESULTS

IV.I INTRODUCTION

This chapter details the results of the research. The first section of the chapter sought to identify PhiC2 and Tn6215 regions within each *C. difficile* genome isolate. Leading to the second section to identify PhiC2-Tn6215 pairings in each isolate. The research hypothesised PhiC2 is necessary for the transfer of erythromycin resistant gene Tn6215. Following identification of PhiC2-Tn6215 pairings the study aimed to show a positive relationship between these variables. The fourth section details the results of the DNA augmentation methods and leads into the fifth section of the research the machine learning and deep learning techniques applied to identify PhiC2 and Tn6215. The research aimed to compare the deep learning techniques to the proposed 'gold standard'.

IV.II IDENTIFICATION AND ANNOTATION OF GENOME REGIONS OF INTEREST

Sequencing methods were performed for the same five *C. difficile* genome isolates. The goal was to identify PhiC2 and Tn6215 regions using the designated software PHASTER and BLAST, respectively. The sequences obtained from sequencing results of all the genomes, and the alignments against them are available in Appendix A.

IV.II.I PHIC2

Figure IV.I shows the circular genome of PhiC2 the genome sequence was obtained from GenBank with accession numbers NC_009231.

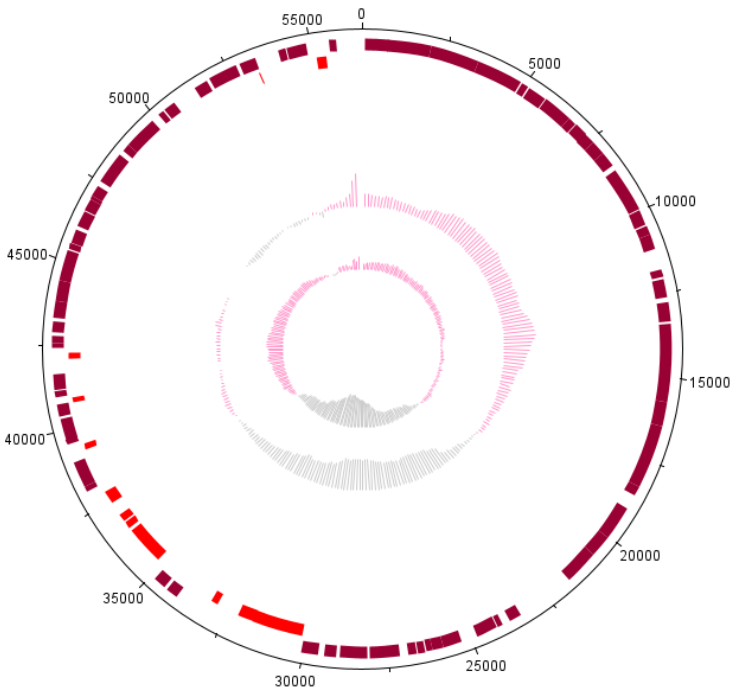


Figure IV.I. *Circular genome of PhiC2 with corresponding GC contents as distribution graphs produced using Artemis.*

The genome of PhiC2 is 56,538 base pairs and is organised into 84 putative open reading frames (ORF). The levels of gene regions confer directionality of the open reading frames, with the outermost level, containing 73 ORF, read 5'-3' and the inner level, comprising of 11 ORF, reading 3'-5'. The GC content of the genomes are highlighted in the 2 central circles, it was observed PhiC2 has an abundance of GC regions at the start of the genome sequence.

IV.II.II TN6215

Figure IV.II shows the circular genome of Tn6215 the genome sequence was obtained from GenBank with accession number KC166248.

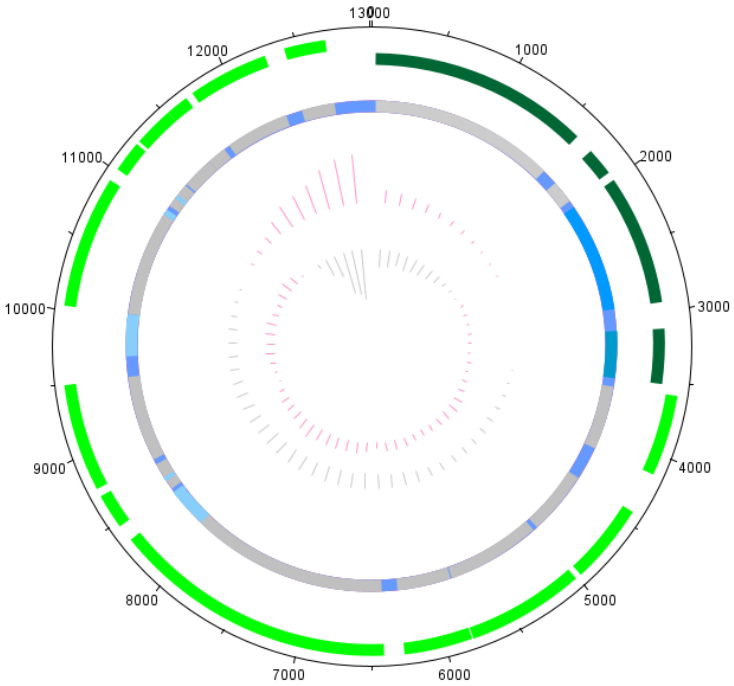


Figure IV.II. *Circular genome of Tn6215 with corresponding GC contents as distribution graphs produced using Artemis.*

The genome of Tn6215 is 13,008 base pairs and contains 17 putative ORF. The levels of gene regions confer directionality of the open reading frames, with the outermost level, comprising of 13 ORF, being read 5'-3' and the inner level, containing 4 ORF, reading 3'-5'. Tn6215 has a third level of genes highlighting the complete genome with grey regions corresponding to introns and blue regions corresponding to exons. The GC content of the genomes are highlighted in the 2 central circles, Tn6215 has an abundance of GC regions at the end of the genome, a mirror image to the characteristic observed in PhiC2.

IV.III IDENTIFICATION OF PHIC2-TN6215 PAIRINGS

IV.III.I QUALITATIVE METHOD

The infancy of previous studies describing mispackaging mechanisms contributing to HGT of transposons in bacteriophage capsids has led to the application of qualitative analysis of genomes. Identification of Tn6215 and PhiC2 regions were performed on 5 *C. difficile* isolates; D133, E185B, E011, S4 1, and S8, visualised using Artemis, are shown in Figure IV.III. PhiC2, red, and Tn6215, green, pairings were only identified in 2 of the 5 genome isolates (B-E185B and C-S8), shown by the symbol '**' in Figure IV.III. The analysis of these pairings incorporated the results in Table V.I as well as qualitative analysis.

The incomplete phage regions, found using PHASTER, were denoted '*' in Figure IV.III. Interestingly, 60% of bacteriophage genome regions that were categorised as incomplete from PHASTER contained phage nucleotide bases significantly similar to PhiC2.

Additionally, PHASTER observed the occurrence of other bacteriophage regions that did not contain PhiC2, these were marked as turquoise on the genome isolates. These bacteriophage regions were not included in Table IV.I because the study focused on PhiC2 regions however, it was interesting to observe the possibility that any bacteriophage could enable HGT. Despite this, the other bacteriophage regions did not significantly suggest capability of Tn6215 via HGT.

The outermost purple circular level in each genome shows the genome cluster consensus' it was only observed D133 and E011, Figure IV.IIIA and Figure IV.IIIE respectively, had more than one consensus in their genomes. Despite having more consensus clusters neither of these clusters contained significant results for PhiC2-Tn6215 pairings.

The GC content of the genomes are highlighted in the 2 central circles to observe significant spikes however, spikes occurred throughout the entire genome of each isolate and was not distinguishable to one area that contained either PhiC2 or Tn6215.

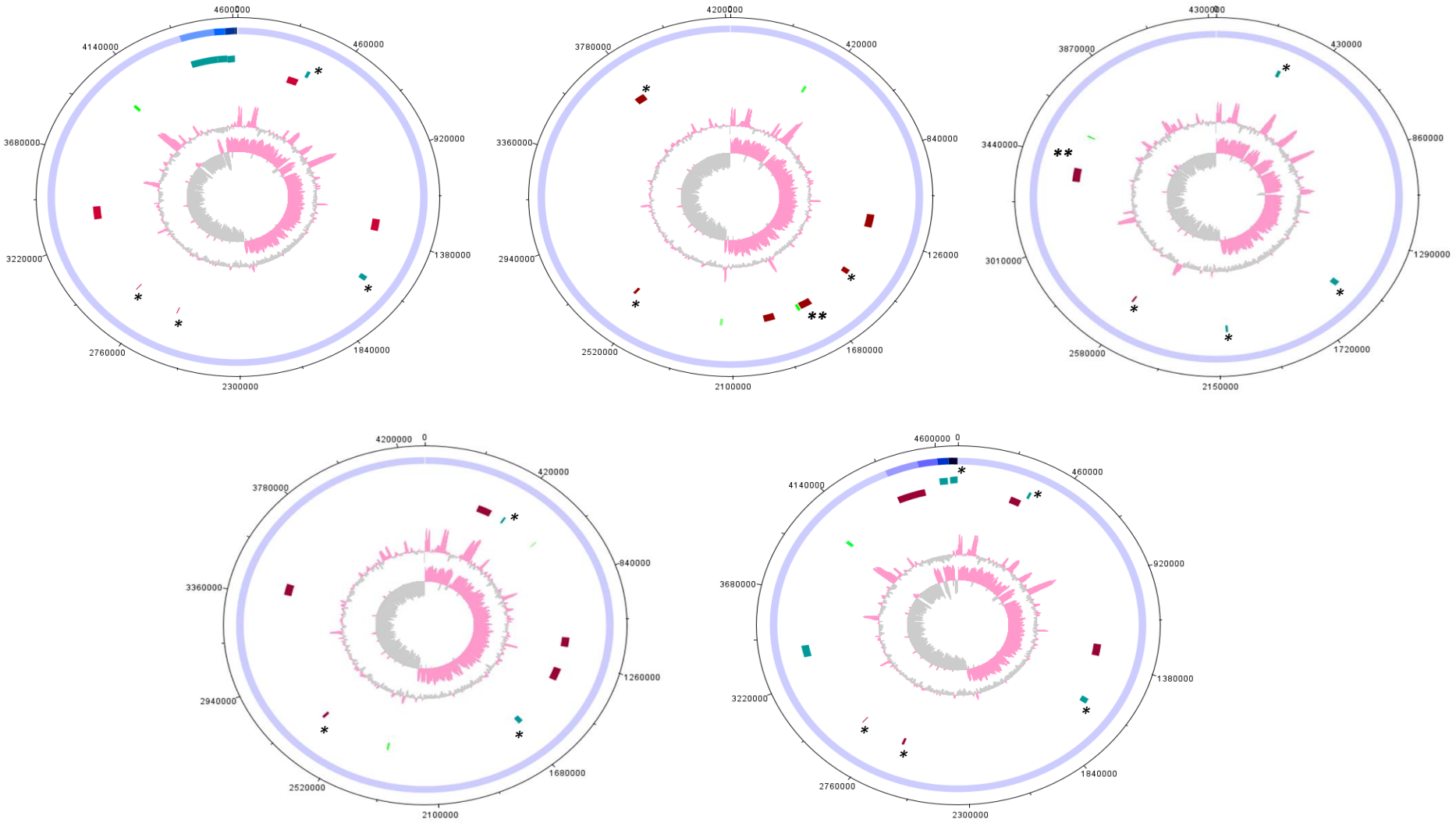


Figure IV.III. *Results from 'gold standard' method applying BLAST and PHASTER for identification of Tn6215 and PhiC2 regions, respectively. Circular genome isolates with corresponding GC contents as distribution graphs produced using Artemis.* ***A*** *D133,* ***B*** *E185B,* ***C*** *S8,* ***D*** *S4 1, and* ***E*** *E011.*

Green areas *highlight Tn6215,* ***Red areas*** *highlight PhiC2,* ***Turquoise areas*** *highlight both other bacteriophage species,* ***Purple areas*** *highlight the different genome isolate clusters,* ***** *mark incomplete bacteriophage, and* ****** *mark potential Tn6215-PhiC2 pairings.*

IV.III.II QUANTITATIVE METHOD

Owing to the difficulty of qualitatively analysing the identification of PhiC2-Tn6215 pairings were quantitatively analysed for significance. However, as shown in Table IV.I the distances between PhiC2 regions and the closest Tn6215 regions have been provided to portray the difficulty in analysis. PhiC2 regions categorised as incomplete were not included in Table IV.I because these regions are non-pathogenic and incapable of excision from *C. difficile* genomes. Additionally, only PhiC2 and Tn6215 regions in the same cluster were analysed. Interestingly, the genome isolates with significant PhiC2-Tn6215 pairings had a nucleotide position distance of less than 200,000 between the two regions.

