PhageScan - Automated viral particle detection

Sousa, R.[1][0009−0000−2558−7176], Dias, O.[1][0000−0002−1765−7178], and Oliveira, H.[1][0000−0001−9564−5714]

Department of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal sac@ceb.uminho.pt
https://www.deb.uminho.pt/

Abstract. Bacteriophages are viewed as a promising solution to fight antibiotic-resistant bacteria. Being the most abundant organism in the biosphere, classifying these viruses is a complex and morose task that can be obviated through application of computational tools. This work proposes PhageScan, a tool to identify phages from Transmission Electron Microscopy images using YOLO algorithms. In this first iteration we have achieved a model capable of identifying siphoviruses with relative precision for a confidence threshold of around 0.4, with a large margin for improvement.

Keywords: Bacteriophage classification · deep learning · convolutional neural networks

1 Background

1.1 What are bacteriophages?

Bacteriophages (phages), regarded the most abundant organisms in the biosphere, are a type of virus that thrives by infecting and replicating within bacterial cells. [1] These organisms are generally species-specific, typically targeting a single bacterial species to maintain their lifecycle and co-evolving with it. [2,3,4,5]

Phages adopt two distinct life styles: lytic or lysogenic (Fig 1 [6]). Lytic phages infect the host cell and hijack its biosynthetic machinery to rapidly generate offspring that exit the cell, bursting and killing it in the process. Phages with a lysogenic life style, by their turn, can be stable for generations, integrating their genome in the host's, sometimes even shaping the host phenotype. In this state, they do not necessarily kill the host, adopting an opportunistic approach, transitioning towards a lytic approach when the host reaches its end of life and / or if a specific trigger occurs. There are also phages that exhibit a a chronic life style, which is generally non-bactericidal. [2,6,7]

Given their natural predating relationship with bacteria, phages are increasingly gaining traction as potential therapeutic agents in the combat of antibiotic-resistant bacteria, a pressing global health challenge.[7,8,9,10] A result of this growing interest in phages creates a pressing need to develop methods to study phages to facilitate identification and classification.

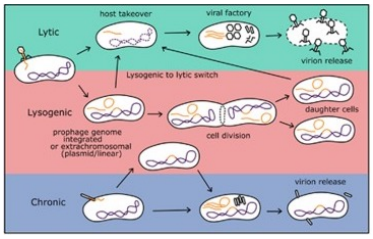


Fig. 1: Simple schematic representation of the different phage life cycles. [6]

1.2 Phage classification

With the advancements in genomics and metavirunomics, phage classification underwent significant changes in recent years, transitioning from a morphological to a genomically coherent taxonomical method. The biggest change in the system was abolishment of the most abundant family of phages, caudovirales (Fig 2 [11]), which was divided into three subfamilies: (i) Podoviridae; (ii) Myoviridae; and (iii) Siphoviridae. In its stead emerged the new class of caudiviricetes, comprised of 4 new orders and 22 new families. [11,12,13,14]

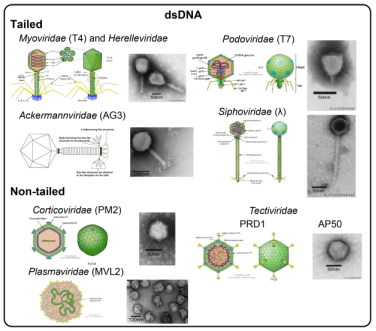
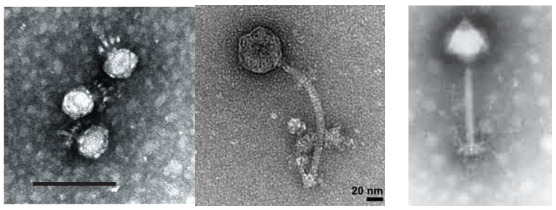


Fig. 2: Schematic representation of tailed phages morphology. [11]

