Phages, Phage-plasmids, and Plasmids Sequence Predictions from Metagenome Sequences Using Machine Learning and Deep Learning Algorithms

Md Karim Uddin

Master’s Thesis

MEng. in Big Data Analytics

2025

Master’s Thesis

Md Karim Uddin

Phages, Phage-plasmids, and Plasmids Sequence Predictions from Metagenome Sequences Using Machine Learning and Deep Learning Algorithms

Arcada University of Applied Sciences: MEng in Big Data Analytics, 2025.

Commissioned by:

N/A

Abstract:

Mobile genetic elements (MGEs) serve as the architects of bacterial evolution and adaptation, with phage-plasmids—a fascinating hybrid class exhibiting dual phage and plasmid characteristics—emerging as particularly significant yet elusive players in antimicrobial resistance spread. This thesis introduces a novel computational framework that surpasses the traditional binary classification of MGEs by employing advanced machine learning techniques to distinguish phages, plasmids, and phage-plasmids from sequence data alone.

Through rigorous analysis of pentamer (k=5) frequency profiles derived from 4,248 carefully curated MGE sequences, I developed and compared three progressively sophisticated models: Logistic Regression, Random Forest, and Convolutional Neural Network (CNN). The CNN achieved remarkable 90% accuracy, revealing the power of deep learning to capture subtle sequence patterns that define these genetic elements. The exceptional precision (93%) for phage-plasmid identification represents a significant advancement in detecting these hybrid elements in complex metagenomic datasets.

Strikingly, my analysis uncovered the distinctive genomic signatures of each MGE class—AT-rich motifs dominating phages, GC-rich patterns characterizing plasmids, and unique sequence compositions marking phage-plasmids. Dimensionality reduction visualizations elegantly confirmed the intermediate evolutionary position of phage-plasmids, while revealing multiple distinct clusters suggesting diverse evolutionary trajectories for these hybrid elements.

Beyond its methodological contributions, this research provides critical biological insights into the sequence-level characteristics that underpin the hybrid functionality of phage-plasmids. The intermediate nucleotide composition and distinctive k-mer patterns observed in phage-plasmids offer computational evidence supporting their proposed role as evolutionary bridges facilitating genetic exchange between different MGE types.

This work creates new possibilities for metagenomic exploration, antimicrobial resistance surveillance, and biotechnological innovation by enabling accurate identification of all three MGE classes without requiring gene annotation or reference databases. By illuminating the genomic nature of these important vehicles of bacterial adaptation, this research advances our fundamental understanding of horizontal gene transfer and provides practical tools to address pressing challenges in infectious disease and microbial ecology.

Keywords: Artificial Intelligence, Machine Learning, Deep learning, Logistic regression Model, Random Forest Model, Convolutional Neural Networks, Neural Networks, K-mers Embedding, Metagenomics Sequence Predictions, Phage-plasmid, Phages, Plasmids

Contents

1. Literature Review

1.1 Introduction

Bacteriophages are viruses that kill bacteria, and plasmids are accessory DNA that replicate in the bacteria. Phage-plasmids are mobile genetic elements with properties of both phages and plasmids, functioning as hybrids (Shanet et al. 2023) Mobile genetic elements contribute to bacterial evolution, spread of antibiotic resistance and functioning of ecosystems (Hilpert et al. 2020; Kerkvliet et al. 2024). The finding of these mobile genetic elements in metagenomic sequences can help in understanding their distribution, diversity, and impact on microbial communities. Nonetheless, this work implies challenging genomic fragments reconstructions, high-throughput annotation, and association of antimicrobial resistance genes (ARGs) with mobile genetic elements (Kerkvliet et al. 2024).

Recently developed bioinformatics tools assist in the identification of plasmids, phages and insertion sequence elements in metagenomes, thus aiding the study of mobile ARGs (mARGs) dissemination. Certain cell types influence the phage and plasmid infection. These cell types are associated with strains that acquire more conjugative plasmids in their natural environments (Haudiquet et al., 2024). The interactions of bacterial cells with mobile genetic elements lead to the evolution of microbial communities. A wide range of machine learning techniques have been employed to tackle the difficulties involved in recognizing and categorization of mobile genetic elements. These are random forest models for major capsid proteins sequences to predict phage capsid architecture (Lee et al., 2022) and alignment-free methods using k-mer frequencies to classify phage lifestyles (Song, 2020). Machine learning methods can obtain a classification accuracy that is much greater than random classification (33%) with several methods. These are Logistic Regression (LR), Random Forest (RF) and Convolutional Neural Networks (CNN). Logistic Regression has been found to be highly accurate in numerous studies. For instance, taking apple varieties' classification, LR achieved accuracy of 99.08 when combined with deep features and PCA (Taner et al., 2024). LR might find it more difficult to model the intricacies in biological data when compared to advanced models. Random Forest, which is an ensemble learning method, was used to classify phage virion proteins with 91.84% accuracy (Zhang & Li, 2023). CNNs enhance the accuracy of classification by automatically extracting features. An example is CNN models achieving accuracies of 93.16% in DNA sequence classification (Gunasekaran et al., 2021). The advantage of CNN is that it learns automatically hierarchically from raw data which is great for recognizing complex patterns.

This literature review discusses what is known about the biology of phages, plasmids and phage-plasmids as well as the computational difficulties to detect them from metagenomic data and machine learning approaches promising for their classification.

1.2 Historical Context and Taxonomic Ambiguity

The identification of phages and plasmids has brought about the tools necessary to study bacterial genomes. Viruses that infect bacteria are called phages discovered in the early 20th century and an extrachromosomal DNA element that can replicate independently within the cellular context. As illustrated recently, phage-plasmids are a new class of genetic element which bears characteristics of both phages and plasmids. Phage-plasmids are extra-chromosomal elements which can behave as plasmids and phages, showcasing a unique eco-evolutionary strategy. These are extra-chromosomal elements which can act as both plasmids and phages, exhibiting a unique eco-evolutionary strategy. The combination of the two allows for rapid transmission of a phage through a bacterial population and stable maintenance either in lytic or lysogenic state.

The classification of phages, plasmids and phage-plasmids is quite difficult owing to their diverse sequences. Traditional taxonomic approaches that involve nucleotide sequence homology are computationally expensive and have failed to keep pace with the increasing numbers of sequenced genomes (Gauthier & Hatfull, 2023). The presence of widespread metagenomics sequence patterns of these mobile genetic elements hinder classification (Smug et al., 2023).

New bioinformatics approaches must be adopted to solve these problems. The new bioinformatics tool PhamClust uses proteomic equivalence quotients to cluster phages based on their inter-genome relatedness (Gauthier & Hatfull, 2023). These techniques seek to provide a more efficient and scalable way of approaching phage taxonomy that would reflect the continuum of diversity evident among these genomic entities.

1.3 Ecological and Functional Roles

Bacteriophages, phage-plasmids and plasmids shape microbial communities and drive bacterial evolution. Phages can affect ecosystem community dynamics as well as biogeochemical cycles. This impacts various ecosystems such as soils and sediments (Shaidullina & Harms, 2022; Zhang et al., 2022). According to Zhang et al. (2022), they can transfer antibiotic resistance genes among and between bacteria via transduction. Phage plasmids could be intermediate between phages and plasmids regarding nucleotide composition and sequence pattern, detectable by k-mer frequency analysis (Song, 2020).

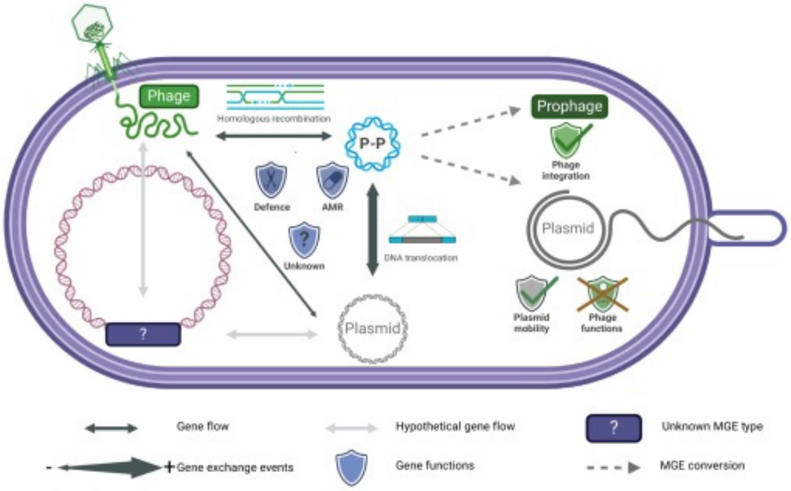


Figure 1. Evolutionary trajectories of phage-plasmids (P-Ps), and gene flow between different mobile genetic element (MGE) types. Adopted from (Figunia et al., 2024)

The potential of phages for biotechnology and therapeutics directed towards bacterial infections is huge (Shaidullina & Harms, 2022). Phage therapy researchers concentrate on lytic phages that can kill bacteria while avoiding temperate phages that can transfer unwanted genes (Grigson et al., 2023). A novel approach termed ‘phage rehabilitation’ has emerged aimed at modifying the bacterial composition or function without necessarily eliminating pathogens (Baaziz et al., 2022). In addition, identifying nonessential genes in phages by CRISPRi assays can help engineer phages for various purposes, including tracking and quantifying in different environments (Piya et al., 2023). In case of Phage-plasmids (PPs), they have wider distribution in the bacteria. In addition, their carriage of phage and plasmid core genes implied that they could promote horizontal gene transfer, and host adaptation between different MGE types (Figunia et al., 2024, Pfeifer & Rocha, 2024)

1.4 Metagenomics and the Phage-Plasmids Detection Problem

Metagenomics enables analysis of microbial communities by sequencing DNA from environmental samples, providing insights into community structure and function (Kim et al., 2024). This technique is vital in medicine, agriculture, environmental science, and forensics (Rahman & Rangwala, 2020). Detecting phages, phage-plasmids, and plasmids in metagenomic data is challenging as these mobile genetic elements spread antimicrobial resistance genes across hosts (Kerkvliet et al., 2024).

To solve this problem, many computation tools have been developed but which vary in performance. Researchers examined 19 metagenomic phage detection tools and observed a significant difference in their findings. Almost 80% of contigs were marked as phage by at least one tool, but the highest overlap between any two tools was limited to only 38.8% (Schackart et al., 2023). Tools using homology-based approaches (e.g., VirSorter, MARVEL) are associated with low false positive rates and are robust against eukaryotic contamination. By contrast, sequence composition-based tools (e.g., VirFinder, DeepVirFinder) exhibit higher sensitivity, including for phages with limited presence in reference databases (Schackart et al., 2023).

This means that the present methodologies and systems which are used to detect have limitations. Kerkvliet et al. (2024) indicate that in short-read metagenomics sequencing experiments the metagenomic assembly process is the main bottleneck for linking ARGs to the identified MGEs. To tackle these issues, the researchers proposed pipelines like MetaMobilePicker, which integrates multiple tools to identify ARGs associated with plasmids, IS elements, and phages (Kerkvliet et al. 2024).

1.5 Feature-Based Classification Strategies

Sequence features can be used to classify mobile genetic elements (MGEs) like phages, phage-plasmids and plasmids. K-mer frequencies are alignment-free strategies for comparing genomes and predicting lifestyles (Song, 2020) and are suitable for metagenomic data. The D2* dissimilarity measure using k-mer frequencies is effective in classifying phage lifestyles. Feature extraction refers to the identification of mobility-related features in plasmids and MGEs. The MetaMobilePicker tool that combines various tools for identifying ARGs associated with plasmids, IS elements and phages (Kerkvliet et al., 2024). Kerkvliet et al. (2024) state that metagenomic assemblage poses the greatest obstacle to linking ARGs to MGEs. The MMPSO algorithm simultaneously ranks features and uses heuristic search techniques for optimal feature selection (Wang et al., 2022). The cABC performs a recursive analysis which lets you use only the most interoperative features (Lötsch & Ultsch, 2023).