Table IV.I. *The nucleotide distances between the closest PhiC2 and Tn6215 regions in each genome isolate.*

Strain	**PhiC2 (end nucleotide position)**	**Tn6215 (start nucleotide position)**	**Distance**
D133	3408314	4037964	629650
E185B	1215952	365576	-850376
E185B	1763158	1768666	5508*
E185B	1949510	2154414	204904
S8	3401850	3583323	181473*
S4 1	330409	597932	267523

The genome isolate E011 did not have any suitable PhiC2 regions flanking either side of the Tn6215 regions.

IV.IV PHIC2 AND TN6215: CORRELATION ANALYSIS

Following identification of the PhiC2 and Tn6215 regions in each of the *C. difficile* genome isolates, a bivariate correlation analysis was undertaken to investigate if a relationship existed between the two variables (Appendix A – Figure 1). The bar chart (Figure IV.IV) illustrates the quantities of Tn6215, PhiC2 and other bacteriophage species in each of the 5 *C. difficile* strains.

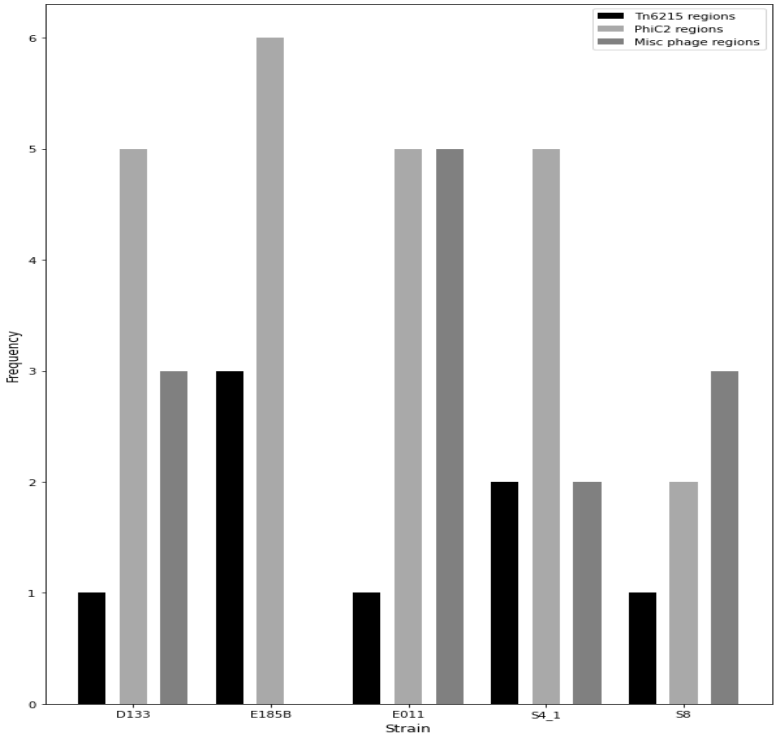


Figure IV.IV. *A bar chart showing the frequency of Tn6215 (black), PhiC2 (dark grey), and miscellaneous other bacteriophage (light grey) regions within each C. difficile genome isolate.*

In opposition with the hypothesis, analysis did not show significant positive correlations between the number of PhiC2 and Tn6215 regions (r=0.587, p>0.05). This highlights the presence of PhiC2 is not correlated with the presence of Tn6215. Following this, in confirmation with the observation that PhiC2 does not have a significant relationship with Tn6215 a linear regression analysis was conducted to identify if any number of PhiC2 could significantly predict the number of Tn6215 in a *C. difficile* genome. It was observed the presence of PhiC2 explained 34.8% of the variance (R^2=0.3478, p>0.05).

IV.V DNA NUCLEOTIDE AUGMENTATION METHODS

The three DNA nucleotide augmentation methods; One Hot encoder, Label encoder, and *k*-mer encoder was used to probe how data nucleotide augmentation methods impact the speed and reproducibility of changing the categorical DNA data to data that can be used as an input for the CNN model.

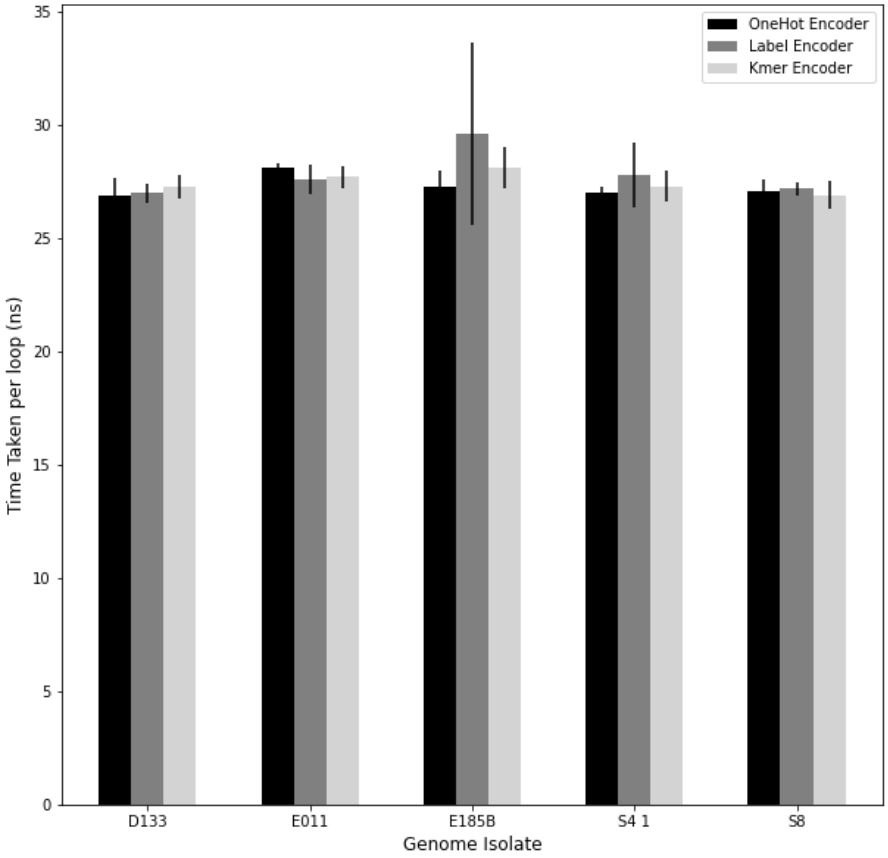


Figure IV.V. *A bar chart showing the time taken per loop in ns of each of the DNA augmentation methods; One Hot (black), Label (dark grey) and K-mer (light grey) encoding in each of the C. difficile genome isolates with calculated standard error bars.*

The bar chart (Figure IV.V) illustrates the mean time taken per loop (ns) of each DNA augmentation method when applied to the five *C. difficile* genome isolates. Visually, there is not a clear most appropriate method because the mean time taken per loop values are very close to one another. However, it would appear One-Hot encoding method takes the shortest amount of time because it is the lowest bar in three out of the five *C. difficile* genomes. Observations of the standard error of the mean (SEM) show that all the DNA augmentation methods overlap in all genome isolates; showing the results lie within experimental error of the two means. Therefore, an overall mean was calculated alongside the overall SEM (Table IV.II) of each DNA augmentation method to find the most appropriate method.

Table IV.II. *Summary of the mean time taken per loop (ns) and standard deviation (ns) of each of the data augmentation methods; One Hot, Label and K-mer encoding.*

Data Augmentation Method	Mean (ns)	SEM (ns)	Rank
One Hot Encoding	27.28	0.4826	1
Label Encoding	27.84	1.37	3
Kmer Encoding	27.46	0.6396	2

Table IV.II displays the mean time taken per loop for DNA augmentation methods. These results clearly show the One Hot encoding DNA augmentation method is the most appropriate because both the mean and SEM are the lowest values, 27.28 ns and 0.4826 ns, respectively. Surprisingly, the Label encoding DNA augmentation method performed the worst, with both the mean and SEM being the highest values, 27.84 and 1.37 ns, respectively.

IV.VI DNA-SEQUENCE ALIGNMENT ALGORTIHMS

The final stage of the study explored the accuracy and speed of identifying Tn6215 and PhiC2 nucleotide regions within the *C. difficile* isolates. A Convolutional Neural Network (CNN) model was used to identify features specific to the sequence query, Tn6215 or PhiC2, in a training set to apply the trained model to an unseen test set to analyse the reproducibility of the model.

IV.VI.I IDENTIFYING PHIC2 REGIONS

Initially, the CNN was trained on the raw imbalanced one-hot encoded nucleotide bases to identify PhiC2 to observe how imbalanced data affects the accuracy and reproducibility of a DL model. It can be seen from Figure IV.VI the accuracy of the data remains stationary at 92.5% despite the loss continuously changing over the 20 epochs. To further investigate these results a confusion matrix was generated, shown in Figure IV.VII. In a confusion matrix the number of correct and incorrect predictions are summarised with count values and broken down by each class (Ruuska et al., 2018). The confusion matrix identifies the model is predicting every class as 0.

