Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ProteoKnight: Image-Based Encoding for Phage Virion Protein Classification and Uncertainty Estimation

Md. Ishrak Khan

md.ishrak.khan@g.bracu.ac.bd

BRAC University

Abir Ahammed Bhuiyan

BRAC University

Samiha Afaf Neha

BRAC University

Md. Khalilur Rahman

BRAC University

Jannatun Noor

United International University

Research Article

Keywords: Phage Virion Proteins, Classification, Deep Learning, Uncertainty, Monte Carlo Dropout, DNA-Walk

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11334404/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ProteoKnight: Image-Based Encoding for Phage Virion Protein Classification and Uncertainty Estimation

Md. Ishrak Khan[1*], Abir Ahammed Bhuiyan[1], Samiha Afaf Neha[1], Md. Khalilur Rahman[1], Jannatun Noor[2]

[1*]Department of Computer Science and Engineering, BRAC University, Kha 224 Pragati Sarani, Merul Badda, Dhaka, 1212, Bangladesh.
[2]C2SG Research Group, Department of Computer Science and Engineering, United International University, United City, Madani Ave, Dhaka, 1212, Bangladesh.

*Corresponding author(s). E-mail(s): md.ishrak.khan@g.bracu.ac.bd;
Contributing authors: abir.ahammed.bhuiyan@g.bracu.ac.bd; samiha.afaf.neha@g.bracu.ac.bd; khalilur@bracu.ac.bd; jannatun@cse.uiu.ac.bd;

Abstract

Accurate prediction of Phage Virion Proteins (PVPs) is essential for genomic studies due to their crucial role as structural elements in bacteriophages. Machine learning has emerged as a promising approach for annotating phage protein sequences obtained via high-throughput sequencing. However, effective annotation requires informative sequence encoding. This paper introduces ProteoKnight, an image-based encoding method that adapts the classical DNA-Walk algorithm using amino-acid-specific pixel colors and adjusted walk distances. It addresses the spatial information loss inherent in the previously reported image-based representation for PVP classification, frequency chaos game representation (FCGR). Moreover, this study bridges a gap in uncertainty analysis for protein sequence classification by evaluating prediction uncertainty in binary PVP classification through the Monte Carlo Dropout technique. In this study, variance and entropy measures were used to assess prediction uncertainty across proteins of various classes and lengths, forming the foundation of this investigation into protein classification and uncertainty quantification. The experiments highlight the efficacy of our approach in binary classification, achieving approximately 89% accuracy, demonstrating competitive performance with previously reported

methods. Nevertheless, multi-class classification accuracy remains suboptimal. Furthermore, the uncertainty analysis unveils variability in prediction confidence influenced by protein class and sequence length, contributing novel insights to protein classification research. Combined with parameter-efficient convolutional neural networks, ProteoKnight enables accurate PVP prediction. Moreover, our uncertainty investigation identifies data points associated with low-confidence predictions, enhancing the comprehensiveness of our analysis.

Keywords: Phage Virion Proteins, Classification, Deep Learning, Uncertainty, Monte Carlo Dropout, DNA-Walk

1 Introduction

Proteins are essential macromolecules responsible for the structural and functional mechanisms of living things. Phage structural proteins are a subclass of the protein family corresponding to the bacteriophages, which are the most prevalent kind of biological creatures (Cobián Güemes et al. 2016). The phage virion proteins (PVPs) are mainly associated with building structural constituents of the bacteriophage (Kabir et al. 2022), such as the baseplate and head as shown in Fig. 1, enabling efficient host-phage binding for genome transfer. The amino acid sequences, particularly those involved in structure synthesis, exhibit significant diversity among phages and their respective groups (Seguritan et al. 2012). Consequently, characterizing these sequences becomes a challenging yet crucial task, as the shortage of annotations for phage proteins has emerged as a hindrance in numerous research studies focused on phage genomics (Henn et al. 2010).

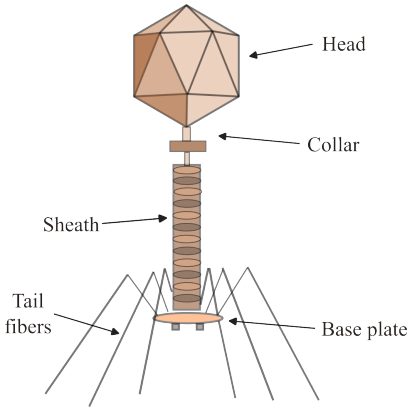


Fig. 1 Structural constituents of Phage

With the advent of high-throughput sequencing technology, rapid additions of sequences are being observed in standard biological databases as shown in Fig. 2,

giving rise to the need for proper annotation of the sequences. Enhancement of annotations for the phage family is vital for exploring effective anti-bacterial drug synthesis (Meng et al. 2020; McCallin et al. 2019), disease diagnosis (Wang and Yu 2004), food production (Brüssow and Desiere 2001), bacterial genome remodeling (Menouni et al. 2015), etc. Traditionally, mass spectrometry and protein array techniques (Lavigne et al. 2009; Yuan and Gao 2016) were utilized for characterizing these proteins. However, these methods are associated with significant time, labor, and computational expenses (Cobián Güemes et al. 2016).

Similarly, alignment-based methods do not always work well with PVP categorization due to the lack of collinearity in viral genomes, resulting from factors such as horizontal gene transfer, high mutation rates, etc., leading to dissimilar sequences having similar structural conformations and vice versa. Thus, alignment-free methods have proved more plausible for approaching the classification task.

In light of these issues, in-silico or computational approaches leveraging machine learning techniques for faster and lower-cost classification are gaining acclaim. To enhance machine readability, amino acid sequences are initially encoded using various schemes. These encodings vary in dimension and feature selection, extracting specific features from the protein sequence necessary for identification. Subsequently, prediction models use these encoded sequences to perform their calculations. Most traditional machine learning methods approach the PVP task as binary classification, categorizing proteins as either PVP or non-PVP (Arif et al. 2020; Zhang and Li 2023), structural or non-structural (Feng et al. 2013), capsid or non-capsid (Galiez et al. 2016), etc. However, there remains room for improvement in classification results, particularly for protein sequences with low sequence conservation. Consequently, deep learning methods have gained popularity, demonstrating superior performance over traditional machine learning approaches across various protein classification tasks. These deep learning models not only showcase better accuracy but also incorporate more PVP classes for the classification task.

1.1 Motivation

Evidently, there is a constant need for developing better and more efficient classification techniques for newly sequenced phage virion proteins. An essential aspect of this task involves refining the sequence representation schemes before training the models to extract valuable local and global information. Previously, feature extraction for phages has primarily focused on one-dimensional data. While these methods yield satisfactory outcomes, the potential for improved prediction through increasing dimensionality remains largely unexplored. Studies analyzing sequence data have demonstrated that when represented as 2-D images, the data exhibits enhanced characteristics due to increased feature space (Akbari Rokn Abadi et al. 2022; Shang et al. 2023). Augmenting dimensionality typically results in improved performance and simplifies the process of feature extraction by applying a standard text-to-image encoding on sequences rather than relying on the collection of several distinct features based on the sequences' physicochemical properties, such as k-mer frequencies. Consequently, numerous studies mention utilizing image encoding techniques for biological sequences (Jung et al. 2023; Zhao et al. 2023).

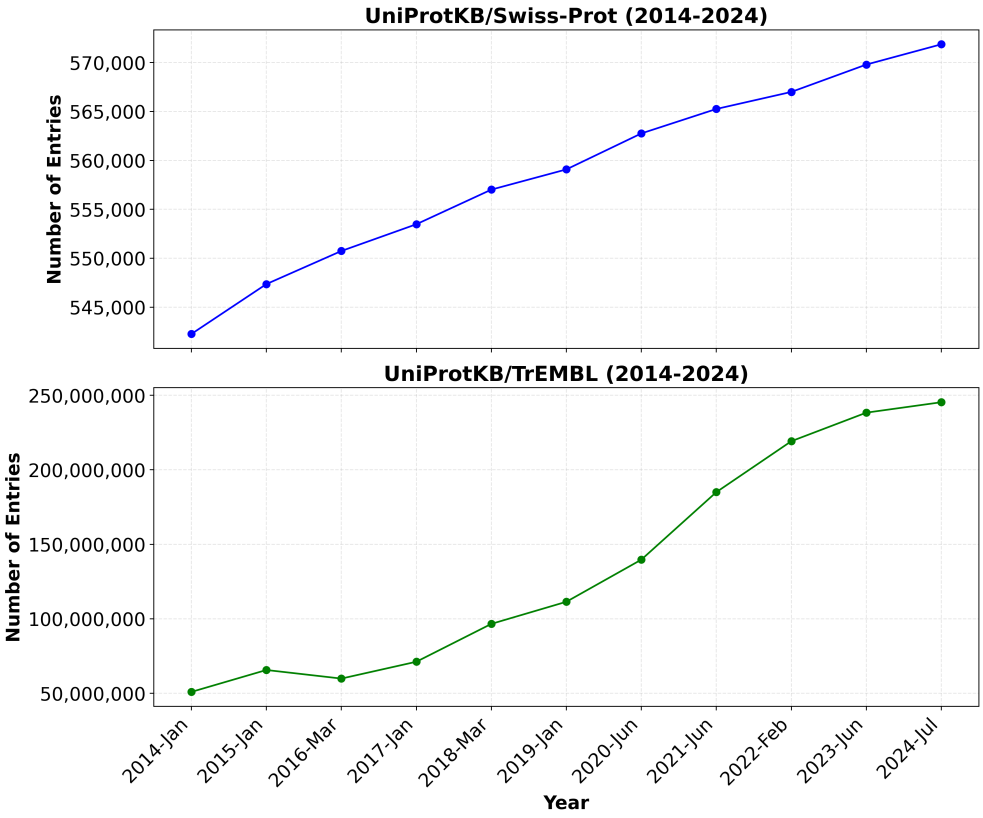


Fig. 2 Growth of protein databases in past 10 years (UniProt Consortium 2024)

To the best of our knowledge, PhaVIP (Shang et al. 2023) is the only previously reported PVP classification method that uses an image-based sequence representation, which employs frequency chaos game representation (FCGR). However, these encoding methods based on k-mer frequency tend to remove the spatial information of the sequences because of their compact transformation (Akbari Rokn Abadi et al. 2022). This highlights the need to devise an alternative effective encoding strategy. Given the constraints of both earlier and contemporary encodings, there is a drive to develop innovative and more effective attributes tailored to specific objectives.