Pentamers (k=5) provide the right balance of sequence context and computational efficiency for genomic classification. These patterns distinguish phages from plasmids (Song, 2020). Phages are usually AT-rich; however plasmids are GC-rich. Phages like Lu221 and Hi226 have AT-rich pentamer signatures (Parra et al., 2023), while conjugative plasmids exhibiting antibiotic resistance genes have a higher GC content (Parra et al., 2023). Phage-plasmids may possess unique pentamer signatures not found in a phage or a plasmid. These hybrid elements could be identified through k-mer patterns, indicating their hybrid origin (Parra et al., 2023; Song, 2020). Patterns like these may allow for quick assignment of novel genomic elements in meta-genomics.

Feature selection methods enhance classification accuracy. By employing feature ranking as well as the heuristic search technique, the MMPSO algorithm has shown to be effective in selecting the best feature subsets (Wang et al, 2022). The recursive computed ABC analysis is a specific method that successfully reduces feature sets that skew the distribution of feature importance in biological data sets (Lötsch & Ultsch, 2023).

1.6 Machine Learning Approaches for Mobile Genetic Element Classification

1.6.1 Logistic Regression

Statistical Modeling and Machine Learning apply the Logistic Regression (LR) as the basic model for a classification problem. Despite its simplicity, logistic regression demonstrates high accuracy across multiple domains, achieving 99.08% accuracy when classifying apple varieties with deep features and principal component analysis (Taner et al., 2024). The main strength of the method lies in its interpretability as the features' coefficients directly reflect their importance and influence in the output, giving us insights into the features that discriminatively separate the two classes. Genomic analysis benefits from this interpretability as it points to biology. When classifying MGE using k-mer frequencies, a logistic regression can handle high dimensional spaces with ease and also provides a measure of confidence for each classification. On the other hand, the model assumes linear relationships between features and classification outcomes. Thus, the model is unable to capture the complex patterns present in sequence data. The model also struggles with the correlated features that are present in k-mer representations. MGE classification may be done with logistic regression, despite its limitation.

1.6.2 Random Forest

Random forest (RF) is a model that builds a number of different decision trees and gives the most frequent predictive class. Random Forest is appropriate for genomic sequence classification because it works well with high-dimensional data. It also has its own feature importance metrics and is not prone to overfitting when used with low label data. Keith et al. (2024) used RF models to predict phage activity, achieving F1 scores over 0.6 for generalist phages. The model is capable of capturing complex relationships between features, which could make it preferable to logistic regression for sequence classification. Random Forest supports feature importance and partial dependence plots and runs well on high-dimension data, thus rendering it promising for MGE classification.

1.6.3 Deep Learning in Genome Prediction

Deep learning (especially Convolutional Neural Networks) is a powerful method to analyse mobile genetic elements (MGE) such as the prediction of phages, phage-plasmids and plasmids in genomic prediction. CNNs have been used with success in computer vision and natural language processing, and are now being applied to genomic sequence analysis (Ding et al., 2023; Sajja & Kalluri, 2021). CNNs have achieved performance as high as 97.48% in apple variety classification (Taner et al., 2024) and 99% in waste plastic bottle classification (Fadlil et al., 2022) in classification tasks. CNNs are useful to analyze genomic sequences having the potential to learn hierarchical features from raw data. The different layers of the architecture consist of convolutional layers for extraction of the local patterns, pooling layers for reducing dimensionality and fully connected layers for prediction. They are great in capturing complex patterns in the genomic sequences (Ding et al., 2023; Krützfeldt et al., 2020). Recently a tool developed for bacteriophage lifestyle prediction using convolutional neural network (CNN) (Zhang et al.2024).

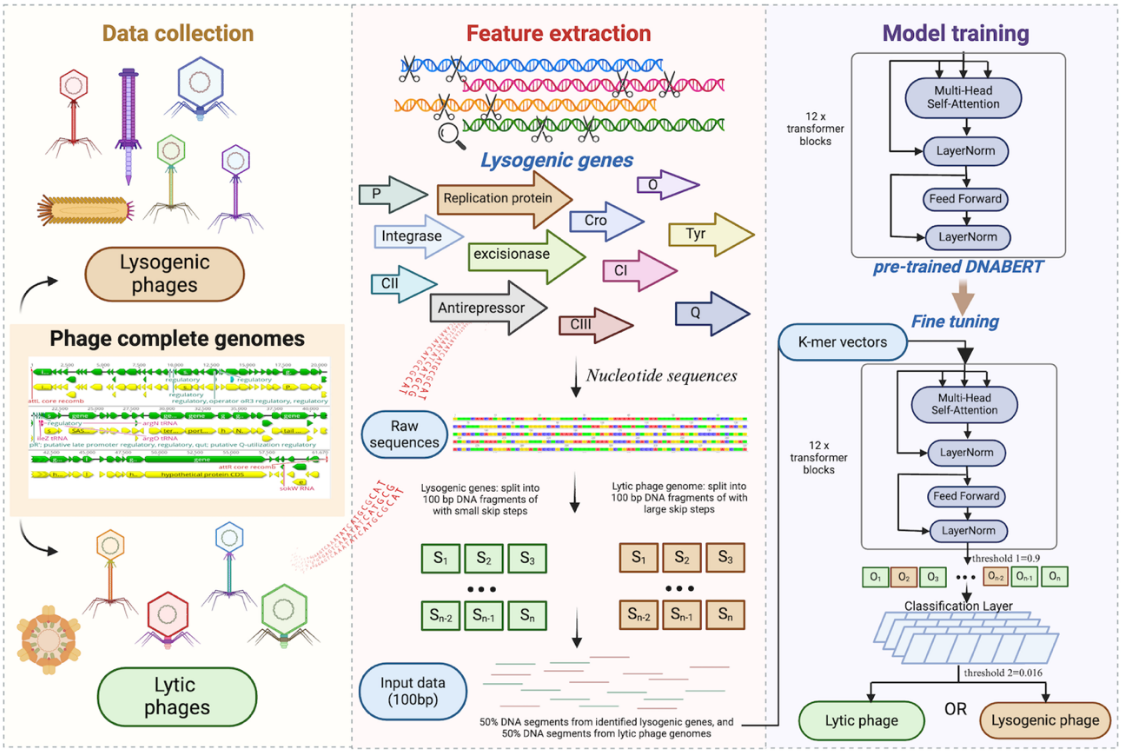


Figure 2. DeepPL: A deep-learning-based tool for the prediction of bacteriophage lifecycle

Adopted from (Zhang et al.2024)

Due to large-scale datasets and complex interrelationships that do not need feature engineering manually, CNNs have become widely accepted. DeepSTF uses different types of neural networks to predict the locations of TFBSwith superior performance (Ding et al., 2023). CNNs can learn discriminative features from sequence data for phage, plasmid and phage-plasmid classification. PlasmidFinder (Carattoli & Hasman, 2020) and PlasFlow (Krawczyk et al, 2018) used for predicting plasmid sequences.

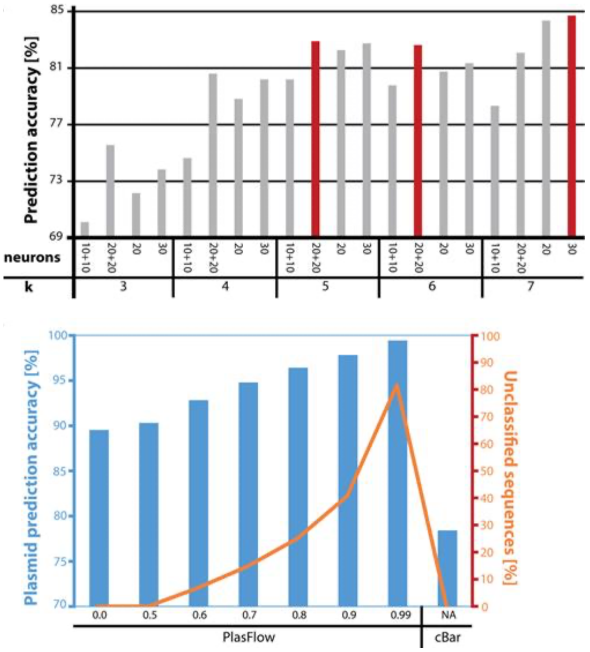


Figure 3. PlasFlow, a novel plasmid sequences prediction tool based on genomic signatures that employs a neural network approach for identification of bacterial plasmid sequences in environmental samples

The feature interpretation of these models is difficult because they are black boxes. Also, their performance changes with training data quality. Providing large training datasets for rare type MGE types housed in the Krützfeldt et al. (2020) is also a requirement.

1.7 Addressing Class Imbalance in MGE Datasets

Datasets related to mobile genetic elements (MGEs), such as phages, phage-plasmids and plasmids often suffer from class imbalance which occurs when one class gets significantly fewer representation than the other classes (Sowah et al., 2021). The MGE data sets have samples of different genetic elements in a disproportionate manner that leads to biased predictions. There are various methods to tackle this problem. The first method is oversampling whereby synthetic minority class samples are created using SMOTE (Malhotra & Lata, 2020). The other way is HCBST that removes instances of majority class (Sowah et al., 2021). Methods which combine both can deal with large imbalance ratios (Wang et al., 2020). According to studies, these resampling methods improve accuracy and other metrics like the F1 score and AUC. For example, it was found that the HCBST had an average AUC of 0.73. A geometric mean of 0.67, and Matthews Correlation Coefficient of 0.35 (Sowah et al., 2021). Still, generating any synthetic data before dataset splitting will lead to leakage, therefore, resampling should only be applied to training data post-splitting (Nieto-Del-Amor et al., 2022).

1.8. Comparative Model Performance and Interpretability

There are many different evaluation metrics that are used to compare diverse machine learning models for MGE classification. This word refers to various scoring metrics including accuracy, precision, recall, F1-score and Area Under the Receiver Operating Characteristic (AUROC) curve, (Albuquerque et al., 2022: Sasaki & Sakata, 2020: Xiao et al., 2024). The (Xiao et al, 2024) is a model that scored 0.7792 in accuracy, 0.7448 in precision, 0.8769 in recall, 0.8055 in F1 score and 0.8387 in AUC on the test set.

Compare logistic regression, random forest and CNN for MGE classification.

a) Logistic Regression: Offers an easy-to-interpret equation that shows the importance of features. Identifying Heavy Metal Spe. in the YS Blase Plant. Though, it may not work well on complex, nonlinear functions.
b) The Random forest model has a moderate interpretability and gives you a feature importance measure. It is also very good at handling high-dimensional data. It can learn complex relationships without much tuning.
c) Deep learning is the technology that powers CNN. CNNs need more data and computing resources, while the way they make classifications is difficult to understand because they're like a black box.

Model quality versus understanding is an important trade-off. Complex models might provide better prediction power, but they lose in terms of interpretability (Arkoudi et al., 2023; Mariotti et al., 2023). Mariotti et al. (2023) propose a Constrainable Neural Additive Model that balances performance and interpretability.

In Sasaki & Sakata (2020), the F1 scores with random forests and support vector machine and logistic regression were over 80%. According to Keith et al. (2024), for the random forest models that predict phage activity, the F1 scores were >0.6 for generalist phages but the performance varied by dataset.