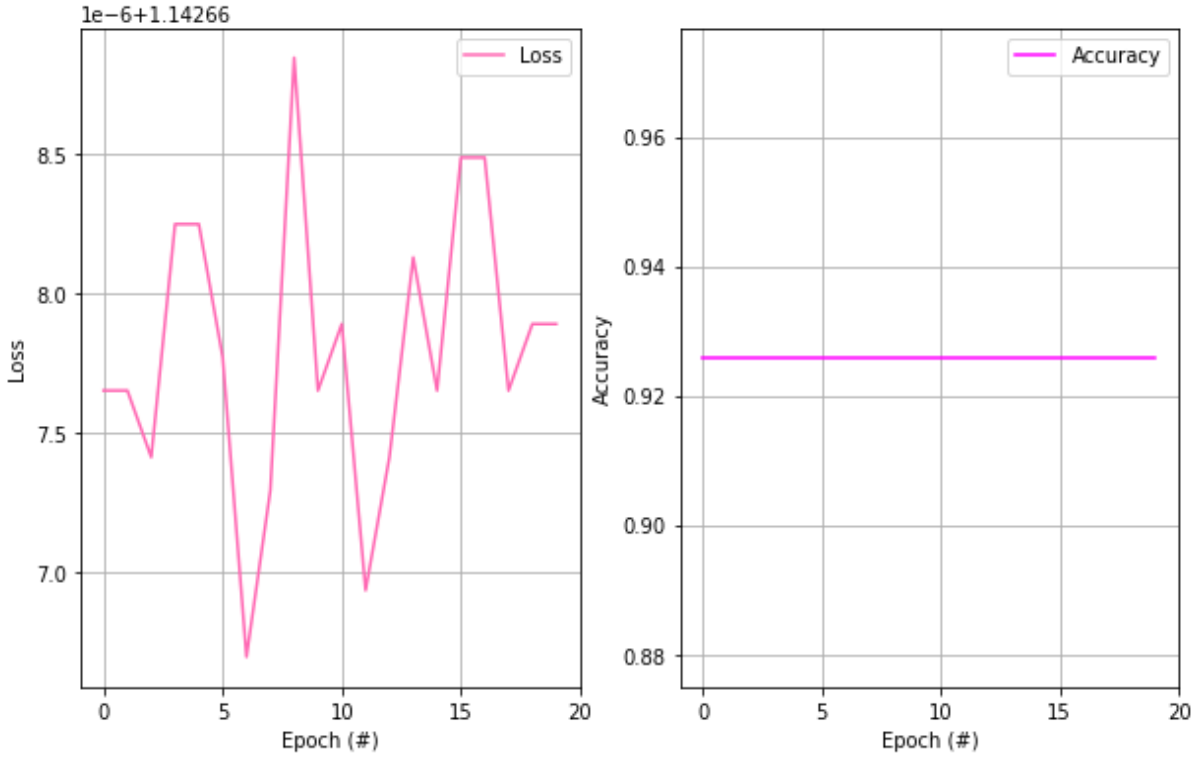


Figure IV.VI. *The loss and accuracy of the CNN model applied to the imbalanced E185B C. difficile genome isolate to predict PhiC2 regions.*

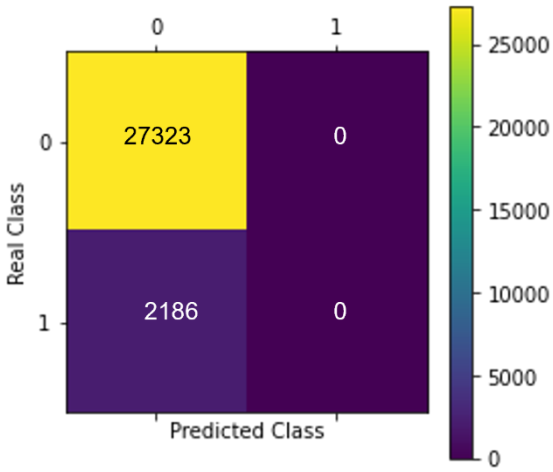


Figure IV.VII. *The correlation matrix of the CNN model applied to the imbalanced raw E185B C. difficile genome isolate to predict PhiC2 regions.*

IV.VI.I.I THE IMPORTANCE OF SMOTE

It is clear from Figure IV.VI and Figure IV.VII the E185B dataset is imbalanced and therefore, affects the training, learning, and predicting binary label of the CNN model. Therefore, Synthetic Minority Oversampling Technique (SMOTE) was employed to handle this problem. As seen from Figure IV.VIII the standard trend of results is seen as described by X. The CNN model loss continuously decreases as epoch training progresses and accuracy continuously increases as epoch training progresses.

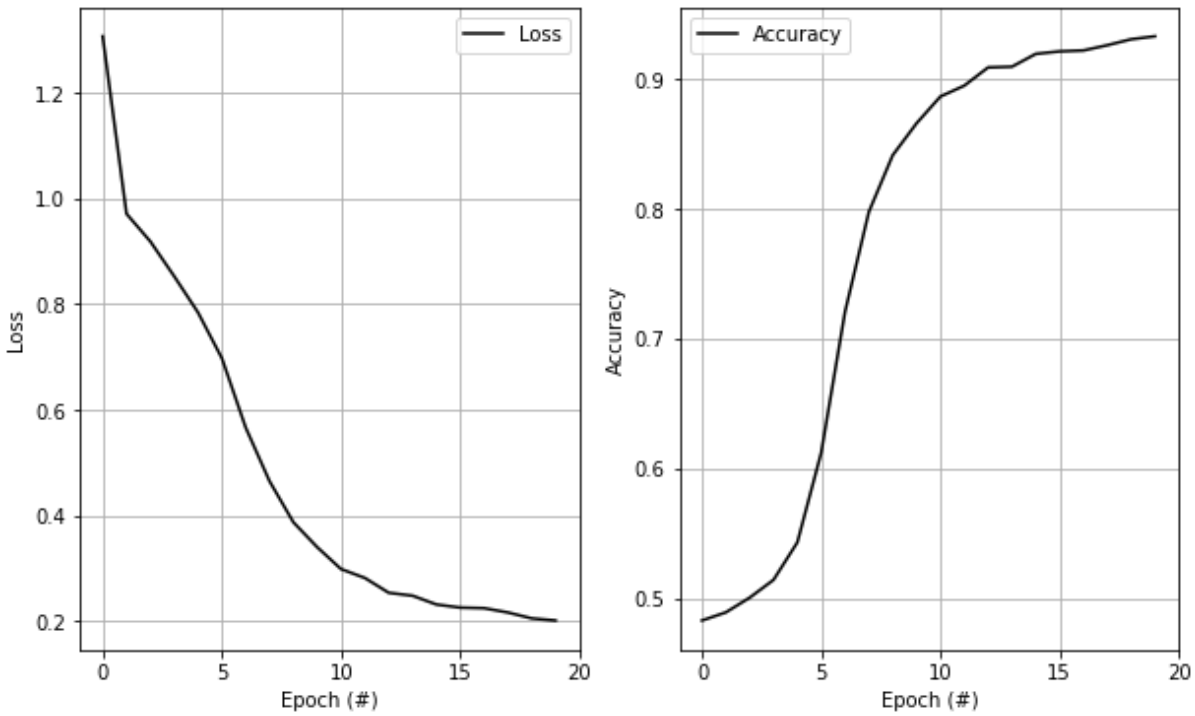


Figure IV.VIII. *The loss and accuracy of the CNN model applied to the balanced E185B C. difficile genome isolate to predict PhiC2 regions.*

To confirm SMOTE has positively affected CNN model label prediction of the balanced *C. difficile* E185B genome the correlation matrices and corresponding ROC curves were produced, as shown in Figure IV.IX.

The correlation matrix results from the training set in Figure IV.IX(A) demonstrate the CNN model successfully predicted 23,306 out of 27,323 training labels as regions containing PhiC2. Furthermore, the CNN model successfully predicted all training labels that did not contain PhiC2. Additionally, the CNN model produced a 'very good' classification outputs shown in Figure IV.IX(B) because the AUC was 0.9663, with this value being so close to a perfect score of 1.0. Therefore, the CNN was appropriate and has successfully categorised 96.63% of the training set.

However, application of the CNN model to the testing set for further analysis, shown in Figure IV.IX(C) observed a decrease in successful prediction of PhiC2 with no labels being correctly predicted. The prediction of regions not containing PhiC2 remained perfect as before. Therefore, when applied to the testing data the CNN predicted all regions as 0. This outcome saw an effect on the AUC in Figure IV.IX(D) with a value of 0.6993.

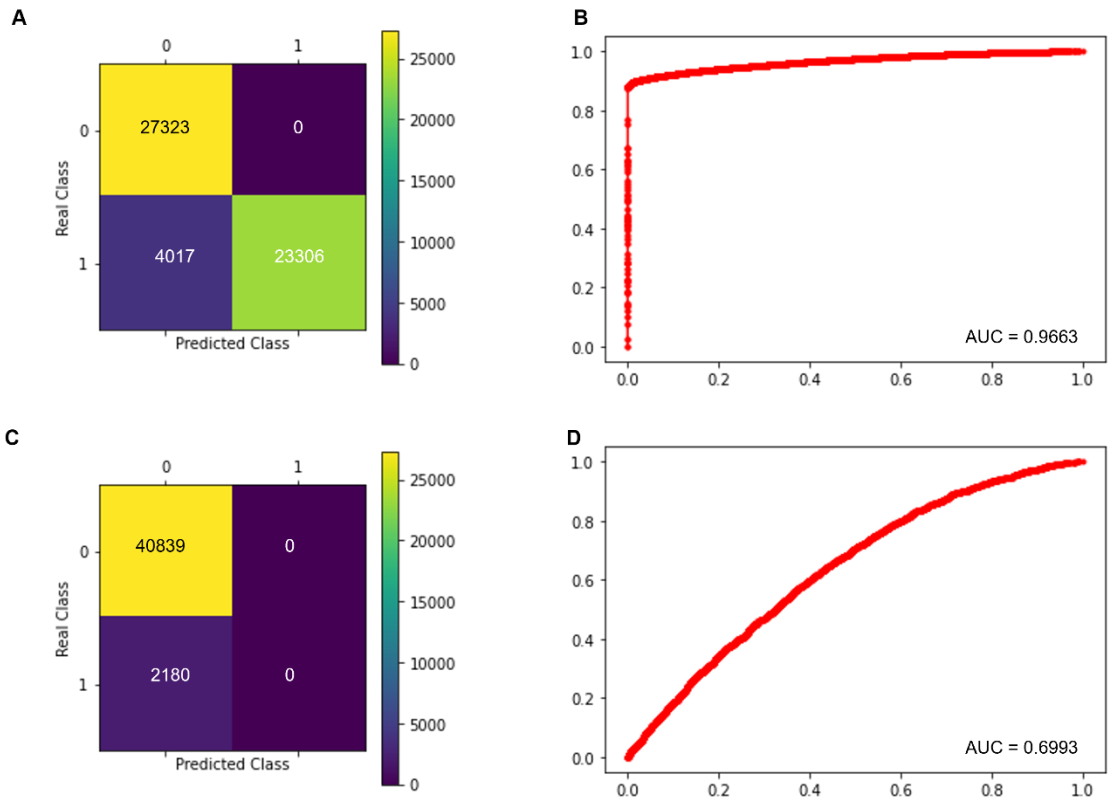


Figure IV.IX. *The correlation matrices and corresponding ROC curves, displaying the AUC, from the CNN model to predict PhiC2 regions in* ***(A,B)*** *of the training set and* ***(C,D)*** *of the testing set. The CNN model was applied to the E185B C. difficile genome isolate.*

IV.VI.I.II IDENTIFICATION OF PHIC2 ON FULL TEST GENOME ISOLATE

The trained CNN model for predicting PhiC2 was applied to the whole genome of S8, an unseen dataset to observe the accuracy. However, as shown from Figure IV.X(A) the CNN model continued to predict every binary label as 0. The corresponding AUC (0.4599) from the ROC was below the y=x line, highlighting this model is an inappropriate fit and is not suitable for further prediction of PhiC2 on unseen test data.

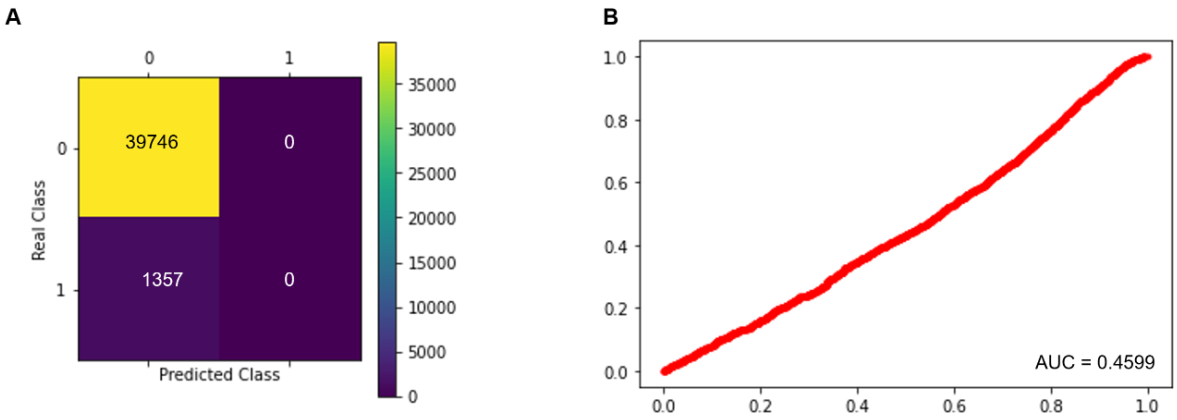


Figure IV.X. *The CNN model was applied to the whole S8 C. difficile genome isolate to predict PhiC2 regions. The following classification evaluation metrics were produced from the CNN predicted output labels* ***A*** *correlation matrix and* ***B*** *ROC curve, displaying the AUC.*

IV.VI.II IDENTIFYING Tn6215 REGIONS

Synthetic Minority Oversampling Technique (SMOTE) was employed to handle the imbalanced classes in the *C. difficile* E185B genome isolate. CNN model loss continuously decreases as epoch training progresses, the loss reached 0.1345. The accuracy continuously increases as epoch training progresses, the accuracy of CNN had an overall value of 96.4%.