Apart from the concern of encoding, a lack of study has been observed on the impact of uncertainty in phage virion classification. Studies show that deep learning methods often give overconfident results (Srivastava et al. 2014), which might lead to high uncertainty in situations previously unseen by the model. Especially for PVP, where the sequence conservation is low and training data goes through a specific set of curations, there is a reasonable possibility of model uncertainty. On a contrasting note, increased accuracy of prediction is at times positively correlated with the model's

uncertainty (Ghoshal and Tucker 2020). These uncertainties are particularly worrisome in safety-essential scenarios, such as medical diagnosis and drug engineering, where result reliability is of utmost importance (Hüllermeier and Waegeman 2021). Therefore, in this study, we design a novel image-based encoding technique inspired by the DNA-walk algorithm used for DNA sequences and adapt it for phage protein sequences. This encoding has the potential to overcome the drawback of spatial encoding loss (Akbari Rokn Abadi et al. 2022), employing pre-trained convolutional neural networks (CNNs) to attain competitive results on the prediction task. Furthermore, we investigate the trend of prediction uncertainty for these protein sequences and quantify the prediction confidence in terms of sequence-varying features, providing a validation standard as to which kinds of sequences tend to give better accuracy but with lower confidence, necessitating further assessment before annotation.

1.2 Research Contribution and Organization

Given the existing limitations on PVP research, our study proposes ProteoKnight, a method for PVP classification via a new encoding technique, named Knight Encoding. The core contributions from our research are as follows:

- We proposed a novel encoding strategy, "Knight Encoding", to convert text-based protein sequences into image-based data.
- We employed state-of-the-art deep learning techniques to conduct comprehensive classification analyses, evaluating the efficacy of the proposed encoding algorithm.
- We explored uncertainty aspects in PVP classification using Monte Carlo Dropout (MCD), identifying sequence types exhibiting high accuracy yet comparatively low confidence, necessitating further assessment before annotation.

The advancements in PVP classification through ProteoKnight hold the potential for far-reaching implications across multiple scientific domains. By offering accurate and efficient classification, our model serves as a versatile tool that bridges gaps in various research and application areas. From accelerating phage therapy development in medicine, host prediction, and taxonomy classification to enhancing metagenomic analyses in environmental microbiology, ProteoKnight's utility spans a broad spectrum of biological sciences. Importantly, ProteoKnight's novel encoding approach, when used in conjunction with existing methods, offers a complementary view of sequence classification. This synergy between different approaches has the potential to capture nuanced features of protein sequences, ultimately leading to more robust and comprehensive classification results. By exploring a new encoding method for PVP classification and addressing prediction uncertainty, ProteoKnight contributes to both computational biology and practical applications, offering a fresh perspective that could be valuable in safety-critical areas like phage therapy and bio-engineering.

To fully explore these contributions and their implications, the remainder of this paper is structured as follows: Section 2 provides the literature review, contextualizing our work within the field of PVP classification. Section 3 details our methods, including the novel Knight Encoding strategy and ProteoKnight implementation. Section 4 presents our experimental results and performance evaluation, while Section 5 offers a comprehensive discussion of our findings and their implications. Finally, Section

6 concludes the paper, summarizing key contributions and outlining future research directions.

1.3 Pre-trained CNNs for Classification

Pre-trained convolutional neural networks (CNNs) have gained significant attention in the field of bioinformatics, particularly for protein sequence classification tasks. These models, which are initially trained on large-scale datasets such as ImageNet (Russakovsky et al. 2015), have demonstrated remarkable performance in various computer vision applications. The success of pre-trained CNNs can be attributed to their ability to learn rich, transferable feature representations that can be fine-tuned for specific tasks, including phage virion protein (PVP) classification.

A major advantage of pre-trained CNNs is their computational efficiency. Training deep learning models from scratch can be time-consuming and resource-intensive, especially when dealing with large datasets. By using pre-trained models, it is possible to significantly reduce the training time and computational resources required, as the models have already learned meaningful features from extensive training on diverse datasets (Kornblith et al. 2019). This efficiency allows researchers to focus on fine-tuning the models for specific tasks and iterating on different architectures and hyperparameters.

Furthermore, the strong generalization capabilities of pre-trained CNNs make them a good choice for phage classification tasks, given their high diversity and low sequence conservation. Pre-trained CNNs have demonstrated their ability to generalize well to unseen data, making them suitable for handling the challenges associated with phage protein classification (Zhuang et al. 2021).

1.4 Uncertainty Estimation in Prediction

Although existing deep learning models have been seen to provide impressive accuracy, no study has yet been conducted on the impact of uncertainty in phage virion classification. Studies show that deep learning methods often give overconfident results (Guo et al. 2017), which might lead to high uncertainty in situations previously unseen by the model. On a contrasting note, increased accuracy of prediction is at times positively correlated to the model's uncertainty (Ghoshal and Tucker 2020). The aforementioned uncertainty can be categorized into two distinct types: Aleatoric Uncertainty (data-centric), and Epistemic Uncertainty (model-centric). The nature of the two uncertainties is illustrated in Figure 3.

These uncertainties are particularly worrisome in safety-critical scenarios, such as medical diagnosis and drug engineering, where result reliability is of utmost importance (Hüllermeier and Waegeman 2021). Rather than having point estimates for a task, uncertainty analysis attempts to deduce the variance over the distribution of the data in consideration. This analysis provides insight into both the reliability and accuracy of the model. Although much work has been done in uncertainty analysis for deep learning frameworks, the tasks dealt with are primarily based on either computer vision, image processing, or natural language processing targeting human language.

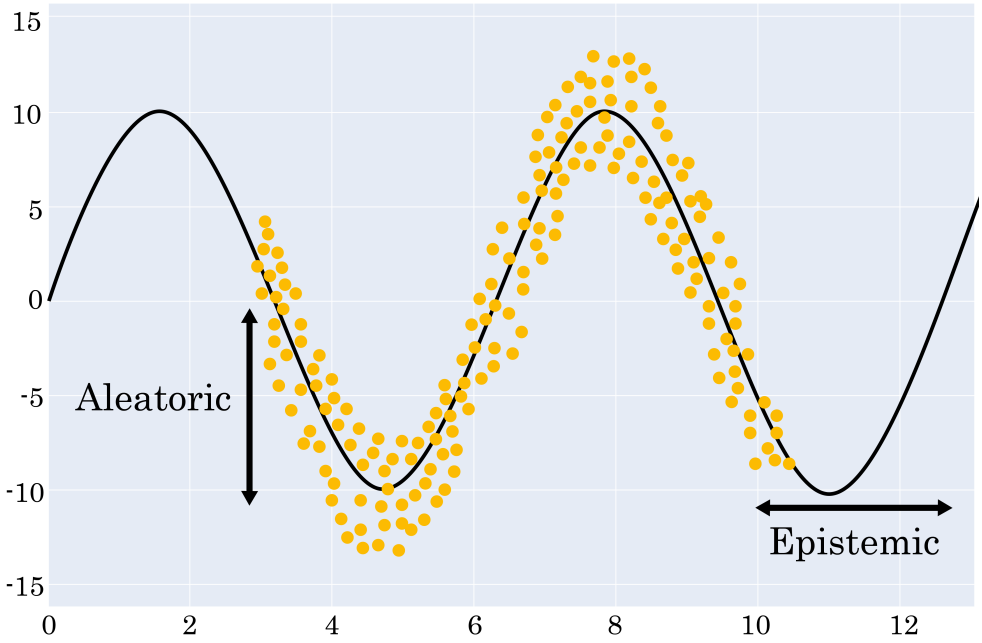


Fig. 3 A schematic view of aleatoric and epistemic uncertainty in prediction (Abdar et al. 2021)

In cases of protein sequences depicting biological language, studies regarding uncertainty have been a field yet to be explored. Especially for PVP, where the sequence conservation is low and training data goes through a specific set of curations, there is a reasonable possibility of model uncertainty. Exploring this uncertainty can establish a validation standard to detect sequence types that offer higher accuracy but lower confidence, necessitating additional evaluation before annotation.

1.4.1 Monte Carlo Dropout

The concept of utilizing dropout was introduced by Gal and Ghahramani (2016), who employed it as a means of approximating probabilistic Bayesian models for deep Gaussian processes. Dropout is a regularization technique that mitigates overfitting. Bayesian neural networks aim to acquire knowledge about the posterior distribution of weights, conditioned on a given input sample. However, computing these posteriors analytically is infeasible, so sampling techniques can be used to estimate the weight posteriors. This involves conducting several stochastic evaluations utilizing distinct weight samples within the model.