Despite morphology no longer being used, it remains relevant towards identification and classification, standing out as a potential approach for automating tailed phage classification from Transmission Electron Microscopy (TEM) images. There are three distinct tailed phage morphotypes (Fig 3), based on their respective tailed structures: podoviruses, siphoviruses and myoviruses. [15] **Podoviruses** typically display a short non-contractile tail and an icosahedral capsid of about 60 nm in diameter. An example of this virus is *Brucealvirus* CP7R as shown below. [16] **Siphoviruses** typically display an icosahedral capsid and a long, flexible but non-contractile tail, spiked at the tip. An example

of this morphotype, BlueFeather, can be seen below. [17] **Myoviruses** display a contractile tail with a base plate and spikes at the tip and an icosahedral capsid. One example of this viruses is *Synechococcus* Phage S-PM2. [18]



(a) Podovirus (b) Siphovirus (c) Myovirus

Fig. 3: Morphotypes of phages present in the caudoviricetes class, the most abundant class of phages in the currently existing knowledge. From left to right are shown specimens of podovirus CP7R, a *Chlostridium perfringens* phage, siphovirus BlueFeather, a *Arthrobacter globiformis* phage, and myovirus S-PM2, a *Synecochus sp.* phage. [16,17,18]

1.3 Automated viral particle detection and classification

One of the main advantages of developing accurate and efficient methods for automatic detection of viral particles is the automation of the fastidious and labour-intensive work associated with microscope analysis - safeguarding the impact that human factors such as fatigue can have on the analysis. [19] This is also the case for automatic recognition of viral particles, where qualitative and quantitative measurements involve processing of thousands of microscopy images to count particles and identify viral morphotypes. [19,20]

It is accepted that in order to be regarded as valuable, automatic recognition systems must attain an accuracy greater than 75% when compared to a human expert. [19] There are several methods in the literature for automatic particle identification using machine learning techniques such as neural networks, Gaussian classifiers, Naive Bayes, and random forests that have achieved this requisite to different degrees. [20,21,22,23] From these methods, we highlight neural networks, specificall convolutional neural networks (CNNs) as particularly adequate for the purpose of automated phage classification given their specialization in processing grid-like data.

1.4 An overview of CNNs

CNNs are a variation of Neural Networks, consisting of a feed forward mathematical model that contain: i) an input layer; ii) N hidden layers each containing

M number of neurons that perform mathematical operations for classification; iii) and output layer representing the outcome of said calculations, usually the classification of the input under the form of a probability. This is achieved due to neurons having distinct weights (W) and biases (b), thus influencing the classification process. [24] A neuron can be represented mathematically as an affine function followed by a non-linearity:

$$M(x) = \phi \cdot (W \cdot x + b) \tag{1}$$

CNNs have emerged as an approach to process grid-like data, such as images, adopting a sweeping approach, enabling weight sharing and local neuron responses instead of activating all the neurons of the network thus enhancing its efficiency. Conversely, CNNs are also constituted by convolutional layers and pooling layers.

Convolutional layers operate on the input data (*i.e.* image) using vector and matrix operations in order to extract features from input data, generating a feature map. The mathematical operation performed by the convolutional layers is called a convolution and can be represented mathematically by

$$(x \cdot w)(a) = \int x(t) \cdot w(a–t)d \cdot a \tag{2}$$

Where a is an n-dimensional vector that iterates over all values in the input, w represents the convolutional kernel, a matricial filter or mask that operates and transforms the data. The integral is replaced by the higher dimensional variant of a. In the context of this work we will only use 2D microscopy images and as such will be working with 2-dimensional convolutions, which can be represented mathematically by

$$(I \cdot K)(i, j) = \sum_m \sum_n (I(m, n)K(i - m, j - n) \tag{3}$$

Pooling layers act as intermediate nodes, performing dimensionality reduction and downsampling operations to the feature map while retaining essential features.