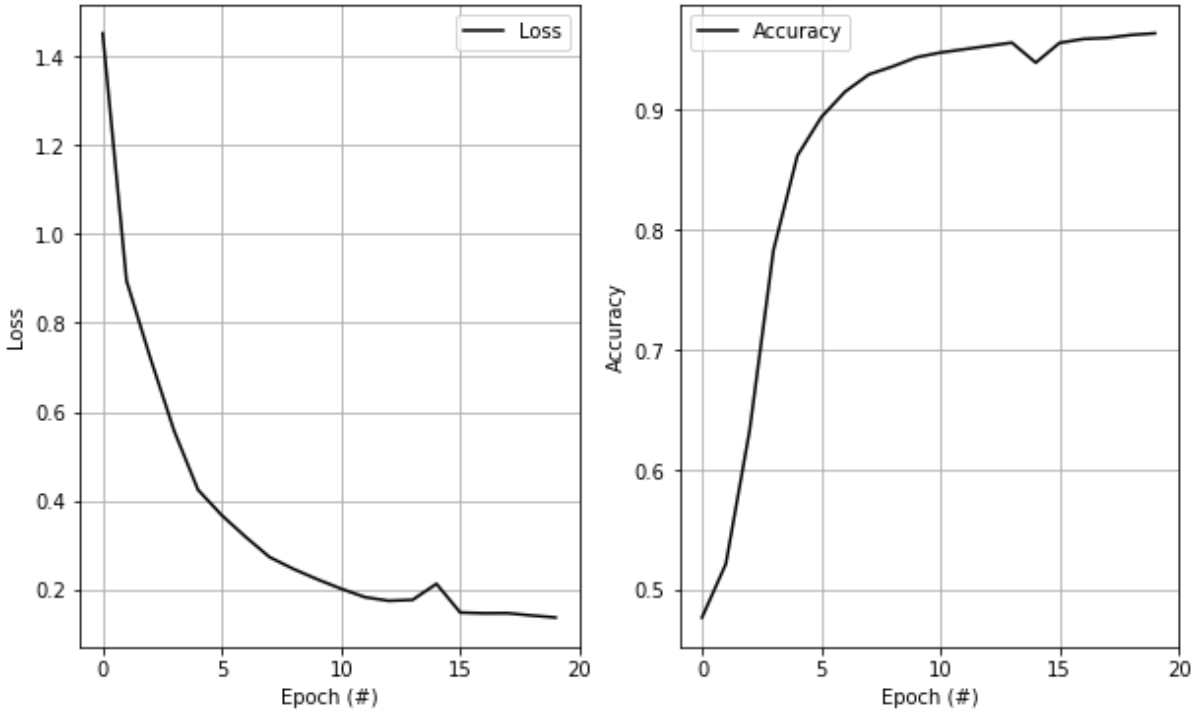


Figure IV.XI. *The loss and accuracy of the CNN model applied to the balanced E185B C. difficile genome isolate to predict Tn6215 regions.*

Figure IV.XII shows the comparison of correlation matrices and corresponding ROC curves from the *C. difficile* E185B genome isolate, applying a CNN to predict Tn6215 regions.

The correlation matrix results from the training set in Figure IV.XII(A) demonstrate the CNN model successfully predicted 27,482 out of 29,295 training labels as regions containing Tn6215. Furthermore, the CNN model successfully predicted 29,154 out of 29,295 training labels as regions not containing Tn6215. Additionally, the CNN model produced a 'very good' classification outputs shown in Figure IV.XII(B) because the AUC was 0.9985 value with this value being so close to 1.0. Therefore, the CNN was appropriate and has successfully categorised 99.85% of the training set.

The CNN model was applied to the testing set for further analysis, it is shown in Figure IV.XII(C) the CNN successfully predicted 29,154 out of 29,295 testing labels as regions not containing Tn6215. However, the accuracy of predicting regions with Tn6215 reduced with only 44 out of 214 regions were successfully labelled as Tn6215. However, the CNN model continued to produce a 'very good' classification of the inputs because the AUC was 0.9679.

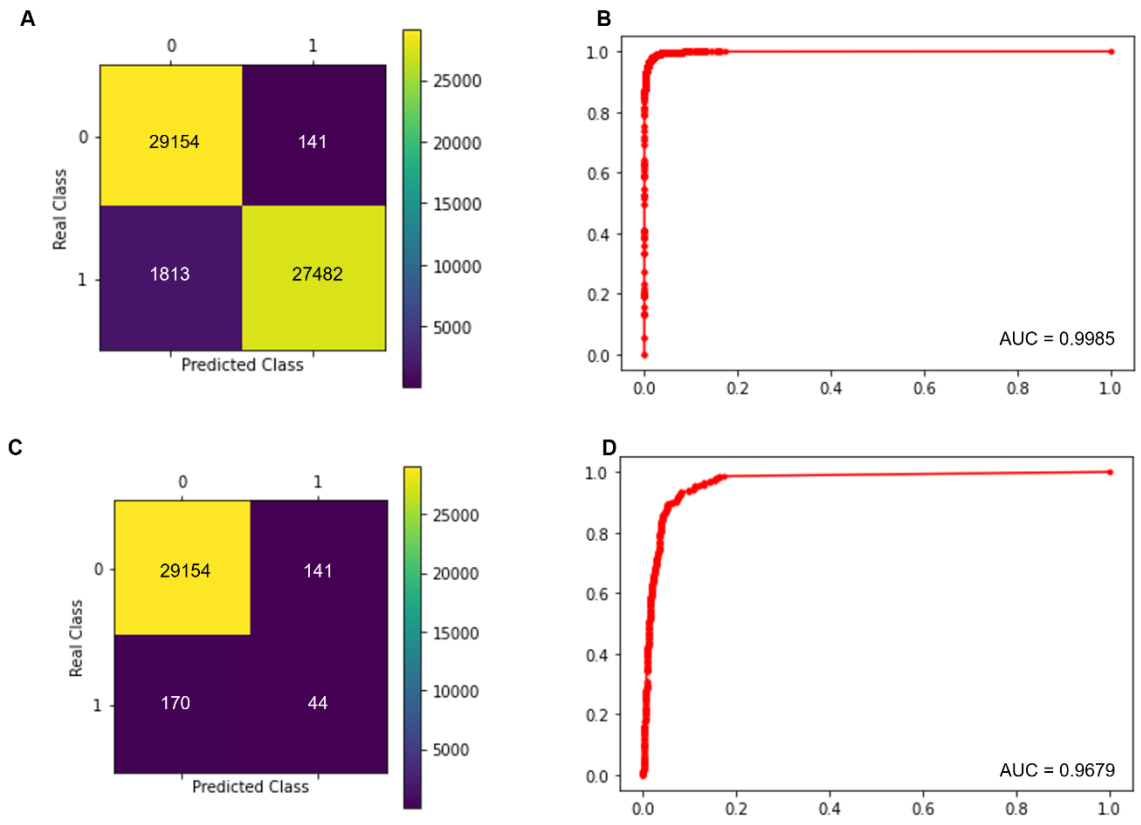


Figure IV.XII. *The correlation matrices and corresponding ROC curves, displaying the AUC, from the CNN model to predict Tn6215 regions in* ***(A,B)*** *of the training set and* ***(C,D)*** *of the testing set.*

IV.VI.II.I IDENTIFICATION OF TN6215 ON FULL TEST GENOME ISOLATE

The trained CNN model for predicting Tn6215 was applied to the whole genome of S8, an unseen dataset to observe the accuracy. The CNN model did not successfully predict any genome regions containing Tn6215. Additionally, the CNN model mis-classified 255 out of 41,094 that did not contain Tn6215 as regions that contained Tn6215. The corresponding AUC (0.6464) highlighting this model is an inappropriate fit and is not suitable for further prediction of Tn6215 on unseen test data.

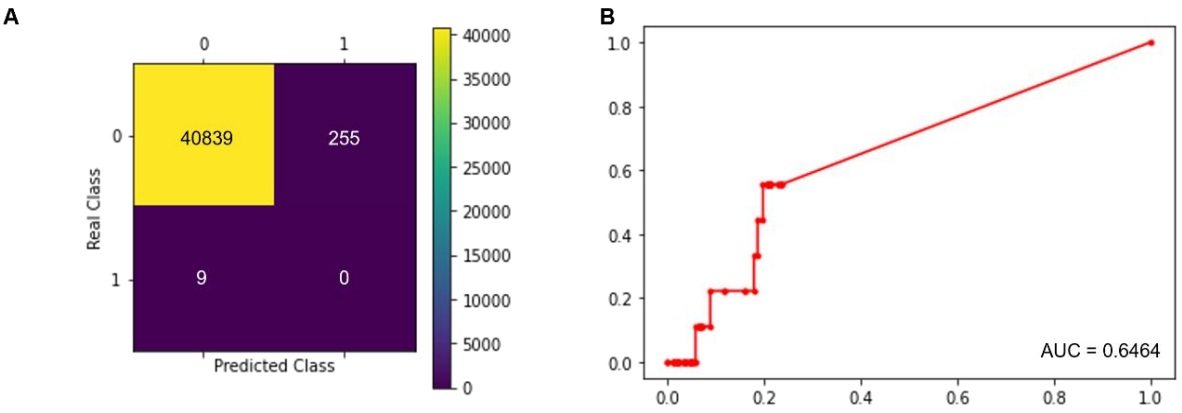


Figure IV.XIII. *The CNN model was applied to the whole S8 C. difficile genome isolate to predict Tn6215 regions. The following classification evaluation metrics were produced from the CNN predicted output labels* ***A*** *correlation matrix and* ***B*** *ROC curve, displaying the AUC.*

CHAPTER V: DISCUSSION

V.I INTRODUCTION

The purpose of this thesis study has been to add to the body of knowledge relating to the potential contribution bacteriophages have on the facilitation of antimicrobial resistant genes. More specifically, it was anticipated that the research would assist in the bioinformatical practice of HGT evaluation to move towards being a more effective medium for machine learning algorithms. This chapter will draw together and summarise the key findings of discussion relating to these central themes and will clearly outline how this research project has met the goals described above. This chapter is divided into four parts.

Part one will discuss the clinical significance the results of this study.

Part two of the chapter, discussion will identify the presence of PhiC2 and Tn6215 pairings within the data. To begin, the sequence identification pipeline encapsulated in Figure III.II in Chapter III will be used to provide an interpretive framework to begin to weave together a conclusion of the hypothesis. The research hypothesised PhiC2 is necessary for the transfer if erythromycin resistance through the transduction of Tn6215 in the bacterial species *C. difficile*. Part two will also revisit the discussion of the pathophysiological properties of PhiC2 as a central dynamic to explain the relationship between PhiC2 and Tn6215.