Monte Carlo Dropout is such a sampling technique used to estimate the weight posterior in a given dataset. During training, it randomly retains or drops nodes inside the neural network based on a Bernoulli random variable's probability distribution P (refer Equation 1). As shown in Figure 4, the dropout approach incorporates this probability parameter. Consequently, when the same input passes through the model,

the output may vary slightly depending on which nodes are activated or dropped. Each of these outputs represents a distinct sample within the network.

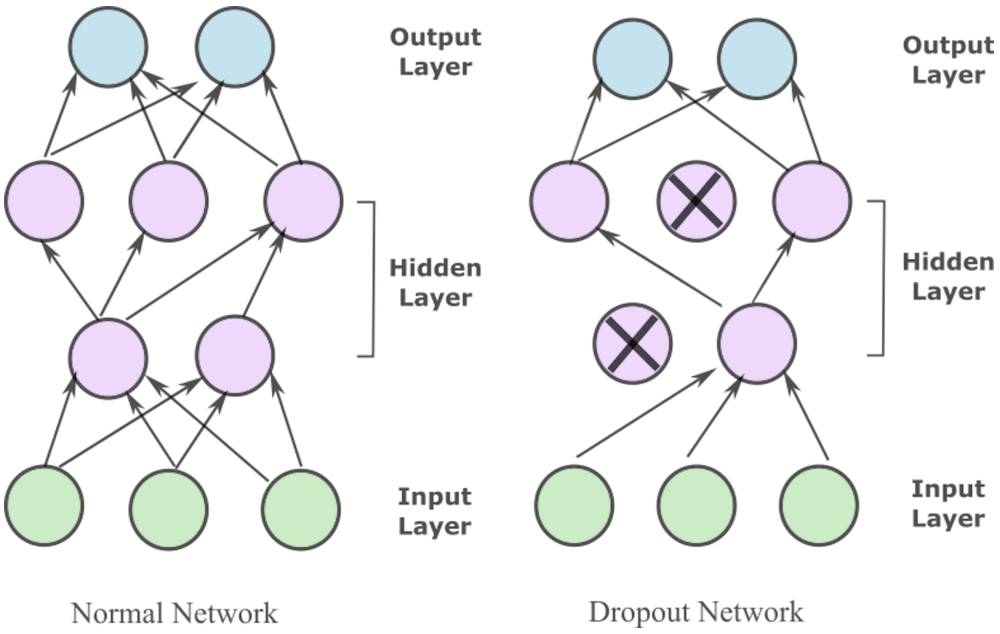


Fig. 4 Dropout mechanism in a neural network

$$Z_{w,t} \sim Bernoulli(p) \; \forall \, w \in \mathbf{W} \tag{1}$$

In this case, T samples are selected from the weight distribution (Equation 2) and used to perform T stochastic forward passes with dropout, where the input is $\mathbf{X}$ and the output is $\mathbf{Y}$. This allows the calculation of the expected value of the prediction Y and the variance of these predictions across iterations (Equations 3 and 4). Thus, using T stochastic forward passes with different weight samples $W_{t\,t=1}^{T}$, dropout acts as stochastic sampling. The observed variability in the T stochastic outputs primarily represents the model's inherent uncertainty or epistemic uncertainty. In other words, when predictions exhibit large variance, it suggests significant epistemic uncertainty in the model.

$$\mathbf{W}_t = \text{train}(f; \mathbf{X}, \mathbf{Y}) \tag{2}$$

$$\mathbb{E}(\hat{\mathbf{Y}} \mid \mathbf{X}) = \frac{1}{T} \sum_{t=1}^{T} f(\mathbf{X} \mid \mathbf{W}_t) \tag{3}$$

$$Var(\hat{\mathbf{Y}} \mid \mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{X})^2 - \mathbb{E}(\hat{\mathbf{Y}} \mid \mathbf{X})^2 \tag{4}$$

2 Related Work

Machine learning and deep learning have significantly benefited healthcare research. As medical data became more complex and accessible, traditional methods proved inadequate, leading to widespread integration of machine learning in healthcare analysis (Habehh and Gohel 2021). Deep learning techniques emerged to handle complex, heterogeneous, and unstructured medical data from various sources, showing impressive results in biomedical research (Miotto et al. 2018). Their applications include studying diseases, interpreting medical imaging, and forecasting health metrics. Consequently, cancer research has particularly benefited, with new studies exploring previously understudied areas like skin cancer (Kavitha et al. 2023), colon cancer (Haq et al. 2023), and viral cancer expressions (Elbasir et al. 2023). Increased computing power has allowed deep learning to excel in computer vision, streamlining medicine and neuroimaging research (Sánchez Fernández and Peters 2023). These techniques are revolutionizing the diagnosis, treatment, and management of neurological conditions (Hussain et al. 2024). Furthermore, some studies have used deep learning to generate synthetic medical data, addressing resource constraints (Hu et al. 2022). In some research studies, evidence has also been found supporting the clinical utility of these methods in decision support tools, laboratory diagnostics, and public health outbreak management (Theodosiou and Read 2023; Badawy et al. 2023). As these technologies evolve, they promise to revolutionize healthcare practices and improve patient outcomes across various medical specialties.

Among the abstract data found in medical analysis, proteins stand out as both vital and prominent. These essential macromolecules are responsible for the structural and functional mechanisms of living organisms and have been extensively studied using various artificial intelligence techniques for protein classification, interaction prediction, and functional analysis (Ghosh and Mitra 2024). Such studies aim to classify and extract information from different protein families to serve various medical sectors. While these studies address specific aspects of protein analysis, recent breakthroughs have pushed the boundaries of AI applications in protein research further. Such as, in the study by Brandes et al. (2022), which introduced ProteinBERT, a deep learning model for protein sequences. Combining language modeling with Gene Ontology annotation prediction, this model achieves near state-of-the-art performance on multiple benchmarks while using a smaller and faster model than competing methods. Additionally, Senior et al. (2020) developed AlphaFold, a groundbreaking protein structure prediction system that uses deep learning to predict highly accurate protein structures by leveraging genetic information and employing neural networks. These advancements demonstrate the growing capabilities of AI-driven approaches in protein research, offering promising tools for various applications in structural biology and drug discovery. More recently, transformer-based architectures have also been explored for extracting biological information directly from nucleotide and protein sequences. For example,

Stefanini et al. (2023) employed the Perceiver architecture to predict gene and protein expression levels from DNA and protein sequences, demonstrating the effectiveness of attention-based models in sequence learning.

Within this vast protein universe, phage structural proteins, also known as phage virion proteins (PVPs), represent another crucial subgroup. As essential components of bacteriophage virions, these proteins are associated with the most abundant biological entities on Earth—bacteriophages (Cobián Güemes et al. 2016). As high-throughput sequencing technologies advance, the need for robust computational frameworks to analyze and characterize newly discovered bacteriophages becomes increasingly important for future researches in this field (Nami et al. 2021). Based on recent literature reviews (Meng et al. 2020; Nami et al. 2021), the classification of phage virions using machine learning techniques has shown promising results. Commonly employed algorithms such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Scoring Card Method (SCM) have demonstrated satisfactory classification accuracy, typically ranging from 70% to 80% on test sets. Building upon these foundational approaches, researchers have developed more sophisticated methods for phage virion classification. Seguritan et al. (2012) introduced iVIREONS, a deep learning approach leveraging amino acid frequency and protein isoelectric points for binary classification of major capsid and tail proteins. Similarly, Fang and Zhou (2021) developed VirionFinder, utilizing biochemical properties and one-hot encoding of sequences to identify complete and partial virion proteins via 1-D CNNs. Expanding on binary classification, Cantu et al. (2020) created PhaANN, achieving 86.2% test accuracy across multiple classes by integrating dipeptide and tripeptide composition features within artificial neural networks (ANNs). Building upon these advancements, Fang et al. (2022) introduced DeePVP, employing one-hot encoding and 1-D CNNs for multiclass classification, demonstrating enhanced performance, particularly in binary classification tasks compared to PhaANN.

Diverging from prior methodologies, Shang et al. (2023) utilized an image-based FCGR encoding on sequences, employing a vision transformer model for classifying 8 subclasses, yielding superior results on standard evaluation metrics. Image representations of such 1-dimensional sequences have been widely adopted in literature for various protein-related tasks (Pang et al. 2017). Additionally, Akbari Rokn Abadi et al. (2022) developed WalkIm, an image-based encoding method enhancing biological sequence classification using CNNs with the DNA-walk algorithm. Collectively, these approaches demonstrate the versatility and effectiveness of image-based representations in protein sequence analysis and classification tasks across various domains.

3 Materials and Methods

The following sections detail our methodology, covering the dataset description, encoding algorithm, pre-trained classifiers for phage classification, and the use of MCD for prediction uncertainty estimation. Figure 5 illustrates the workflow, showcasing key stages of data processing, feature extraction, classification, and uncertainty quantification.

