MACHINE LEARNING TO HUNT FOR PHAGE PROTEINS TO CATCH *Klebsiella*

George Wright[1,*], **Slawomir Michniewski**[2,*], **Eleanor Jameson**[2,3$], **Fayyaz Minhas**[1,$]
[*] Joint First Authors, [$] Joint Last Authors
[1]Department of Computer Science,[2] School of Life Sciences, University of Warwick, Coventry, UK.
[3]School of Natural Sciences, Bangor University, UK.
Correspondence: fayyaz.minhas@warwick.ac.uk.

ABSTRACT

Antimicrobial resistance (AMR) has been declared a global threat by the World Health Organization. Development of novel and effective therapies against microbes is an active research area of ever-growing importance. One of the leading threats are *Klebsiella* species, which cause virulent AMR infections with high death rates, particularly in hospital settings. *Klebsiella* species are particularly problematic because they produce a thick sticky polysaccharide capsule that protects them from antimicrobials and allows them to build highly resistant biofilms - defensive layers of cells. A natural solution to eradicate *Klebsiella* capsules and biofilms are depolymerase proteins that can target and neutralize polysaccharide capsules of specific *Klebsiella* species, often found in bacteriophages. However, machine learning guided discovery of depolymerase proteins in such phages is an unexplored area.

In this work, we use machine learning to help identify proteins in phage proteomes that can act as depolymerases against *Klebsiella*. Specifically, we utilize a dataset of phages, containing depolymerase proteins, that can target and neutralize polysaccharide capsules of specific *Klebsiella* species. We train a ranking model to rank proteins in an input phage proteome based on their predicted ability to act as a depolymerase. We use a non-redundant validation protocol to evaluate the predictive accuracy of the proposed model. Our analysis shows that for all test proteomes containing at least one depolymerase, the depolymerase protein was ranked within the top scoring 5% of proteins. We expect that the proposed approach (called DepoRanker) will be useful in accelerating the discovery of such antibacterial proteins in the wet lab.

1 INTRODUCTION

Antibiotic or antimicrobial resistance (AMR) represents a current threat to global health and security, making routine surgery and bacterial infections risky. Of particular concern are multidrug-resistant Enterobacteriaceae infections that are becoming increasingly difficult to treat with existing antibiotics available in clinical settings [6]. The ability of the Enterobacteriaceae to gain and transfer an increasing number of antimicrobial-resistance genes, particularly in health-care settings represents a critical threat to human health [20]. The World Health Organisation has named the Enterobacteriaceae *Klebsiella* as a priority pathogen that requires the urgent development of new antimicrobials due to high levels of resistance [25]. Another treatment for AMR infections is phage therapy: the use of bacteriophages (phages for short), to kill bacterial pathogens in a targeted manner [5].

Phages are natural killers of bacteria and phage therapy is emerging as a potential weapon against multi-drug resistant bacterial infections [4]. The commercial development of phages faces a number of challenges, including (but not limited to) small host range compared to antibiotics, potential adverse reactions in patients, toxins co-purified with phages and proteins of unknown functions encoded by phages. Ultimately, these hurdles can be overcome with an improved understanding and characterisation of phages.

Phages have very narrow host ranges, which is usually limited to one bacterial genera, species or subspecies that they can infect [12]. The host range of *Klebsiella* phages is further restricted by the wide range of polysaccharide capsules their hosts express, making the host bacteria intrinsically resistant to both antibiotic treatment and phage infections. However, some *Klebsiella* phages have been shown to be specific to particular host capsule types. This is frequently linked to phage encoded sugar-degrading enzymes called depolymerases that target specific capsule types [23, 11]. A small number of depolymerases have previously been characterised in wet lab studies, but the process is very time consuming [9, 16, 21, 13].

Klebsiella phage depolymerases are difficult to identify using standard annotation pipelines based on homology searches, such as BLAST. *Klebsiella* phage annotations usually identify at least one putative "tail-fibre" or "tail-spike" protein. Structural predictions suggest these proteins have beta-helical structures, a common protein architecture of both tail-spike proteins and the depolymerase enzymes that are suggested to have evolved from these [24, 14]. This allows us to carry out more targeted homology searches to identify putative depolymerase genes. However, we have identified a number of lytic phages in our collection that display depolymerase-like activity in wet lab experiments, namely the formation of semi-translucent halos around a clear phage plaque in plaque assays on solid agar medium, but lack identifiable depolymerase genes. We suggest that this is either because conventional depolymerases are not essential for the phages to permeate the capsule layer, or because of short-comings in the sequence-based annotation of phage genomes [17, 18].