Part three of the chapter revisits the objective to compare machine learning and deep learning algorithms to the 'gold standard' to identify PhiC2 and Tn6215 genomic sequences within *C. difficile*. The categorical data required augmentation to enable statistical and mathematical processes to be applied in machine learning algorithms. Therefore, the speed and accuracy of these methods were compared using a small training set prior to being applied to the full dataset. The 'gold standard' method will continue to be used as ground truth for comparisons and contextualise discussion.

The final part of this chapter presents limitations that have been regarded during the research project and reflections of the results that might play a role in future research.

V.II PART ONE: CLINICAL SIGNIFICANCE OF THE PROJECT

The pathogenicity of *C. difficile* continues to grow rapidly owing to acquisition of AMR genes, a direct result from the development, widespread use, and poor prescribing practices of antibiotics (Spigaglia, 2016). Consequently, this has led to horizontal gene transference of AMR through various mechanisms; most novel being transduction through bacteriophages. Despite the devastating effects *C. difficile* infection causes to individuals and their families, private and government budgets have restricted the understanding and implication of AMR genes passed on via transduction. This emphasises the paucity of research utilising bacteriophages for the genetic manipulation of *C. difficile* and applied in drug discovery.

Furthermore, since the onset of the COVID-19 pandemic in 2019, approximately 72% of COVID-19 patients were treated with antibiotics, including erythromycin, to prevent co-infections of bacterial species (Sheikh et al., 2021). In turn, *C. difficile* infections and recurrence are projected to increase. Additionally, *C. difficile* co-infection in patients with COVID-19 has been demonstrated to worsen the course and prognosis of COVID-19 (Wee et al., 2021) as well as being novel risk factors for more serious co-morbidities including acute portal vein thrombosis.

V.III PART TWO: ADDRESSING THE NOVEL HYPOTHESIS PHIC2 FACILITATES HORIZONTAL GENE TRANSFER OF TN6215

V.III.I INVESTIGATING THE DEPENDENCE OF TN6215 ACQUISITON VIA PHIC2 IN *Clostridium difficile*

A 'gold standard' method was implemented in this research to investigate the primary hypothesis that PhiC2 mediates transduction of Tn6215 between *C. difficile* strains. The results of the study do not support this hypothesis because only 2 out of the 5 genome isolates (E185B and S8) had significant PhiC2-Tn6215 pairings (Figure III.III). Even though Tn6215 and PhiC2 regions were identified in all five genome isolates. However, it is important to evaluate the findings in the wider context of the developing field.

Despite Goh et al. (2013), being the first report of gene transfer by bacteriophage transduction between *C. difficile* strains only one of the four strains tested as recipients acquired the AMR gene *ermB,* despite all four strains being susceptible to PhiC2 infection. A similar result was reported by Mazaheri et al. (2011) and Battaglioli et al. (2011), showing bacterial cell susceptibility to infection from a transducing phage does not necessarily result in transduction. However, it should be highlighted these results were observed in *Escherichia coli* species and did not characterise PhiC2 as a bacteriophage capable of infecting this bacteria. Hence, these previous studies findings support the current research findings that Tn6215 is not dependent on transduction of PhiC2 between *C. difficile* strains.

Previous bioinformatic studies suggest the role that bacteriophages play in transduction of AMR genes remains controversial. Enault et al. (2017) applied several bioinformatic approaches to identify the presence of AMR genes within bacteriophage genomes caused by generalised transduction. The study concluded AMR genes are rarely encoded in bacteriophages owing to low homologies as well as matches to proteins unrelated to AMR. Additionally, Allen et al. (2011) used metagenomic to evaluate the effects of two antibiotics on intestinal swine bacteriophage viromes. They observed the frequency of bacteriophages was increased in swine cells medicated with antibiotics as opposed to non-medicated. Therefore, this suggests that whilst antibiotics did induce the over-expression of bacteriophage genes, did not confer HGT of AMR genes from the bacteriophage to bacteria in swine intestines.

Klumper et al. (2014) evaluated the possibility of HGT of MGEs by a conjugation-like mechanism, without the requirement of bacteriophages for transduction. The study found several mobilizable plasmids in the IncQ group were capable of conjugation by filter mating in *Pseudomonas putida* causing susceptibility to antibiotics. Similarly, Gupta et al. (2003) described the *ermB* element can be transferred between *Bacteroides* species through a conjugation, resulting in *ermB* presence and

absence of bacteriophages. However, it should be noted, this research was performed only on *C. difficile* species but conjugation-like mechanisms would explain the cause of the presence of Tn6215 in all five of the genome isolates but at a much lower frequency than PhiC2. Further reinforcing this statement, Goh et al., (2013) proposed the inability of PhiC2 to transfer Tn6215 to all the other *C. difficile* strains may be owing to the absence of a transposon integration site. However, the transposon integration site in these laboratory strains was confirmed using PCR. Therefore, this theory may need to be confirmed through further research to indicate whether the integration site remains intact to enable transduction to occur.

Although, previous studies suggest a direct association between bacteriophages mediating transduction of AMR genes. Goh et al. (2013) previously described erythromycin resistance contained within Tn6215 was transferred by PhiC2 from *C. difficile* strain CD80 to CD062. Similarly, Marinus and Poteete (2014) reported Stx-converting transducing bacteriophage mediated transfer of tetracycline resistance genes by transduction in from *Escherichia coli* (*E. coli*) 0157:H7 to laboratory strain *E. coli* K-12. Furthermore, Bearson et al. (2014) investigated a mixture of bacteriophages transduced SGI1 to *Salmonella typhimurium* DT104 strain with sensitivity to ampicillin and tetracycline owing to an internal deletion of SGI1. This study observed the transduction of tetracycline and ampicillin resistance. Therefore, despite this research provides reasonable evidence not supporting the concept PhiC2-mediated transduction of Tn6215 is a contributing factor to the dissemination of antibiotic resistance in *C. difficile*. Given the ubiquity, abundance and resilience of phages and the role of the environment as a reservoir, the contribution of transduction to the spread of AMR must be considered.

Unlike previous studies that have investigated bacteriophages facilitating AMR gene transduction the current research applied a 'gold standard' method of utilising the sequence similarity algorithms PHASTER and BLAST. Ecovoiu et al. (2016) applied an algorithm harnessing a multi-step implementation of a Smith-Waterman algorithm to compute the mapping alignments of transposons within *Drosophila melanogaster* (the fruit fly). However, the accuracy of the Smith-Waterman algorithm was affected by mutations and sequencing artefacts. Therefore, this approach lacked robustness when compared to BLAST, applied in this research. Accordingly, the findings of the current study from the PHASTER and BLAST alignments generated rich, valid data on the genomic regions of PhiC2 and Tn6215.

V.III.II EXPLORING THE PATHOPHYSIOLOGICAL ROLES OF PHIC2 CONTRIBUTING TO THE TRANSDUCTION OF TN6215

Even though this current study has not appeared to have determine the transduction of Tn6215 is not dependent on PhiC2. Despite the results from this research, the next logical step in understanding these results was the exploration of the pathophysiological roles of PhiC2 and the role PhiC2 has in transducing Tn6215.

Studies regarding the identification of PhiC2 in *C. difficile* and the associated pathophysiological properties can aid understanding of transduction described in this current research. PhiC2 was partially sequenced and characterised to show an increase of toxin B levels in *C. difficile* lysogens (Goh et al., 2005). This previous study supports the current research findings that the current role of PhiC2 may not be in generating genetic diversity amongst bacterial species but perhaps in other areas of the host genome related to virulence. However, Goh et al. (2007) demonstrated there was no correlation between the presence of toxin A or toxin B gene overexpression and PhiC2-related prophage genes in *C. difficile*. Despite this, Goh et al. (2007) established PhiC2 can influence the expression of gene transcriptional regulator *gnt*R-like which can alter *C. difficile* cell fitness.

The development of PHASTER has improved the identification of complete bacteriophages in *C. difficile* bacterial genomes. PHASTER also identified decaying, incomplete bacteriophage remnants that have lost their conserved gene components. The research identified 60% of the incomplete bacteriophage regions contained nucleotide bases statistically significant to PhiC2. This is relevant because these remnants may remain to potentially influence *C. difficile* host cells despite their inability to replicate. Kirk et al., (2017) demonstrated Diffocins, bacteriophage tail-like particles, are able to kill their bacterial host following induction and cell lysis but also kill other competing cells in the surrounding environment, preventing the further production of virulence particles. Although, similar theories remain to be drawn including remnant PhiC2 bacteriophage regions, but they may possibly provide a competitive advantage to *C. difficile* strains carrying them.

The various pathophysiological properties of bacteriophages described by previous studies emphasise the potential application of bacteriophages in medicine that have manipulated genomes to cause immediate bacterial cell lysis.

V.III.III UNDERSTANDING THE RELATIONSHIP BETWEEN PHIC2 AND TN6215

Whilst the current study did not focus on the direct effects of the varying of transposon regions, an unexpected relationship was potentially observed between the number of Tn6215 and PhiC2 regions. The research explored this correlation through a correlation analysis however the results ($r=0.587$, $p>0.05$) did not support the theory of a positive correlation between these two variables. However, Goh et al. (2013) first proposed this theory; suggesting Tn6215 and PhiC2 have a positive correlation, regardless of nucleotide position in the genome. This is because transposons have a unique 'jumping' characteristic being able to translocate to another position within *C. difficile* genomes. Owing to the results of this study, despite the above proposed theory, this relationship does not occur because, as highlighted in the previous section, transduction of MGEs appear to be dependent on optimal conditions such as the presence of integration sites.

V.IV PART THREE: COMPARISON OF DEEP LEARNING ALGORITHMS

V.IV.I UTLISING SEVERAL NUCLEOTIDE AUGMENTATION METHODS TO CHANGE CATEGORICAL DATA

Data augmentation methods were employed in this study to evaluate the most appropriate method; one hot encoding, label encoding, and *k*-mer encoding, to alter the categorical DNA data into numerical data to be applied in DL algorithms. Specifically, the encoded sequences were meant for CNN learning which expect the input samples to be in vector or matrix format. The results of the study established One Hot encoding was the most appropriate method because this method was able to augment the DNA into a suitable format in the quickest amount of time in three out of the five *C. difficile* genome isolates. Furthermore, the mean time taken per loop was also the shortest amount of time to execute the augmentation action (mean time taken per loop = 27.28 ns, SEM = 0.4826 ns). In confirmation with the previous research, there is a consensus of thought that One Hot encoding is the most innovative and easily reproducible method for DNA sequences (Kircher et al., 2022; Lee et al., 2021; Pezoulas et al., 2021).

Unfortunately, the research did not have enough time allocated to apply each individual DNA augmentation method to the CNN to address the potential change in accuracy. However, Choong and Lee (2017) showed that label encoding and one hot encoding methods had almost the same running time in a CNN model (89 minutes and 38.2 seconds vs 87 minutes 46.9 seconds) because the number of computations required at both pooling and coevolutionary layers grows with the length of the vector. However, since the implementation of *k*-mer encoding requires an additional word embedding layer for CNN implementation, it is hard to compare the running time for this tool (Gunasekaren et al., 2021). Additionally, this layer requires excess memory storage because all the *k*-mers generated are concatenated to form a sentence.

Although, it has not always been observed One Hot encoding is the most appropriate method for application to all machine learning algorithm. This is because the DNA augmentation method applied is dependent on the DNA evaluation problem. Ghandi et al. (2014) discovered that *k*-mer augmentation applied to generate short sequences were useful for the construction of a classification model. This was similar to the resulting binary classifier produced in this research because the data was split into *k*-mers of size 100 after one hot encoding. Whereas, Lee et al. (2011) found the application of label encoding to be important for nucleotide base location identification. Furthermore, Colbran, Chen & Capra (2017) explained this phenomena is caused by the conversion function applies hierarchal weighting, affecting algorithm training.