Each operation is followed by the application of an activation function which adds non-linearity to the system, aiding in the classification of the input.There are several activation functions, but for the sake of this work, we'll detail SiLU (Sigmoid-Weighted Linear Units), which is the activation function we will be using during the training of YOLOv8 [24].

SiLU delivers smooth profiles by combining linear and non-linear functions, thus increasing feature detection sensitivity. SiLU can be represented mathematically as:

$$SiLU(x) = x \cdot \sigma(x) \tag{4}$$

where $\sigma(x)$ is the sigmoid function given by:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

and x is the input [25].

Finally, fully connected layers or dense layers take in the output of the previous layers and flattens it, transforming the representation of all feature maps into a single vector. The final fully connected layer is the output layer that gives the final probabilities for each considered class. Figure 4 depicts the architecture of a CNN. [26]

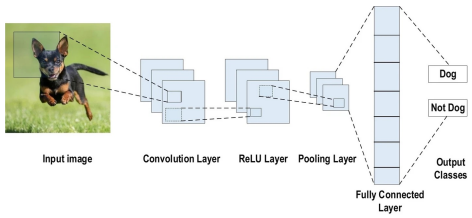


Fig. 4: Schematic representation of a CNN. [26]

2 Methods/approach

This work aimed to explore and evaluate the employment of CNNs in automated phage particle recognition given their versatility and high level of discernment for image classification with high accuracy. [21] Moreover, we explored and evaluated the application of You Only Look Once (YOLO) algorithms, specifically YOLOv8. YOLO was created as an approach for object detection and has been demonstrating growing accuracy and real-time performance and have already been employed in the field of biological images processing. [27,28,29]

It is important to consider several challenges that this project entails, including: (i) the lack of publicly available data sets, (ii) high noise levels in available TEM micrographs, (iii) a lack of variety of phages and hosts on a sample, (iv) podoviruses have small tails that can be difficult to detect, possibly being confounded with vesicles, and myoviruses have contracted/relaxated tails, adding complexity to labeling and classification as they can be confounded with siphoviruses, (v) recently the first-ever observation of phages attaching to another viruses was made, further adding complexity to this task [30]. Considering these challenges, siphoviruses, with their consistent capsid shape and long non-contractile tails, are a good starting point and, in future steps, the application of more complex image processing techniques, such as edge detection/filtering and pixel intensity interpolation, are needed.

2.1 Data set creation

Data gathering and pre-processing The data set was created through sourcing 119 siphoviruses TEM micrographs from and PhagesDB together with TEM images obtained from the CEB-UMINHO phage research lab.

The data was pre-processed by: i) resizing images to a fixed size (640 x 640 pixels), ensuring all scales are correctly represented; ii) adjusting image quality (brightness, contrast, noise and artifact removal).

Annotation Finally, to make the data set processable by the model, it's fundamental to annotate the images in order to specify what is what. This process is achieved through specifying object classes and location within the images with bounding boxes. The data set images were anottated using Roboflow. The annotation process for phages followed the following approach: (i) draw a bounding box directly on top of the phage, classifying phages in micrographs where there are other objects present, increasing context in micrographs where could exist other objects with strong contrast resembling the head such as bacteria.

Augmentation The data set will be augmented through random rotations and flips in order to create scenarios where the target objects of classification and identification can be identified independently from image rotation angle, being this a common data augmentation approach for object detection algorithms. [29]

2.2 Training parameters

The training was performed using YOLOv8, on each pass through the data set (epoch) the data set was subdivided in batches of 16 images. The data set contained 271 images, corresponding to approximately 17 batches. The model was trained for 10/25/50 epochs corresponding to 170/425/850 optimization steps respectively, using SiLU activation function and a training rate of 10e-3 (Adam).

3 Results

In this section, we analyze and compare the performance metrics of various trained models, focusing on box, class, and distribution focal losses, as well as precision, recall, and F1 score.