Here, for the first time, we propose the use of a machine learning approach to improve on the standard bioinformatics pipelines to solve the problem of depolymerase identification as no such methods are currently available. Phage proteomes containing known depolymerases were used as learning and test datasets with characterised depolymerase proteins providing labels. Using amino acid composition features, our method ranks whole phage proteomes by their ability to act as a depolymerase. This was rigorously tested using non redundant data sets for cross validation and shows that known depolymerases rank highly in the proteomes.

2 Materials and Methods

We model the problem of predicting depolymerase proteins in a phage as a supervised machine learning problem in which the goal is to rank all proteins in a phage based on their ability to act as a depolymerase. Below, we describe different steps in the development of the proposed method.

2.1 Data Collection and Preprocessing

We collected information from previously published data consisting of depolymerase proteins in phages. The set of all depolymerases in this data constituted the positive class of our dataset. To get a negative dataset, we used the phage proteomes to which these proteins belonged. All proteins were taken and any already in the positive class were removed with the remaining proteins used to create the negative dataset. This created a final dataset which comprised of 24 phage proteomes with a total of 39 characterized depolymerase protein sequences and 2,601 non-depolymerase protein sequences.

2.2 Machine Learning Model

2.2.1 Feature Extraction

In this work, a simple amino acid composition of a protein is used as its feature representation [26, 22, 1]. Since the alphabet for a protein sequence has length 20, this gives a 20-dimensional non-normalised feature vector representation for a given protein sequence.

2.2.2 Gradient Boosting Model

We have used a ranking model based on Extreme Gradient Boosting (XGBoost) [3]. XGBoost is a fast and effective learning algorithm and has been widely used in machine learning, including predicting AMP in bacteria [26] and predicting protein function [27].

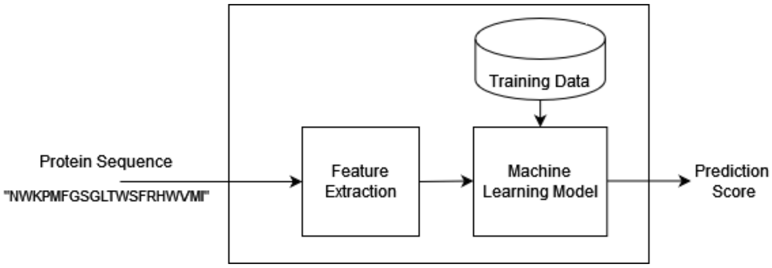

Figure 1: Proposed model framework

In this paper we have developed a ranking implementation of XGBoost which operates on proteomes in a pairwise fashion. The XGBoost ranking model is trained using multiple training proteomes to produce high scores for known depolymerases and low scores for non-depolymerases in each input phage. In testing, the same approach is used for generation of a ranked list of proteins in the test phage based on predicted deopolymerase activity. Stated formally, our objective is to obtain a prediction function from XGBoost, $f(x, \theta)$, with learnable parameters θ for a protein sequence represented in terms of its feature vector x. We require the model to learn the optimal parameters θ^* such that the score $f(x_i, \theta^*)$ for positive examples $x_i \in P_g$ (known depolymerases in phage g) is higher than $f(x_j, \theta^*)$ for $x_j \in N_g$ (non-depolymerase from the same phage) across all training phages. The hyper parameters of the learning model are selected through the cross-validation performance evaluation and the optimal results obtained are a learning rate of 0.1, a subsample of 0.9 and max tree depth of 3. The same mode has previously been utilized for the discovery of novel anti-CRISPR proteins [7].

As a baseline we have used a kernelized support vector machine (SVM) and BLAST. For the BLAST baseline, all test proteins in a phage were searched against the dataset of known depolymerases and sorted based on their e-values to rank proteins based on their expected depolymerase activity.

2.2.3 Performance Evaluation

To evaluate the performance of our machine learning model, we have performed non-redundant cross-validation on the data. Since we do not want to train the model on depolymerases with a similar sequence to those in the test data, we use the non-redundant training sets gained by clustering the proteins from CD-HIT as the folds for cross validation. All known depolymerases were clustered at a sequence threshold of 10%. If a depolymerase has a similarity score greater than 10% then it may contain the same structure and function [2]. We used CD-HIT [10] to create non-redundant sets based on sequence similarity of the 39 known depolymerases. The sets were then translated into sets of the phage proteomes to which the depolymerase belonged. If a phage contained multiple depolymerases separated into different clusters, then this phage would appear in multiple training sets. To ensure the non-redundancy of our training sets, in this scenario the sets were merged. We set aside all proteomes in the fold to be used as a validation set and any proteomes not removed were used as the training set.