1.9. Sequence Embedding and Representation Learning

Sequence embedding and representation learning are important techniques for genomic data analysis. The patterns in biological sequences are captured in lower-dimensional space for downstream analysis. Genomic sequences can make use of various embedding methods like k-mer-based and learned. B cell receptor BCR sequences were evaluated with models ranging from BCR-specific embeddings to general protein language models to predict the sequence properties (J. Wang et al., 2023). BCR-specific data embeddings perform marginally better than general protein models in predicting specificity. It seems that having domain-specific knowledge is beneficial for enhanced performance in certain applications of machine learning. According to J Wang et al. (2023), structural choice of embedding strategy affects model performance with better prediction performance on full-length heavy chains and paired light chain sequences for BCR analysis. It is clear that the input sequence length and paired-chain information are important for embedding models. The FoldHSphere approach also shows that learning the discriminative embedding can bridge the gap between protein fold recognition performances. Sequence embedding is important for genomic analysis, and different techniques yield different advantages. The ideal technique will thus depend on the task and data. K-mer File Format that stores set of k-mer is presented in (Dufresne et al. 2022), where pentamers used for MGE classification. The authors of (R. Liu et al., 2022) developed an algorithm called KTU, which relies on k-mer based algorithms to utilize the frequencies of other tetra-nucleotide segments present in the genome to cluster sequence variants. Though not specifically focused on pentamer optimization for MGE classification, these papers do provide k-mer insights that may be applicable to MGE pentamers. To detect mobile genetic elements, the identification of an optimal k-mer length and the use of an appropriate classification algorithm is required.

1.10. Real-World Applications and Knowledge Gaps

A correct annotation of mobile genetic elements (MGEs), such as phages, phage-plasmids, and plasmids has several applications in microbiology, ecology and medicine. In microbiology and ecology, these datasets help us better understand bacterial evolution, adaptation, and community dynamics (Arredondo-Alonso et al., 2023; Silva et al., 2022). Microbial genome sequencing is important for tracking antimicrobial resistance genes (ARGs) through hosts and environments whose insights may help us understand resistance dissemination (Kerkvliet et al., 2024; Mitchell et al., 2021). A precise classification of MGE is essential in medicine for the development of targeted phage therapies against multidrug-resistant bacterial infections and to optimise their dosing (Bosco et al. 2023; Nguyen et al. 2023). However, several challenges and knowledge gaps remain.

Kerkvliet et al. (2024) found that the principal bottleneck in linking ARGs to MGEs in short-read metagenomic sequencing is not ARG and MGE identification but rather assembly. According to Partridge et al. (2021), benchmarking datasets ought to be more representative, and biases ought to be recognized. Research indicates that the hindgut of the horse is a significant reservoir of ARGs. Future research must work to enhance the sensitivity and specificity of tools identifying MGE (Kerkvliet et al., 2024). To critically evaluate these tools, comprehensive benchmarking datasets must be developed soon (Partridge et al., 2021). Biological insights must be integrated into analytical tools for accurate MGE-ARG classification (Partridge et al., 2021). Investigating the role of MGEs in the expansions of environmental niches could help bacteria adapt. Advancing predictive algorithms for phage-host specificity has potential benefits for phage therapy applications (Bosco et al., 2023; Gaborieau et al., 2024).

2. Objectives and Hypotheses

2.1 Objectives

The primary objectives of this research are as follows.

1) To develop and evaluate three distinct machine learning approaches (Logistic Regression, Random Forest, and Convolutional Neural Network) for classifying mobile genetic elements into phages, phage-plasmids, and plasmids using pentamer (k=5) frequency profiles.
2) To determine the effectiveness of pentamer (k=5) frequency profiles as feature representations for mobile genetic element classification, balancing discrimination power with computational efficiency.
3) To investigate the compositional and sequence characteristics that distinguish phage-plasmids from their parent elements (phages and plasmids).
4) To determine which machine learning approach provides the highest accuracy and most balanced performance across all three mobile genetic element classes.

2.2 Hypotheses

The hypotheses of this research is based on existing literature and preliminary observation.

1) **H1**: Machine learning models trained on pentamer (k=5) frequency vectors can effectively distinguish between phages, phage-plasmids, and plasmids.
2) **H2**: Phage-plasmids will display intermediate sequence characteristics between phages and plasmids in terms of nucleotide composition, GC content, and pentamer frequency patterns, reflecting their hybrid evolutionary origins.
3) **H3**: Deep learning models (CNN) will outperform traditional machine learning approaches (Logistic Regression and Random Forest) in classification accuracy, particularly for phage-plasmids, due to their ability to capture complex non-linear patterns in pentamer frequency distributions.
4) **H4**: A balanced training dataset with equal representation of all three classes will produce models with uniform performance across phages, phage-plasmids, and plasmids.

3. Materials and Methods

3.1 Data Acquisition and Description

3.1.1 Sources of Data and Selection Criteria

The dataset used in this study was derived from the research "Phage-plasmids promote recombination and emergence of phages and plasmids" published in Nature Communications (2024). This publication identified and cataloged a substantial number of mobile genetic elements from open-source databases, specifically 3,585 phages, 1,146 phage-plasmids, and 20,274 plasmids. All sequences were extracted from the NCBI database using their respective accession numbers.

To create a balanced dataset suitable for machine learning applications, I extracted equal numbers of sequences from each class. This might be optimal approach for developing effective classification models, as it reduces the overfitting tendency of the model. By following way, we balance our dataset for model building:

a) All 1416 accessible phage-plasmid sequences were incorporated.
b) A total of 1416 phage sequences were selected randomly from a total of 3585.
c) From a total of 20,274 plasmid sequences, a random set of 1,416 was selected.

This balanced approach ensures that model training is not biased toward any particular class, which is essential for developing effective classification models.

The sequence extraction process involved.

a) Compiling accession numbers for each mobile genetic element category from the source study.
b) Include all available sequences for the phage-plasmids.
c) For the phages and the plasmids: Randomly chosen 1,416 sequences from either.
d) Retrieving the complete genomic sequences from NCBI using the Entrez API.
e) Storing sequences in FASTA format in dedicated directories for subsequent analysis.

3.1.2 Data Organization and Management

The sequences were placed in separate folders for each class. Every sequence record had an identifier, a header with metadata and nucleotide sequence.

3.2 Preprocessing and Feature Extraction

3.2.1 Controlling Sequence Quality and Standardization

All sequences underwent a standardized preprocessing workflow:

1. FASTA Parsing: Extraction of sequence identifiers, descriptions, and nucleotide sequences.
2. Quality Assessment: Evaluation of completeness, length, and nucleotide composition, including calculation of sequence length, GC content, and presence of non-standard bases.
3. Sequence Standardization: Conversion to uppercase and linearization at a standardized position (typically the origin of replication or start of a major structural gene).
4. Calculation of Basic Properties:
    - Sequence length (total number of nucleotides)
    - GC content using the formula: $GC\% = (G + C) / (A + T + G + C) \times 100$
    - Individual nucleotide frequencies (A%, T%, G%, C%)
    - Dinucleotide frequencies for all 16 possible combinations

To ensure the quality of the data collected from all sequence classes, the data underwent pre-processing. This was done to minimize biases that could occur as a result of lack of quality of the sequences or having different formats.

3.2.2 Collection of K-Mer Features and Their Vectorization

K-mer frequency analysis was employed to transform nucleotide sequences into numerical vectors suitable for machine learning. After evaluating different k values (k = 2, 3, 4, 5) in preliminary experiments, k=5 (pentamers) was selected as it provided the optimal balance between discriminative power and computational efficiency. With k=5, we capture sufficient sequence context to distinguish between MGE classes while maintaining a manageable feature space (4^5 = 1,024 dimensions). Shorter k-mers lacked the necessary specificity, while longer k-mers exponentially increased computational requirements without proportional performance gains. This choice aligns with previous findings by Song (2020) who also found pentamers effective for MGE classification.

The k-mer extraction process involved.

1) K-mer enumeration in this study refers to the systematic enumeration of all possible k-mers. This is done through the sliding window method with the help of a step size of 1
2) Frequency calculation for every sequence, the occurrence frequency of all possible 5-mers (4^5 = 1,024 distinct k-mers), was calculated.
3) Normalization of raw k-mer counts for sequence length. We used the following formula:

   x_ij = count(kmer_j in sequence_i) / sum(count(all k-mers in sequence_i))

   where:

   - x_ij is the normalized frequency of k-mer j in sequence i
   - count(kmer_j in sequence_i) is the number of occurrences of k-mer j in sequence i
   - sum(count(all k-mers in sequence_i)) is the total count of all k-mers in sequence i

4) Transformation of each sequence into a 1,024 dimensional vector for normalized frequency of all possible 5-mers.

The resultant feature matrix X had size 4,248×1,024, where rows corresponded to sequences and the columns corresponded to a k-mer frequency.

3.3 Dataset Partitioning and Preprocessing

3.3.1 Train-Test Split Strategy

Using stratified sampling, the feature matrix was divided into 80% training set and 20% testing set.

a) Training set: 3,398 sequences.
b) Testing set: 850 sequences.

3.3.2 Feature Scaling and Transformation

Two scaling approaches were evaluated.

1. Min-Max Scaling: Feature values were scaled to the range [0, 1] using:

   x_scaled = (x - min(x)) / (max(x) - min(x))

2. Standardization (Z-score normalization): Features were transformed to have zero mean and unit variance using:

   x_standardized = (x - μ) / σ

   where μ is the mean and σ is the standard deviation of the feature.

3.3.3 Label Encoding and Transformation

Class labels were encoded according to model requirements.

a) For LR, and RF models, we used integer encoding (0 for phages, 1 for phage-plasmids, 2 for plasmids).
b) One-hot encoding for the deep learning models
   - Phages: [1, 0, 0].
   - Phage-Plasmids: [0, 1, 0].
   - Plasmids: [0, 0, 1].

c) The labels were converted to binary for ROC curve analysis on a one-vs-rest basis.

3.4 Machine Learning Models

3.4.1 Logistic Regression

We used multinomial logistic regression as our baseline model. For a multi-class problem with K classes, the probability that a sample x belongs to class k is modeled as:

$$P(y = k|x) = \frac{e^{\beta_k \cdot x}}{\sum_{j=1}^{K} e^{\beta_j \cdot x}}$$

Where β_k represents the weight vector for class k. The model was set to 'liblinear' solver using one-vs-rest multi-class strategy and C=1.0 to prevent overfitting.

3.4.2 Random Forest with Hyperparameter Tuning

Random Forests combine multiple decision trees through bagging, with final prediction given by:

$$\hat{y} = \mathrm{mode}\{\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_T}\}$$

Where $\hat{y}_t$ is the prediction of the t-th tree, and T is the total number of trees.

A grid search was conducted through 108 parameter combinations with 5-fold cross-validation.

The model was trained using the following hyper-parameter grid search:

- Number of trees: 100, 200, 300
- Maximum tree depth: None, 10, 20, 30
- Minimum samples for split: 2, 5, 10
- Minimum samples per leaf: 1, 2, 4

The optimum configuration (max_depth=20, min_samples_leaf=1, min_samples_split=2, n_estimators=300) was identified by maximizing accuracy.

$$\mathrm{accuracy} = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y_\imath} = y_i)$$

In this equation for each term n is the numbers of samples, $\hat{y}_i$ is predicted class of the i-th sample and y_i is true class. Feature importance scores were calculated using mean decrease in Gini impurity.

$$\mathrm{Importance}(j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n \in N_t} \Delta\mathrm{Gini}(n, j)$$

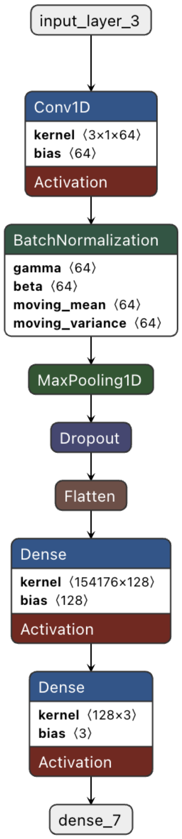
Where.

