Journal of Automation and Intelligence 4 (2025) 115–124

Contents lists available at ScienceDirect

Journal of Automation and Intelligence

Research article

DPCIPI: A pre-trained deep learning model for predicting cross-immunity between drifted strains of Influenza A/H3N2

Yiming Du [a], Zhuotian Li [b], Qian He [c], Thomas Wetere Tulu [c,d], Kei Hang Katie Chan [c,e], Lin Wang [f], Sen Pei [g], Zhanwei Du [h], Zhen Wang [i,j], Xiao-Ke Xu [k,l,*], Xiao Fan Liu [m,*]

[a] Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, 999077, China
[b] Division of Medical Science, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, 999077, China
[c] Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, 999077, China
[d] Beijing Key Laboratory of Topological Statistics and Applications for Complex Systems, Beijing Institute of Mathematical Sciences and Applications(BIMSA), Beijing, 101408, China
[e] Department of Electrical Engineering, City University of Hong Kong, Hong Kong, 999077, China
[f] Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, United Kingdom
[g] Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, 10027, NY, US
[h] Division of Epidemiology and Biostatistics, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, 999077, China
[i] School of Cybersecurity, Northwestern Polytechnical University, Xi'an, 710129, China
[j] School of Artifcial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, 710129, China
[k] The School of Journalism and Communication, Beijing Normal University, Beijing, 100875, China
[l] Computational Communication Research Center, Beijing Normal University, Zhuhai, 519087, China
[m] Department of Media and Communication, City University of Hong Kong, Hong Kong, 999077, China

ARTICLE INFO

Keywords:
Cross-immunity prediction
Pre-trained model
Deep learning
Influenza strains
Hemagglutination inhibition

ABSTRACT

Predicting cross-immunity between viral strains is vital for public health surveillance and vaccine development. Traditional neural network methods, such as BiLSTM, could be ineffective due to the lack of lab data for model training and the overshadowing of crucial features within sequence concatenation. The current work proposes a less data-consuming model incorporating a pre-trained gene sequence model and a mutual information inference operator. Our methodology utilizes gene alignment and deduplication algorithms to preprocess gene sequences, enhancing the model's capacity to discern and focus on distinctions among input gene pairs. The model, i.e., DNA Pretrained Cross-Immunity Protection Inference model (DPCIPI), outperforms state-of-the-art (SOTA) models in predicting hemagglutination inhibition titer from influenza viral gene sequences only. Improvement in binary cross-immunity prediction is 1.58% in F1, 2.34% in precision, 1.57% in recall, and 1.57% in Accuracy. For multilevel cross-immunity improvements, the improvement is 2.12% in F1, 3.50% in precision, 2.19% in recall, and 2.19% in Accuracy. Our study showcases the potential of pre-trained gene models to improve predictions of antigenic variation and cross-immunity. With expanding gene data and advancements in pre-trained models, this approach promises significant impacts on vaccine development and public health.

1. Introduction

Seasonal influenza infects up to a billion people worldwide annually, causing millions of severe cases and up to 650,000 deaths [1]. Vaccination before epidemics remains the most effective prevention strategy [2,3]. A critical factor in effective vaccination is cross-immunity, where immune responses to a vaccine strain provide protection against other circulating strains [4]. However, influenza viruses frequently undergo antigenic drift [5,6], enabling them to evade immune detection and limiting the cross-immunity from previous infections or vaccinations [4]. The efficacy of previous influenza vaccines to protect against drifting strains depends on the antigenic similarity

between the vaccine and epidemic strains [7], as well as the level of cross-immunity [4].

Hemagglutinin (HA), a surface glycoprotein of the influenza virus, is primarily responsible for triggering the immune response, specifically its subunit, HA1. The measurement of Hemagglutination Inhibition (HI) titer, obtained from the assay, evaluates the degree of antigenic similarity [8,9], and can also indicate the potential for cross-immunity among different strains. Traditional HI experiments involve preparing and diluting antibodies, reacting them with antigens, and testing the reactions with red blood cells [10]. Given the pairwise nature of these reactions and the annual emergence of thousands of influenza strains—such as the H3N2 subtype alone—conducting millions of HI tests is labor-intensive and time-consuming. Therefore, it is imperative to explore computational models that can assess antigenic variation and predict cross-immunity without relying solely on extensive HI testing.

Genetic alterations, including point mutations or deletions in the HA1 gene sequence, can modify antigenic epitopes and reduce vaccine efficacy [11–14]. These changes affect the ability of antibodies from one strain to neutralize others, as quantified by HI assays. Machine learning approaches, particularly neural networks, have been widely applied to various gene-related tasks. For example, LSTM-based models demonstrated the capability to forecast mutation probabilities [15], predict DNA-protein binding [16], and produce new reasonable molecules [17]. CNN-based models [18] exhibited proficiency in variant detection [19], cancer type prediction [20], and gene expression dynamic profiles classification [21]. BiLSTM-based models [22] showcased the ability to predict viral escape [23], cancer types [24], and genetic disorders [25]. However, no existing neural network models are specifically designed to predict HI titers from influenza viral gene sequences.

Moreover, applying existing models to predict HI titers presents significant challenges [26]. Firstly, these models rely heavily on large datasets [27], but the labor-intensive and time-consuming nature of HI assays [28] limits data availability, hindering the models' ability to learn complex gene patterns. Secondly, current models [29] typically concatenate the two input gene sequences, treating them as a single sequence during training. This approach fails to capture mutual information and the unique characteristics of each gene, thereby limiting the models' ability to effectively distinguish differences between reference and test viruses.

To address these challenges, we propose the DNA Pretrained Cross Immunity Protection Inference Model (DPCIPI), a novel framework designed to predict cross-immunity with enhanced accuracy and efficiency. As illustrated in Fig. 1, DPCIPI consists of four key procedures, including (1) converting influenza gene sequences into k-mers through a sequence preprocessing step, (2) encoding these k-mers using a pre-trained model-based encoding layer, (3) capturing relationships between reference and test viruses via a mutual information inference layer, and (4) inferring cross-immunity labels through a classification layer.

Our work introduces a novel approach for accurately predicting cross-immunity between reference and test viruses by integrating adaptive pre-trained embeddings and a mutual information inference layer. Specifically, we leverage DNABERT, a pre-trained model on human genes, to initialize embeddings for gene sequences converted into k-mers. For k-mers absent in DNABERT's vocabulary, we avoid random initialization by averaging their neighboring k-mer embeddings, ensuring improved representational quality. Inspired by methods in natural language processing [30], we then incorporate a mutual information inference operator that performs arithmetic operations on the embeddings of the reference and test sequences, thereby capturing subtle genetic variations and preserving mutual information. Through these combined strategies, our model—DPCIPI—surpasses current state-of-the-art methods, delivering more precise and robust predictive performance.