3.1 Training Results - 10 Epochs

Figure 5 illustrates the training outcomes for the model trained over 10 epochs. Analyzing these results reveals a consistent improvement across the epochs, with a notable decline in box, class, and distribution focal losses (DFL) for

both training and validation sets. This trend indicates an enhanced model capability in inferring and identifying correct bounding box coordinates, class attributions, and discerning underrepresented classes, specifically distinguishing between siphoviruses and the background.

Moreover, despite the decreasing trend in these metrics, they have not yet reached a plateau, suggesting that further training could yield additional improvements. This potential for enhancement is further corroborated by a continuous increase in the model's precision, average precision at a confidence threshold of 0.5, average precision across the range of 0.5-0.95, and recall. These improvements indicate that the model is progressively better at correctly identifying true instances of objects and assigning the appropriate class.

The consistent upward trend in these performance metrics, coupled with the absence of plateau behavior, strongly suggests that the model's performance could be further enhanced with additional training epochs, as will be demonstrated in subsequent sections.

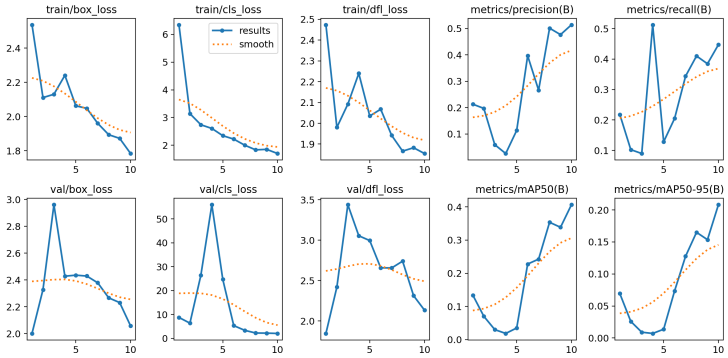
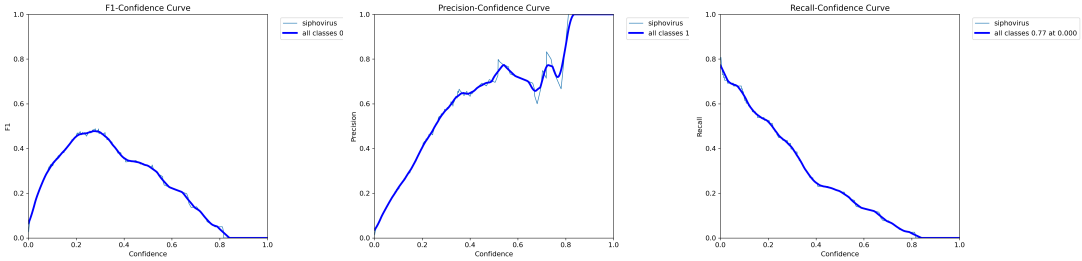


Fig. 5: Model training results - 10 epochs

Figure 6 depicts the distribution of the F1 score, precision, and recall metrics across various confidence thresholds ranging from 0 to 1. Analyzing the F1 score curve reveals that the highest F1 score occurs at a confidence threshold of 0.275, indicating the optimal balance between precision and recall at this threshold. This observation is supported by the precision and recall curves, which show an inverse relationship: precision increases with higher confidence levels, while recall decreases.

Additionally, the data clearly indicate over fitting around a confidence threshold of 0.837, where precision reaches a perfect score of 1. Conversely, recall significantly drops at this threshold. This over fitting is likely due to the model being overly familiar with siphovirus occurrences and the lack of variability in the dataset.



(a) F1 Score (b) Precision (c) Recall

Fig. 6: Performance metrics vs confidence threshold (10 epochs)

Figure 7 demonstrates the model's ability to infer and classify siphoviruses in micrographs. Considering the previously analyzed performance metrics and results, it is evident that the model can accurately identify and classify siphoviruses under various contrast settings (e.g., dark-stained and white-stained capsids). Although the model occasionally makes incorrect detections and partial classifications, these results indicate a solid starting point. As we will demonstrate in the following section, further training with additional epochs yields improved results.