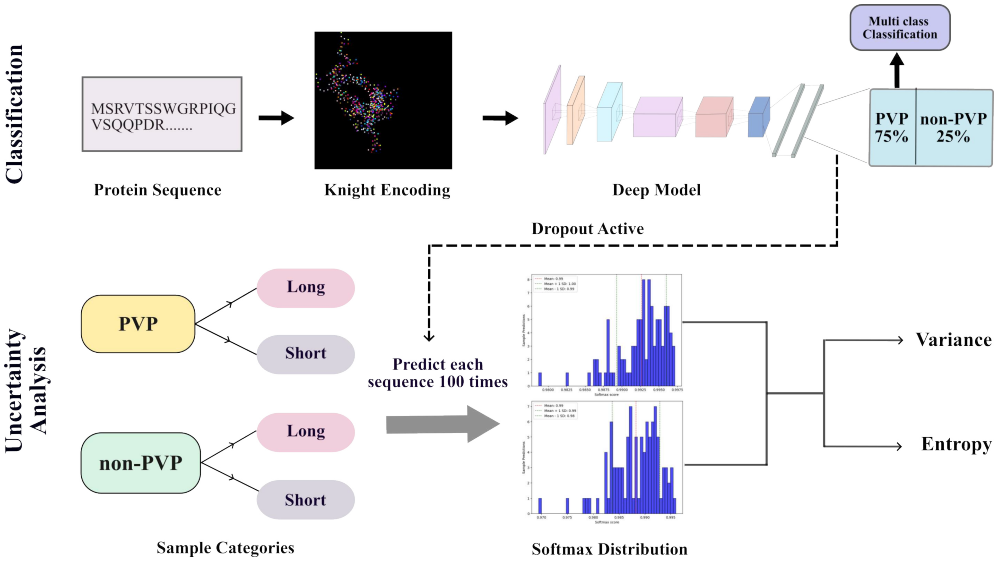


Fig. 5 PVP classification and uncertainty analysis framework

3.1 Dataset Description

For this study, we utilized a benchmark dataset (Shang et al. 2023). Sequences with low-confidence labels, which introduced ambiguity, were removed from the dataset. Subsequently, a search was conducted using diverse keywords to identify and extract structural proteins from the remaining sequences. Non-structural proteins were collected by identifying enzymes with names ending in the suffix 'ase'. The CD-HIT algorithm (Li and Godzik 2006) was then applied to identify clusters of sequences with a similarity threshold of 90%. The longest sequence within each cluster was chosen as the representative sequence. The final database comprises a total of **35,204** PVP sequences and **46,883** non-PVP sequences of varying lengths. The PVP sequences are further categorized into eight groups, curated for multi-class classification purposes. Figure 6 illustrates the overall class breakdown, while Table 1 provides a summary of the multiclass data counts.

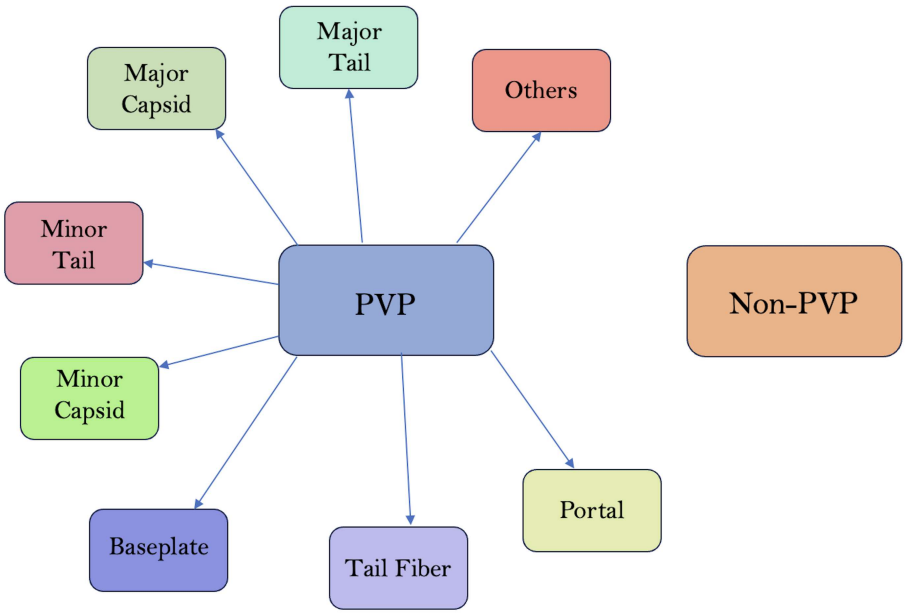


Fig. 6 Illustration of dataset classes

3.2 'Knight' Encoding

Protein sequences lack image-based encoding compared to DNA sequences. For DNA sequences, the random walk or DNA walk method has been widely used for encoding for sequence analysis (Berger et al. 2004), achieving high prediction accuracy. The study by Akbari Rokn Abadi et al. (2022) utilizes DNA-walk to classify viral genomes using CNNs, suggesting it as an alternative to k-mer-based encodings like FCGR, which may lose spatial information of encoded sequences. They propose its usability for other sequence types such as RNA and proteins, although no such implementation has been reported. Based on this, we introduce a novel walk-based encoding technique specifically for protein sequences, aiming to evaluate its effectiveness in characterizing phage structural proteins.

This encoding method employs polar coordinates to interpret protein sequences. Each amino acid is assigned an angle value, totaling 360 degrees, representing its directional movement. Unique color-coded markers or circular points depict the amino acids, positioned on the image based on their angle and distance of movement (radius). Equations 5 and 6 are used to calculate the x and y direction displacements for encoding residues from their current positions.

$$x = r \times \cos(\theta) \tag{5}$$

$$y = r \times \sin(\theta) \tag{6}$$

The entire encoding algorithm can be divided into three major parts:

1. **Definitions**
   A list is defined that includes the letter representation of each amino acid, along with a complementary dictionary that maps each amino acid to a corresponding color. These are subsequently utilized for the computation of angles and placement of points.

Table 1 PVP multi-class data entries

PVP Classes	Number of entries
Baseplate	3362
Portal	2770
Tail Fiber	2305
Major Capsid	2443
Minor Capsid	398
Major Tail	5083
Minor Tail	1458
Others	17385

List of amino acids for the encoding,

```
self.amino_acids = ['A', 'C', 'D', 'E', 'F', 'G',
                    'H', 'I', 'K', 'L', 'M', 'N',
                    'P', 'Q', 'R', 'S', 'T', 'V',
                    'W', 'Y']
```

Amino acid to color mapping,

```
self.colors = {
    'A': (255, 0, 0), 'C': (255, 255, 0),
    'D': (0, 234, 255), 'E': (170, 0, 255),
    'F': (255, 127, 0), 'G': (191, 255, 0),
    'H': (0, 149, 255), 'I': (255, 0, 170),
    'K': (237, 185, 185), 'L': (185, 215, 237),
    'M': (231, 233, 185), 'N': (220, 185, 237),
    'P': (185, 237, 224), 'Q': (143, 35, 35),
    'R': (35, 98, 143), 'S': (143, 106, 35),
    'T': (107, 35, 143), 'V': (115, 237, 155),
    'W': (204, 204, 204), 'Y': (0, 64, 255)
}
```

2. **Angle Calculation**
   (a) We have employed the structure of a 20-sided polygon called 'Icosagon' to represent the 20 amino acids. The amino acids are evenly distributed throughout the 20 vertices of the Icosagon. Each point is separated from the others by a constant angle of 18° ($360°/20 = 18°$), as seen in Figure 7.
   (b) The angles for each letter in the amino acid list are determined based on the index position of the amino acids in the list mentioned previously. For instance, C is at index 1, and G is at index 5.

   So for 'C', the associated angle is,

$$\text{index of C} = 1$$

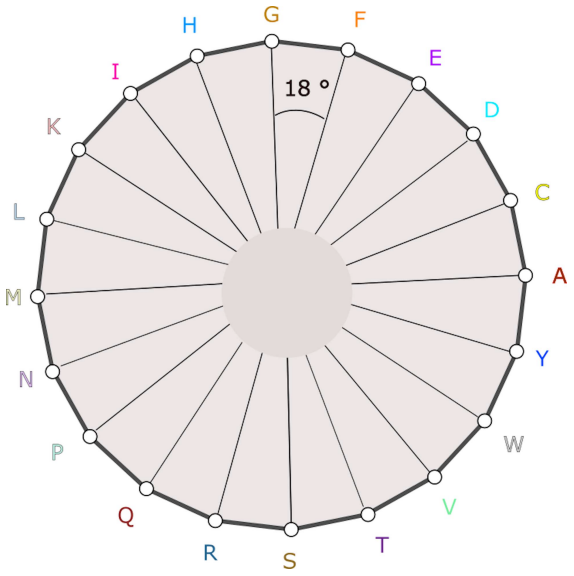


Fig. 7 Icosagon with color-coded vertices and 18° angular assignments for amino acids

$$\theta = 1 \times 18^\circ = 18^\circ$$

Similarly, for 'G', the angle will be,

$$\text{index of G} = 5$$

$$\theta = 5 \times 18^\circ = 90^\circ$$

As we shall be using polar coordinate formulas in our study, the angles are converted from degrees to radians.

3. **Placement**

(a) If the dimensions of the image are defined as $M \times M$, then the encoding will start from the origin, which is located at the coordinates $(x, y) = (M/2, M/2)$, or in other words, from the middle of the image.

(b) For the first amino acid (or the first letter of our sequence), $(x, y) = (M/2, M/2)$ or the center of our image will act as the starting point, i.e., current coordinates.