- Nt is the set of nodes in tree t that use feature j for splitting
- ΔGini(n, j) is the decrease in Gini impurity achieved by splitting node n on feature j
- T is the total number of trees in the forest

3.4.3 Convolutional Neural Network (CNN)

A 1D Convolutional Neural Network was designed with the following architecture:

1. **Input Layer**: Accepts the reshaped k-mer frequency vector (1024, 1)
2. **Convolutional Block**:
    - Conv1D layer with 64 filters, kernel size 3, and ReLU activation
    - BatchNormalization layer
    - MaxPooling1D layer with pool size 2
    - Dropout layer with rate 0.5
3. **Feature Processing**:
    - Flatten layer
4. **Fully Connected Layers**:
    - Dense layer with 128 neurons and ReLU activation
    - Output layer with 3 neurons and softmax activation



The CNN was trained using categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})$$

Where:

- y_{i,c} is 1 if sample i belongs to class c and 0 otherwise
- p_{i,c} is the predicted probability for class c of sample i
- N is the number of samples
- C is the number of classes

The Adam optimizer was used with default parameters (learning rate = 0.001, β1 = 0.9, β2 = 0.999). Training used 50 epochs, batch size of 32, and a validation split of 20%.

3.5 Analysis of Performance Using Evaluation Metrics

The performance of the model was evaluated using a combination of multiple complementary metrics.

1) The ratio of instances which are classified correctly
   Accuracy = (TP + TN) / (TP + TN + FP + FN).
2) Measures the ratio of true positive to total positive predictions.
   Precision = TP / (TP + FP).
3) The true positive ratio to the total positive is called recall.
   Recall = TP / (TP + FN).
4) F1-score is the average of precision and recall.
   F1-score = 2 × (Precision × Recall) / (Precision + Recall).
5) Area Under the ROC Curve (AUC) tells us how well our model differentiates between classes.
6) The number of false positive over the number of true negative.
   FPR = FP / (FP + TN).
7) False Negative Rate (FNR): the ratio of false negatives to actual positives.
   FNR = FN / (FN + TP).

Visualising the distributions of predictions across classes are done via confusion matrices, while plotting the ROC curves visualize the trade-off between true positives and false positives at different thresholds.

4. Results

4.1 Dataset Characteristics

4.1.1 Sequence Length Distribution

The sequencing results of the balanced dataset (1,416 sequences per class) showed a characteristic difference in the distribution of length between phages, phage-plasmids, and plasmids (Table 1, Figure 4). Phages had a mean length of 70,059.8 bp (σ = 53,513.8 bp), ranging from 3,405 to 490,380 bp. Phage-plasmids had lengths that were intermediate of that of phage and plasmids with a mean of 81,046.7 bp (σ = 48,542.7 bp), and range 10,545-290,957 bp. The plasmids showed the widest variation in length (σ = 212,574.9 bp). Plasmids sized from very small elements (619 bp) to extremely large (2,249,899 bp) were detected. Mean length was 106,071.6 bp.

Table 1. Sequence Length Statistics by MGE Class

Statistic	Phages	Phage-Plasmids	Plasmids
Count	1,416	1,416	1,416
Minimum (bp)	3,405	10,545	619
Maximum (bp)	490,38	290,957	2,249,899
Mean (bp)	70,059.8	81,046.7	106,071.6
Median (bp)	50,699.5	74,298.0	51,310.5
Standard Deviation (bp)	53,513.8	48,542.7	212,574.9

The median lengths revealed marked differences; phage-plasmids have a higher median (74,298.0 bp) compared with phages (50,699.5 bp) and plasmids (51,310.5 bp). The observation seen in Figure 4 indicates that phage-plasmids may have a more consistent size distribution. This could point towards selective constraints that must preserve both phage structural genes and plasmid maintenance functions.

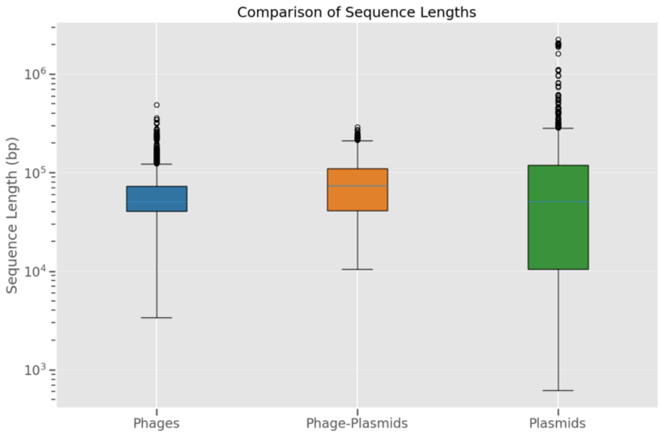


Figure 4. Sequence Length Distribution of Phages, Phage-Plasmids, and Plasmids. Box plots showing the distribution of sequence lengths across the three mobile genetic element classes, with outliers indicated as points outside the whiskers.

4.1.2 Nucleotide Composition Analysis

The analysis of GC content in different MGE types had shown their distribution. The mean GC value for the phages was 46.0% with a significant variability (σ = 10.2%); the plasmids exhibited a higher mean GC value of 53.2% (σ = 8.4%). Interestingly, phage-plasmids had a mean GC content of 45.9% (σ = 9.3%), intermediate between that of their parent elements (Figure5).

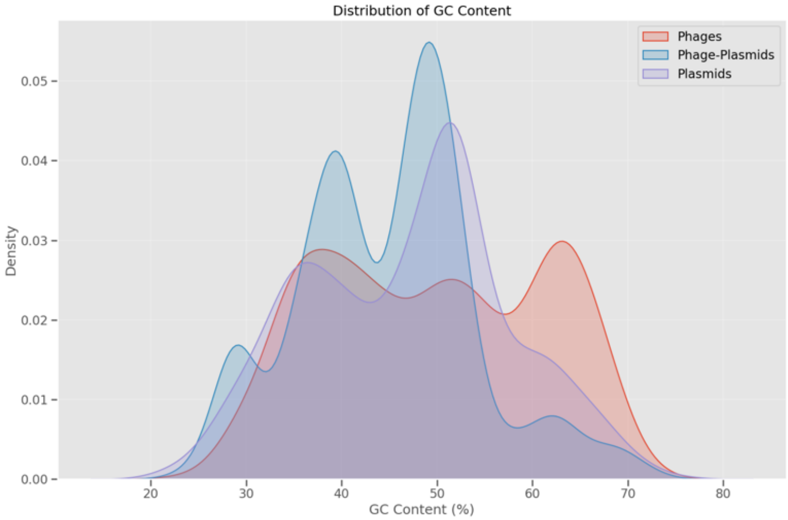


Figure 5. GC Content Distribution Across MGE Classes. Density plots showing the distribution of GC content across phages (blue), phage-plasmids (orange), and plasmids (green), with vertical lines indicating the mean values for each class.

Statistical testing confirmed significance for all pairwise comparisons (p-value < 0.05s, two-tailed t-test).

- Phages vs. Phage-Plasmids: t = 12.12, p = 5.26 × 10^{-33}
- Phages vs. Plasmids: t = 6.34, p = 2.66 × 10^{-10}
- Phage-Plasmids vs. Plasmids: t = -5.72, p = 1.17 × 10^{-8}

By analyzing the frequency of nucleotides, we were able to gain further insight into the biasing composition of each MGE class as shown in figure 6. Phages and phage-plasmids had similar percent composition of A and T. Phages values were about 26.6 and 25.5 respectively whereas phage-plasmids values were 27.1 and 27.0. Unlike that, the distribution of nucleotides in plasmids was much more even, with G and C (26.6% and 26.6%, respectively) being present in higher proportions than A and T (23.4% and 23.4%)(Figure 6).

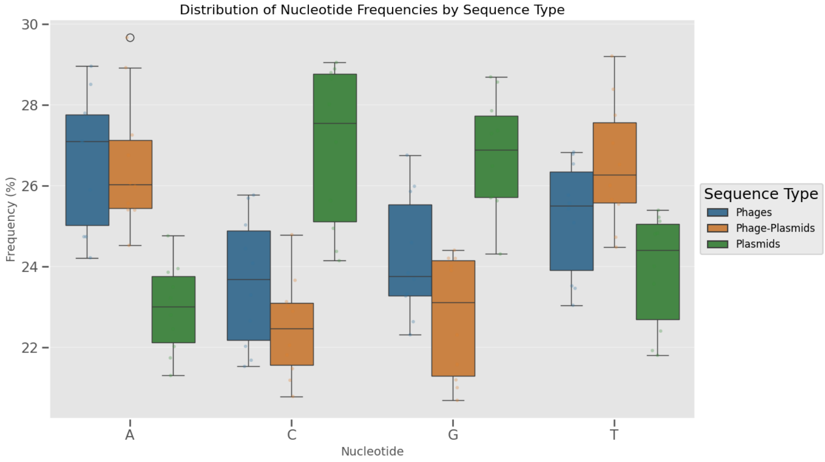


Figure 6. Nucleotide Frequencies (%) by MGE Class

4.1.3 K-mer Frequency Patterns

Discriminative K-mer frequency patterns were detected across the three MGE classes. The dimer with the most frequency in phages is AA/TT whose average counts were

5508.7/5070.9. In plasmids, it is GC*CG that has A maximum frequency with 9072.7/8635.4 as average counts.

Phage-plasmids, notably, showed intermediate k-mer frequencies for many motifs, consistent with their hybrid nature (Table 2).

Table 2. Top 10 K-mers (k=2) by Average Frequency for Each MGE Class

Rank	Nucleotide	Phages	Phage-Plasmids	Plasmids
1	AA	5508.72	6943.21	9072.67
2	TT	5070.92	6908.62	8635.41
3	AT	4793.36	6005.50	7010.75
4	GA	4657.40	4980.71	6984.29
5	TG	4563.79	5287.60	6704.17
6	CA	4502.38	5315.06	6665.27
7	GC	4384.51	4951.62	6845.48
8	CG	4370.77	4531.50	6822.71
9	AC	4179.18	5176.64	7020.86
10	AG	4129.53	4689.84	6989.13

Using PCA and t-SNE for dimensionality reduction of the data also visually confirms the differences in composition of the three classes (Figure 7). The first two principal components accounted for 98.92% of the variance in the PCA plot. Furthermore, the phage-plasmids' position along PC1 was between phages and plasmids. t-SNE showed stronger clustering with several phage-plasmids. Hybrid class may contain different subgroups arising from more than one plasmid or evolutionary origin.

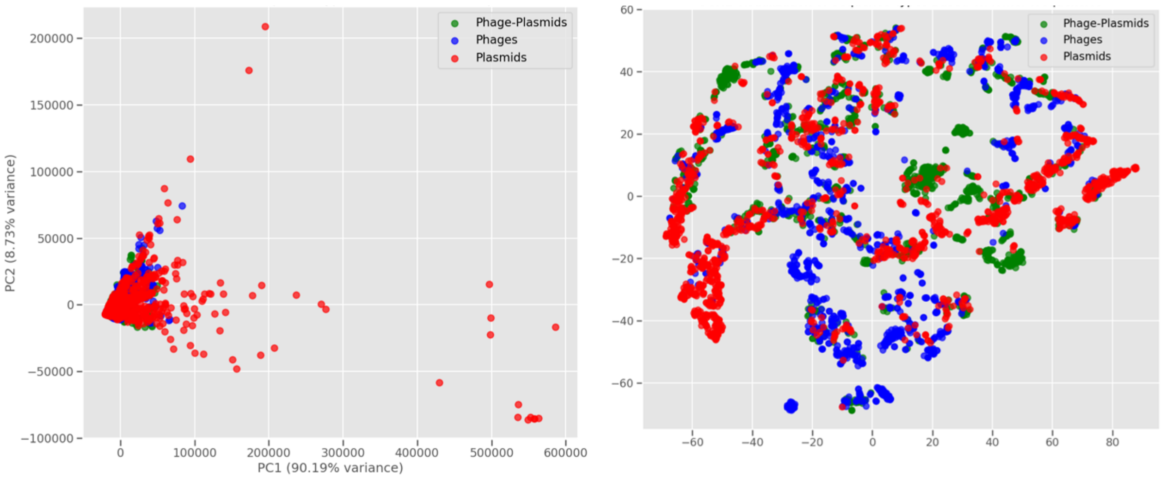


Figure 7. Dimensionality Reduction and Feature Visualization of Mobile Genetic Element Classes.