It is noteworthy, however, that the model exhibits signs of overfitting and bias due to the extensive training on siphoviruses. Incorporating training with additional classes will likely produce different results, especially given the similarities between siphoviruses and myoviruses in their relaxed states.

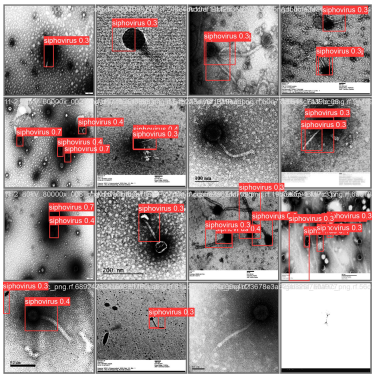


Fig. 7: Performance for validation batch - 10 epochs

3.2 Training results - 25 Epochs

Figure 8 presents the main training metrics for the model trained over 25 epochs. Similar to the previous results, there is a clear improvement in model performance with increased training epochs. Compared to the model trained for 10 epochs, all metrics show enhanced performance, including a slight reduction in box and distribution focal losses (DFL) and a significant decline in class loss. Precision, average precision, and recall have also improved.

The upward trend and the absence of plateau behavior are evident, suggesting that the model could continue to improve with further training beyond 25 epochs.

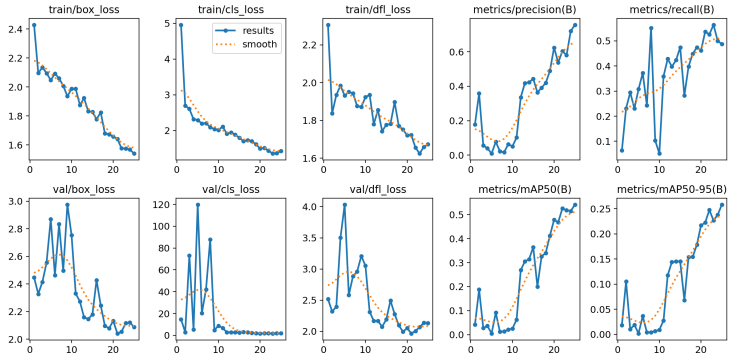
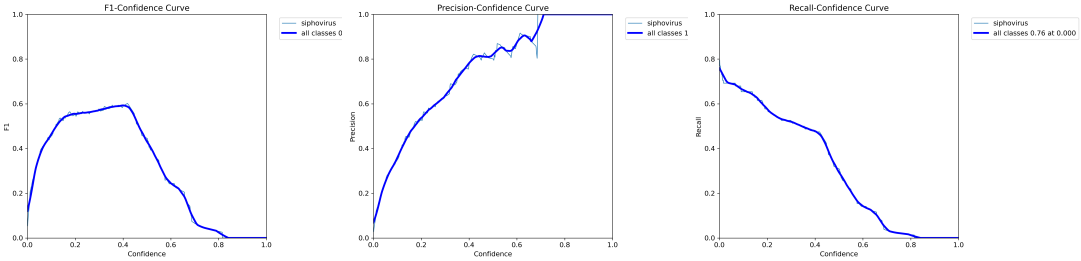


Fig. 8: Model training results - 25 epochs

Figures 9 depict the distribution of the F1 score, precision, and recall metrics across various confidence thresholds ranging from 0 to 1. There is a noticeable improvement in the F1 score, which peaks at 0.6 around a confidence threshold of 0.4, compared to the previous iteration's peak of 0.5 around a threshold of 0.3. This improvement is supported by the corresponding distributions of precision and recall. However, the overfitting behavior is still present and even more pronounced, further confirming our concerns.



(a) F1 Score (b) Precision (c) Recall

Fig. 9: Performance metrics vs confidence threshold (25 epochs)

Figure 10 demonstrates these improvements, showing more accurate classifications and higher confidence levels for correct identifications. Despite these advancements, it is crucial to remain aware of the persistent overfitting and biases, which are likely due to the extensive number of siphovirus micrographs used during training.