(c) Using the fixed radius and corresponding angle of an amino acid, the horizontal and vertical displacement $(x\prime, y\prime)$ of the encoding is determined using equations 5 and 6, relative to the current coordinates (x, y), as shown in equations 7 and 8.

$$x = x + x\prime \tag{7}$$

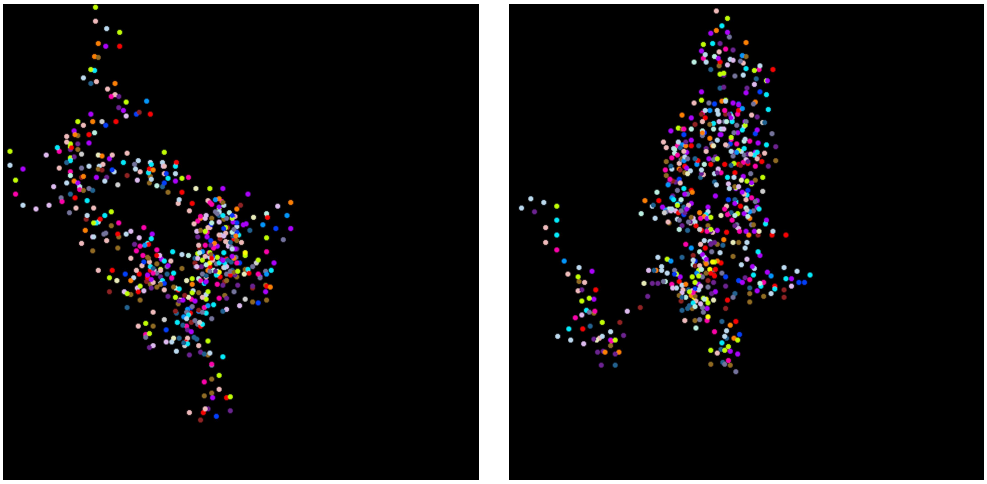
$$y = y - y\prime \tag{8}$$

A circular point will be positioned at the coordinates (x, y) to represent the amino acid being encoded, with its corresponding color. The polar coordinate

radius and point size remain consistent at 15 and 2, respectively, for all amino acids.

(d) Subsequently, the next character of the protein sequence will utilize the coordinates (x, y) of the previous character as its starting point. A new horizontal and vertical shift $(x\prime, y\prime)$ will be generated based on the amino acid's associated angle. Using these new shift values $(x\prime, y\prime)$, a new set of coordinates (x, y) are calculated for placing the current amino acid character.

(e) The remaining characters in the sequence will continue to follow the previously specified technique until the entire sequence is encoded, as shown in Figure 8. If a point hits the boundary of the image during encoding, it will begin encoding from the center of the image at $(x, y) = (M/2, M/2)$.



(a) **(b)**

Fig. 8 Encoding of two example proteins: (a) PVP (YP_009900749.1 minor capsid protein [Lactococcus phage 62503]); (b) non-PVP (YP_009847768.1 (NAD(+)) DNA ligase [Vibrio phage USC-1])

In summary the Knight encoding algorithm (1) is given as follows:

3.3 Pre-trained CNN Models

CNNs have demonstrated remarkable efficacy across various domains of computer vision tasks. To extend the utility of such powerful models to fields like medicine and bioinformatics, where large datasets are often limited, transfer learning is employed to train models (Salama and Aly 2021). Besides circumventing the need for extensive datasets, transfer learning facilitates the construction of models in a resource-efficient and time-saving manner (Rawat and Wang 2017). Pre-trained CNNs have exhibited competitive performance compared to popular pre-trained transformer models, further highlighting their relevance in image classification tasks (Liu et al. 2022). Therefore, for our study, we have utilized existing pre-trained CNN variants such as GoogLeNet (Szegedy et al. 2015), EfficientNetV2 (Tan and Le 2021), and MobileNetV3 (Howard

Algorithm 1: Encoding a Protein Sequence into an $M \times M$ Image

Input: Protein sequence S of length N; image size M; step radius $r = 15$; residue ordering *amino_acids* (20 residues); color map *colors*

Output: Encoded RGB image I of size $M \times M$

$(x, y) \leftarrow \big(\lfloor M/2 \rfloor,\ \lfloor M/2 \rfloor\big)$
for $i \leftarrow 0$ **to** $N - 1$ **do**
  $residue \leftarrow S[i]$
  $\theta \leftarrow 18^\circ \times amino_acids.\text{index}(residue)$
  $x \leftarrow x + \text{trunc}(r \cos\theta)$
  $y \leftarrow y - \text{trunc}(r \sin\theta)$
  if $x \notin (0, M)$ **or** $y \notin (0, M)$ **then**
    $(x, y) \leftarrow \big(\lfloor M/2 \rfloor,\ \lfloor M/2 \rfloor\big)$
  $color \leftarrow colors[residue]$
  Plot a point of size 2 at (x, y) in I using $color$

et al. 2019), focusing on models with fewer parameters yet capable of achieving superior results.

3.4 Estimating Uncertainty using MCD

To assess uncertainty in binary classification, we activate the model's dropout layer with a 0.2 (20%) dropout rate during the test phase. Consequently, we develop a method to classify data into distinct groups based on sequence properties to analyze uncertainty for each category. Unlike text data, protein sequences do not reveal unique variability apart from class labels. The only noticeable diversity is sequence length, ranging from hundreds to thousands of residues (Figures 9 and 10). Therefore, investigating the model's robustness to sequence length can determine its vulnerability to uncertainty. The training data (70% of the full dataset) were divided based on an equilibrium sequence length δ, ensuring nearly equal data with lengths less than or equal to δ and greater than δ to mitigate length bias (Table 2).

Table 2 Data Distribution for PVP and non-PVP δ-values

Class	**Sequences** $\leq \delta$	**Sequences** $> \delta$	**Total Sequences**
PVP ($\delta = 350$)	12589	12060	24649
non-PVP ($\delta = 275$)	16502	16316	32818

The data are separated into four categories: PVP (< 350), PVP (> 350), non-PVP (< 275), and non-PVP (> 275). For each category, 100 randomly chosen sequences are predicted 100 times using the dropout model, providing a prediction distribution. The predictions are passed through the softmax activation, transforming raw logits into

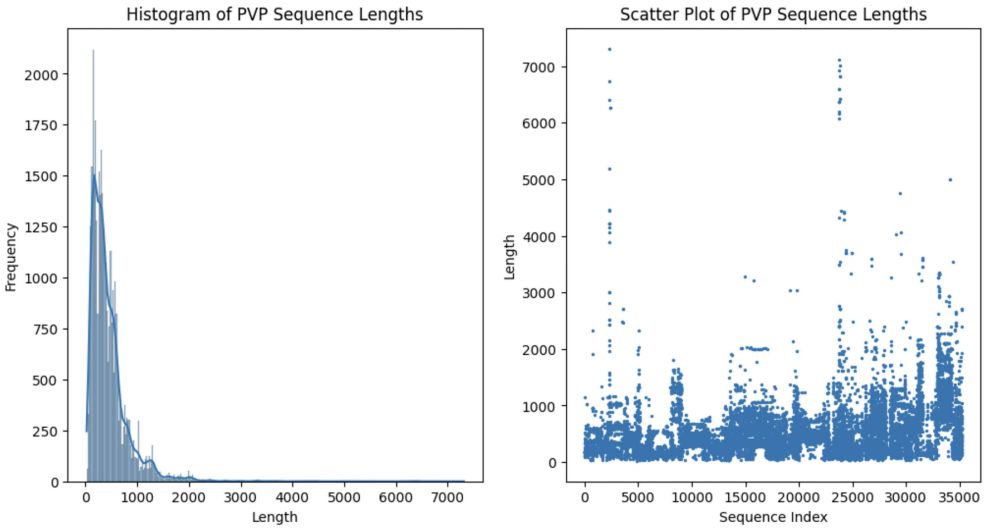


Fig. 9 Length Distribution of PVP sequences

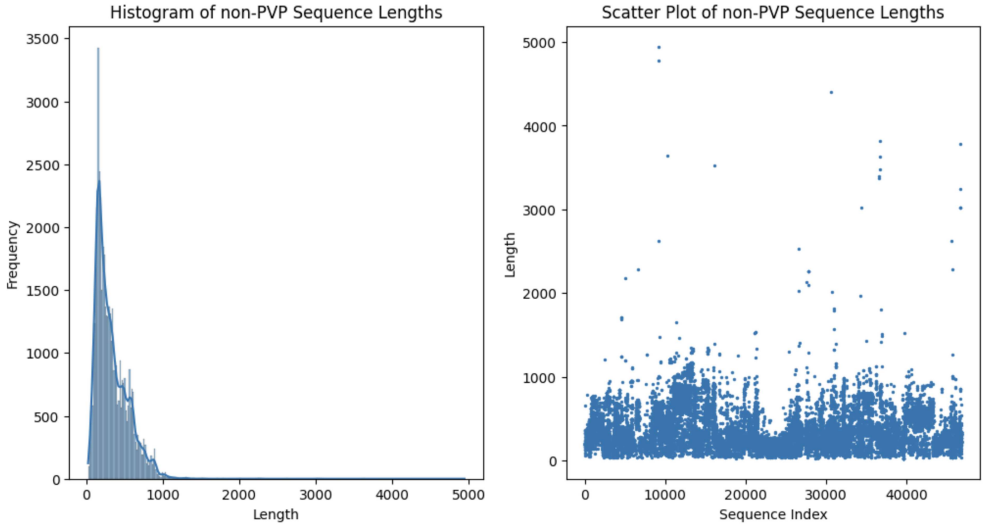


Fig. 10 Length Distribution of non-PVP sequences

probabilities. The mean and variance of these softmax probabilities from each category are determined to compare uncertainty across categories. As shown in Figures 11 and 12 for a sample, we expect a probability distribution, with the x-axis representing softmax scores and the y-axis indicating the number of predictions for each score.

