Heliyon 10 (2024) e32951

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A BERT-based approach for identifying anti-inflammatory peptides using sequence information

Teng Xu [a], Qian Wang [b], Zhigang Yang [a,*], Jianchao Ying [c,d,**]

[a] *Institute of Translational Medicine, Baotou Central Hospital, Baotou, China*
[b] *Department of Clinical Laboratory, Wenzhou People's Hospital, The Third Clinical Institute Affiliated to Wenzhou Medical University, Wenzhou, China*
[c] *Wenzhou Key Laboratory of Emergency, Critical Care, and Disaster Medicine, Department of Emergency, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China*
[d] *Central Laboratory, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China*

ARTICLE INFO

Keywords:
Anti-inflammatory peptide
Protein function prediction
Feature extraction
Model development
Deep learning

ABSTRACT

The use of anti-inflammatory peptides (AIPs) as an alternative therapeutic approach for inflammatory diseases holds great research significance. Due to the high cost and difficulty in identifying AIPs with experimental methods, the discovery and design of peptides by computational methods before the experimental stage have become promising technology. In this study, we present BertAIP, a bidirectional encoder representation from transformers (BERT)-based method for predicting AIPs directly from their amino acid sequence without using any other information. BertAIP implements a BERT model to extract features of a protein, and uses a fully connected feed-forward network for AIP classification. It was constructed and evaluated using the AIP datasets that were reconstructed from the latest Immune Epitope Database. The experimental results showed that BertAIP achieved an accuracy of 0.751 and a Matthews correlation coefficient of 0.451, which were higher than other commonly used methods. The results of the independent test suggested that BertAIP outperformed the existing AIP predictors. In addition, to enhance the interpretability of BertAIP, we explored and visualized the amino acids that the model considered important for AIP prediction. We believe that the BertAIP proposed herein will be a useful tool for large-scale screening and identifying novel AIPs for drug development and therapeutic research related to inflammatory diseases.

1. Introduction

Inflammation is part of the innate defense mechanism of the body against infectious or non-infectious etiologies, which is non-specific and immediate [1]. Inflammation can divide into three types, acute, subacute, and chronic, according to the time of the process that responds to the injurious cause [2,3]. Acute inflammation begins after a specific injury that will cause soluble mediators such as cytokines, acute phase proteins, and chemokines to promote the migration of neutrophils and macrophages to the area of inflammation, with the objective of removing the inflammatory stimulus or cells damaged by injury and initiate healing [4]. If this

* Corresponding author.

** Corresponding author. Wenzhou Key Laboratory of Emergency, Critical Care, and Disaster Medicine, Department of Emergency, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China.

E-mail addresses: 1971yangzhigang@sina.com (Z. Yang), yingjc@wmu.edu.cn, yingjc@spaces.ac.cn (J. Ying).

https://doi.org/10.1016/j.heliyon.2024.e32951
Received 15 May 2024; Accepted 22 May 2024
Available online 13 June 2024

inflammation does not resolve, this will cause the acute inflammation to develop from subacute to the chronic form of inflammation with the migration of T lymphocytes and plasma cells to the site of inflammation [2]. Chronic inflammation is linked to a range of diseases, such as arthritis, asthma, autoimmune diseases, diabetes, cancers, and aging [3]. Presently, nonspecific anti-inflammatory drugs and immunosuppressive therapy are the primary treatment for inflammatory conditions, but they come with challenges such as drug resistance and multiple adverse effects [5,6]. Many efforts are directed toward developing alternative and more selective anti-inflammatory therapies, several of which involved the use of bioactive peptides [7]. Endogenous peptides produced during the inflammatory responses have shown anti-inflammatory activities through the inhibition, reduction, and/or modulation of mediators' expression and activity [7,8]. The therapy based on anti-inflammatory peptides (AIPs) under normal conditions has high specificity and minimal toxicity, and may offer a new alternative therapy for inflammation treatment [5,9]. Accurately identifying AIPs is crucial due to their significance.