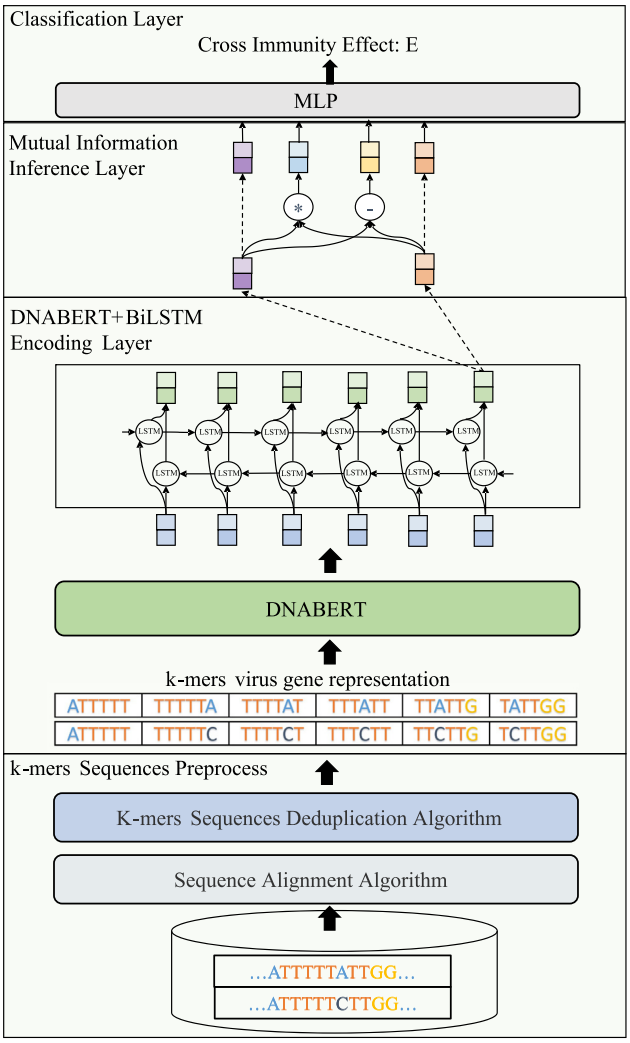


Fig. 1. The DNA Pretrained Cross-immunity Protection Inference-Model (DPCIPI) framework.

The remainder of this study is structured as follows: Section 2 reviews traditional and neural network models, pre-trained model methods, and co-expression methods for gene sequence analysis. Section 3 lays out the problem definition, data construction, and metrics. Section 4 elaborates on the proposed DPCIPI framework. Section 5 presents experimental evaluations. Section 6 concludes and outlines future research directions.

2. Related work

2.1. Traditional and neural network methods

2.1.1. Statistical learning methods

Statistical learning methods are essential in analyzing genetic sequences, particularly in elucidating the resemblances between genes to discern their interconnected functionalities. Hooper et al. [31] introduced a two-stage logistic regression approach to predict genetic structures within eukaryotic DNA. Yang et al. [32] employed a k-mer mixture logistic regression model, delineating the susceptibility of DNA methylation across diverse cell types. Nevertheless, the intricate regulatory interplays among genes often exhibit nonlinear or nonmonotonic characteristics, thereby posing a challenge for their explication through linear models. To surmount this challenge, the perceptron models have been advocated to capture the intricate regulatory relationships

inherent in genes derived from single-cell RNA-seq data [33]. In a different vein, decision trees have also been harnessed to prognosticate the multifaceted functionalities of open reading frames [34]. However, it is noteworthy that these methodologies tend to impose constrictive assumptions concerning gene expression dynamics, thereby limiting their capacity to effectively capture complex gene interactions.

2.1.2. Neural network methods

Various neural network approaches have been proposed to address the intricacies of gene sequence analysis, encompassing recurrent neural networks (RNNs) [35], Long Short-Term Memory (LSTM) networks [36], Gated Recurrent Unit (GRU) networks [37], convolutional neural networks (CNNs) [18], and Bidirectional LSTMs (BiLSTMs) [22]. Notably, the Recurrent Neural Network-based Gene Regulatory Network (RNN-GRN) [38] integrates the generalized extended Kalman model to introduce non-linear features for gene analysis. A recent approach EvoLSTM [15] leverages LSTM architecture to simulate gene sequence evolution, capturing intricate mutational context dependencies within genes. In gene variation analysis, the DeepVariant model [19], a deep CNN-based architecture, has demonstrated proficiency in calling genetic variations within aligned sequencing gene data. This model exhibits generalizability across diverse genome builds and mammalian species [19]. This work draws an analogy between mutations and word alterations in a sentence, both preserving grammaticality while altering the meaning. To address these mutations, a BiLSTM-based model [23] has been proposed to identify escape mutations. Nonetheless, it is imperative to acknowledge the inherent limitations of these methodologies, primarily due to their sensitivity to the scale of accessible data. The intricate nature of certain gene patterns continues to present challenges for comprehensive assimilation within the current framework of these approaches.

2.2. Pre-trained model methods

Pre-trained models such as BERT [39], GPT-2 [40], and RoBERTa [41] have made significant strides in the field of natural language processing, showcasing their proficiency in capturing intricate patterns from training corpora and aptly generalizing to specific tasks. These strengths have also been recognized and harnessed in genetics to create specialized models. DNABERT [42], originally designed for pre-training on human DNA sequences, shows impressive adaptability, allowing it to be used for various tasks such as gene prediction, variant calling, and sequence alignment. Particularly noteworthy is DNABERT's capacity for seamless adaptation to diverse genomes. This adaptability is highlighted in its outperformance, as observed in a comprehensive evaluation involving 78 mouse ENCODE ChIP-seq datasets. Remarkably, even when pre-trained on the human genome [43], DNABERT surpasses the efficacy of CNN, CNN + LSTM, CNN + GRU, and randomly initialized DNABERT. This robust and superior performance underscores its inherent cross-domain prowess. Expanding its utility further, DNABERT demonstrates competence in handling cross-linking and immunoprecipitation (CLIP-seq) data, thereby facilitating predictions pertaining to RNA-binding protein (RBP) binding preferences [44]. This wide-ranging applicability extends its potential usage within the viral domain.

2.3. Co-expression methods

Co-expression methods represent widely employed techniques in analyzing gene expression data, classifiable into two principal categories: correlation coefficients and mutual information measures. Among correlation coefficients, approaches such as Weighted Gene Co-expression Network Analysis (WGCNA) [45] stand out for their ability to discern potential biomarkers or therapeutic targets. Pearson correlation [46], for instance, is leveraged to distill gene features from microarray gene expression data characterized by high dimensionality and limited samples [47]. However, correlation coefficient methods encounter challenges rooted in multicollinearity, particularly when variables exhibit pronounced interdependence. This can complicate the disentanglement of individual contributions within gene analysis. Conversely, mutual information measures like Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [48] possess the capability to capture non-linear gene expressions. Yet, they grapple with issues encompassing discretization, sample size, and computational intensity during gene analysis. Notably, conventional co-expression methodologies encounter impediments when confronted with high-dimensional vector spaces. An effective strategy to infer mutual information, drawn from the realm of natural language processing, pertains to utilizing Natural Language Inference (NLI) technology [49,50]. This technology can be applied to delve into the mutual information inherent in gene representation vectors engendered by neural network models. By employing operations such as multiplication, subtraction, and preservation on sentence-level vectors, this approach captures intricate non-linear gene expressions. Given the inherent analogies shared between gene sequences and textual constructs, the transference of this methodology for the analysis of influenza genes is poised to yield insights of significance.

3. Preliminaries

3.1. Problem definition

We frame cross-immunity prediction as a machine learning classification task using HA1 gene sequences as inputs. Variations in the HA1 region can modify antigenic epitopes, reducing antibody binding and consequently lowering Hemagglutination Inhibition (HI) titers, which act as indicators of cross-immunity. By identifying genetic patterns that correlate with HI titer changes, we can predict the protective effect (E) of antibodies from a reference virus (S_R) on a test virus (S_T). Our formulation considers both a binary classification (presence or absence of E) and a multi-level classification four discrete levels of E, ultimately estimating the probability $P(E|S_R, S_T)$:

$$P(E|S_R, S_T) = M(S_R, S_T), \tag{1}$$

where M represents the machine learning model. The DPCIPI model is introduced as a specific instance within the family of models denoted by M, highlighting its role as the primary contribution of this research.