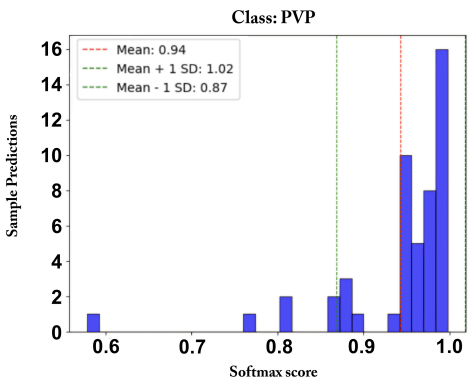


Fig. 11 Softmax Distribution of PVP

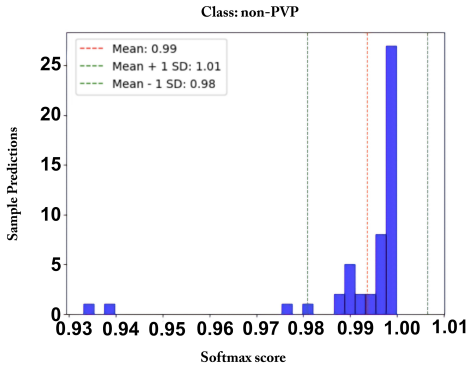


Fig. 12 Softmax Distribution of non-PVP

Typically, correct predictions should cluster on the right side of the axis, as a softmax score greater than 0.5 indicates the data belongs to that class. Conversely, predictions toward the left denote incorrect predictions. The histogram's dispersion, or prediction variance, signifies the model's confidence level. Reduced variance indicates higher confidence, while higher variance suggests greater prediction uncertainty for that specific data.

4 Experimental Setup and Results

4.1 Experimental Configuration

The training and testing procedures for the experiments were developed and implemented using a comprehensive suite of Python libraries, including TensorFlow, Keras, and PyTorch. The computational resources utilized for model training and evaluation consisted of a high-performance computer equipped with an AMD Ryzen 7 3700X 8-core CPU, 16GB of DDR4 RAM, and an NVIDIA GeForce RTX 3060 Ti graphics processing unit. The experiments were conducted on the Windows 11 Pro operating system, leveraging the Python 3.11 interpreter as the primary programming environment.

4.2 Training and Evaluation

To evaluate the pre-trained CNN models, each was fine-tuned on the encoded image dataset for 25 epochs, with a batch size of 32. Accuracy was the primary metric, and 'binary_crossentropy' was the optimization loss function. Accuracy was considered the primary metric, as correctly identifying positive and negative samples is equally important. Additionally, recall/sensitivity, specificity, precision, and F1-score were calculated for a comprehensive performance overview. By evaluating diverse metrics, a thorough understanding of the CNN models' performance was obtained, facilitating

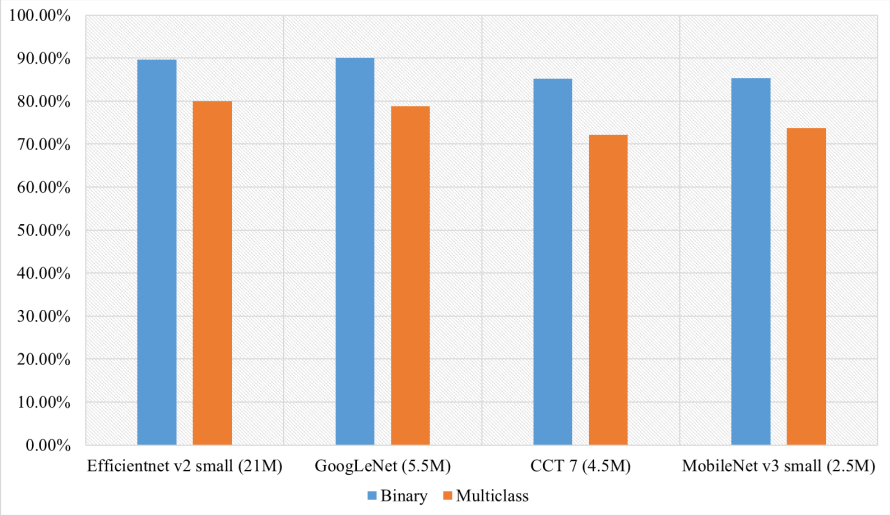


Fig. 13 Model comparison

optimal model selection. These additional metrics were derived from the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

4.3 Experimental Results

4.3.1 Model Predictions

As outlined in the literature review section, the PVP classification task encompasses both binary and multiclass categorization challenges. To comprehensively evaluate the efficacy of our proposed encoding approach, we employed deep learning techniques, which have been demonstrated to outperform traditional machine learning methods in terms of prediction accuracy and feature extraction simplicity. Rather than constructing an entirely new model from scratch, we leveraged several high-performing pre-trained CNNs and fine-tuned them on our dataset. Our particular interest lies in assessing the encoding efficiency under resource-constrained conditions, which motivated the selection of the specific CNN models employed in this study. To this end, we

conducted a comparative analysis of various pre-trained CNN architectures to identify the most suitable model for our classification task.

Several pre-trained models from the PyTorch library were utilized. Among these, the top four pre-trained architectures that demonstrated satisfactory performance on our dataset, along with their respective parameters (provided in brackets) and prediction accuracy are illustrated in Figure 13. All these CNNs exhibited competitive performance in terms of binary and multiclass classification, with EfficientNet v2 yielding the best results, surpassing GoogLeNet by a slight margin. However, GoogLeNet gained an advantage due to its significantly smaller parameter count, roughly one-fourth that of EfficientNet v2. Ultimately, the GoogLeNet architecture achieved an optimal trade-off between computational efficiency and classification quality.

Based on the evaluation results, the GoogLeNet architecture emerged as the most promising candidate model for further analysis. This analysis includes comparative assessments with previous works and in-depth uncertainty quantification. Table 3 presents the GoogLeNet model's performance across various evaluation metrics, while Table 4 outlines the optimal hyperparameters used in training. Here, "Image Size" refers to the original size of the images generated from protein sequences after encoding, while "Image Input Size" refers to the final resized dimensions of the images as fed into the classification models'. The layer-wise parameters of GoogLeNet are the same as those provided in torchvision (2017). Evaluating on the test set, our GoogLeNet-based method achieved an accuracy of approximately 89%, which is competitive with state-of-the-art methods. Notably, the precision, sensitivity, and F1-score of our GoogLeNet model surpassed most previous methods, as shown in Table 5. For comparison, we also tested EfficientNet V2, which achieved a slightly higher accuracy of around 91.2%. Table 5 provides a comprehensive comparison of the results of our GoogLeNet-based approach with existing state-of-the-art tools for the PVP classification task.

Table 3 Evaluation results of the optimal model (GoogLeNet)

Task	**Params**	**Accuracy**	**F1 Score**	**Recall**
Binary	5.5M	88.60%	89.87%	88.54%
Multiclass	5.5M	76.37%	76.37%	76.37%

Table 4 Optimal hyperparameters used for model training

Image Size	**Input Size**	**Epochs**	**Batch Size**	**Learning Rate**
512×512	224×224	25	32	0.001

Table 5 Comparative analysis for binary classification

	PVP			non-PVP		
Work	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
PhANN (Cantu et al. 2020)	0.81	0.87	0.84	0.86	0.80	0.83
DeePVP (Fang et al. 2022)	0.99	0.51	0.68	0.67	1.00	0.80
VirionFinder (Fang and Zhou 2021)	0.77	0.92	0.84	0.90	0.73	0.81
Meta-iPVP (Charoenkwan et al. 2020b)	0.76	0.84	0.80	0.82	0.73	0.77
PVPred-SCM (Charoenkwan et al. 2020a)	0.78	0.57	0.66	0.66	0.84	0.74
PhaVIP (Shang et al. 2023)	0.93	0.95	0.94	0.95	0.93	0.94
ProteoKnight	**0.83**	**0.89**	**0.90**	**0.94**	**0.89**	**0.90**

4.3.2 Uncertainty Analysis

Uncertainty estimations of deep learning models serve as a measure of the reliability of their predictions. As stated in the methodology, the variance of the prediction distribution during dropout passes is regarded as a standard uncertainty quantification metric. Entropy (H) is another commonly utilized metric in prior research focused on uncertainty quantification, as demonstrated by Milanés-Hermosilla et al. (2021) and Kendall and Gal (2017). In the context of machine learning, entropy is characterized as the disorder or uncertainty of the prediction model, or simply the level of 'surprise' exhibited by the model when encountering particular data samples (Starmer 2021). The lower the value of entropy, the better. Equation 14 is utilized in this study to calculate the entropy of our binary classification samples, where P is the prediction probability for the particular data under consideration.

$$Entropy = \left(P \times \log_2(\frac{1}{P})\right) + \left((1-P) \times \log_2 \frac{1}{(1-P)}\right) \tag{14}$$

In our uncertainty analysis, we employed both variance and entropy metrics. Variance was computed across all four sequence categories, while entropy was determined for specific sequence encodings exhibiting higher variance compared to others. The top-performing GoogLeNet model was selected for prediction analysis. Images from four distinct categories were collected: PVP short sequences, PVP long sequences, non-PVP short sequences, and non-PVP long sequences. The average and variance of the prediction distribution were graphed and compared for each category. It was observed that the model exhibited higher confidence (lower prediction variance) when predicting non-PVP sequences compared to PVP sequences, and showed greater certainty for shorter sequences compared to longer ones. This uncertainty pattern was validated by employing various dropout rates (0.1, 0.2, 0.3) and randomly shuffling the sequences within each category. The figures presented utilize a dropout value of 0.2 for illustration purposes. Subsequently, Table 6 lists the variance and corresponding MCD values for each category, indicating lower variance for shorter sequences and non-PVP predictions. Samples with the highest and lowest variance were selected from each category,

and their mean prediction entropy was calculated, which is listed in Table 7. It was found that entropy values were lower for shorter sequences and non-PVP predictions, corroborating the observed uncertainty patterns.