(A) PCA plot showing the distribution of phages (blue), phage-plasmids (orange), and plasmids (green) based on k-mer frequencies. (B) t-SNE visualization of k-mer frequency profiles showing clustering of the three MGE classes.

4.2 Model Performance

4.2.1 Logistic Regression Results

The Logistic Regression model, serving as our baseline classifier, achieved an overall accuracy of 85% on the test dataset. Class-specific performance metrics revealed varying effectiveness across the three MGE classes (Table 3, Figure 8).

Table 3. Logistic Regression Classification Performance by Class

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Phages	0.91	0.86	0.89	313
Phage-Plasmids	0.81	0.82	0.82	288
Plasmids	0.83	0.87	0.85	249
Macro Average	0.85	0.85	0.85	850
Weighted Average	0.85	0.85	0.85	850

The model worked best for classifying phages (F1 = 0.89) and worst for phage-plasmids (F1 = 0.82), showing that the latter is relatively harder to identify. Error rate analysis showed relatively high false negative rates for the phage-plasmids (FNR=0.1806), indicating that a substantial number of hybrid elements were misclassified as phages or plasmids.

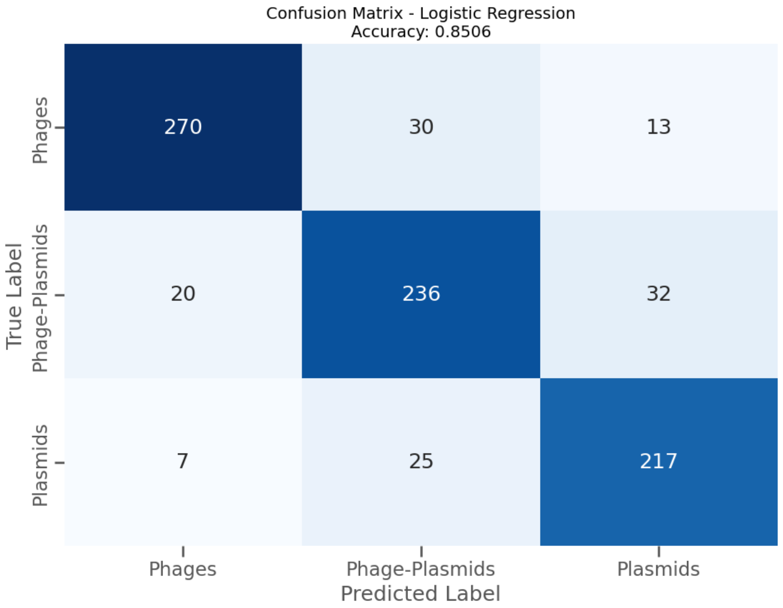


Figure 8: Logistic Regression Classification Performance Visualization.

Confusion matrix showing the distribution of true vs. predicted classes for the logistic regression model, with color intensity reflecting prediction counts.

This pattern is also seen in the AUC scores. Which gives AUC scores of 0.9623, 0.9095, and 0.9442 to phages, phage-plasmids and plasmids respectively.

4.2.2 Random Forest Results

The Random Forest classifier, optimized through extensive grid search, achieved an overall accuracy of 89% on the test dataset, representing a substantial improvement over Logistic Regression. As shown in Table 4, class-specific metrics are presented, and these metrics are visualized in Figure 9.

Table 4. Random Forest Classification Performance by Class

Class	Precision	Recall	F1-Score	Support
Phages	0.92	0.93	0.93	313
Phage-Plasmids	0.91	0.83	0.87	288
Plasmids	0.84	0.90	0.87	249
Macro Average	0.89	0.89	0.89	850
Weighted Average	0.89	0.89	0.89	850

The Random Forest model showed marked improvements in both precision and recall for phage-plasmids compared to Logistic Regression, increasing the F1-score from 0.82 to 0.87. This enhancement likely reflects the model's ability to capture non-linear relationships in the feature space that better distinguish hybrid elements.

As per error analysis across all classes, the false positive rates (FPR) were significantly lower (average FPR = 0.0548) as compared to logistic regression (average FPR = 0.0743). But phage-plasmids had a relatively high false negative rate (FNR = 0.1667) indicating persistent challenges in hybrid element detection.

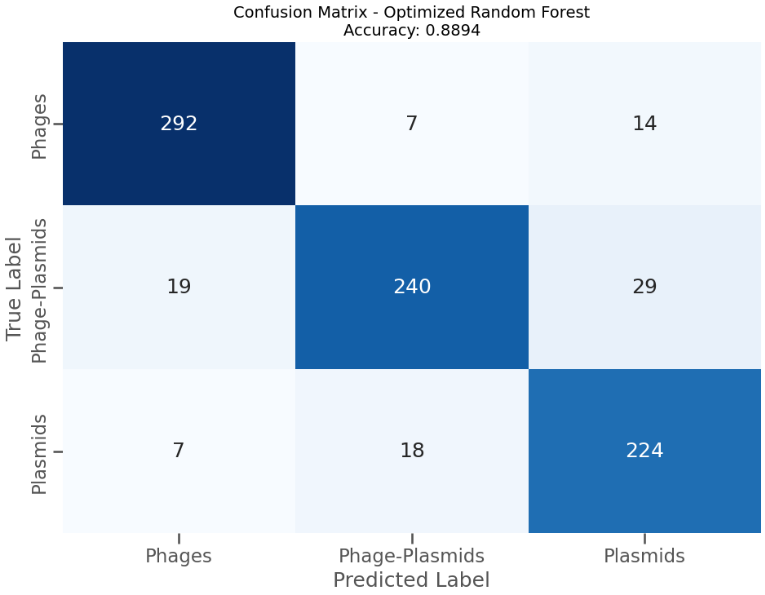


Figure 9. Random Forest Classification Performance Visualization.

Confusion matrix showing the distribution of true vs. predicted classes for the Random Forest model, with color intensity reflecting prediction counts.

The AUC scores for the Random Forest model were found to be very high (above 0.95) for the three classes. The AUC score was 0.9848 for phages, 0.9585 for phage-plasmids, 0.9740 for plasmids and their average was 0.9724.

4.2.3 CNN Results

The Convolutional Neural Network model achieved the highest overall accuracy at 90%, outperforming both Logistic Regression and Random Forest. Table 5 summarizes the performance metrics for each class, and Figure 6 provides a visual representation of the same.

Table 5. CNN Classification Performance by Class

Class	Precision	Recall	F1-Score	Support
Phages	0.91	0.94	0.93	313
Phage-Plasmids	0.93	0.82	0.87	288
Plasmids	0.86	0.95	0.91	249
Macro Average	0.90	0.90	0.90	850
Weighted Average	0.90	0.90	0.90	850

However, the recall for phage-plasmids (0.82) remained comparable to other models, suggesting that the CNN was highly selective but still missed a similar proportion of true hybrid elements.

The CNN achieved the highest F1-score (0.91) of any model-class combination for plasmids, fueled by excellent recall (0.95); thus, illustrating the model's strength in identifying plasmid sequences across a wide range of sizes and compositions. (Figure 10).

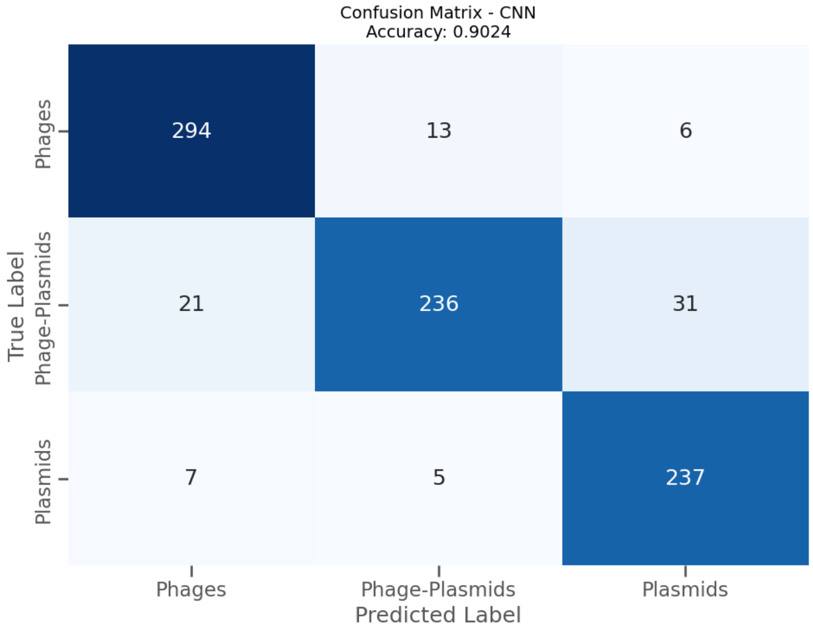


Figure 10. CNN Classification Performance Visualization.
Confusion matrix for the CNN model demonstrating improved classification accuracy across all three classes compared to the Random Forest model.

Error analysis revealed that the CNN achieved the lowest average false positive rate (0.0486) across all classes, representing a 35% reduction compared to Logistic Regression and 11% reduction compared to Random Forest. The false negative rates were also lowest with the CNN (average FNR = 0.0965), indicating superior overall sensitivity.

The CNN model yields the best AUC scores of all models: 0.9851 for phages, 0.9664 for phage-plasmids, 0.9781 for plasmids, and average AUC score of 0.9765.

4.3 ROC examination and Comparative Model Performance

Figure 11 shows the ROC curves for each model and for each class. The ROC curves plot the true positive against the false positive using different classification thresholds. with the area under the curve (AUC) providing a quantitative measure of classification performance.