The XGBoost model is trained on all proteins in the training set and the prediction scores were computed on a proteome-by-proteome basis. To report the prediction performance of the machine learning model, we have used Area under the Receiver Operating Characteristic Curve (AUROC) as well as Area under the Precision Recall Curve (AUPRC). An optimal predictor will rank all depolymerases higher than all other proteins in a test proteome.

For the background context of this problem, we want to find depolymerase proteins in as few wet-lab experiments as possible. To evaluate the predictive performance of the proposed models, we utilize a metric called the Rank of First Positive Prediction (RFPP). RFPP is a biologist-centric performance metric which is based on the fact that a perfect machine learning model should give the highest score to the known depolymerases in a phage proteome and therefore rank them lower in a sorted list compared with non-depolymerases [19]. Therefore the perfect machine learning model would have an RFPP of 1 for each proteome. RFPP is computed across different percentiles in the test set to get an idea of the overall performance on the test set.

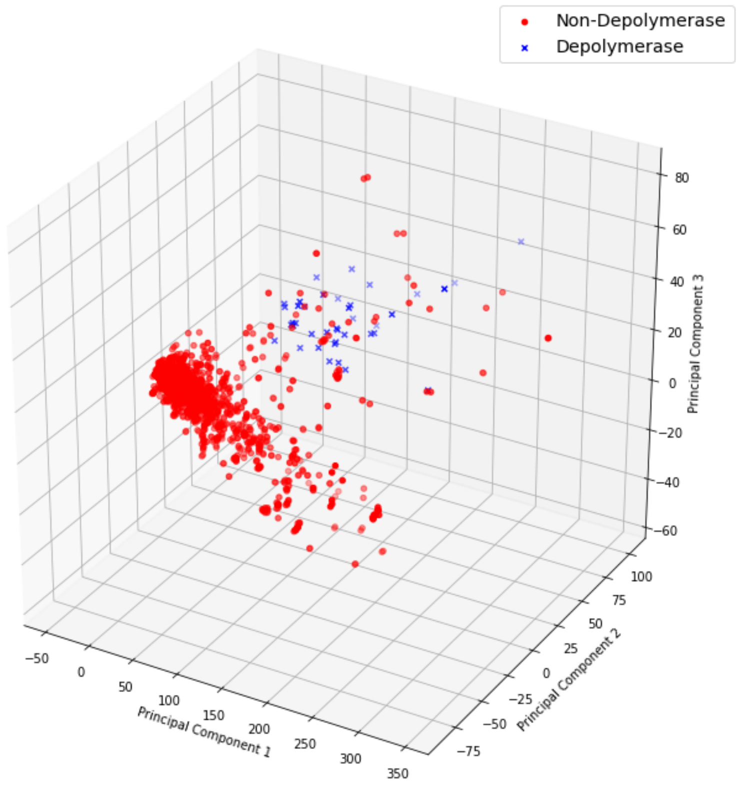

Figure 2: 3D scatter plot of the top 3 principal components of the protein features for known depolymerases (blue) and other proteins (red).

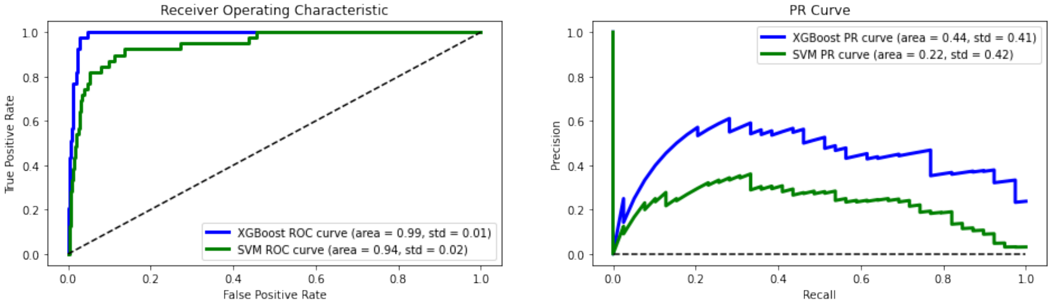

Figure 3: ROC-curve plotted for 10% threshold (left). PR-curve plotted for 10% threshold (right) for both SVM and proposed XGBoost based ranking model (Deporanker).