Table 6 Variance for PVP and non-PVP with different dropout rates

Categories	**MCD 0.1**	**MCD 0.2**	**MCD 0.3**
PVP (short)	0.06801	0.08558	0.05536
non-PVP (short)	0.05914	0.07555	0.04860
PVP (long)	0.07276	0.09119	0.05926
non-PVP (long)	0.05994	0.07663	0.04900

5 Discussion

The primary objective of this study was to investigate the viability of employing a highly efficient DNA sequence encoding technique for the more intricate task of phage protein classification and to assess whether the results would be equally effective as those obtained with DNA sequences. We leveraged a variety of state-of-the-art deep learning convolutional neural networks to classify the encoded images. Additionally, we conducted an uncertainty analysis on the most efficient model to evaluate its strengths and limitations. Traditional machine learning methods for phage virion protein classification, such as Meta-iPVP (Charoenkwan et al. 2020b) and PVPred-SCM (Charoenkwan et al. 2020a), rely on algorithms like random forest, genetic algorithms, and support vector machines. These approaches require manual feature curation, such as k-mer frequencies and probabilistic features, and generally yield lower accuracy compared to newer deep learning methods.

Table 7 Entropy for PVP and non-PVP samples with high and low variance

Categories	**Entropy (High Var)**	**Entropy (Low Var)**
PVP (short)	0.1122	0.1259
non-PVP (short)	0.1369	0.1398
PVP (long)	0.2111	0.3943
non-PVP (long)	0.06539	0.1716

Deep learning techniques, where hidden features are learned by neural networks, have shown significant improvements in PVP classification. For instance, PhANN (Cantu et al. 2020) demonstrates superior performance compared to earlier methods, but it still relies on inputting thousands of features into its predictive artificial neural

Table 8 Qualitative comparison of existing methods and ProteoKnight

Work	Approach	Model	Feature	Image Data
PhANN	Deep learning	ANN	K-mer frequency	✗
DeePVP	Deep learning	1-D CNN	One-hot encoding	✗
VirionFinder	Deep learning	CNN	One-hot encoding, biochemical properties	✗
Meta-iPVP	Machine learning	SVM, RF, NB, ANN	Probabilistic features	✗
PVPred-SCM	Statistical	SCM, Genetic Algorithm	Dipeptide composition	✗
PhaVIP	Deep learning	ViT	FCGR	✓
ProteoKnight	**Deep learning**	**CNN**	**Knight encoding**	✓

network architecture, which can be computationally intensive and potentially less efficient. Additionally, the study by Fang et al. (2022) mitigates the need for extensive feature extraction by using one-hot encoding and employing a more powerful 1-D CNN architecture for classification. While this method shows good accuracy, the results are inconsistent across other evaluation metrics. Moreover, Shang et al. (2023) introduce the classification of FCGR-based encoded sequences using a transformer architecture, achieving high prediction performance on the dataset. However, this FCGR encoding tends to result in spatial feature loss (Akbari Rokn Abadi et al. 2022).

Addressing these concerns, our approach, ProteoKnight, explores an image-based encoding method based on the DNA-walk algorithm, which potentially mitigates the issue of spatial loss (Akbari Rokn Abadi et al. 2022) in the image-based representation of biological sequences. We chose to work with pre-trained CNN architectures, specifically GoogleNet, as they have demonstrated superior performance compared to pre-trained transformers for many classification tasks (Tay et al. 2021). This, in turn, offers the advantage of leveraging existing knowledge, significantly reducing training time and computational needs, while also providing consistent results across various evaluation metrics by using a proven, high-efficiency DNA encoding method adapted for proteins. A qualitative comparison between various methods and our approach is highlighted in Table 8. Our proposed technique for encoding and model selection yielded binary classification performance that approaches current state-of-the-art methods, narrowing the gap with the leading transformer-based approach (PhaVIP) while relying on a substantially simpler, parameter-efficient CNN pipeline.

However, in the multi-class prediction task, our model exhibited room for further improvement, producing moderate outcomes with an accuracy range of 72% to 78%. This can be attributed to a characteristic of our suggested encoding approach, where dot points may converge at the same location when amino acids align, potentially leading to overlap, particularly when one dot is precisely positioned over another. While this characteristic did not significantly impact the less complex binary classification task, it posed a challenge for the more intricate multi-class classification scenario. Nevertheless, our encoding technique demonstrated promising potential, and with further refinements, it could potentially address the multi-class classification task more effectively. The class-based uncertainty analysis revealed that our model demonstrated greater confidence when applied to non-PVP data compared to PVP data. This disparity may stem from the larger volume of training data available for the non-PVP

category versus the PVP category, or it could be related to their distinctive underlying sequence compositions. These insights into model performance and uncertainty not only inform our understanding of the current approach but also guide our future research directions, setting the stage for more robust and versatile protein classification methodologies.

6 Conclusion

Bacteriophages play vital ecological roles and present promising therapeutic avenues against bacterial pathogens. Accurate classification of their structural proteins, i.e., phage virion proteins (PVPs), is therefore crucial across disciplines. This study introduces ProteoKnight, a novel image encoding methodology for PVP sequences. Overcoming spatial constraints of existing techniques like frequency chaos game representation, ProteoKnight enables high-accuracy binary PVP classification (GoogleNet - 89% & EfficientNet V2 - 91%) via pre-trained convolutional neural networks on a benchmark dataset. Furthermore, our uncertainty quantification through Monte Carlo dropout elucidates how prediction confidence varies across PVP classes and sequence lengths, pioneering insights for phage protein classification research. Collectively, ProteoKnight, coupled with efficient neural models, offers an accurate way to annotate important phage proteins, thereby setting the stage for exploiting the great medical and environmental benefits of bacteriophages through careful analysis of their viral proteins.

While these results are promising, our research has also uncovered areas for further investigation and improvement. During the analysis, an atypical trend emerged for PVP sequences, which showed higher prediction variance (i.e., lower confidence) under MCD sampling than non-PVP sequences, despite comparable point-estimate accuracy. This phenomenon serves as an anchor for further investigation of uncertainty quantification for biological sequences. Regarding length-based uncertainty, the encoding limitation mentioned in the discussion section may elucidate the increased uncertainty observed in longer sequences compared to shorter ones, as longer sequence encodings are more susceptible to point overlaps. To address this issue, future research could focus on enhancing the knight encoding approach by extending it to a higher dimension through the segregation of frames for each amino acid or by optimizing the encoding's hyperparameters, such as radius, sphere/dot size, and so forth. This would enable the model to more effectively process the point representation of each amino acid residue, ensuring that no sequence information is compromised.

Acknowledgements. The authors would like to thank all individuals who provided valuable discussions and support during the course of this research.

Statements and Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. Not applicable. This study did not involve human participants or animals.

Consent for publication. Not applicable.

Data availability. This study utilizes the benchmark dataset established by Shang et al. (2023). The dataset was constructed by retrieving the latest sequence annotations up to December 2022 from the RefSeq viral protein database, followed by necessary data reconditioning procedures. The dataset is publicly available at https://phage.ee.cityu.edu.hk/download.

Code availability. The source code for ProteoKnight — including the Knight Encoding scheme, dataset construction, model training and evaluation, and the Monte Carlo Dropout uncertainty analysis — is publicly available on GitHub at https://github.com/eniac00/ProteoKnight.

Author contributions. **Md. Ishrak Khan:** Conceptualization, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Abir Ahammed Bhuiyan:** Conceptualization, Software, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Samiha Afaf Neha:** Visualization, Methodology, Investigation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Md. Khalilur Rahman:** Supervision, Project administration, Writing – review & editing. **Jannatun Noor:** Supervision, Project administration, Writing – review & editing.

References

Abdar M, Pourpanah F, Hussain S, et al (2021) A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 76:243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Akbari Rokn Abadi S, Mohammadi A, Koohi S (2022) WalkIm: Compact image-based encoding for high-performance classification of biological sequences using simple tuning-free CNNs. PLoS ONE 17(4):e0267106. https://doi.org/10.1371/journal.pone.0267106

Arif M, Ali F, Ahmad S, et al (2020) Pred-BVP-Unb: Fast prediction of bacteriophage virion proteins using un-biased multi-perspective properties with recursive feature elimination. Genomics 112(2):1565–1574. https://doi.org/10.1016/j.ygeno.2019.09.006

Badawy M, Ramadan N, Hefny HA (2023) Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology 10(1):40. https://doi.org/10.1186/s43067-023-00108-y

Berger JA, Mitra SK, Carli M, et al (2004) Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute 341(1-2):37–53. https://doi.org/10.1016/j.jfranklin.2003.12.002

Brandes N, Ofer D, Peleg Y, et al (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38(8):2102–2110. https://doi.org/10.1093/bioinformatics/btac020

Brüssow H, Desiere F (2001) Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Molecular Microbiology 39(2):213–223. https://doi.org/10.1046/j.1365-2958.2001.02228.x

Cantu VA, Salamon P, Seguritan V, et al (2020) PhANNs, a fast and accurate tool and web server to classify phage structural proteins. PLoS Computational Biology 16(11):e1007845. https://doi.org/10.1371/journal.pcbi.1007845

Charoenkwan P, Kanthawong S, Schaduangrat N, et al (2020a) PVPred-SCM: improved prediction and analysis of phage virion proteins using a scoring card method. Cells 9(2):353. https://doi.org/10.3390/cells9020353