3.2. Dataset construction

In this study, we compile a revised dataset sourced from Smith et al. [13], subsequently redesignated as the Virus Hemagglutination Inhibition Dataset (VHID).[1] This compilation encompasses a total of 2472 hemagglutination inhibition (HI) titer outcomes derived from an assemblage of 240 reference viruses and 43 test viruses. We retrieved the gene sequences of the viruses from the GenBank of the NCBI database using Accession Numbers. Our dataset contains fewer viruses than reported [13] due to duplicated accessions and missing gene data. Out of the 10,320 possible reference-test virus combinations (240×43), we identified 2472 valid combinations with HI titer values, leaving 7848 samples without HI titer values.

To classify the samples into positive and negative groups, we used an HI titer threshold of 40 [51], widely recognized as corresponding to a 50% reduction in the risk of influenza [52]. As a result, the 2472 valid samples were divided into 1733 positive and 739 negative samples. For the multi-level cross-immunity prediction task, we divided the 2472 valid samples into four intervals based on their HI titer values: [0, 40), [40, 100), [100, 1000), and [1000, 10240]. These intervals contained 693, 372, 839, and 568 data examples, respectively. Each interval was

[1] https://github.com/Elvin-Yiming-Du/DPCIPI_cross-immunity_prediction/tree/main/VHID

assigned a label (0, 1, 2, or 3) representing the level of cross-immunity. In this experiment, we used the cross-entropy loss function instead of a binary classification approach. We use strains reported prior to 1995 as the training set and those after 1995 as the test set. This approach ensures that the model considers cross-immunity between historical strains and strains that may emerge in the future, rather than only focusing on cross-protection among historical strains.

4. Method

4.1. The DPCIPI model

4.1.1. k-Mers sequence preprocess

We preprocess the gene sequences in VHID to obtain representations suitable for the encoding layer, as illustrated in Fig. 2.

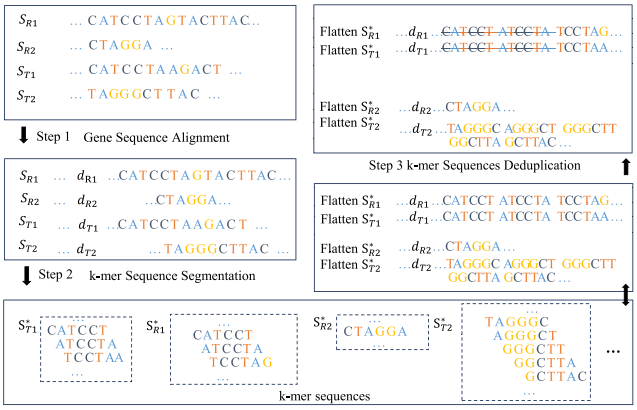


Fig. 2. A visual guide to preprocessing gene sequences: Input reference virus S_R and test virus S_T gene sequences, Step 1. Gene Sequence Alignment, Step 2. k-mer Segmentation, Step 3. k-mer Sequence Deduplication, output aligned deduplicated reference and test k-mer sequences $S_R{}^*$ and $S_T{}^*$.

Step 1. Gene Sequence Alignment. Given the high similarity among influenza virus gene sequences, we align all functional gene sequences in VHID to a common template to identify sequence differences. Using the Sequence Alignment Algorithm (Algorithm 1), we determine the starting alignment positions (D) for each sequence relative to the leftmost endpoint.

To achieve this, the algorithm first identifies a reference sequence, s_{max}, which is the most extensive typical homologous sequence. It then employs the `FindStartPosition` function to calculate the start alignment positions dictionary (D) by identifying common sites among the sequences.

Step 2. k-mer segmentation. Subsequently, we extract the reference virus gene sequence S_R, and the test virus gene sequence S_T, from VHID. These sequences are then converted into representations known as k-mers as depicted in Fig. 2. k-mers are DNA segments consisting of consecutive nucleotides, each segment having a length of k. In our practical implementation, we set k to 6, a choice validated for its high performance by [42]. The resulting k-mer sequences for the reference and test viruses are labeled as $S_R{}^*$ and $S_T{}^*$, respectively, where '$*$' signifies the k-mer format used in this paper.

Step 3. k-mer sequences deduplication. To eliminate the influence of identical locus k-mers on the prediction of cross-immunity, we design a k-mer Sequences Deduplication Algorithm (Algorithm 2) to remove duplicate k-mer from reference virus $S_R{}^*$ and test virus $S_T{}^*$ at the same locus. The algorithm consists of two main functions:

- `DeduplicationPairSequences` aligns and fills the pair sequences while identifying and recording common positions between them. It then removes the common k-mers from both sequences, resulting in modified sequences ($S^*_{R_d}$ and $S^*_{T_d}$) without duplicate k-mers, where subscript d indicates the k-mers after deduplicatoin.
- `AlignmentAndPadding` aligns and fills the sequences with placeholders, based on the start alignment positions in D. Finally, the algorithm outputs the pre-processed reference and test virus k-mer sequences.

Algorithm 1 Sequence Alignment Algorithm

Input: S (A dictionary of unique gene sequences: key is virus names, value is sequences.)

Output: D (A dictionary of start alignment positions: key is virus names, value is the distance between the farthest starting position and the current sequence starting position.)

```
Function AlignSequences(S):
    Find the longest gene sequence s_max and the corresponding length
     l from S;
    Initialize D ← {};
    for each virus, sequence in S do
        d ← FindStartPosition(sequence, s_max);
        D[virus] ← d;
    return D;
Function FindStartPosition(s.value, s_max.value):
    max_common_length ← 0;
    best_start_pos ← 0;
    s_max_length ← length of s_max;
    seq_length ← length of sequence;
    for i ← 0 to s_max_length − seq_length do
        common_length ← CalculateCommonLength(sequence,
         s_max[i :]);
        if common_length > max_common_length then
            max_common_length ← common_length;
            best_start_pos ← i;
    return best_start_pos;
Function CalculateCommonLength(sequence1, sequence2):
    length ← 0;
    for j ← 0 to length of sequence1 do
        if sequence1[j] == sequence2[j] then
            length ← length + 1;
    return length;
```

4.1.2. The DNABERT+BiLSTM encoding layer

DNABERT [42], pre-trained on human genomic data, effectively learns the contextual relationships among k-mers (subsequences of length k). Leveraging this prior knowledge, we initialize the embeddings of influenza HA1 k-mers (with $k = 6$) using DNABERT's pre-trained weights. This approach alleviates data scarcity issues and enhances the model's ability to capture meaningful genetic patterns.

We define the k-mer sequences for the reference and test viruses as:

$$S^*_{R_d} = (k^1_{R_d}, k^2_{R_d}, \ldots, k^m_{R_d}), \tag{2}$$

$$S^*_{T_d} = (k^1_{T_d}, k^2_{T_d}, \ldots, k^n_{T_d}), \tag{3}$$

where m and n denote their respective sequence lengths, d denotes the deduplicated k-mer sequences as shown in Algorithm 2.