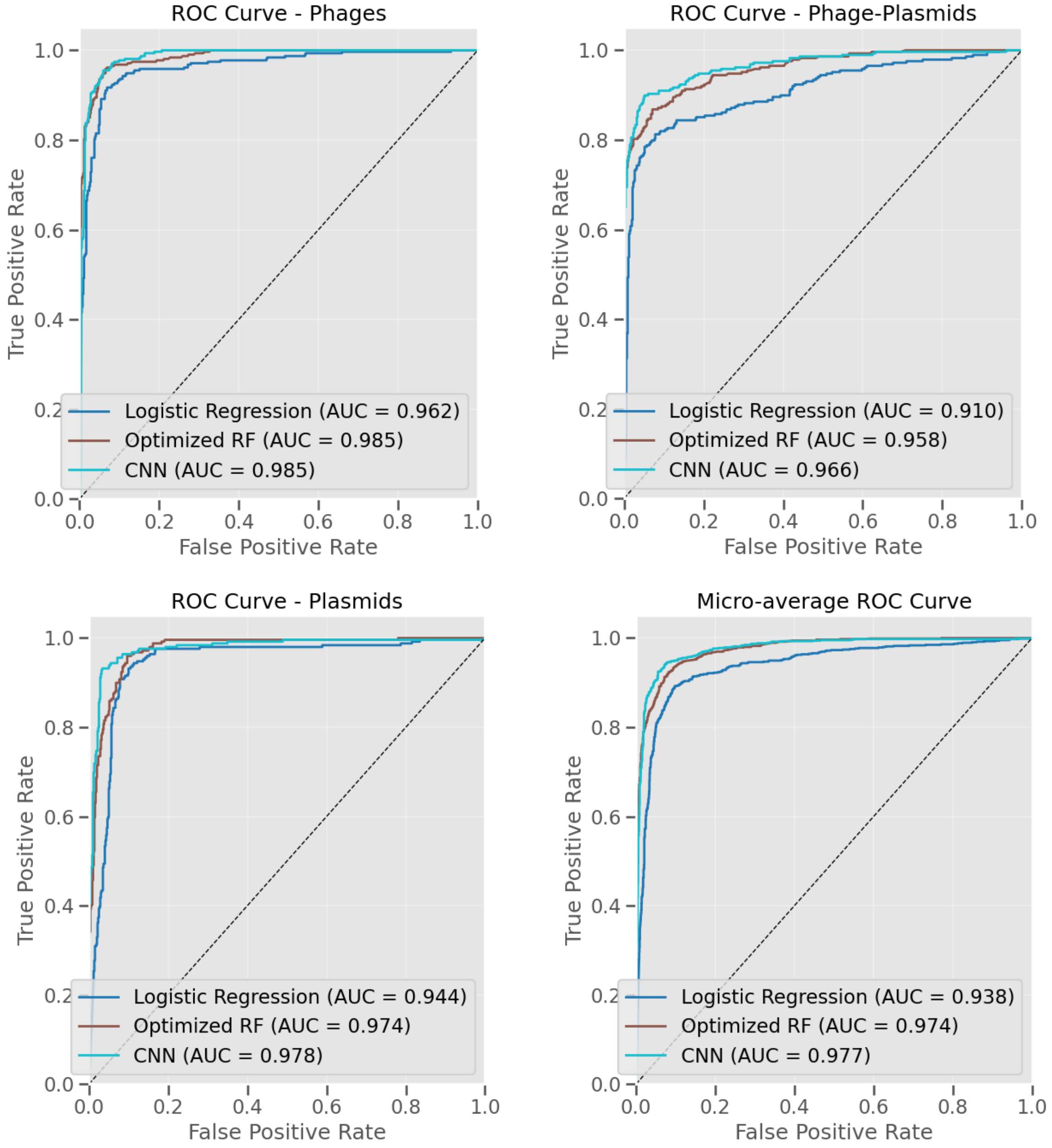


Figure 11. ROC Curves for Classification Models

(A) ROC curves for phage classification. (B) ROC curves for phage-plasmid classification. (C) ROC curves for plasmid classification. Different colors represent different models: Logistic Regression (blue), Random Forest (orange), and CNN (green).

As visualized in Figure 11, all models performed well in distinguishing phages and plasmids (AUC > 0.94), but showed comparatively lower performance for phage-plasmids. This pattern is consistent across all three models, indicating that hybrid elements present inherent classification challenges regardless of the algorithm employed. However, the CNN model demonstrated the most balanced performance across all three classes, with the highest AUC for phage-plasmids (0.9664).

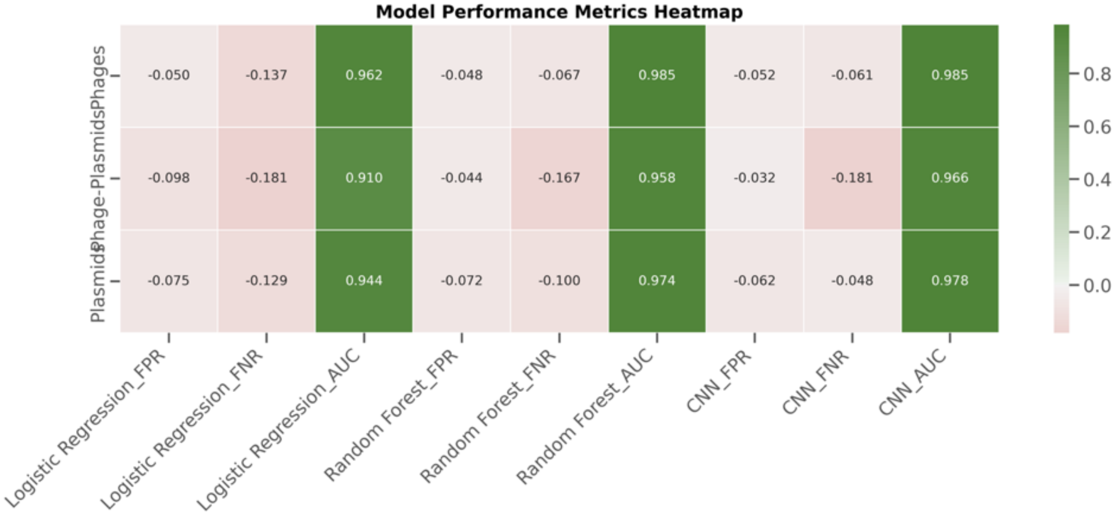


Figure 12. Comparative Model Performance

Heatmap comparing key performance metrics across the three models, showing progressive improvement from Logistic Regression to Random Forest to CNN.

The comparison of three models in terms of various parameters is shown in figure 12. It is evident that the performance of random forest and CNN is better than that of logistic regression. The CNN outperformed both Logistic Regression and Random Forest, while Random Forest model outperformed Logistic Regression model.

5. Discussion

5.1 Interpretation of Key Findings

This study aimed to classify bacteriophages (phages), plasmids, and phage-plasmid hybrids using three machine learning approaches. All models performed significantly above random chance (~33% for three classes), with the baseline logistic regression attaining 85% accuracy. The random forest reached 89% accuracy, and the CNN achieved 90% accuracy, confirming that more complex, non-linear models better capture distinguishing patterns in sequence data.

The CNN outperformed other classifiers, demonstrating superior ability to capture complex sequence patterns. CNN performance was particularly strong for plasmid identification (F1-score of 0.91, recall of 0.95), suggesting it effectively learned subtle motifs that simpler models missed. The CNN model's high AUC scores (0.9851 for phages, 0.9664 for phage-plasmids, and 0.9781 for plasmids) further confirm its robust discriminative ability across all classes.

Our k-mer frequency approach (k=5) provided rich information for uncovering class-specific sequence signatures, validating our choice of feature representation. Pentamers offered an optimal balance between capturing sufficient sequence context and maintaining computational efficiency. By balancing the training dataset across classes, we ensured the classifiers did not become biased toward majority classes, leading to relatively uniform performance across categories. The CNN achieved F1-scores of 0.93 (phages), 0.87 (phage-plasmids), and 0.91 (plasmids), supporting Hypothesis 5 that balanced training produces models with more uniform performance across classes.

5.2 Model Performance by Class

The results across all models are consistent: phages and plasmids were classified more accurately than phage-plasmids. The best performance observed with phages as Precision 0.91, Recall 0.86, F1 = 0.89 and the worst with phage-plasmids with Precision 0.81, Recall 0.82, F1 = 0.82 which misclassifies a lot of hybrid sequences. Lower performance can be expected for hybrid elements since they are intermediate and linear models have limitations.

The Random Forest classifier showed marked improvement for the phage-plasmid class, raising its F1-score to 0.87 (from 0.82), largely through a boost in precision (91%) while maintaining recall around 83%. The model captured the feature interactions that define phage-plasmids well according to Ma et al. (2023). The CNN classified all input samples to their correct classes. The accuracy in identifying phages was very high - with a precision of 0.91, recall of 0.94 and F1 = 0.93. The classification of plasmids had an F1 score of 0.91. Fang *et al. (2019)* developed PPR-Meta, a deep learning tool that was among the first to simultaneously classify sequences as phage, plasmid, or chromosomal, illustrating the promise of CNNs in this domain. Likewise, a recent random forest-based classifier, SourceFinder, by (Aytan-Aktug et al., 2022) achieved high accuracy (AUC ≈0.94) for distinguishing chromosomal DNA, plasmids, and phages using k-mer distributions. Our CNN's performance (average AUC ~0.977) is comparable or superior to these state-of-the-art approaches, albeit our task omits chromosomal sequences and explicitly targets the hybrid phage-plasmid class.

The plasmid recall soared to 0.95, meaning the CNN missed very few plasmids. This high recall suggests the CNN was able to generalize across the great diversity of plasmid sizes and compositions (which ranged from kilobase miniature plasmids to megaplasmids). By scanning the k-mer composition with convolutional filters, the CNN captured subtle sequence patterns that are hallmarks of plasmids but were not fully leveraged by simpler models. This ability to automatically learn hierarchical features aligns with findings from Fadlil et al. (2022), who demonstrated CNN superiority in capturing complex patterns in classification tasks.

For phage-plasmids, the CNN achieved an F1 of 0.87, matching the Random Forest. Interestingly, the CNN traded a slight decrease in recall (82%, similar to logistic regression's 82% and just below the Random Forest's 83%) for a further increase in precision (93%). In practical terms, when the CNN predicts a sequence to be a phage-plasmid, it is very likely correct (only a 7% false discovery rate). This high precision is valuable for database curation or experimental follow-up, where one might prioritize high-confidence predictions for validation. The persistent recall around ~82% across all models for phage-plasmids suggests a subset of hybrid elements consistently evade detection, likely because they closely resemble conventional phages or plasmids. Kerkvliet et al. (2024) similarly observed that some phage-plasmids have lost many characteristic genes, blurring distinctions with classical elements.

Based on analysis of errors, the CNN showed the overall lowest false positive rate at 0.0486. This is at least a 35% improvement versus Logistic Regression and an 11% versus Random Forest. The average FNR, which comes out to be 0.0965, was also lowest with CNN, which indicated it had better sensitivity as well. This lower error rate allows for better classifications, especially for metagenomic use cases. It was seen at the class-level of performance that phages were easiest to classify, then plasmids, then the phage-plasmid which was hardest. The classification difficulty aligns with previous reports that variation in genetic modes creates problems. The application of sophisticated algorithms like CNN helped this performance gap in identifying hybrid MGEs.

5.3 Distinct sequence and unique genomic signature

The nucleotide composition and k-mer frequency analysis revealed clear differences corresponding to their hybrid nature and identifiable genomic signatures for each class.

5.3.1 Genome Size and Composition

The median genome length of phage-plasmids was found to be ~74 kb, which is longer than that of typical phages (median ~50 kb) and plasmids (median ~51 kb). However, plasmids were found to show higher length variability than phage-plasmids. It would seem that the phage-plasmids will need to encode elements required for phage structure, as well as functions which allow for plasmid maintenance, which imposes an upper size limit. Higher median length than parents denotes selective constraints requiring essential genes from both phage and plasmid. The minimum size of the phage-plasmid was 10,545 bp and was higher than the minimum of phage 3,405 bp and that of plasmid 619 bp. This means that a size threshold exists below which elements cannot house both functions. Shan and colleagues (2023) state that the need for phage-plasmids to encode both phage and plasmid functions reflects their hybrid nature.

5.3.2 Nucleotide Composition

Statistical analysis of nucleotide composition revealed distinctive patterns across the three MGE classes. Plasmids had the highest average GC content (~53.2%), whereas phages averaged around 46.0% GC. Interestingly, phage-plasmids showed an average GC of ~45.9%, mirroring phages more closely. Statistical testing confirmed the significance of these differences ($p < 0.05$, two-tailed t-test) for all pairwise comparisons, despite the similarity between phage and phage-plasmid means.

At the individual nucleotide level, phages and phage-plasmids showed similar proportions of adenine (A) and thymine (T), with values of approximately 26.6% and 25.5% for phages, and 27.1% and 27.0% for phage-plasmids, respectively. In contrast, plasmids exhibited more balanced nucleotide distributions, with slightly higher proportions of guanine (G) and cytosine (C) (26.6% each) compared to A and T (23.4% each).