V.IV.II THE APPLICATION OF THE CONVOLUTIONAL NEURAL NETWORK TO IDENTIFY SPECIFIC REGIONS WITHIN THE GENOME

A deep learning-based convolutional neural network (CNN) classifier was trained using the *C. difficile* strain E185B because a PhiC2-Tn6215 pairing was identified in this dataset using the 'gold standard' DNA identification pipeline. The CNN was employed in this study because this DL method is widely recognised and has been shown to outperform the traditional BLAST classification in the current literature (Fadja et al., 2021; Onimaru et al., 2019). Experimental results show that imbalanced datasets affected feature identification of PhiC2 in training sets. The CNN model had an accuracy of 92.5% however, further evaluation matrices highlight the CNN was predicting every class as 0. Hence, the accuracy was high because the majority of the training class was 0. Therefore, SMOTE was applied to both the Tn6215 and PhiC2 classification datasets to increase the number of minority classes. The CNN model appeared to determine patterns in the E185B training set for both Tn6215 and PhiC2, with both training sets applied achieving high accuracies 93.3% and 96.4%, respectively. Additionally, both models had high AUC 0.9663 and 0.9985, respectively. However, the trained CNN model unsuccessfully classified any of Tn6215 or PhiC2 regions in the unseen test data, *C. difficile* S8 genome. The corresponding AUCs also highlighted the model is an inappropriate fit and is not suitable for further prediction of Tn6215 or PhiC2 on unseen test data.

It is important to explore the issues in training a CNN model for genome region identification. Hinton et al. (2012) study overfitting problems in a CNN where the model fits too well to the training set. Overfitting indicates the model is too complex for the research problem because the model fits too well to the training set (Hedyehzadeh et al., 2020). This causes the CNN model to have difficulty generalising new examples that were not in the training set (Hinton et al., 2012). This can be shown to have occurred in this research because the CNN model has recognised specific nucleotide base sequences instead of general nucleotide patterns. Therefore, random dropout following each feature selection layer (25%) and the first dense layer (50%) was employed to prevent complex co-adaptations, so each neuron only learns to detect a feature that is generally useful to predict the correct binary class.

Several studies have shown class imbalance can affect CNN performances (Bria et al., 2020; Qu et al., 2020; Wu et al., 2020). This is because the CNN assumes data is equally distributed among classes therefore the classifier has bias towards the majority class, causing bad classification of the minority class. Additionally, Hinton et al. (2012) suggest overfitting in a large feedforward neural network arises when a network is trained on a small training set. The training set in this research was X and X for PhiC2 and Tn6215, respectively. Thus, the research applied SMOTE to balance the minority class by generating synthetic data points rather than duplication of the original dataset.

Previous bioinformatic studies suggest the properties and characteristics of the DNA sequences can contribute to the effectiveness of training a CNN model. Goh et al. (2013) observed PhiC2 has a mosaic genome structure; meaning each genome is a unique composite of gene modules interspaced with non-coding spacer regions. This highlights the CNN will have difficulty extracting features specific to PhiC2 alone. Additionally, Zheng et al. (2019) demonstrated datasets containing nucleotide-only bases can affect CNN optimisation. This is because the datasets only contain four nucleotide bases: A, T, G and C. Therefore, the CNN may not be able to derive abstract features to form hierarchical relationships between regions.

Despite the results of this research, several previous studies have proposed CNNs are able to predict and identify genome regions accurately. However, owing to the novelty of bacteriophage identification techniques there is an absence of direct research utilising CNNs to identify bacteriophage regions. Fadja et al. (2021) investigated whether a genomic region shows a pattern which is more consistent with a natural selection, or a neutral model using a CNN and achieved an accuracy of almost 90%. Furthermore, Onimaru et al. (2019) successfully predicted gene regulatory regions with a proposed CNN model (AUC = 0.9623). The CNN employed in these studies used a logarithmic loss function was used to classify the regions. This is because the research used datasets with several classes whereas this research only had two outcomes thus a binary crossentropy loss function was applied. However, Medium (2022) demonstrates the loss function is independent of the CNN architecture and is instead dependent on the goal of the network.

V.IV.III THE CONVOLUTIONAL NEURAL NETWORK IN COMPARISON TO THE 'GOLD STANDARD' METHOD

The final aim of the research was to compare the CNN model to the 'gold standard' method in the identification of PhiC2 and Tn6215 regions in the *C. difficile* genomes. The experimental results have shown the 'gold standard' method remains to have the best accuracy compared to the CNN model. However, it is important to evaluate other metrics that can contribute to the most appropriate method including reproduction.

Regarding reproducibility, the CNN model can be applied to any future dataset and is not dependent on the size of the genomes only the balance of the class labels. Although, the CNN requires user training owing to utilisation in Python. The web-interface of BLAST limits the query and input sequence to 100,000 nucleotide bases; hence, the software was downloaded externally. This research only required one library to be downloaded however, future research may require numerous libraries for genome identification and thus, large computer storage is required. Whereas the PHASTER API and web-interface are not labour intensive or require user training.

Overall, the research has found the CNN model has greater efficiency and reproducibility however as a result the accuracy is reduced. Whereas the 'gold standard' method has high accuracy because it is computationally heavy, but causes an increase in time and decreased reproducibility for large genomes. Therefore, the size of the genome depends on the method used.

V.V PART FOUR: LIMITATIONS AND IMPLICATIONS FOR FUTURE DIRECTIONS

V.V.I LIMITATIONS AND DELIMINATORS

Whilst the current study did not observe PhiC2 facilitates HGT of Tn6215 through transduction in *C. difficile* genomes, there are several limitations that need to be addressed before the hypothesis can be rejected.

It should be noted the research was completed only on Australian/ New Zealand *C. difficile* bacterial strains. Therefore, other *C. difficile* strains from other continents e.g., European or North American, were not measured to test this hypothesis. Additionally, owing to the novelty documentation of PhiC2 complete genome it is undetermined whether the bacteriophage is limited to solely *C. difficile* or capable of infecting other bacterial species that reside in gut microflora such as *Escherichia spp.* or *Pseudomonas spp.* Thus, these may impact the results found and would need to be accounted for when comparisons are made between studies that have recorded bacteriophages enabling HGT of mobile genetic elements.

Although the current study did not find significant correlations between the number of PhiC2 and Tn6215 regions, owing to the correlation design adopted, the direction of causality cannot be inferred. For example, it cannot be assumed that increased numbers of PhiC2 did not *cause* a greater increase in Tn6215; this highlights a central limitation of this aim. This is because as previous research has shown, Tn6215 is a transposon capable of 'jumping' to a new position in the genome (Goh et al., 2013).

Choosing the E185B dataset introduced restricted performance outcomes from the CNN. This is because the dataset had unbalanced classes. It is well-known that CNN models cannot be effective until there are sufficient number of elements in a particular class. Therefore, initially the model did not yield optimal results for the identification of either PhiC2 or Tn6215, the unbalanced class, because the CNN did not identify sufficient features for separation. However, this research applied SMOTE to balance the classes to strive for maximum overall performance.

V.V.II RECOMMENDATIONS FOR FUTURE RESEARCH

There are many directions that future research might take, and as a qualitative researcher, it is important to recognise that readers will have most certainly identified opportunities that have not been considered. However, the discussions that have taken place have signalled several key opportunities for further research relating to the development of bacteriophages facilitating transference of antimicrobial resistance genes and different machine learning algorithms to identify nucleotide sequences in genomes. Therefore, the key opportunities for further research are described below.

This research only conducted tests on the identification of PhiC2-Tn6215 pairings within Australian/ New Zealand *C. difficile* strains. Repeating the results of the identification of PhiC2-Tn6215 pairings from this thesis on other *C. difficile* donor strains from different regions around the globe to observe if the phenomenon was the same or changed. This relationship would be interesting to apply to Asian *C. difficile* strains because this region is a close neighbour to Australia and therefore, proximity may contribute to transduction of antimicrobial resistance between neighbouring geolocated species.

Despite the promising results obtained by the CNN in chapter V, bridging the gap between DNA sequence identification and machine learning remains an open problem for sequence algorithm tools. For example, sequence algorithms targeting Python code need to support a wider range of input types such as, strings and categorical data. While work has been done to support categorical inputs in testing, mostly outside of the DNA field using proteomics, no previous work in the field of pairwise sequence algorithms addresses the problem of DNA nucleotide bases in Python code, even though they are commonly used programming paradigms in real-world code.

Additionally, this research only performed tests on CNN. The same tests for identifying Tn6215 and PhiC2 could be conducted with a Long short-term memory (LSTM), a recurrent neural network, capable of continuous learning and retaining of information owing to loops in the network (Olah, 2015).

CHAPTER VI: CONCLUSION

This chapter presents concluding thoughts and reflections that have been gained throughout the preliminary research into the newly emerging theory of PhiC2 mediating the transduction of the erythromycin resistant gene Tn6215 in *C. difficile* species and therefore is limited in some respects.

The experimental results utilising the 'gold standard' DNA identification pipeline, comprising of PHASTER API and BLAST, did not support the primary hypothesis of this research that PhiC2 mediates the transduction of Tn6215, conferring erythromycin resistance. Owing to the paucity of research regarding the transducing ability of PhiC2, further work is required for evaluation of this novel hypothesis.

Utilising the number of Tn6215 and PhiC2 regions in each *C. difficile* genome provided a useful insight to determine the secondary aim, that there is a positive relationship between the number of Tn6215 regions and PhiC2. Therefore, a corelation analysis was applied, unfortunately the analysis did not show a significant relationship between these variables.

Furthermore, the research explored the application of convolutional neural networks, an automated biological deep learning algorithm, to predict regions of *C. difficile* containing PhiC2 or Tn6215. Three DNA augmentation methods; one hot encoding, label encoding, and *k*-mer encoding, were compared for speed and reproducibility. The research found that one hot encoding was the most appropriate method. The CNN with imbalanced datasets affects the accuracy of predicting unseen test data owing to overfitting the training data. Thus, SMOTE was applied to handle this problem. The dataset could not be evaluated only with accuracy metrics. Other metrics including, sensitivity, and specificity have also been considered. The CNN for PhiC2 and Tn6215 performed well in training but surprisingly; the testing accuracies were low. Additionally, it was observed the PhiC2 CNN when applied to test data continued to categorise each class as 0; this was shown to be because bacteriophages are mosaic, meaning each genome is a unique composite. Despite the CNN results, application of a CNN will be useful for future directions exploring automation of biological sequence identification.

As a concluding statement, all the aspects investigated in this research have further reinforced the requirement to understand both the pathological relationship between PhiC2 and Tn6215 but also, the constituents involved and initiated in transduction. As well as adapting methods to apply machine learning and deep learning algorithms for the prediction and identification of potentially virulent genome regions.

BIBLIOGRAPHY

Abatángelo, V., Peressutti Bacci, N., Boncompain, C. A., Amadio, A. F., Carrasco, S., Suárez, C. A., & Morbidoni, H. R. (2017). Broad-range lytic bacteriophages that kill Staphylococcus aureus local field strains. PloS one, 12(7), e0181671.

Abbas, F., & Vinberg, F. (2021). Transduction and Adaptation Mechanisms in the Cilium or Microvilli of Photoreceptors and Olfactory Receptors From Insects to Humans. Frontiers in cellular neuroscience, 15, 662453.

Allen H. K., Looft T., Bayles D. O., Humphrey S., Levine U. Y., Alt D., et al. (2011). Antibiotics in feed induce prophages in swine fecal microbiomes. mBio 2:e260-11. 10.1128/mBio.00260-11

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed Gish, W. & States, D.J. (1993)

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), W16–W21.

Aspland, E., Harper, P. R., Gartner, D., Webb, P., & Barrett-Lee, P. (2021). Modified Needleman-Wunsch algorithm for clinical pathway clustering. Journal of biomedical informatics, 115, 103668.

Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I., & Koonin, E. V. (2020). Seeker: alignment-free identification of bacteriophage genomes by deep learning. Nucleic acids research, 48(21), e121.

Bai, Z., Zhang, Y. Z., Miyano, S., Yamaguchi, R., Fujimoto, K., Uematsu, S., & Imoto, S. (2022). Identification of bacteriophage genome sequences with representation learning. Bioinformatics (Oxford, England), btac509. Advance online publication.

Battaglioli, E. J., Baisa, G. A., Weeks, A. E., Schroll, R. A., Hryckowian, A. J., & Welch, R. A. (2011). Isolation of generalized transducing bacteriophages for uropathogenic strains of Escherichia coli. Applied and environmental microbiology, 77(18), 6630–6635.

Bayat A. (2002). Science, medicine, and the future: Bioinformatics. BMJ (Clinical research ed.), 324(7344), 1018–1022.

Bearson, B. L., Bearson, S. M., & Kich, J. D. (2016). A DIVA vaccine for cross-protection against Salmonella. Vaccine, 34(10), 1241–1246.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. Nucleic acids research, 41(Database issue), D36–D42.

Boeckaerts, D., Stock, M., Criel, B. et al. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep 11, 1467 (2021).

Bria, A., Marrocco, C., & Tortorella, F. (2020). Addressing class imbalance in deep learning for small lesion detection on medical images. Computers in biology and medicine, 120, 103735.

Brittain D. C. (1987). Erythromycin. The Medical clinics of North America, 71(6), 1147–1154.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., & Rajandream, M. A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics (Oxford, England), 24(23), 2672–2676.

Chaudhary N, Sharma AK, Agarwal P, et al. 16S Classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. PLoS ONE. 2015;10:e0116106.

Che, Y., Yang, Y., Xu, X., Břinda, K., Polz, M. F., Hanage, W. P., & Zhang, T. (2021). Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. Proceedings of the National Academy of Sciences of the United States of America, 118(6), e2008731118.

Chiang, Y. N., Penadés, J. R., & Chen, J. (2019). Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLoS pathogens, 15(8), e1007878.

Choong, Allen & Lee, Nung Kion. (2017). Evaluation of Convolutionary Neural Networks Modeling of DNA Sequences using Ordinal versus one-hot Encoding Method.

Cieplak, T., Soffer, N., Sulakvelidze, A., & Nielsen, D. S. (2018). A bacteriophage cocktail targeting Escherichia coli reduces E. coli in simulated gut conditions, while preserving a non-targeted representative commensal normal microbiota. Gut microbes, 9(5), 391–399.

Colavecchio, A., Cadieux, B., Lo, A., & Goodridge, L. D. (2017). Bacteriophages Contribute to the Spread of Antibiotic Resistance Genes among Foodborne Pathogens of the Enterobacteriaceae Family - A Review. Frontiers in microbiology, 8, 1108.

Dasari, C.M., Bhukya, R. (2022). Explainable deep neural networks for novel viral genome prediction. Appl Intell 52, 3002–3017.

Daubin, V., & Szöllősi, G. J. (2016). Horizontal Gene Transfer and the History of Life. Cold Spring Harbor perspectives in biology, 8(4), a018036.

Ecovoiu, A. A., Ghionoiu, I. C., Ciuca, A. M., & Ratiu, A. C. (2016). Genome ARTIST: a robust, high-accuracy aligner tool for mapping transposon insertions and self-insertions. Mobile DNA, 7, 3.

Enault F., Briet A., Bouteille L., Roux S., Sullivan M. B., Petit M.-A. (2017). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. ISME J. 11 237–247. 10.1038/ismej.2016.90

Fadja, A., Riguzzi, F., Bertorelle, G., & Trucchi, E. (2021). Identification of natural selection in genomic data with deep convolutional neural network. BioData mining, 14(1), 51.

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. GigaScience, 8(6), giz066.

Farrow, K. A., Lyras, D., & Rood, J. I. (2001). Genomic analysis of the erythromycin resistance element Tn5398 from Clostridium difficile. Microbiology (Reading, England), 147(Pt 10), 2717–2728.

Feng, T., Leptihn, S., Dong, K., Loh, B., Zhang, Y., Stefan, M. I., Li, M., Guo, X., & Cui, Z. (2021). JD419, a Staphylococcus aureus Phage With a Unique Morphology and Broad Host Range. Frontiers in microbiology, 12, 602902.

Flück, B., Mathon, L., Manel, S. et al. Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem. Sci Rep 12, 10247 (2022).

Fortier L. C. (2018). Bacteriophages Contribute to Shaping Clostridioides (Clostridium) difficile Species. Frontiers in microbiology, 9, 2033.

Ghandi, D. Lee, M. Mohammad-Noori, M. A. Beer, T. Manolio, M. Maurano, R. Humbert et al., (2014). "Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features," PLoS Computational Biology, vol. 10, no. 7, p. e1003711.

Goh, S., Riley, T. V., & Chang, B. J. (2005). Isolation and characterization of temperate bacteriophages of Clostridium difficile. Applied and environmental microbiology, 71(2), 1079–1083.

Goh, S., Ong, P. F., Song, K. P., Riley, T. V., & Chang, B. J. (2007). The complete genome sequence of Clostridium difficile phage phiC2 and comparisons to phiCD119 and inducible prophages of CD630. Microbiology (Reading, England), 153(Pt 3), 676–685.

Goh, S., Hussain, H., Chang, B. J., Emmett, W., Riley, T. V., & Mullany, P. (2013). Phage φC2 mediates transduction of Tn6215, encoding erythromycin resistance, between Clostridium difficile strains. mBio, 4(6), e00840-13.

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Computational and mathematical methods in medicine, 1835056.

Gupta, V., & Oliver, B. (2003). Drosophila microarray platforms. Briefings in functional genomics & proteomics, 2(2), 97–105.

Haneef, D.T.K., Jazir (n.d.). Difference between Global and Local Sequence Alignment. [online] Major Differences. Available at: https://www.majordifferences.com/2016/05/difference-between-global-and-local.html.

Hedyehzadeh, M., Maghooli, K., MomenGharibvand, M., & Pistorius, S. (2020). A Comparison of the Efficiency of Using a Deep CNN Approach with Other Common Regression Methods for the Prediction of EGFR Expression in Glioblastoma Patients. Journal of digital imaging, 33(2), 391–398.

ICON plc. (n.d.). C diff and COVID-19, is this a health risk that can be dealt with sooner than later? [online] Available at: https://www.iconplc.com/insights/blog/2020/11/19/c-diff-and-covid-19/ [Accessed 3 Sep. 2022].

Imwattana, K., Knight, D. R., Kullin, B., Collins, D. A., Putsathit, P., Kiratisin, P., & Riley, T. V. (2014). Antimicrobial resistance in Clostridium difficile ribotype 017. Expert review of anti-infective therapy, 18(1), 17–25.

Imwattana, K., Rodríguez, C., Riley, T. V., & Knight, D. R. (2021). A species-wide genetic atlas of antimicrobial resistance in Clostridioides difficile. Microbial genomics, 7(11), 000696.

Isa Irawan, M., Mukhlash, I., Rizky, A., & Ririsati Dewi, A. (2019). Application of Needleman-Wunch Algorithm to identify mutation in DNA sequences of Corona virus. Journal of physics. Conference series, 1218(1), 012031.

Kattenborn, Teja & Leitloff, Jens & Schiefer, Felix & Hinz, Stefan. (2021). Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. ISPRS Journal of Photogrammetry and Remote Sensing. 173. 24-49.

Kirk, J. A., Gebhart, D., Buckley, A. M., Lok, S., Scholl, D., Douce, G. R., Govoni, G. R., & Fagan, R. P. (2017). New class of precision antimicrobials redefines role of Clostridium difficile S-layer in virulence and viability. Science translational medicine, 9(406), eaah6813.

Klümper, U., Droumpali, A., Dechesne, A., & Smets, B. F. (2014). Novel assay to measure the plasmid mobilizing potential of mixed microbial communities. Frontiers in microbiology, 5, 730.

Knight, D. R., Squire, M. M., Collins, D. A., & Riley, T. V. (2017). Genome Analysis of Clostridium difficile PCR Ribotype 014 Lineage in Australian Pigs and Humans Reveals a Diverse Genetic Repertoire and Signatures of Long-Range Interspecies Transmission. Frontiers in microbiology, 7, 2138.

Kyrkou, I., Carstens, A. B., Ellegaard-Jensen, L., Kot, W., Zervas, A., Djurhuus, A. M., Neve, H., Franz, C., Hansen, M., & Hansen, L. H. (2020). Isolation and characterisation of novel phages infecting Lactobacillus plantarum and proposal of a new genus, "Silenusvirus". Scientific reports, 10(1), 8763.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Lee, R. Karchin, and M. A. Beer. (2017). “Discriminative prediction of mammalian enhancers from dna sequence,” Genome Research, vol. 21, no. 12, pp. 2167–2180, 2011. [21] L. L. Colbran, L. Chen, and J. A. Capra, “Short dna sequence patterns accurately identify broadly active human enhancers,” BMC Genomics, vol. 18, p. 536.

Lee, C. T., & Peng, S. L. (2017). A Pairwise Alignment Algorithm for Long Sequences of High Similarity. Information and Communication Technology : Proceedings of ICICT 2016, 625, 279–287.

Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics (Oxford, England), 34(18), 3094–3100.

Liimatainen, K., Huttunen, R., Latonen, L., & Ruusuvuori, P. (2021). Convolutional Neural Network-Based Artificial Intelligence for Classification of Protein Localization Patterns. Biomolecules, 11(2), 264.

Lim, S. C., Knight, D. R., & Riley, T. V. (2020). Clostridium difficile and One Health. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 26(7), 857–863.

Liu, K. L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A., & Xie, G. (2012). Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. Applied and environmental microbiology, 78(5), 1523–1533.

Luhmann, N., Holley, G. & Achtman, M. BlastFrost: fast querying of 100,000s of bacterial genomes in Bifrost graphs. Genome Biol 22, 30 (2021).

Maslennikov R, Ivashkin V, Ufimtseva A, Poluektova E, Ulyanin A. Clostridioides difficile co-infection in patients with COVID-19. Future Microbiol. 2022;17:653-663.

Marinus, M. G., Poteete, A., & Arraj, J. A. (1984). Correlation of DNA adenine methylase activity with spontaneous mutability in Escherichia coli K-12. Gene, 28(1), 123–125.

Mazaheri Nezhad Fard, R., Barton, M. D., & Heuzenroeder, M. W. (2011). Bacteriophage-mediated transduction of antibiotic resistance in enterococci. Letters in applied microbiology, 52(6), 559–564.

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research, 32(Web Server issue), W20–W25.

Medium. 2022. Common Loss functions in machine learning. [online] Available at: <https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23#:~:text=Machines%20learn%20by%20means%20of%20a%20loss%20function.,function%20would%20cough%20up%20a%20very%20large%20number.> [Accessed 4 September 2022].

Mohammad, N. S., Nazli, R., Zafar, H., & Fatima, S. (2022). Effects of lipid based Multiple Micronutrients Supplement on the birth outcome of underweight pre-eclamptic women: A randomized clinical trial. Pakistan journal of medical sciences, 38(1), 219–226.

Mullan L. (2006). Pairwise sequence alignment--it's all about us!. Briefings in bioinformatics, 7(1), 113–115.

Muhamad, F., Ahmad, R., Asi, S. and Murad, M. (2018). Performance Analysis Of Needleman-Wunsch Algorithm (Global) And Smith-Waterman Algorithm (Local) In Reducing Search Space And Time For Dna Sequence Alignment. Journal of Physics: Conference Series, 1019, p.012085.