We first encode these sequences using DNABERT:

$$X^*_{R_d} = \text{DNABERT}(S^*_{R_d}) = (x^{*1}_{R_d}, x^{*2}_{R_d}, \ldots, x^{*m}_{R_d}), \tag{4}$$

$$X^*_{T_d} = \text{DNABERT}(S^*_{T_d}) = (x^{*1}_{T_d}, x^{*2}_{T_d}, \ldots, x^{*n}_{T_d}). \tag{5}$$

Algorithm 2 k-mer Sequences Deduplication Algorithm

Input: S_R^* (Reference virus k-mer sequences), S_T^* (Test virus k-mer sequences), R_{name} (Virus name of S_R^*), T_{name} (Virus name of S_T^*), D (Start alignment position dictionary)

Output: $S_{R_d}^*$ (Modified S_R^* without common k-mers), $S_{T_d}^*$ (Modified S_T^* without common k-mers)

Function `DeduplicationPairSequences`$(S_R^*, S_T^*, R_{name}, T_{name}, D)$:

  $alignedR, alignedT \leftarrow$ `AlignmentAndPadding`$(S_R^*, S_T^*, R_{name}, T_{name}, D)$;
  $l \leftarrow \text{Min}(|alignedR|, |alignedT|)$;
  $O \leftarrow []$// Record common positions between $alignedR$ and $alignedT$
  $S_{R_d}^* \leftarrow []$, $S_{T_d}^* \leftarrow []$;
  for $i \leftarrow 0$ **to** l **do**
    if $alignedR[i] == alignedT[i]$ **then**
      O.add(i);
  $S_{R_d}^* \leftarrow$ delete $m[o]$ and # from $alignedR$ where $o \in O$;
  $S_{T_d}^* \leftarrow$ delete $n[o]$ and # from $alignedT$ where $o \in O$;
  return $S_{R_d}^*, S_{T_d}^*$;

Function `AlignmentAndPadding`$(S_R^*, S_T^*, R_{name}, T_{name}, D)$:

  $alignedR, alignedT \leftarrow []$;
  for $i \leftarrow 0$ **to** $D[R_{name}]$ **do**
    $alignedR$.add(#);
  for $j \leftarrow 0$ **to** $D[T_{name}]$ **do**
    $alignedT$.add(#);
  $kmerListR \leftarrow$ split S_R^* into k-mers;
  $kmerListT \leftarrow$ split S_T^* into k-mers;
  $alignedR \leftarrow alignedR + kmerListR$;
  $alignedT \leftarrow alignedT + kmerListT$;
  return $alignedR, alignedT$;

For any k-mer not present in DNABERT's vocabulary, we initialize its embedding by averaging its neighboring k-mers:

$$x_{\text{unknown}} = \frac{x_{\text{left}} + x_{\text{right}}}{2}. \tag{6}$$

After obtaining $X_{R_d}^*$ and $X_{T_d}^*$, we feed these sequences into a BiLSTM to capture contextual dependencies:

$$E_R^* = \text{BiLSTM}(X_{R_d}^*), \tag{7}$$

$$E_T^* = \text{BiLSTM}(X_{T_d}^*). \tag{8}$$

Here, E_R^* and E_T^* represent the encoded sequential embeddings for the reference and test virus sequences, respectively.

4.1.3. The mutual information inference layer

A **mutual information inference operator** is used to fuse the information of the two sequence-level embeddings using techniques from natural language inference [50]. In natural language inference, the task is to predict whether a hypothesis can be inferred by a premise, essentially using the inference results labeled in advance to analyze some similarity between two sentences. Similarly, we can use annotated HI titer markers to analyze whether the antibody generated by the reference virus to stimulate the immune system could cross-protect another test virus.

The mutual information inference operator extracts similar information between sequences by performing dot product operations and subtraction on vectors [53]. Dot product operations help in measuring the similarity between two vectors by computing the sum of the products of their corresponding components. The splicing vector q is obtained from the mutual information inference operator:

$$E_{combined} = [E_R^*; E_R^* \cdot E_T^*; E_R^* - E_T^*; E_R^*], \tag{9}$$

where $E_R^* \cdot E_T^*$ represents the multiplication operation, $E_R^* - E_T^*$ represents the subtraction operation in the mutual information inference operator, which have a strong enhancement effect on the salient features and difference features in the k-mer representation vector, and the original hidden mode information of the two original k-mers vectors E_R^* and E_T^* are preserved. The final interaction information will be fed into the full neural network model to predict cross-protection.

4.1.4. The classification layer

In the hemagglutination inhibition test [51], the highest dilution of hemagglutinin working fluid at which red blood cells are not completely agglutinated is used as the endpoint of determination. Since the working solution was used in twofold dilution, the experimental results showed discrete characteristics, so we used classification machine learning methods to make inferences about the similarity of cross-immunity. Put the splicing vector $E_{combined}$ obtained by the mixing layer into a multi-layer perceptron neural network (MLP) to get the classification result y':

$$P(y'|E_{combined}) = \text{MLP}(E_{combined}). \tag{10}$$

4.2. k-Mer embedding initialized with DNABERT

When working with gene sequences, there can be various k-mers that were not encountered during the pre-training of DNABERT. These unknown k-mers might correspond to specific genetic variations or rare sequences not included in the original training data. If these unknown k-mers are not appropriately initialized, it can lead to several problems. Firstly, it may result in incorrect representations, where the model assigns random or arbitrary embeddings to these k-mers, leading to inaccurate representations and difficulty in proper interpretation. Secondly, there can be a loss of valuable genetic information encoded in the unknown k-mers, hindering the model's ability to capture essential patterns and relationships in the gene sequences. Additionally, improper initialization may introduce bias in the model predictions since it lacks sufficient information about these k-mers, potentially leading to predictions that do not reflect the true characteristics of the gene sequences.

To mitigate these problems, DNABERT employs the mean value of the k-mer vector at the adjacent position for initializing unknown k-mers, providing a reasonable approximation based on the context of neighboring k-mers. This strategy enables the model to masterfully handle previously unseen k-mers, preserving the integrity of the gene sequence information. Consequently, proper initialization ensures that the model can generalize expertly and make more accurate predictions on a wide range of gene sequences, including those containing new or rare k-mers. In summary, the initialization of k-mers embedding with DNABERT is crucial in capturing hidden patterns and reducing predictive performance drops due to sparse data.

4.3. Mutual information inference operators

Drawing inspiration from natural language inference, the mutual information inference operator fuses information from two sequences by examining their similarity. Similar to how natural language inference predicts the inferability of a hypothesis from a premise, the mutual information inference operator can assess whether the antibody produced by a reference virus can cross-protect another test virus using the annotated HI titer.

The operator employs vector dot production and subtraction to extract similar information between sequences. The resulting splicing vector, denoted as $E_{combined}$, is obtained through a combination of operations: multiplication ($E_R^* \cdot E_T^*$) and subtraction ($E_R^* - E_T^*$). These operations enhance the salient features and differences in the k-mer representation vector while preserving the original hidden-mode information of the two original k-mer vectors (E_R^* and E_T^*). The interaction information derived from this process is then fed into a full neural network model to predict cross-immunity.

5. Experiments

5.1. Baseline models

Statistical methods use gene sequences (k-mers) for cross-protective predictions. Logistic Regression (LR) is for categorical classification and uses similarity scores to predict cross-immunity. The Perceptron, a classic binary classifier, employs gene sequence similarity for cross-immunity prediction. Decision Tree (DTree) classifies similarity scores for both binary and multi-label tasks. Changing from statistical learning-based similarity to vector-based similarity using gene sequence embeddings (GSE) can enhance cross-immunity prediction performance. GSE is the cumulative average of k-mer vectors in the sequence, representing sequence-level features:

$$E_R^{*\prime} = \frac{1}{n}\sum_{i=1}^{n} X_{R_d}{}^{*}, \tag{11}$$

$$E_T^{*\prime} = \frac{1}{m}\sum_{i=1}^{m} X_{T_d}{}^{*}. \tag{12}$$

Neural network methods use gene vectors derived from k-mers to predict cross-immunity. We employ classic models: 1-Nearest Neighbor (1NN) [54], Convolutional Neural Network (CNN) [55], and Bidirectional Long Short-Term Memory (BiLSTM). As a baseline model, BiLSTM directly combines two gene sequences to predict cross-immunity.