The fact that phage-plasmids did not simply average the GC content of phages and plasmids but leaned towards the lower GC/AT-rich side could be explained by several factors. Parra et al. (2023) noted that AT-richness in certain phage regions (packaging sites, regulatory regions) might explain why phage-plasmids lean toward lower GC content despite carrying some plasmid genes.

5.3.3 K-mer Signatures

MGE Classes show different sequence characteristics according to k-mer. Phages were enriched for AT dimers (AA and TT: 5,508.7 and 5,070.9 average counts), while plasmids were enriched for GC-rich dimers (GC and CG: 9,072.7 and 8,635.4 average counts). Phage-plasmids were intermediate in frequency demonstrating mixed genotypes. Analysis via PCA and t-SNE revealed those compositional differences. The first two principal components explained 98.92% of variance. T-SNE showed different clusters, suggesting that they have different origins. Haudiquet et al. (2024) observed the AT-rich motifs also enable genetic exchanges in mobile elements which may help to preserve hybrid status. These pentamers that have been enriched could serve as signature motifs for hybrid elements.

5.4 Implications for Mobile Genetic Element Classification

Creating very accurate classification models for phages, phage-plasmids, and plasmids has important consequences for MGE study and the overall classification strategy.

5.4.1 Beyond Binary Classification

Traditional MGE classification used binary frameworks categorizing elements as phages or plasmids, leading to misclassification of hybrid elements like phage-plasmids. Our three-way classification system addresses this by modeling phage-plasmids as a distinct class. This approach recognizes MGEs exist on a continuum rather than discrete categories, as noted by Shan et al. (2023). Our models' high performance proves this three-way classification is both biologically appropriate and computationally feasible. Kerkvliet et al. (2024) proposed integrated approaches using key genetic markers for phage and plasmid functions, allowing elements on a continuum. The identification of class-specific patterns provides a foundation for nuanced classification systems, with potential to expand beyond three classes to recognize additional hybrid categories.

5.4.2 K-mer Embedding Effectiveness

The effectiveness of k-mer frequency embeddings for MGE classification, particularly with k=5, confirms the value of this approach. Unlike methods relying on gene annotation or homology, k-mer-based approaches capture statistical patterns without requiring extensive prior knowledge, making them valuable for identifying novel elements. Our findings align with studies demonstrating the utility of k-mer embeddings. Lötsch & Ultsch (2023) used tetranucleotide frequency vectors to classify metagenomic contigs with over 90% accuracy, while Song (2020) found that k=5 provides optimal balance between specificity and computation. The effectiveness of pentamer frequencies aligns with Mariotti et al. (2023), who found they provide robust discrimination while maintaining efficiency. This approach addresses a limitation of existing tools that depend on protein family databases or reference genomes. By focusing on sequence composition rather than specific genes, our models are more robust to MGE genetic diversity and mosaicism. This aligns with Schackart et al. (2023)

and Kerkvliet et al. (2024), who found composition-based approaches can identify sequences missed by homology-based methods.

5.5 Biological Significance and Applications

The results from this study have various biological significance and application prospects for a range of microbiological fields.

5.5.1 Evolutionary Insights

The intermediate compositional characteristics of phage-plasmids provide computational evidence supporting their role as evolutionary intermediates between phages and plasmids. The differences in sequence length, GC content, and k-mer distributions align with their proposed role as bridges facilitating genetic exchange between MGEs. These findings complement work by Kerkvliet et al. (2024), who showed phage-plasmids can promote recombination between phages and plasmids. Our identification of sequence motifs enriched in phage-plasmids may provide insights into mechanisms enabling this recombination. The multiple phage-plasmid clusters observed suggest diverse evolutionary origins, with some more closely related to phages and others to plasmids, reflecting different stages in the evolutionary continuum. This aligns with findings from Hilpert et al. (2020) on prophage domestication and Shan et al. (2023) on hybrid elements in bacterial evolution.

5.5.2 Metagenomic Applications

The models produced in this work which have high accuracy can be applied immediately for metagenomic analysis, which is currently lacking in bioinformatic pipelines. The models allow for reliable identification of phage-plasmids in complex metagenomic datasets, offering a more comprehensive characterization of mobile genetic elements in microbial communities. This feature is useful for examining horizontal gene transfer across habitats ranging from human microbiomes to environmental specimens. Understanding the transmission of antibiotic resistance genes and metabolic genes is improved by the identification of phage-plasmids. Common metagenomic tools usually analyse phages or plasmids, not both. VIBRANT (Schackart et al., 2023), VirSorter (Schackart et al., 2023) and DeepVirFinder (Rahman & Rangwala, 2020) are tools targeting sequences from viruses, PlasmidFinder (Carattoli & Hasman, 2020) and PlasFlow (Krawczyk et al, 2018) target plasmids.

We explicitly model phage-plasmids as a separate class to tackle this issue. Due to its accuracy for phage-plasmid identification, the CNN model can be useful in metagenomic screening, as it can generate low false positives and demands less computing resources compared to prophage prediction or gene annotation.

5.5.3 Clinical and Biotechnological Relevance

Improved detection of phages and plasmids can help to monitor the transmission of antibiotic resistance genes. As shown by Kerkvliet et al. (2024), phage-plasmids can act as vectors for the spread of resistance genes in different bacterial populations and host ranges. The models developed in this work could potentially be used in surveillance systems to interrogate clinical samples for the presence of these hybrid vectors and monitor their frequency over time, aligning with comments by Haudiquet et al. (2024) staging multifunctional mechanisms for monitoring resistance gene transmission vectors. The phage-plasmids can be identified so that new genetic tools can be developed which will have the properties of both. Shan and colleagues (2023) utilized engineered hybrids of phage and plasmid towards applications in biotechnology where targeted delivery of a phage and stability of a plasmid are required. The found motifs can be applied to create novel vectors with new properties.

5.6 Limitations

This study advances phage, plasmid, and phage-plasmid classification, but has important limitations.

Dataset Composition and Bias: The balanced training sets do not reflect natural abundance, where plasmids outnumber phage-plasmids. This may lead to under-detection of rare elements in metagenomic data. Public databases contain taxonomic biases favoring certain hosts, potentially limiting generalizability. Phage-plasmid annotation subjectivity could introduce classification errors despite curation efforts.

Dataset Size and Diversity: Although including 4,000 genomes, the dataset may not fully represent all subgroups, creating uncertainty for novel sequences in metagenomic applications. We selected 1416 phages, and plasmid sequences from 3585 phage, and plasmid sequences respectively, which might lead to generalization problem of the classification models and for unknown sequences.

Fragmented Sequence Classification: Models trained on complete genomes may underperform on short metagenomic contigs that incompletely capture k-mer profiles.

Category Scope: The three-class classifier excludes bacterial chromosomes and other mobile elements, risking misclassification of chromosomal sequences. A broader classification framework may be needed for metagenomic use.

Computational Constraints: K-mer frequency extraction and CNN training require significant resources, though models remain lightweight. Hyperparameter optimization was time-intensive.

Interpretability: While CNNs perform well, their black-box nature limits understanding of classification decisions, despite k-mer importance analysis.

6. Conclusion

This research evaluated three machine learning approaches for classifying mobile genetic elements into phages, phage-plasmids, and plasmids, with the Convolutional Neural Network achieving 90% accuracy. The study shows phage-plasmids represent a distinct class with characteristics intermediate between phages and plasmids. We identified specific pentamer motifs for each MGE class, with AT-rich patterns in phages, GC-rich in plasmids, and unique signatures in phage-plasmids. The k-mer frequency approach proved effective for MGE classification without requiring gene annotation or homology. While all models performed well, the CNN's precision for phage-plasmid identification (93%) demonstrates deep learning's value, though recall challenges (~82%) persist across models. These findings impact horizontal gene transfer and microbial evolution understanding. Future work could enhance feature representations, expand model applications to metagenomic datasets, develop larger reference collections, and investigate the biological significance of identified motifs. The methodological framework combining k-mer frequency analysis with machine learning provides a template for addressing similar classification challenges in genomics. By modeling intermediate categories rather than forcing binary classifications, this approach acknowledges the continuous nature of biological diversity. This research demonstrates the effectiveness of machine learning for distinguishing MGEs based on sequence composition. The high performance validates k-mer frequency profiles for classification and supports phage-plasmids as hybrid entities bridging phages and plasmids.

7. References

Albuquerque, J., Bourbon, M., Medeiros, A. M., Alves, A. C., & Antunes, M. (2022). Comparative study on the performance of different classification algorithms, combined with pre- and post-processing techniques to handle imbalanced data, in the diagnosis of adult patients with familial hypercholesterolemia. *PLOS ONE*, *17*(6), e0269713. https://doi.org/10.1371/journal.pone.0269713

Arkoudi, I., Krueger, R., Azevedo, C. L., & Pereira, F. C. (2023). Combining discrete choice models and neural networks through embeddings: Formulation, interpretability and performance. *Transportation Research Part B: Methodological*, *175*, 102783. https://doi.org/10.1016/j.trb.2023.102783

Arredondo-Alonso, S., Corander, J., Samuelsen, Ø., Lanza, V. F., Pöntinen, A. K., Johnsen, P. J., Gama, J. A., Tonkin-Hill, G., Schürch, A. C., & Gladstone, R. A. (2023). Mge-cluster: a reference-free approach for typing bacterial plasmids. *NAR Genomics and Bioinformatics*, *5*(3). https://doi.org/10.1093/nargab/lqad066

Aytan-Aktug D, Grigorjev V, Szarvas J, Clausen PT, Munk P, Nguyen M, Davis JJ, Aarestrup FM, Lund O. SourceFinder: A machine-learning-based tool for identification of chromosomal, plasmid, and bacteriophage sequences from assemblies. Microbiology Spectrum. 2022 Dec 21;10(6):e02641-22. https://doi.org/10.1128/spectrum.02641-22

Baaziz, H., Baker, Z. R., Franklin, H. C., & Hsu, B. B. (2022). Rehabilitation of a misbehaving microbiome: phages for the remodeling of bacterial composition and function. IScience, 25(4), 104146. https://doi.org/10.1016/j.isci.2022.104146.

Bosco, K., Khatami, A., Lynch, S., & Sandaradura, I. (2023). Therapeutic Phage Monitoring: A Review. *Clinical Infectious Diseases*, *77*(Suppl 5), S384–S394. https://doi.org/10.1093/cid/ciad497

Carattoli, A., & Hasman, H. (2020). PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Horizontal gene transfer: methods and protocols*, 285-294.

Da Silva, G. C., Rosa, J. N., França, K. C., Gonçalves, O. S., Bossé, J. T., Bazzolli, D. M. S., Langford, P. R., & Santana, M. F. (2022). Mobile Genetic Elements Drive Antimicrobial Resistance Gene Spread in Pasteurellaceae Species. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.773284

Ding, P., Gao, X., Yu, B., Zhang, X., Wang, Y., & Liu, G. (2023). DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Briefings in Bioinformatics*, *24*(4). https://doi.org/10.1093/bib/bbad231

Dufresne, Y., Rahman, A., Marijon, P., Lemane, T., Chikhi, R., Kokot, M., Peterlongo, P., Medvedev, P., & Deorowicz, S. (2022). The K-mer File Format: a standardized and compact disk representation of sets of k-mers. *Bioinformatics*, *38*(18), 4423–4425. https://doi.org/10.1093/bioinformatics/btac528

Fadlil, A., Nugroho, A. S., Sunardi, S., & Umar, R. (2022). Comparison of Machine Learning Approach for Waste Bottle Classification. *Emerging Science Journal*, *6*(5), 1075–1085. https://doi.org/10.28991/esj-2022-06-05-011

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. https://doi.org/10.1093/gigascience/giz066.