Murata M. (1990). Three-way Needleman--Wunsch algorithm. Methods in enzymology, 183, 365–375.

Nielsen, A.A.K., Voigt, C.A. Deep learning to predict the lab-of-origin of engineered DNA. Nat Commun 9, 3135 (2018)

Nguyen, N. , Tran, V. , Ngo, D. , Phan, D. , Lumbanraja, F. , Faisal, M. , Abapihi, B. , Kubo, M. and Satou, K. (2016) DNA Sequence Classification by Convolutional Neural Network. Journal of Biomedical Science and Engineering, 9, 280-286.

Olah, C. (2015). Understanding LSTM Networks -- colah’s blog. [online] Github.io. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Oliveira, F. F., Dias, L. A., & Fernandes, M. (2022). Proposal of Smith-Waterman algorithm on FPGA to accelerate the forward and backtracking steps. PloS one, 17(6), e0254736.

Olsen, R., Hwa, T., & Lässig, M. (1999). Optimizing Smith-Waterman alignments. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 302–313.

Onimaru, K., Nishimura, O., & Kuraku, S. (2020). Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. PloS one, 15(7), e0235748.

Pacífico, C., Hilbert, M., Sofka, D., Dinhopl, N., Pap, I. J., Aspöck, C., Carriço, J. A., & Hilbert, F. (2019). Natural Occurrence of Escherichia coli-Infecting Bacteriophages in Clinical Samples. Frontiers in microbiology, 10, 2484.

Patel, P. H., & Hashmi, M. F. (2022). Macrolides. In StatPearls. StatPearls Publishing.

Phothichaisri, W., Ounjai, P., Phetruen, T., Janvilisri, T., Khunrae, P., Singhakaew, S., Wangroongsarb, P., & Chankhamhaengdecha, S. (2018). Characterization of Bacteriophages Infecting Clinical Isolates of Clostridium difficile. Frontiers in microbiology, 9, 1701.

Pohl, J. F., Patel, R., Zobell, J. T., Lin, E., Korgenski, E. K., Crowell, K., Mackay, M. W., Richman, A., Larsen, C., & Chatfield, B. A. (2011). Clostridium difficile Infection and Proton Pump Inhibitor Use in Hospitalized Pediatric Cystic Fibrosis Patients. Gastroenterology research and practice, 2011, 345012.

Portin P. (2014). The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. Journal of genetics, 93(1), 293–302.

Pratama, A. A., Bolduc, B., Zayed, A. A., Zhong, Z. P., Guo, J., Vik, D. R., Gazitúa, M. C., Wainaina, J. M., Roux, S., & Sullivan, M. B. (2021). Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. PeerJ, 9, e11447.

Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J., & Tyrrell, P. N. (2020). Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. International journal of computer assisted radiology and surgery, 15(12), 2041–2048.

Robertson, C., Pan, J., Kavanagh, K., Ford, I., McCowan, C., Bennie, M., Marwick, C., & Leanord, A. (2020). Cost burden of Clostridioides difficile infection to the health service: A retrospective cohort study in Scotland. The Journal of hospital infection, 106(3), 554–561.

Rokkam, V., Kutti Sridharan, G., Vegunta, R., Vegunta, R., Boregowda, U., & Mohan, B. P. (2021). Clostridium Difficile and COVID-19: Novel Risk Factors for Acute Portal Vein Thrombosis. Case reports in vascular medicine, 2021, 8832638.

Rucci, E., Garcia, C., Botella, G., De Giusti, A., Naiouf, M., & Prieto-Matias, M. (2018). SWIFOLD: Smith-Waterman implementation on FPGA with OpenCL for long DNA sequences. BMC systems biology, 12(Suppl 5), 96.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behavioural processes, 148, 56–62.

Sandhu, A., Tillotson, G., Polistico, J., Salimnia, H., Cranis, M., Moshos, J., Cullen, L., Jabbo, L., Diebel, L., & Chopra, T. (2020). Clostridioides difficile in COVID-19 Patients, Detroit, Michigan, USA, March-April 2020. Emerging infectious diseases, 26(9), 2272–2274.

Sarıgül, M., Ozyildirim, B. M., & Avci, M. (2019). Differential convolutional neural network. Neural networks : the official journal of the International Neural Network Society, 116, 279–287.

Schroeder, M. R., & Stephens, D. S. (2016). Macrolide Resistance in Streptococcus pneumoniae. Frontiers in cellular and infection microbiology, 6, 98.

Sheikh, A., Sheikh, A. B., Shah, I., Khair, A. H., Javed, N., & Shekhar, R. (2021). COVID-19 and Fulminant Clostridium difficile Colitis Co-Infection. European journal of case reports in internal medicine, 8(8), 002771.

Smith, E. (n.d.). Library Guides: NCBI Bioinformatics Resources: An Introduction: BLAST: Compare & identify sequences. [online] guides.lib.berkeley.edu. Available at: https://guides.lib.berkeley.edu/ncbi/blast#:~:text=Basic%20Local%20Alignment%20Search%20Tool%20(BLAST)&text=The%20program%20compares%20nucleotide%20or.

Sechaud, L., Cluzel, P. J., Rousseau, M., Baumgartner, A., & Accolas, J. P. (1988). Bacteriophages of lactobacilli. Biochimie, 70(3), 401–410.

Sothiselvam, S., Liu, B., Han, W., Ramu, H., Klepacki, D., Atkinson, G. C., Brauer, A., Remm, M., Tenson, T., Schulten, K., Vázquez-Laslop, N., & Mankin, A. S. (2014). Macrolide antibiotics allosterically predispose the ribosome for translation arrest. Proceedings of the National Academy of Sciences of the United States of America, 111(27), 9804–9809.

Spigaglia P. (2016). Recent advances in the understanding of antibiotic resistance in Clostridium difficile infection. Therapeutic advances in infectious disease, 3(1), 23–42.

Taxt, A.M., Avershina, E., Frye, S.A. et al. (2020). Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. Sci Rep 10, 7622

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261–5267.

Wang, M., Tai, C., E, W., & Wei, L. (2018). DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic acids research, 46(11), e69.

Wang, J., Yang, C., Zhang, C., Mao, X., & Lizhe, A. (2021). Complete genome sequence of the Clostridium difficile LCL126. Bioengineered, 12(1), 745–754.

Wang, S., Jiang, K., Du, X., Lu, Y., Liao, L., He, Z., & He, W. (2021). Translational Attenuation Mechanism of ErmB Induction by Erythromycin Is Dependent on Two Leader Peptides. Frontiers in microbiology, 12, 690744.

Wasels, F., Spigaglia, P., Barbanti, F., Monot, M., Villa, L., Dupuy, B., Carattoli, A., & Mastrantonio, P. (2015). Integration of erm(B)-containing elements through large chromosome fragment exchange in Clostridium difficile. Mobile genetic elements, 5(1), 12–16.

Wee, L., Conceicao, E. P., Tan, J. Y., Magesparan, K. D., Amin, I., Ismail, B., Toh, H. X., Jin, P., Zhang, J., Wee, E., Ong, S., Lee, G., Wang, A. E., How, M., Tan, K. Y., Lee, L. C., Phoon, P. C., Yang, Y., Aung, M. K., Sim, X., … Ling, M. L. (2021). Unintended consequences of infection prevention and control measures during COVID-19 pandemic. American journal of infection control, 49(4), 469–477.

Weldon (2021). Difference Between Local And Global Sequence Alignment. [online] CHEMPINSTA. Available at: https://chempinsta.com/difference-between-local-and-global-sequence-alignment/.

Wu, Q., Gao, T., Lai, Z., & Li, D. (2020). Hybrid SVM-CNN Classification Technique for Human-Vehicle Targets in an Automotive LFMCW Radar. Sensors (Basel, Switzerland), 20(12), 3504.

Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995

Vázquez-Laslop, N., & Mankin, A. S. (2018). How Macrolide Antibiotics Work. Trends in biochemical sciences, 43(9), 668–684.

Naidu, V. and Narayanan, A. (2016) "Needleman-Wunsch and Smith-Waterman Algorithms for Identifying Viral Polymorphic Malware Variants," 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), pp. 326-333.

Xia, Z., Cui, Y., Zhang, A. et al. A Review of Parallel Implementations for the Smith–Waterman Algorithm. Interdiscip Sci Comput Life Sci 14, 1–14 (2022).

Xiong, X., Zhu, T., Zhu, Y. et al. Molecular convolutional neural networks with DNA regulatory circuits. Nat Mach Intell 4, 625–635 (2022)

Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics (Oxford, England), 32(12), i121–i127.

Zheng, X., Xu, S., Zhang, Y., & Huang, X. (2019). Nucleotide-level Convolutional Neural Networks for Pre-miRNA Classification. Scientific reports, 9(1), 628.

APPENDICES

Appendix A: Raw Data

The Raw *C. difficile* genome isolates are stored on an external DISC, owing to the size of the files.

The Python code including the corresponding output files and figures are stored in a GitHub repository accessible at: github.com/sarahjaynebyrne/MSc_DataScienceProject

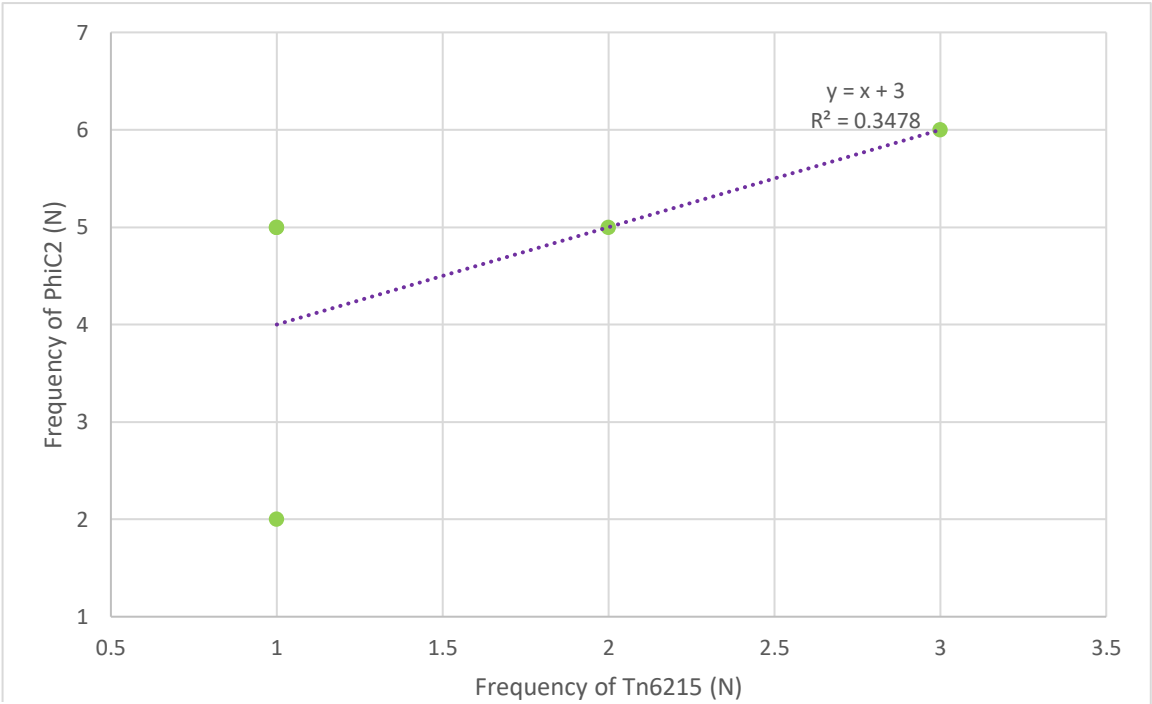


Figure 1 - *There was a moderate (r=0.587) but not a statistically significant (p<0.05) correlation between the presence of PhiC2 and Tn6215.*