5.2. Metrics and implementation

We use Accuracy, Weighted Precision, Weighted Recall, and Weighted F1 [56] for evaluating the classifications. The weighting mechanism accounts for the varying ratio of HI titer samples under our classification task settings. All experiments were implemented on our server with 512 G memory and 2 Nvidia 3090 graphics cards. PyTorch [57], PyTorch geometric [58], and the DNABERT library are used to conduct the experiments. We train DPCIPI for 50 epochs under the settings batch size = 10 and learning rate = 0.0001.

5.3. Results

5.3.1. Binary cross-immunity prediction

Table 1 presents the binary cross-immunity prediction results obtained from statistical learning methods, neural network models, and DPICPI. The table shows that DPCIPI achieves 90. 40% in the precision metric, indicating that the model has a high degree of confidence in predicting cross-immunity. The model has achieved 1.59%, 2.34%, 1.57%, and 1.57% improvements in Weighted F1, Weighted Precision, Weighted Recall, and Accuracy, respectively, to the best-performing baseline model BiLSTM. It also achieves a performance improvement of 22. 92%, 15. 83%, and 7. 76% over statistical learning methods (logistic regression, perceptron, and decision tree) in the Weighted F1 metric, respectively.

We also found that when compared to the Eulerian distance-based similarity score calculation, gene sequence embedding (GSE) achieves the worse performance in Logistic Regression, Perceptron, and Decision tree models. It indicates that the direct use of gene sequence embedding on the similarity calculation is risky. Differently, the neural network-based models, such as DPCIPI, improved tremendously in all metrics compared to the conventional methods such as logistic regression and perceptron. Besides, we also found that using a decision tree with a max depth of 5 can achieve comparable performance to NN.

5.3.2. Multi-level cross-immunity classification

Table 1 additionally displays the results of multi-level cross-immunity prediction obtained through statistical learning methods, neural network models, and DPICPI. DPCIPI again achieves the best performance on the metric of Accuracy (64.69%), Weighted F1 (64.71%), Weighted Precision (67.25%), and Weighted Recall (64.69%) which far surpasses other models. Compared to the best-performing BiLSTM, DPCIPI achieves a 2.12%, 3.50%, 2.19%, and 2.19% improvement in the Weighted F1, Weighted Precision, Weighted Recall, and Accuracy metrics.

The confusion matrix in Fig. 3 provides a straightforward depiction of the outcome. The vertical axis signifies the actual HI titer values, while the horizontal axis denotes the predicted HI titer values. The numerical entries within the heatmap cells reflect the extent of alignment in the model's predictions, ranging on a scale from 0 to 1. The outcomes showcase a notable consistency between the predicted and actual HI titer values. Unlike LR, LR-GSE, and NN models, which have achieved poor performance (less than 30% in all metrics), the decision tree (DTree) again achieves comparable performance to the convolutional neural network model (CNN) (nearly 50%). This indicates that the cut points in decision trees perform well for multi-classification tasks.

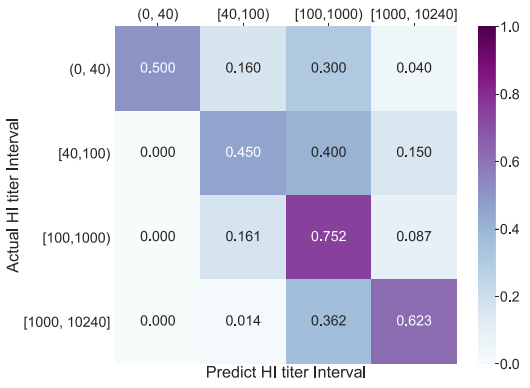


Fig. 3. The Confusion Matrix of the cross-immunity prediction on DPCIPI.

5.3.3. DNABERT embedding ablation

Pre-trained models, such as DNABERT, have shown significant progress in understanding the complex polysemy and semantic relationship between genes [59,60]. However, DNABERT was trained on the human genome. To evaluate the validity of DNABERT embeddings, we replaced the DNABERT embedding initialization with random embedding initialization in the model CNN, BiLSTM, and DPCIPI.

Table 1
Comparison of the performance across statistic learning-based models (LR, Perceptron, DTree), neural network-based models (NN, CNN, BiLSTM), and DPCIPI models. 'Improvement' indicates the relative improvement against the best baseline performance.

Task	Metric	LR	LR-GSE	Perceptron	Perceptron-GSE	DTree	DTree-GSE	NN	CNN	BiLSTM	DPCIPI	Improvement
Binary cross-immunity prediction	Weighted F1	65.22	60.82	72.31	69.39	80.38	79.08	80.35	81.88	86.56	**88.14**	**+1.58%**
	Weighted Precision	75.71	71.27	65.39	65.31	81.45	78.64	87.13	81.47	88.06	**90.40**	**+2.34%**
	Weighted Recall	61.11	56.17	80.86	74.38	79.63	79.63	84.69	82.50	88.12	**89.69**	**+1.57%**
	Accuracy	61.11	56.17	80.86	74.38	79.63	79.63	84.69	82.50	88.12	**89.69**	**+1.57%**
Multi-level cross-immunity prediction	Weighted F1	29.32	10.34	–	–	50.89	41.41	31.26	42.93	62.59	**64.71**	**+2.12%**
	Weighted Precision	14.32	18.23	–	–	51.75	51.06	24.85	50.07	63.75	**67.25**	**+3.50%**
	Weighted Recall	29.32	18.23	–	–	51.54	43.21	31.26	45.62	62.50	**64.69**	**+2.19%**
	Accuracy	29.32	18.23	–	–	51.54	43.21	31.26	45.62	62.50	**64.69**	**+2.19%**

Table 2 shows that after initialization with DNABERT, CNN, BiLSTM, and DPCIPI have each demonstrated varying degrees of performance enhancement. In particular, the CNN model exhibits the most significant improvement, achieving an increase in precision of more than 10% in the prediction of binary cross-immunity. On the contrary, in the multilevel cross-immunity prediction, the performance of CNN witnesses a notable decline following initialization with DNABERT but BiLSTM and DPCIPI demonstrate performance enhancements. Overall, the results demonstrate that DNABERT has captured hidden genetic patterns of virus strains and can make a significant impact on the prediction of the model.

5.3.4. Mutual information inference operator ablation

The incorporation of a mutual information inference (MII) operator within the model draws inspiration from the notion of a hybrid layer in natural language inference. This technique involves the fusion of two distinct word sequences to ascertain the veracity of a hypothesis whether it is true (entailment), false (contradiction), or inconclusive (neutral) in relation to a given premise. To evaluate the efficacy of the MII operator within CNN, BiLSTM, and DPCIPI models, we conducted comprehensive comparative experiments. These experiments encompassed scenarios both with and without the MII operator, spanning binary cross-immunity prediction and multi-level cross-immunity prediction tasks.