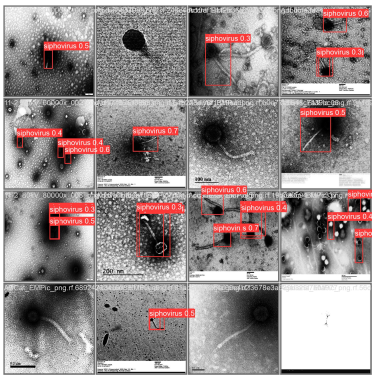


Fig. 10: Performance for validation batch - 25 epochs

3.3 Training results - 50 Epochs

When the model is trained for 50 epochs, Figure 11 shows a further improvement in box, class, and distribution focal losses (DFL). Precision and recall also improve, reaching 0.8 and 0.6, respectively. However, signs of overtraining become apparent after around 40 epochs, with precision beginning to decline. This indicates overfitting, where continued training beyond 40 epochs deteriorates

performance. Implementing early stopping strategies could have addressed this issue by halting the training process when no further improvements are detected.

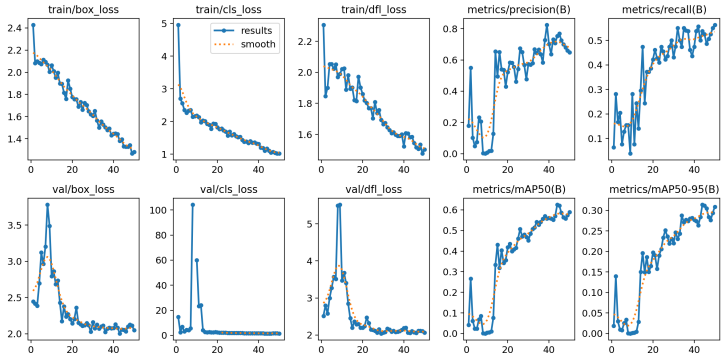
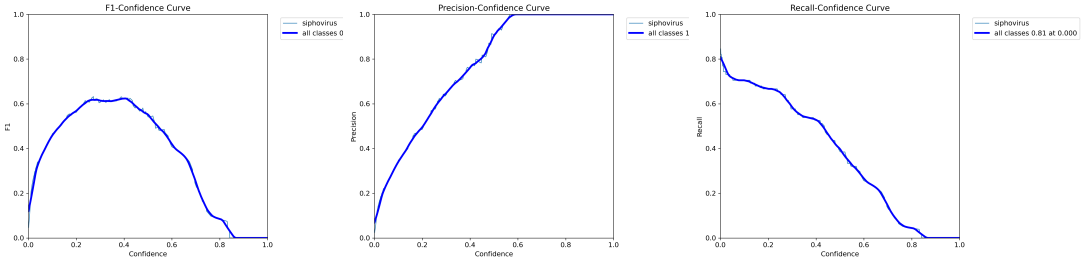


Fig. 11: Model training results - 50 epochs

Regarding the F1 score, precision, and recall (Figure 12), the highest F1 score remains close to that achieved with 25 epochs, but its decline is less steep and it remains at peak level across a broader confidence threshold range. However, it is important to note that overfitting manifests much sooner, as evidenced by precision reaching a value of 1 around a 0.5 confidence threshold. Nonetheless, recall is more consistent and hits a value of 0 later, indicating improvements in this metric.



(a) F1 Score (b) Precision (c) Recall

Fig. 12: Performance metrics vs confidence threshold (50 epochs)

By analyzing Figure 13, we observe these improvements in the form of an increased number of predictions across different scenarios and higher confidence levels for true positives. However, the model still misidentifies some objects as

siphoviruses (as seen in the second image), underscoring the need to fine-tune hyperparameters and use a more diverse set of training images.