3 Results

3.1 Principal Component Analysis

To explore the relationship between the two classes in our data, we performed Principal Component Analysis (PCA) on our features [15]. We took the top 3 principal components to compare visually, (fig. 2) shows these in a 3D scatter plot. There is a clear separation between the two classes in our data showing that machine learning will be able to accurately classify these in this feature space.

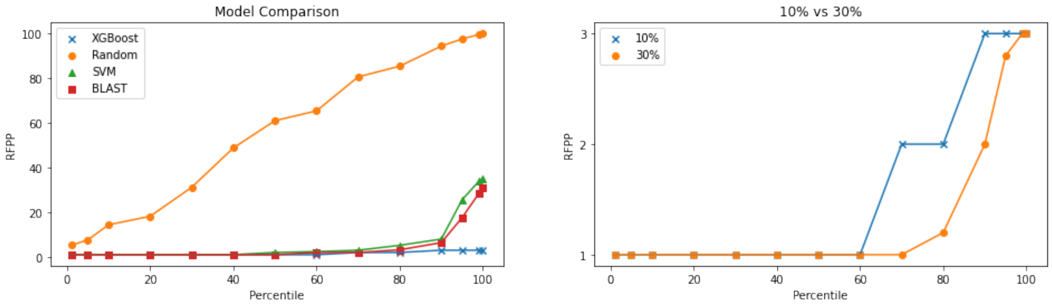

Figure 4: Comparison between different models (left) using percentiles of Rank of First Positive Prediction (RFPP). An ideal model should have an RFPP of 1.0 for all test phages. The curves show RFPP percentiles for the proposed model, a random baseline, an SVM model, and BLAST. The proposed model is able to rank known depolymerases within the top 3 predictions for all test phages. Comparison between 10% similarity threshold and 30% similarity threshold (right) for non-redundancy analysis.

3.2 Non-redundant Cross-Validation Performance

To evaluate our machine learning model, we computed the receiver operating characteristic (ROC) curve which is a common analysis method in machine learning that demonstrates the discriminative ability of a prediction model [27, 8]. The area under the curve is the empirical metric which can be observed from this analysis method. An AUC-ROC score of 1 would indicate that a classifier can perfectly separate the depolymerase and non-depolymerase proteins. Fig. 3 shows the ROC curve plotted for our ranking model. The AUC-ROC score for the proposed model is 0.99. In comparison, a kernelized support vector machine (SVM) with the polynomial kernel and BLAST both give an AUROC of 0.94.

The precision recall curves plotted in fig. 3 show a similar trend with the proposed model giving an improved performance in comparison to BLAST (with AUC-PR of 0.37) and SVM (AUC-PR of 0.42).

3.3 Ranking Depolymerases in Phage Proteomes

Fig. 4 shows the RFPP for all proteomes for different redundancy thresholds, and different models with a random predictor for an experimental control. For a random ranking model which produces a random score for any given example, the median RFPP is 52. So for each proteome tested during cross validation 50% would have a known depolymerase in the top 52 predictions by the model. This is greatly improved upon by the proposed model and a kernelised SVM which both have median RFPP of 1. However, for the 95th percentile, our model shows an RFPP of 3 whereas for SVM this is 25.6. Out of the 24 phage proteomes tested, the ranking model had an RFPP of 3 or lower for all proteomes, which is in the top 5% in their respective proteome. Full results are shown in table 1. This clearly shows the effectiveness of our ranking model for identifying depolymerases in phage proteomes.

To see the effect of increasing the similarity thresholds of training sets, we compare our results with our ranking model trained on a non-redundant dataset created from a 30% similarity threshold for depolymerase sequences. As shown in (fig. 4), a more redundant set of data creates better results. However, this is to be expected as the model will be trained on more similar depolymerases, we have opted to use the 10% threshold for the final model as it better represents the challenge of identifying depolymerases in proteomes independent of our dataset. These ranking results also show that the proposed model gives better performance in comparison to BLAST and SVM baselines.