Table 3 presents the performance of the MII operator across CNN, BiLSTM, and DPCIPI for both binary and multi-level cross-immunity prediction tasks. The DPCIPI with MII achieves an improvement of 1.2%, 2.06%, 1.25%, and 1.25% in the binary classification task, and 1.57%, 1.74%, 0.63%, and 0.63% in the multi-level classification task in the metrics of Weighted F1, Weighted Precision, Weighted Recall, and Accuracy, compared to DPCIPI without MII. Furthermore, the CNN model with MII shows a significant increase (more than 10%) in all four metrics compared to CNN without MII in the binary classification task. In the multi-level classification task, CNN with MII shows a decrease in Weighted Recall score (2.51%) compared to CNN without MII. However, there is a significant increase in precision, resulting in a significant improvement in the Weighted F1 score.

5.4. Parameter sensitivity analysis

To further evaluate the robustness of DPCIPI, we examined its sensitivity to the number of training epochs and the learning rate (Fig. 4). Performance metrics improved steadily as training epochs increased but plateaued around the fifth epoch, indicating diminishing returns and potentially unnecessary computational costs beyond this point. Similarly, a learning rate of 0.0001 achieved both stability and accuracy, while higher rates converged prematurely and lower rates prolonged training without benefit. These observations highlight the necessity of prudent hyperparameter selection to ensure robust and efficient model performance.

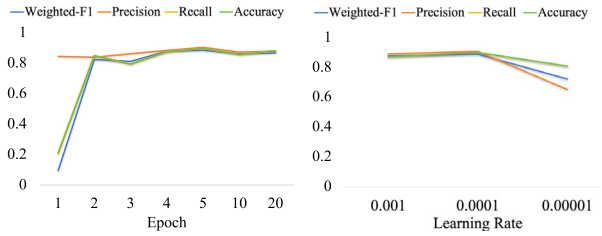


Fig. 4. Parameter sensitivity analysis of DPCIPI under different epochs and learning rates.

6. Conclusion and future work

The study introduces the DNA Pretrained Cross-Immunity Protection Inference (DPCIPI) model to predict cross-immunity between influenza virus strains using virus gene sequences. DPCIPI outperforms existing models in both binary and multi-level cross-immunity prediction tasks. In the binary task, DPCIPI shows significant improvements over BiLSTM and traditional statistical learning methods, such as logistic regression, perceptron, and decision tree. The 90.40% precision indicates a high level of confidence in DPCIPI's cross-immunity predictions. In the multilevel cross-immunity classification task, DPCIPI again achieved the highest Accuracy, Weighted F1, Weighted Precision, and Weighted Recall. Confusion matrix analysis reveals consistent

Table 2
Comparison of the performance of DNABERT Initialization across CNN, BiLSTM and DPCIPI models. 'Improvement' indicates the relative improvement against the model without DNABERT Initialization. '@' indicates a concatenation operation.

Task	Metric	CNN			BiLSTM			DPCIPI		
Initialization settings		@ -	@DNABERT	Improvement	@ -	@DNABERT	Improvement	@ -	@DNABERT	Improvement
Binary cross-immunity prediction	Weighted F1	72.41	81.88	**+9.474%**	86.56	86.94	**+0.38%**	86.97	88.14	**+1.17%**
	Weighted Precision	65.61	81.47	**+15.86%**	88.06	88.34	**+0.28%**	88.60	90.40	**+1.8%**
	Weighted Recall	80.94	82.50	**+1.56%**	88.12	88.44	**+0.32%**	89.06	89.69	**+0.63%**
	Accuracy	80.94	82.50	**+1.56%**	88.12	88.44	**+0.32%**	89.06	89.69	**+0.63%**
Multi-level cross-immunity prediction	Weighted F1	54.38	46.85	**−7.53%**	62.59	63.14	**+0.55%**	62.59	64.71	**+2.12%**
	Weighted Precision	54.65	50.07	**−4.58%**	63.75	65.61	**+1.86%**	63.75	67.25	**+3.5%**
	Weighted Recall	55.63	45.62	**−10.01%**	62.50	64.06	+1.56%	62.50	64.69	+2.19%
	Accuracy	55.63	45.62	**−10.01%**	62.50	64.06	**+1.56%**	62.50	64.69	**+2.19%**

Table 3
Comparison of the performance of MII operators across CNN, BiLSTM and DPCIPI models. 'Improvement' indicates the relative improvement against the model without MII operator. '@' indicates a concatenation operation.

Task	Metric	CNN			BiLSTM			DPCIPI		
Operator Settings		@ -	@MII	Improvement	@ -	@MII	Improvement	@ -	@MII	Improvement
Binary cross-immunity prediction	Weighted-F1	69.54	81.88	**+12.34%**	86.56	87.12	**+0.56%**	86.94	88.14	**+1.20%**
	Weighted-Precision	66.82	81.47	**+14.65%**	88.06	89.26	**+1.32%**	88.34	90.40	**+2.06%**
	Weighted-Recall	72.81	82.50	**+9.69%**	88.12	88.75	**+0.63%**	88.44	89.69	**+1.25%**
	Accuracy	72.81	82.50	**+9.69%**	88.12	88.75	**+0.63%**	88.44	89.69	**+1.25%**
Multi-level cross-immunity prediction	Weighted-F1	38.96	42.93	**+3.97%**	61.61	62.59	**+0.98%**	63.14	64.71	**+1.57%**
	Weighted-Precision	33.63	50.07	**+16.44%**	65.45	63.75	**+2.3%**	65.51	67.25	**+1.74%**
	Weighted-Recall	48.13	45.62	**−2.51%**	60.94	62.50	**+1.56%**	64.06	64.69	**+0.63%**
	Accuracy	48.13	45.62	**−2.51%**	60.94	62.50	**+1.56%**	64.06	64.69	**+0.63%**

predictions with actual types. Additionally, pre-trained models, specifically DNABERT, proved essential in capturing hidden genetic patterns. Replacing DNABERT embeddings with random embeddings led to performance drops in all tested models, highlighting the critical role of DNABERT initialization. Incorporating a mutual information inference (MII) operator consistently improved results, demonstrating its ability to enhance both binary and multi-level cross-immunity predictions.

However, the study still has some limitations: it did not involve animal experiments, relying solely on authoritative datasets (e.g., from Smith et al. and WHO vaccine recommendations). While this approach reduces costs and leverages well-validated data, it does not provide experimental verification through controlled biological studies. Another technical limitation arises from DNABERT's input length constraint, leading to fragmented representations rather than complete gene sequences.

In future work, it is a promise direction to train larger models that can handle complete gene sequences without fragmentation. We also aim to integrate additional domain knowledge, conduct controlled biological validations, and explore diverse genomic data sources. These efforts can further improve DPCIPI's interpretability, accuracy, and real-world impact, ultimately guiding vaccine composition, disease surveillance, and global public health preparedness.

CRediT authorship contribution statement

Yiming Du: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhuotian Li:** Writing – review & editing. **Qian He:** Writing – review & editing. **Thomas Wetere Tulu:** Writing – review & editing. **Kei Hang Katie Chan:** Writing – review & editing. **Lin Wang:** Writing – review & editing. **Sen Pei:** Writing – review & editing. **Zhanwei Du:** Writing – review & editing. **Zhen Wang:** Writing – review & editing. **Xiao-Ke Xu:** Writing – review & editing, Supervision. **Xiao Fan Liu:** Writing – review & editing, Supervision, Funding acquisition.

Remarks

During the preparation of this work, the author(s) used GPT4 to polish the revised paragraph and enhance its readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is part of the Grand Challenges ICODA pilot initiative, delivered by Health Data Research UK. This work has been supported by the Bill & Melinda Gates Foundation and the Minderoo Foundation.