Charoenkwan P, Nantasenamat C, Hasan MM, et al (2020b) Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. Journal of Computer-Aided Molecular Design 34(10):1105–1116. https://doi.org/10.1007/s10822-020-00323-z

Cobián Güemes AG, Youle M, Cantú VA, et al (2016) Viruses as winners in the game of life. Annual Review of Virology 3:197–214. https://doi.org/10.1146/annurev-virology-100114-054952

Elbasir A, Ye Y, Schäffer DE, et al (2023) A deep learning approach reveals unexplored landscape of viral expression in cancer. Nature Communications 14(1):785. https://doi.org/10.1038/s41467-023-36336-z

Fang Z, Zhou H (2021) VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. Frontiers in Microbiology 12:615711. https://doi.org/10.3389/fmicb.2021.615711

Fang Z, Feng T, Zhou H, et al (2022) DeePVP: Identification and classification of phage virion proteins using deep learning. GigaScience 11:giac076. https://doi.org/10.1093/gigascience/giac076

Feng PM, Ding H, Chen W, et al (2013) Naive Bayes classifier with feature selection to identify phage virion proteins. Computational and Mathematical Methods in Medicine 2013(1):530696. https://doi.org/10.1155/2013/530696

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, PMLR, pp 1050–1059

Galiez C, Magnan CN, Coste F, et al (2016) VIRALpro: a tool to identify viral capsid and tail sequences. Bioinformatics 32(9):1405–1407. https://doi.org/10.1093/bioinformatics/btv727

Ghosh S, Mitra P (2024) MaTPIP: A deep-learning architecture with eXplainable AI for sequence-driven, feature mixed protein-protein interaction prediction. Computer Methods and Programs in Biomedicine 244:107955. https://doi.org/10.1016/j.cmpb.2023.107955

Ghoshal B, Tucker A (2020) Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:2003.10769, https://doi.org/10.48550/arXiv.2003.10769

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330, URL https://proceedings.mlr.press/v70/guo17a.html

Habehh H, Gohel S (2021) Machine learning in healthcare. Current Genomics 22(4):291–300. https://doi.org/10.2174/1389202922666210705124359

Haq I, Mazhar T, Asif RN, et al (2023) A deep learning approach for the detection and counting of colon cancer cells (HT-29 cells) bunches and impurities. PeerJ Computer Science 9:e1651. https://doi.org/10.7717/peerj-cs.1651

Henn MR, Sullivan MB, Stange-Thomann N, et al (2010) Analysis of high-throughput sequencing and annotation strategies for phage genomes. PLoS ONE 5(2):e9083. https://doi.org/10.1371/journal.pone.0009083

Howard A, Sandler M, Chu G, et al (2019) Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1314–1324, https://doi.org/10.1109/ICCV.2019.00140

Hu N, Zhang T, Wu Y, et al (2022) Detecting brain lesions in suspected acute ischemic stroke with CT-based synthetic MRI using generative adversarial networks. Annals of Translational Medicine 10(2):35. https://doi.org/10.21037/atm-21-4056

Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning 110:457–506. https://doi.org/10.1007/s10994-021-05946-3

Hussain I, Nazir MB, et al (2024) Mind matters: Exploring AI, machine learning, and deep learning in neurological health. International Journal of Advanced Engineering Technologies and Innovations 1(4):209–230. URL https://ijaeti.com/index.php/

Journal/article/view/270

Jung F, Frey K, Zimmer D, et al (2023) DeepSTABp: a deep learning approach for the prediction of thermal protein stability. International Journal of Molecular Sciences 24(8):7444. https://doi.org/10.3390/ijms24087444

Kabir M, Nantasenamat C, Kanthawong S, et al (2022) Large-scale comparative review and assessment of computational methods for phage virion proteins identification. EXCLI Journal 21:11–29. https://doi.org/10.17179/excli2021-4411

Kavitha P, Ayyappan G, Jayagopal P, et al (2023) Detection for melanoma skin cancer through ACCF, BPPF, and CLF techniques with machine learning approach. BMC Bioinformatics 24(1):458. https://doi.org/10.1186/s12859-023-05584-7

Kendall A, Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems 30

Kornblith S, Shlens J, Le QV (2019) Do better ImageNet models transfer better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2661–2671, https://doi.org/10.1109/CVPR.2019.00277

Lavigne R, Ceyssens PJ, Robben J (2009) Phage proteomics: applications of mass spectrometry. In: Bacteriophages: Methods and Protocols, Volume 2 Molecular and Applied Aspects. Springer, p 239–251, https://doi.org/10.1007/978-1-60327-565-1_14

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liu Z, Mao H, Wu CY, et al (2022) A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11976–11986, https://doi.org/10.1109/CVPR52688.2022.01167

McCallin S, Sacher JC, Zheng J, et al (2019) Current state of compassionate phage therapy. Viruses 11(4):343. https://doi.org/10.3390/v11040343

Meng C, Zhang J, Ye X, et al (2020) Review and comparative analysis of machine learning-based phage virion protein identification methods. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1868(6):140406. https://doi.org/10.1016/j.bbapap.2020.140406

Menouni R, Hutinet G, Petit MA, et al (2015) Bacterial genome remodeling through bacteriophage recombination. FEMS Microbiology Letters 362(1):1–10. https://doi.org/10.1093/femsle/fnu022

Milanés-Hermosilla D, Trujillo Codorniú R, López-Baracaldo R, et al (2021) Monte Carlo dropout for uncertainty estimation and motor imagery classification. Sensors 21(21):7241. https://doi.org/10.3390/s21217241

Miotto R, Wang F, Wang S, et al (2018) Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics 19(6):1236–1246. https://doi.org/10.1093/bib/bbx044

Nami Y, Imeni N, Panahi B (2021) Application of machine learning in bacteriophage research. BMC Microbiology 21(1):193. https://doi.org/10.1186/s12866-021-02256-5

Pang S, Yu Z, Orgun MA (2017) A novel end-to-end classifier using domain transferred deep convolutional neural networks for biomedical images. Computer Methods and Programs in Biomedicine 140:283–293. https://doi.org/10.1016/j.cmpb.2016.12.019

Rawat W, Wang Z (2017) Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation 29(9):2352–2449. https://doi.org/10.1162/neco_a_00990

Russakovsky O, Deng J, Su H, et al (2015) ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115:211–252. https://doi.org/10.1007/s11263-015-0816-y

Salama WM, Aly MH (2021) Deep learning in mammography images segmentation and classification: Automated CNN approach. Alexandria Engineering Journal 60(5):4701–4709. https://doi.org/10.1016/j.aej.2021.03.048

Sánchez Fernández I, Peters JM (2023) Machine learning and deep learning in medicine and neuroimaging. Annals of the Child Neurology Society 1(2):102–122. https://doi.org/10.1002/cns3.5

Seguritan V, Alves N, Arnoult M, et al (2012) Artificial neural networks trained to detect viral and phage structural proteins. PLoS Computational Biology 8(8):e1002657. https://doi.org/10.1371/journal.pcbi.1002657

Senior AW, Evans R, Jumper J, et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577(7792):706–710. https://doi.org/10.1038/s41586-019-1923-7

Shang J, Peng C, Tang X, et al (2023) PhaVIP: Phage VIrion Protein classification based on chaos game representation and Vision Transformer. Bioinformatics 39(Supplement 1):i30–i39. https://doi.org/10.1093/bioinformatics/btad229

Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1):1929–1958

Starmer J (2021) Entropy (for data science). YouTube video, URL https://youtu.be/YtebGVx-Fxw

Stefanini M, Lovino M, Cucchiara R, et al (2023) Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. Computer Methods and Programs in Biomedicine 234:107504. https://doi.org/10.1016/j.cmpb.2023.107504

Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–9, https://doi.org/10.1109/CVPR.2015.7298594

Tan M, Le QV (2021) EfficientNetV2: Smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning, pp 10096–10106, URL https://proceedings.mlr.press/v139/tan21a.html

Tay Y, Dehghani M, Gupta J, et al (2021) Are pre-trained convolutions better than pre-trained transformers? arXiv preprint arXiv:2105.03322, https://doi.org/10.48550/arXiv.2105.03322

Theodosiou AA, Read RC (2023) Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. Journal of Infection https://doi.org/10.1016/j.jinf.2023.07.006

torchvision (2017) Source code for torchvision.models.googlenet. URL https://pytorch.org/vision/main/_modules/torchvision/models/googlenet.html#GoogLeNet_Weights, accessed 27 July 2024

UniProt Consortium (2024) UniProt.org. URL https://www.uniprot.org/, accessed 15 September 2024

Wang LF, Yu M (2004) Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. Current Drug Targets 5(1):1–15. https://doi.org/10.2174/1389450043490668

Yuan Y, Gao M (2016) Proteomic analysis of a novel Bacillus jumbo phage revealing glycoside hydrolase as structural component. Frontiers in Microbiology 7:745. https://doi.org/10.3389/fmicb.2016.00745

Zhang Y, Li Z (2023) RF_phage virion: Classification of phage virion proteins with a random forest model. Frontiers in Genetics 13:1103783. https://doi.org/10.3389/fgene.2022.1103783

Zhao J, Yan W, Yang Y (2023) DeepTP: a deep learning model for thermophilic protein prediction. International Journal of Molecular Sciences 24(3):2217. https://doi.org/10.3390/ijms24032217

Zhuang F, Qi Z, Duan K, et al (2021) A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1):43–76. https://doi.org/10.1109/JPROC.2020.3004555