We have developed a machine learning model which was able to identify the characterised depolymerases from the test dataset to a good accuracy. The principal component analysis shows that a number of genes that were not characterised depolymerases (non-depolymerase proteins) clustered with the depolymerases (fig. 2). There are two possibilities for these non-depolymerase proteins; 1. they are false positives, 2. they are true depolymerases that were not characterised, present in the proteomes of phages that also contained a characterised depolymerase. These will be further investigated in the future.

4 Conclusions and Future Work

Any novel depolymerase groups identified from the machine learning approach will need to be structurally and biochemically validated to confirm their function. Once a robust machine learning approach has been validated it will enable phage researchers to determine the usefulness of phages for therapeutics. Phage depolymerases also have the added value of being applied as therapeutic enzymes in their own right which circumvents some of the obstacles to using phages as antimicrobials. Enabling accurate identification of phage proteins will be critical for the future optimisation of phage therapeutics.

Phage Accession Number	Cluster	Proteome Size	BLAST	SVM	DepoRanker
NC027399	7	540	1	35	1
AB797215	7	203	20	8	1
MW655991	5	76	3	1	1
MW672037	5	74	3	3	1
MF663761	6	71	1	8	1
OU509534	7	56	4	1	1
OU509533	3	54	2	1	1
AB716666	7	53	4	1	1
KX712070	7	53	1	2	1
KU666550	7	52	2	3	1
KU183006	7	51	1	5	1
GQ413937	7	49	1	3	1
MK903728	7	49	1	30	1
MT966872	7	48	1	2	1
KY389315	7	48	1	1	1
LC413194	7	47	1	1	1
JF501022	2	77	31	6	2
KY385423	3	54	2	1	2
KT964103	4	54	2	1	2
KY389316	7	47	1	1	2
OU509535	1	528	8	2	3
LC413195	7	53	1	1	3
MH587638	7	50	1	4	3
LC413193	7	48	1	1	3

Table 1: Results for non-redundant cross-validation in terms of ranks of first positive prediction (RFPP). Each phage is referred by its accession number and is a part of a cluster. Depolymerases of phages in different clusters have less than 10% sequence similarity (computed using CD-HIT). The ranks of the top scoring depolymerase protein from different methods for each test phage is shown together with the size of the proteome for that phage after training the method on phages from all other clusters. These results clearly show that the proposed model (DepoRanker) gives superior performance in comparison to BLAST and SVM baselines.

References

[1] Pratiti Bhadra et al. “AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest”. In: *Scientific Reports* 8.1 (Jan. 2018), p. 1697. ISSN: 2045-2322. DOI: 10.1038/s41598-018-19752-w. URL: https://doi.org/10.1038/s41598-018-19752-w.

[2] Junjie Chen et al. “A comprehensive review and comparison of different computational methods for protein remote homology detection”. In: *Briefings in Bioinformatics* 19.2 (Nov. 2016), pp. 231–244. ISSN: 1477-4054. DOI: 10.1093/bib/bbw108. eprint: https://academic.oup.com/bib/article-pdf/19/2/231/25524173/bbw108.pdf. URL: https://doi.org/10.1093/bib/bbw108.

[3] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *CoRR* abs/1603.02754 (2016). arXiv: 1603.02754. URL: http://arxiv.org/abs/1603.02754.

[4] R. M. Dedrick et al. “Prophage-mediated defence against viral attack and viral counter-defence”. In: *Nat Microbiol* 2.3 (2017), p. 16251.

[5] Lin DM, Koskella B, and Lin HC. “Phage therapy: An alternative to antibiotics in the age of multi-drug resistance”. In: *World J Gastrointest Pharmacol Ther 2017; 8(3): 162-173* ().

[6] European Centre for Disease Prevention {and} Control ECDC. *Surveillance Atlas of Infectious Diseases*. 2019. URL: https://atlas.ecdc.europa.eu/public/index.aspx.

[7] Simon Eitzinger et al. “Machine learning predicts new anti-CRISPR proteins”. In: *Nucleic Acids Research* 48.9 (Apr. 2020), pp. 4698–4708. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa219. eprint: https://academic.oup.com/nar/article-pdf/48/9/4698/33221311/gkaa219.pdf. URL: https://doi.org/10.1093/nar/gkaa219.

[8] Ali Ghulam et al. “Accurate prediction of immunoglobulin proteins using machine learning model”. In: *Informatics in Medicine Unlocked* 29 (2022), p. 100885. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2022.100885. URL: https://www.sciencedirect.com/science/article/pii/S2352914822000387.