Data availability

The data that support the findings of this study are available upon request. Interested researchers can apply for access to the data by contacting E-mail: ydu@se.cuhk.edu.hk.

References

[1] B.L. Bullard, J. DeBeauchamp, M.J. Pekarek, E. Petro-Turnquist, P. Vogel, R.J. Webby, E.A. Weaver, An epitope-optimized human H3N2 influenza vaccine induces broadly protective immunity in mice and ferrets, Npj Vaccines 1 (1) (2022) 1–14.
[2] F.M. Davenport, Current knowledge of influenza vaccine, JAMA 182 (1) (1962) 11–13.
[3] D.J. Earn, J. Dushoff, S.A. Levin, Ecology and evolution of the flu, Trends Ecol. Evol. 17 (7) (2002) 334–340.
[4] S.L. Epstein, G.E. Price, Cross-protective immunity to influenza A viruses, Expert. Rev. Vaccines 9 (11) (2010) 1325.
[5] B.B. Finlay, G. McFadden, Anti-immunology: evasion of the host immune system by bacterial and viral pathogens, Cell 124 (4) (2006) 767–782.
[6] M.T. Vossen, E.M. Westerhout, C. Söderberg-Nauclér, E.J. Wiertz, Viral immune evasion: a masterpiece of evolution, Immunogenetics 54 (8) (2002) 527–542.
[7] R.N. Thompson, C.P. Thompson, O. Pelerman, S. Gupta, U. Obolski, Increased frequency of travel in the presence of cross-immunity may act to decrease the chance of a global pandemic, Philos. Trans. R. Soc. Lond. B Biol. Sci. 374 (1775) (2019) 20180274.
[8] G.A. Kirchenbaum, G.A. Sautto, R.A. Richardson, J.W. Ecker, T.M. Ross, A competitive hemagglutination inhibition assay for dissecting functional antibody activity against influenza virus, J. Virol. 95 (23) (2021) e02379–20.
[9] C.W. Potter, J.S. Oxford, Determinants of immunity to influenza infection in man, Br. Med. Bull. 35 (1) (1979) 69–75.
[10] S. Sawant, S.A. Gurley, R.G. Overman, A. Sharak, S.V. Mudrak, T. Oguin III, G.D. Sempowski, M. Sarzotti-Kelsoe, E.B. Walter, H. Xie, et al., H3N2 influenza hemagglutination inhibition method qualification with data driven statistical methods for human clinical trials, Front. Immunol. 14 (2023) 1155880.
[11] S.M. Asaduzzaman, J. Ma, P. van den Driessche, Estimation of cross-immunity between drifted strains of influenza A/H3N2, Bull. Math. Biol. 80 (3) (2018) 657–669.
[12] A.A. Fulvini, A. Tuteja, J. Le, B.A. Pokorny, J. Silverman, D. Bucher, HA1 (Hemagglutinin) quantitation for influenza a H1N1 and H3N2 high yield reassortant vaccine candidate seed viruses by RP-UPLC, Vaccine 39 (3) (2021) 545–553.
[13] D.J. Smith, A.S. Lapedes, J.C. De Jong, T.M. Bestebroer, G.F. Rimmelzwaan, A.D. Osterhaus, R.A. Fouchier, Mapping the antigenic and genetic evolution of influenza virus, Science 305 (5682) (2004) 371–376.
[14] E.M. Nyang'au, W.D. Bulimo, V. Mobegi, S. Opanda, E. Magiri, Genetic analysis of HA1 domain of influenza A/H3N2 viruses isolated in Kenya during the 2007–2013 seasons reveal significant divergence from WHO-recommended vaccine strains, Int. J. Infect. Dis. 95 (2020) 413–420.
[15] D. Lim, M. Blanchette, EvoLSTM: context-dependent models of sequence evolution using a sequence-to-sequence LSTM, Bioinformatics 36 (Supplement_1) (2020) i353–i361.
[16] Y. Zhang, S. Qiao, S. Ji, Y. Li, DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding, Int. J. Mach. Learn. Cybern. 11 (2020) 841–851.
[17] F. Wang, X. Feng, X. Guo, L. Xu, L. Xie, S. Chang, Improving de novo molecule generation by embedding LSTM and attention mechanism in CycleGAN, 2021, Front. Genet. 12: 709500.
[18] Y. Kim, Convolutional neural networks for sentence classification, 2014, EMNLP. 1746–1751.
[19] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P.T. Afshar, et al., A universal SNP and small-indel variant caller using deep neural networks, Nat. Biotechnol. 36 (10) (2018) 983–987.
[20] M. Mostavi, Y.-C. Chiu, Y. Huang, Y. Chen, Convolutional neural network models for cancer type prediction based on gene expression, BMC Med. Genom. 13 (2020) 1–13.
[21] R. Mitra, A.L. MacLean, RVAgene: generative modeling of gene expression time series data, Bioinformatics 37 (19) (2021) 3252–3262.
[22] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015, arXiv preprint arXiv:1508.01991.
[23] B. Hie, E.D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape, Science 371 (6526) (2021) 284–288.
[24] P. Metipatil, P. Bhuvaneshwari, S.M. Basha, S. Patil, An efficient framework for predicting cancer type based on microarray gene expressions using CNN-BiLSTM technique, SN Comput. Sci. 4 (4) (2023) 381.
[25] K. Nandhini, G. Tamilpavai, An optimal stacked ResNet-BiLSTM-based accurate detection and classification of genetic disorders, Neural Process. Lett. (2023) 1–22.
[26] C. Xu, K. Shen, H. Sun, Supplementary features of BiLSTM for enhanced sequence labeling, 2023, arXiv preprint arXiv:2305.19928.
[27] W.S. Alharbi, M. Rashid, A review of deep learning applications in human genomics using next-generation sequencing data, Hum. Genom. 16 (1) (2022) 1–20.
[28] E. Spackman, I. Sitaras, Hemagglutination inhibition assay, Anim. Influ. Virus Methods Protoc. (2020) 11–28.