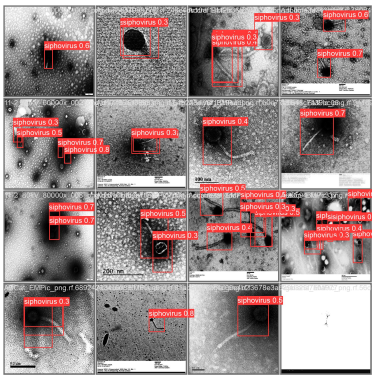


Fig. 13: Performance for validation batch - 50 epochs

4 Conclusions and Future Work

In this work we gave the first steps towards the development of a phage viral particle classifier using YOLO algorithms, having attained in this first iteration an encouraging F1 score of *circa* 0.6 for a confidence threshold around 0.4 despite the occurrence of over-fitting and model conservativeness.

The next steps in the PhageScan roadmap involve enhancing our dataset with greater diversity and incorporating additional classes for training. This initiative aims to fine-tune hyperparameters to achieve a minimum accuracy of 75%, ensuring the model can accurately identify and differentiate between various virus types.

Following this milestone, we plan to explore the implementation of automated measurements and feature identification techniques. This includes experimenting with methods such as OCR for scale identification and automated measurement, as well as edge contour detection for isolating features like tails and capsids. Ultimately, these efforts will enable us to estimate genome sizes based on these parameters, advancing our understanding and capabilities in virus characterization.

References

1. S. Batinovic, F. Wassef, S.A. Knowler, D.T.F. Rice, C.R. Stanton, J. Rose, J. Tucci, T. Nittami, A. Vinh, G.R. Drummond, C.G. Sobey, H.T. Chan, R.J. Seviour, S. Petrovski, and A.E. Franks. Bacteriophages in natural and artificial environments. *Pathogens*, 8(3):100, 2019.
2. M.R. Clokie, A.D. Millard, A.V. Letarov, and S. Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, 2011.
3. L.M. Kasman and L.D. Porter. Bacteriophages. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-., 2022. Available from: url-https://www.ncbi.nlm.nih.gov/books/NBK493185/.
4. B. Koskella and M.A. Brockhurst. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev*, 38(5):916–931, 2014.
5. C. Suttle. Viruses in the sea. *Nature*, 437:356–361, 2005.
6. C. Howard-Varona, K.R. Hargreaves, S.T. Abedon, and M.B. Sullivan. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J*, 11(7):1511–1520, 2017.
7. P. Ioannou, S. Baliou, and G. Samonis. Bacteriophages in infectious diseases and beyond—a narrative review. *Antibiotics*, 12:1012, 2023.
8. I.U. Haq, W.N. Chaudhry, and M.N. et al. Akhtar. Bacteriophages and their implications on future biotechnology: a review. *Virol J*, 9:9, 2012.
9. S.B. Gamachu and M. Debalo. Review of bacteriophage and its applications. *Int J Vet Sci Res*, 8(3):133–147, 2022.
10. A. Sulakvelidze, Z. Alavidze, and J.G. Jr. Morris. Bacteriophage therapy. *Antimicrob Agents Chemother*, 45(3):649–659, 2001.
11. M.B. Dion, F. Oechslin, and S. Moineau. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol*, 18(3):125–138, 2020. Epub 2020 Feb 3.
12. D. Turner, A.M. Kropinski, and E.M. Adriaenssens. A roadmap for genome-based phage taxonomy. *Viruses*, 13(3):506, 2021.
13. D. Turner, A.N. Shkoporov, and C. et al. Lood. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ictv bacterial viruses subcommittee. *Arch Virol*, 168:74, 2023.
14. H.W. Ackermann. 5500 phages examined in the electron microscope. *Arch Virol*, 152(2):227–243, 2007.
15. E. Maffei, A. Shaidullina, M. Burkolter, V. Druelle, L. Willi, F. Estermann, S. Michaelis, H. Hilbi, D. Thaler, and A. Harms. Systematic exploration of escherichia coli phage-host interactions with the basel phage collection. 2021. doi: 10.1101/2021.03.08.434280.
16. N.V. Volozhantsev, B.B. Oakley, C.A. Morales, V.V. Verevkin, V.A. Bannov, V.M. Krasilnikova, A.V. Popova, E.L. Zhilenkov, J.K. Garrish, K.M. Schegg, R. Woolsey, D.R. Quilici, J.E. Line, K.L. Hiett, G.R. Siragusa, E.A. Svetoch, and B.S. Seal. Molecular characterization of podoviral bacteriophages virulent for clostridium perfringens and their comparison with members of the picovirinae. *PLoS One*, 7(5):e38283, 2012.
17. S. Demo, A. Kapinos, A. Bernardino, K. Guardino, B. Hobbs, K. Hoh, E. Lee, I. Vuong, K. Reddi, A. Freise, and J. Parker. Bluefeather, the singleton that wasn't: Shared gene content analysis supports expansion of arthrobacter phage cluster fe. *PLOS ONE*, 16:e0248418, 2021.
18. N. Mann. The third age of phage. *PLoS biology*, 3:e182, 2005.