[9] Chun-Ru Hsu et al. “Isolation of a bacteriophage specific for a new capsular type of Klebsiella pneumoniae and characterization of its polysaccharide depolymerase”. In: *PloS one* 8.8 (2013).

[10] Ying Huang et al. “CD-HIT Suite: a web server for clustering and comparing biological sequences”. en. In: *Bioinformatics* 26.5 (Mar. 2010), pp. 680–682.

[11] K. A. Hughes, I. W. Sutherland, and M. V. Jones. “Biofilm susceptibility to bacteriophage attack: the role of phage-borne polysaccharide depolymerase”. In: *Microbiology* 144 (Pt 11).11 (1998), pp. 3039–47.

[12] Paul Hyman and Stephen T. Abedon. “Bacteriophage Host Range and Bacterial Resistance”. In: *Advances in Applied Microbiology*. Vol. 70. Elsevier, 2010, pp. 217–248. (Visited on 03/11/2022).

[13] Leandra E. Knecht, Marjan Veljkovic, and Lars Fieseler. “Diversity and Function of Phage Encoded Depolymerases”. In: *Frontiers in Microbiology* 10 (2020). ISSN: 1664-302X. DOI: 10.3389/fmicb.2019.02949. URL: https://www.frontiersin.org/article/10.3389/fmicb.2019.02949.

[14] Agnieszka Latka et al. “Modelling the architecture of depolymerase-containing receptor binding proteins in Klebsiella phages”. In: *Frontiers in microbiology* 10 (2019), p. 2649.

[15] Yi Ma, Rene Vidal, and Shankar Sastry. *Generalized principal component analysis*. 1st ed. Interdisciplinary Applied Mathematics. New York, NY: Springer, Apr. 2016.

[16] Grażyna Majkowska-Skrobek et al. “Capsule-targeting depolymerase, derived from Klebsiella KP36 phage, as a tool for the development of anti-virulent strategy”. In: *Viruses* 8.12 (2016), p. 324.

[17] Katelyn McNair et al. “Phage genome annotation using the RAST pipeline”. In: *Bacteriophages*. Springer, 2018, pp. 231–238.

[18] Katelyn McNair et al. “PHANOTATE: a novel approach to gene identification in phage genomes”. In: *Bioinformatics* 35.22 (2019), pp. 4537–4542.

[19] Fayyaz ul Amir Afsar Minhas, Brian J Geiss, and Asa Ben-Hur. “PAIRpred: partner-specific prediction of interacting residues from sequence and structure”. en. In: *Proteins* 82.7 (July 2014), pp. 1142–1155.

[20] S. Navon-Venezia, K. Kondratyeva, and A. Carattoli. “Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance”. In: *FEMS Microbiol Rev* 41.3 (2017), pp. 252–275.

[21] Y. J. Pan et al. “Klebsiella Phage PhiK64-1 Encodes Multiple Depolymerases for Multiple Host Capsular Types”. In: *J Virol* 91.6 (2017), e02457–16.

[22] Mayank Sharma. “A CNN-Based K-Mer Classification of Anti-Microbial Peptide Sequences”. In: *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2020, pp. 819–823. DOI: 10.1109/ICRITO48877.2020.9198006.

[23] E. V. Solovieva et al. “Comparative genome analysis of novel Podoviruses lytic for hypermucoviscous Klebsiella pneumoniae of K1, K2, and K57 capsular types”. In: *Virus Res* 243 (2018), pp. 10–18.

[24] Flavia Squeglia et al. “Structural and Functional Studies of a Klebsiella Phage Capsule Depolymerase Tailspike: Mechanistic Insights into Capsular Degradation”. In: *Structure* (2020).

[25] E Tacconelli et al. “Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics”. en. In: *World Health Organ.* 27 (2017).

[26] Taha ValizadehAslani et al. “Amino Acid k-mer Feature Extraction for Quantitative Antimicrobial Resistance (AMR) Prediction by Machine Learning and Model Interpretation for Biological Insights”. In: *Biology* 9.11 (2020). ISSN: 2079-7737. DOI: 10.3390/biology9110365. URL: https://www.mdpi.com/2079-7737/9/11/365.

[27] Jiancheng Zhong et al. “XGBFEMF: An XGBoost-Based Framework for Essential Protein Prediction”. In: *IEEE Transactions on NanoBioscience* 17.3 (2018), pp. 243–250. DOI: 10.1109/TNB.2018.2842219.