[29] A. Al-Ajlan, A. El Allali, CNN-MGP: convolutional neural networks for metagenomics gene prediction, Interdiscip. Sci. Comput. Life Sci. 11 (2019) 628–635.
[30] S. Wang, F. Tian, Y. Qiu, X. Liu, Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences, J. Theoret. Biol. 265 (2) (2010) 194–201.
[31] P.M. Hooper, H. Zhang, D.S. Wishart, Prediction of genetic structure in eukaryotic DNA using reference point logistic regression and sequence alignment, Bioinformatics 16 (5) (2000) 425–438.
[32] Y. Yang, A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters, in: Proc. 2nd ACM Conf. Bioinform. Comput. Biol. Biomed., 2011, pp. 366–370.
[33] Q. Luo, Y. Yu, X. Lan, SIGNET: single-cell RNA-seq-based gene regulatory network prediction using multiple-layer perceptron bagging, Brief. Bioinform. 23 (1) (2022) bbab547.
[34] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, S. Džeroski, Predicting gene function using hierarchical multi-label decision tree ensembles, BMC Bioinformatics 11 (2010) 1–14.
[35] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, 2014, arXiv preprint arXiv:1409.2329.
[36] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[37] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[38] K. Raza, M. Alam, Recurrent neural network based hybrid model for reconstructing gene regulatory network, Comput. Biol. Chem. 64 (2016) 322–334.
[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, NAACL-HLT 2019. 4171–4186.
[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[42] Y. Ji, Z. Zhou, H. Liu, R.V. Davuluri, DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome, Bioinformatics 37 (15) (2021) 2112–2120.
[43] J.A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D.M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, et al., An encyclopedia of mouse DNA elements (mouse ENCODE), Genome Biol. 13 (8) (2012) 1–5.
[44] S. Gerstberger, M. Hafner, T. Tuschl, A census of human RNA-binding proteins, Nat. Rev. Genet. 15 (12) (2014) 829–845.
[45] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinformatics 9 (1) (2008) 1–13.
[46] I. Cohen, Y. Huang, J. Chen, J. Benesty, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, 2009, pp. 1–4.
[47] Y. Zhang, C. Zhou, Gene feature selection method based on relieff and pearson correlation, in: 2021 3rd Int. Conf. Appl. Mach. Learn., ICAML, 2021, pp. 15–19.
[48] S. Barman, Y.-K. Kwon, A novel mutual information-based Boolean network inference method from time-series gene expression data, PLoS One 12 (2) (2017) e0171097.
[49] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, D. Inkpen, Enhanced LSTM for natural language inference, 2016, ACL 1657–1668.
[50] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, Z. Jin, Natural language inference by tree-based convolution and heuristic matching, 2015, ACL 130–136.
[51] L. Kaufmann, M. Syedbasha, D. Vogt, Y. Hollenstein, J. Hartmann, J.E. Linnik, A. Egli, An optimized hemagglutination inhibition (HI) assay to quantify influenza-specific antibody titers, J. Vis. Exp. (130) (2017) e55833.
[52] S. Black, U. Nicolay, T. Vesikari, M. Knuf, G. Del Giudice, G. Della Cioppa, T. Tsai, R. Clemens, R. Rappuoli, Hemagglutination inhibition antibody titers as a correlate of protection for inactivated influenza vaccines in children, Pediatr. Infect. Dis. J. 30 (12) (2011) 1081–1085.
[53] P.D. Turney, Domain and function: A dual-space model of semantic relations and compositions, J. Artificial Intelligence Res. 44 (2012) 533–585.
[54] J. Vohradsky, Neural network model of gene expression, FASEB J. 15 (3) (2001) 846–854.
[55] Y. Chen, Convolutional Neural Network for Sentence Classification (Ph.D. thesis), Univ. Waterloo, 2015.
[56] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020, arXiv preprint arXiv:2008.05756.
[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[58] M. Fey, J.E. Lenssen, Fast graph representation learning with PyTorch Geometric, 2019, arXiv preprint arXiv:1903.02428.
[59] H. Iuchi, T. Matsutani, K. Yamada, N. Iwano, S. Sumi, S. Hosoda, S. Zhao, T. Fukunaga, M. Hamada, Representation learning applications in biological sequence analysis, Comput. Struct. Biotechnol. J. 19 (2021) 3198–3208.
[60] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, Pre-trained language models in biomedical domain: A systematic survey, ACM Comput. Surv. 56 (2021) 1–52.

Yiming Du obtained his B.Sc. in Software Engineering from Southeast University in 2019 and his M.Sc. in Software Engineering from the same institution in 2022. In 2021, he joined the Department of Media and Communication at City University of Hong Kong as a Research Assistant. Subsequently, in 2022, he commenced his Ph.D. studies at the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. His current research interests primarily focus on dialogue systems and retrieval-augmented generation (RAG).

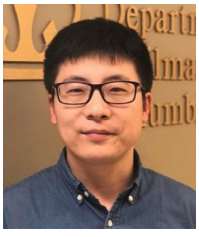
Zhuotian Li is a postgraduate student from the division of Medical Science, Faculty of medicine, The Chinese University of Hong Kong. Her current research interests focus on tumor immunotherapy and microbial pathogenesis.

Qian He holds a Bachelor's degree in Preventive Medicine from Anhui Medical University. She also obtained a Master's degree in Public Health with a specialization in Epidemiology and Biostatistics from the same university. Currently, she is pursuing her Doctor of Philosophy in Epidemiology, focusing on Genetic and Molecular Epidemiology, at City University of Hong Kong under the supervision of Prof. Kei Hang Katie Chan. Her research interests encompass Population Epidemiology and Bioinformatics.

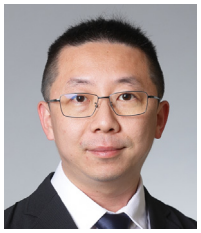
Thomas Wetere Tulu holds a Bachelor of Science in Mathematics from Mekelle University, Ethiopia, awarded in 2006, a Master of Science in Computing Science from Addis Ababa University, completed in 2011, and a Ph.D. degree in Mathematics from Harbin Institute of Technology, awarded in 2017. He is also affiliated with the Beijing Institute of Mathematical Sciences & Applications, Tsinghua University, Beijing, China. His research interests encompass a diverse array of fields, including Machine Learning, Biomathematics, Engineering Mathematics, Numerical Differential Equations, Epidemiology, Applied Analysis of PDE, Bioinformatics, Biostatistics, Applied Stochastic Methods, Scientific Computing, Image Processing, Analysis, Optimization, and High-Performance Computing (HPC).

Kei Hang Katie Chan obtained her Bachelor of Information Engineering degree from the University of Hong Kong, her Master of Public Health in Epidemiology and Biostatistics from the University of Southern California, and her Doctor of Philosophy in Epidemiology from the University of California, Los Angeles (UCLA). Her research interests include genetic and molecular epidemiology, systems biology, computational biology, and bioinformatics.

Lin Wang is a Research Associate in the Department of Genetics and an elected member of the Global Young Academy. He was previously a charge de recherche at the Institut Pasteur and a postdoctoral fellow in the School of Public Health at the University of Hong Kong. His research includes Infectious disease dynamics, computational biology, immune dynamics, serology, vaccinology, pathogen genomics, Bayesian methods, machine learning, and synthesis of heterogeneous data sources.



Sen Pei is an Assistant Professor in the Department of Environmental Health Sciences at Mailman School of Public Health, Columbia University. He received his PhD in Mathematics from Beihang University in 2015. Before joining the faculty at Columbia, he worked as a postdoctoral research scientist and associate research scientist at Columbia University. His research interests include infectious disease modeling, computational epidemiology, and data science applications in public health.



Zhanwei Du received the PhD degree in computer architecture from Jilin University, in 2015. Before joining HKU, he was a postdoc fellow with the Hong Kong Baptist University during 2015 and 2016, the University of Texas, Austin during 2016–2020, and a research associate with the University of Texas at Austin, in 2020. He is an Assistant Professor (research) with the School of Public Health, the University of Hong Kong (HKU). His general research area is computational epidemiology, machine learning, and data science.

Zhen Wang received the Ph.D. degree from Hong Kong Baptist University, Hong Kong, in 2014. He is a Professor with the School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an, China. His current research interests include complex networks, evolutionary game, and data science.

Xiao-Ke Xu received the Ph.D. degree from the College of Information and Communication Engineering, Dalian Maritime University, in 2008. He was the Post-Doctoral Fellow with Hong Kong Polytechnic University and a Visiting Scholar with the City University of Hong Kong. He is currently a Professor with the College of Information and Communication Engineering, Dalian Minzu University. His current research interests are in complex networks, big data of social networks, nonlinear time series analysis, and data mining.

Xiao Fan Liu received the B.Sc. (Hons.) and Ph.D. degrees in electronic and information engineering from the Hong Kong Polytechnic University, Hong Kong, in 2008 and 2012, respectively. He is currently an Assistant Professor with the City University of Hong Kong, Hong Kong. His research interests include cryptocurrency, online collaborative behaviors, and computational social science methods.