19. R.M. Glaeser. Historical background: why is it important to improve automated particle selection methods? *Journal of Structural Biology*, 145(1–2):15–18, 2004.
20. A. Gelzinis, A. Verikas, E. Vaiciukynas, M. Bacauskiene, S. Sulcius, E. Simoliunas, J. Staniulis, and R. Paskauskas. Automatic detection and morphological delineation of bacteriophages in electron microscopy images. *Computers in Biology and Medicine*, 64:101–116, 2015.
21. Toshihiko Ogura and Chikara Sato. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis. *Journal of Structural Biology*, 145(1–2):63–75, 2004.
22. C.O.S. Sorzano, E. Recarte, M. Alcorlo, J.R. Bilbao-Castro, C. San-Martín, R. Marabini, and J.M. Carazo. Automatic particle selection from electron micrographs using machine learning techniques. *Journal of Structural Biology*, 167(3):252–260, 2009.
23. César A.B. Castañón, Jane S. Fraga, Sandra Fernandez, Arthur Gruber, and Luciano da F. Costa. Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus eimeria. *Pattern Recognition*, 40(7):1899–1910, 2007.
24. Jonas Teuwen and Nikita Moriakov. Chapter 20 - convolutional neural networks. In S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger, editors, *Handbook of Medical Image Computing and Computer Assisted Intervention*, The Elsevier and MICCAI Society Book Series, pages 481–501. Academic Press, 2020.
25. Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.
26. Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-dujaili, Ye Duan, Omran Al-Shamma, José I. Santamaría, Mohammed Abdulraheem Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 2021.
27. Xiyao Li, Jingwen Chen, Yong He, Guofeng Yang, Zhongren Li, Yimin Tao, Yanda Li, Yu Li, Li Huang, and Xuping Feng. High-through counting of chinese cabbage trichomes based on deep learning and trinocular stereo microscope. *Computers and Electronics in Agriculture*, 212:108134, 2023.
28. R. Zhu, Y. Cui, J. Huang, E. Hou, J. Zhao, Z. Zhou, and H. Li. Yolov5s-sa: Light-weighted and improved yolov5s for sperm detection. *Diagnostics (Basel)*, 13(6):1100, 2023.
29. D.G. Gonzalez et al. Evaluating rotation invariant strategies for mitosis detection through yolo algorithms. In A. Cunha, M. Garcia, N. Marx Gómez, and S. Pereira, editors, *Wireless Mobile Communication and Healthcare*, volume 484 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Cham, 2023. Springer.
30. T. de Carvalho, E. Mascolo, S. M. Caruso, and et al. Simultaneous entry as an adaptation to virulence in a novel satellite-helper system infecting *Streptomyces* species. *ISME J*, 17:2381–2388, 2023.