Gauthier, C. H., & Hatfull, G. F. (2023). PhamClust: a phage genome clustering tool using proteomic equivalence. *MSystems*, *8*(5). https://doi.org/10.1128/msystems.00443-23

Grigson, S. R., Giles, S. K., Edwards, R. A., & Papudeshi, B. (2023). Knowing and Naming: Phage Annotation and Nomenclature for Phage Therapy. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *77*(Suppl 5), S352–S359. https://doi.org/10.1093/cid/ciad539

Gunasekaran, H., Ramalakshmi, K., Rex, A., Arokiaraj, M., Kanmani, S. D., Venkatesan, C., Suresh, C., & Dhas, G. (2021). Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, Article ID 1835056. https://doi.org/10.1155/2021/1835056

Haudiquet, M., Le Bris, J., Nucci, A., Bonnin, R. A., Domingo-Calap, P., Rocha, E. P. C., & Rendueles, O. (2024). Capsules and their traits shape phage susceptibility and plasmid conjugation efficiency. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-46147-5

Hilpert, C., Bricheux, G., & Debroas, D. (2020). Reconstruction of plasmids by shotgun sequencing from environmental DNA: which bioinformatic workflow? *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/bib/bbaa059

Igler, C. (2022). Phenotypic flux: The role of physiology in explaining the conundrum of bacterial persistence amid phage attack. *Virus Evolution*, *8*(2). https://doi.org/10.1093/ve/veac086

Jiang, Z., Liu, Y., & Yang, J. (2021). Imbalanced Learning with Oversampling based on Classification Contribution Degree. *Advanced Theory and Simulations*, *4*(5), 2100031. https://doi.org/10.1002/adts.202100031

Keith, M., Park De La Torriente, A., Chalka, A., Vallejo-Trujillo, A., Mcateer, S. P., Paterson, G. K., Low, A. S., & Gally, D. L. (2024). Predictive phage therapy for Escherichia coli urinary tract infections: Cocktail selection for therapy based on machine learning models. *Proceedings of the National Academy of Sciences*, *121*(12). https://doi.org/10.1073/pnas.2313574121

Kerkvliet, J. J., Meneses, R., Bossers, A., Schürch, A. C., Willems, R., & Kers, J. G. (2024). Metagenomic assembly is the main bottleneck in the identification of mobile genetic elements. *PeerJ*, *12*, e16695. https://doi.org/10.7717/peerj.16695

Kim, N., Ma, J., Kim, W., Kim, J., Belenky, P., & Lee, I. (2024). Genome-resolved metagenomics: a game changer for microbiome medicine. *Experimental & Molecular Medicine*, *56*(7), 1501–1512. https://doi.org/10.1038/s12276-024-01262-7

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic acids research, 46(6), e35-e35. https://doi.org/10.1093/nar/gkx1321

Krützfeldt, L.-M., Kircher, M., & Schubach, M. (2020). The impact of different negative training data on regulatory sequence predictions. *PLOS ONE*, *15*(12), e0237412. https://doi.org/10.1371/journal.pone.0237412

Lee, D. Y., Bartels, C., Mcnair, K., Edwards, R. A., Swairjo, M. A., & Luque, A. (2022). Predicting the capsid architecture of phages from metagenomic data. *Computational and Structural Biotechnology Journal*, *20*(1), 721–732. https://doi.org/10.1016/j.csbj.2021.12.032

Liu, P., Yang, S., & Yang, S. (2021). KTU: K-mer Taxonomic Units improve the biological relevance of amplicon sequence variant microbiota data. *Methods in Ecology and Evolution*, *13*(3), 560–568. https://doi.org/10.1111/2041-210x.13758

Liu, Z., Zhou, Z., Gu, Y., Liu, X., Li, L., Jiang, C., Drewniak, L., Li, X., Yang, Z., Jiang, Z., Meng, D., Liang, Z., Liu, Y., Liu, T., Yin, H., Wang, J., Huang, Y., Tao, J., & Liu, S. (2021). Mobile genetic elements mediate the mixotrophic evolution of novel Alicyclobacillus species for acid mine drainage adaptation. *Environmental Microbiology*, *23*(7), 3896–3912. https://doi.org/10.1111/1462-2920.15543

Lötsch, J., & Ultsch, A. (2023). Recursive computed ABC (cABC) analysis as a precise method for reducing machine learning based feature sets to their minimum informative size. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-32396-9

Ma, Y., Yuan, F., Chen, H., & Lu, Q. (2023). Comparison of the effectiveness of different machine learning algorithms in predicting new fractures after PKP for osteoporotic vertebral compression fractures. *Journal of Orthopaedic Surgery and Research*, *18*(1). https://doi.org/10.1186/s13018-023-03551-9

Malhotra, R., & Lata, K. (2020). An empirical study on predictability of software maintainability using imbalanced data. *Software Quality Journal*, *28*(4), 1581–1614. https://doi.org/10.1007/s11219-020-09525-y

Mariotti, E., Alonso Moral, J. M., & Gatt, A. (2023). Exploring the balance between interpretability and performance with carefully designed constrainable Neural Additive Models. *Information Fusion*, *99*, 101882. https://doi.org/10.1016/j.inffus.2023.101882

Mitchell, S., Bull, M., Muscatello, G., Chapman, B., & Coleman, N. V. (2021). The equine hindgut as a reservoir of mobile genetic elements and antimicrobial resistance genes. *Critical Reviews in Microbiology*, *47*(5), 543–561. https://doi.org/10.1080/1040841x.2021.1907301

Nguyen, H. M., Wannigama, D. L., Watanabe, S., Kawaguchi, T., Tan, X.-E., Sharmin, S., & Cui, L. (2023). RNA and Single-Stranded DNA Phages: Unveiling the Promise from the Underexplored World of Viruses. *International Journal of Molecular Sciences*, *24*(23), 17029. https://doi.org/10.3390/ijms242317029

Nieto-Del-Amor, F., Diago-Almela, V. J., Ye-Lin, Y., Monfort-Ortiz, R., Hao, D., Diaz-Martinez, A., Garcia-Casado, J., & Prats-Boluda, G. (2022). Combination of Feature Selection and Resampling Methods to Predict Preterm Birth Based on Electrohysterographic Signals from Imbalance Data. *Sensors*, *22*(14), 5098. https://doi.org/10.3390/s22145098

Parra, B., Lutz, V. T., Cockx, B., Brøndsted, L., Dechesne, A., & Smets, B. F. (2023). Isolation and characterization of novel plasmid-dependent phages infecting bacteria carrying diverse conjugative plasmids. *Microbiology Spectrum*, *12*(1). https://doi.org/10.1128/spectrum.02537-23

Partridge, S. R., Enne, V. I., Grohmann, E., Hall, R. M., Rood, J. I., Roy, P. H., Thomas, C. M., & Firth, N. (2021). Classifying mobile genetic elements and their interactions from sequence data: The importance of existing biological knowledge. *Proceedings of the National Academy of Sciences*, *118*(35), e2104685118. https://doi.org/10.1073/pnas.2104685118

Piya, D., Nolan, N., Mutalik, V. K., Arkin, A. P., Cress, B. F., Young, R., Moore, M. L., & Ramirez Hernandez, L. A. (2023). Systematic and scalable genome-wide essentiality mapping to identify nonessential genes in phages. *PLOS Biology*, *21*(12), e3002416. https://doi.org/10.1371/journal.pbio.3002416

Rahman, M. A., & Rangwala, H. (2020). IDMIL: an alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data. *Bioinformatics*, *36*(Suppl 1), i39–i47. https://doi.org/10.1093/bioinformatics/btaa477

Sajja, T. K., & Kalluri, H. K. (2021). Image classification using regularized convolutional neural network design with dimensionality reduction modules: RCNN–DRM. *Journal of Ambient Intelligence and Humanized Computing*, *12*(10), 9423–9434. https://doi.org/10.1007/s12652-020-02663-y

Sasaki, H., & Sakata, I. (2020). Business partner selection considering supply-chain centralities and causalities. *Supply Chain Forum: An International Journal*, *22*(1), 74–85. https://doi.org/10.1080/16258312.2020.1824531

Schackart, K. E., Graham, J. B., Ponsero, A. J., & Hurwitz, B. L. (2023). Evaluation of computational phage detection tools for metagenomic datasets. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1078760

Shaidullina, A., & Harms, A. (2022). Toothpicks, logic, and next-generation sequencing: systematic investigation of bacteriophage-host interactions. *Current Opinion in Microbiology*, *70*, 102225. https://doi.org/10.1016/j.mib.2022.102225

Shan, X., Cordero, O. X., & Szabo, R. E. (2023). Mutation-induced infections of phage-plasmids. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-37512-x

Smug, B. J., Szczepaniak, K., Rocha, E. P. C., Dunin-Horkawicz, S., & Mostowy, R. J. (2023). Ongoing shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial hosts. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-43236-9

Song, K. (2020). Classifying the Lifestyle of Metagenomically-Derived Phages Sequences Using Alignment-Free Methods. *Frontiers in Microbiology*, *11*(1261498). https://doi.org/10.3389/fmicb.2020.567769

Sowah, R. A., Mills, G. A., Kuditchar, B., Twum, R. A., Acakpovi, A., Buah, G., & Agboyi, R. (2021). HCBST: An Efficient Hybrid Sampling Technique for Class Imbalance Problems. *ACM Transactions on Knowledge Discovery from Data*, *16*(3), 1–37. https://doi.org/10.1145/3488280

Taner, A., Mengstu, M. T., Ungureanu, N., Gür, İ., Selvi, K. Ç., & Duran, H. (2024). Apple Varieties Classification Using Deep Features and Machine Learning. *Agriculture*, *14*(2), 252. https://doi.org/10.3390/agriculture14020252

Villegas-Morcillo, A., Sanchez, V., & Gomez, A. M. (2021). FoldHSphere: deep hyperspherical embeddings for protein fold recognition. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04419-7

Wang, M., Patsenker, J., Li, H., Kluger, Y., & Kleinstein, S. H. (2023). Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic Acids Research*, *52*(2), 548–557. https://doi.org/10.1093/nar/gkad1128

Wang, X., Jing, L., Cao, S., Yu, J., & Liu, B. (2020). Important sampling based active learning for imbalance classification. *Science China Information Sciences*, *63*(8). https://doi.org/10.1007/s11432-019-2771-0

Wang, Y., Wang, J., Ru, X., Gao, X., & Sun, P. (2022). A hybrid feature selection algorithm and its application in bioinformatics. *PeerJ Computer Science*, *8*, e933. https://doi.org/10.7717/peerj-cs.933

Xiao, X., Zou, Y., Huang, J., Luo, X., Yang, L., Li, M., Yang, P., Ji, X., & Li, Y. (2024). An interpretable model for landslide susceptibility assessment based on Optuna hyperparameter optimization and Random Forest. *Geomatics, Natural Hazards and Risk*, *15*(1). https://doi.org/10.1080/19475705.2024.2347421

Zhang, Y., & Li, Z. (2023). RF_phage virion: Classification of phage virion proteins with a random forest model. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2022.1103783

Zhang, Y., Gao, M., Qiu, T., Guo, Y., & Wang, X. (2022). Bacteriophages: Underestimated vehicles of antibiotic resistance genes in the soil. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.936267

Zhang, Y., Mao, M., Zhang, R., Liao, Y. T., & Wu, V. C. (2024). DeepPL: A deep-learning-based tool for the prediction of bacteriophage lifecycle. PLOS Computational Biology, 20(10), e1012525. https://doi.org/10.1371/journal.pcbi.1012525