Currently, the identification methods of AIPs are mainly divided into wet experiment methods and computational methods. The wet experiment utilizes biochemical experiments to characterize unknown AIPs, which is complicated to design, difficult and time-consuming to operate. Consequently, this becomes challenging and inefficient in meeting the demands of large-scale batch predictions due to the significant investment of both time and resources. On the other hand, the computational method applies machine learning (ML) techniques to predict the likelihood of an AIP. ML has emerged as a robust approach to identifying critical proteins in biology [10], and its utilization is anticipated to expedite the process of AIP discovery. Several studies have developed ML tools for AIP prediction based on sequence information. Table S1 summarizes the available ML-based AIP prediction methods starting from the earliest AntiInflam [11] to the recently proposed IF-AIP [12]. There are ten methods that incorporate various feature encoding schemes and popular conventional ML models. Among them, AntiInflam [11], AIPpred [14], PreAIP [15], PEPred-Suite [16], AIEpred [17], and iAIPs [18] adopt a single algorithm. The other four methods, including PreTP-EL [19], AIPStack [20], PreTP-Stack [13] and IF-AIP [12], are constructed by integrating multiple ML algorithms. Random forest (RF) is the most popular algorithm, followed by support vector machine (SVM). There is no doubt that these ML-based predictors have made great progress in the identification of AIPs. However, these proposed methods are mainly based on conventional ML techniques for prediction, which are limited in their ability to process natural data in their raw form [21]. They need to rely on domain expertise to select a feature encoding scheme that converts the raw data into a suitable internal representation, such as sequence composition features and physicochemical properties of biological sequences [21,22]. Thus, there is a considerable opportunity to enhance the recognition performance of AIPs beyond the conventional ML techniques.

Biological sequences, such as DNA and protein sequences, share some similarities with human languages, as they can be considered textual information. Natural language processing (NLP) techniques can therefore be employed to learn useful features from this data. Rather than systematically designing and selecting feature encodings, feature descriptors can be automatically generated based on the NLP analysis [23]. The Bidirectional Encoder Representation from Transformers (BERT) [24] is a state-of-the-art language model that excels in several NLP tasks. It adopts a pre-training strategy with a self-attention mechanism as the core. Some research groups have successfully employed the BERT method to obtain useful features from biological sequences for function identification, and achieved impressive results. For instance, Le et al. proposed a transformer architecture that utilizes BERT and 2D convolutional neural network to identify DNA enhancers from sequence information [25]. In another study, Zhang et al. leveraged the pre-training strategy in the field of antibacterial peptide prediction, developing a novel method for antibacterial peptide recognition based on BERT [26]. Taju et al. used the contextualized word embeddings from BERT and the support vector machine classifier to identify efflux proteins [27]. Moreover, Charoenkwan et al. introduced a BERT-based model, BERT4Bitter, that enhances the prediction of bitter peptides solely based on their amino acid sequence [23].

Given the effectiveness and efficiency of the BERT-based method in the field of sequence identification, it is reasonable to suggest that it holds great potential for AIP recognition. Accordingly, this study aimed to explore the possibility of using the BERT method to identify potential AIPs. The experimental results indicated that the fine-tuned BERT language model can effectively distinguish AIPs from non-AIPs. As such, we developed BertAIP, the first BERT-based AIP predictor using the sequence information of peptides only. The performance of BertAIP was compared with a variety of commonly used methods and existing AIP predictors. In addition, we applied the layer attribution algorithm to interpret the reasoning process of BertAIP in predicting AIPs with several peptides as experimental subjects. It is hoped that our method will complement the established AIP identification approaches and assist in follow-up research on inflammation therapy.

2. Materials and methods

2.1. Data preparation

For the purpose of developing an effective and reliable prediction model, it is imperative to have a well-curated and unambiguous dataset. Although several AIP datasets have been proposed in previous studies, given the ever-increasing number of identified AIP sequences, we reconstructed the dataset using the latest database. As described in prior studies [14], we retrieved the experimentally validated positive and negative linear peptides or epitopes from the Immune Epitope Database (IEDB, http://www.iedb.org/home_v3.php) [28] released in September 2022. IEDB catalogs experimental data on antibody and T cell epitopes studied in humans and other animal species in the context of infectious disease, allergy, autoimmunity and transplantation. It is a real and reliable source of data that has been used by all previous ML studies of AIP prediction. A peptide was classified as AIP if it could induce any one of the anti-inflammatory cytokines (IL-4, IL-10, IL-13, IL-22, TGF-β, and IFN-α/β) in T-cell assays of human and mouse. Conversely, linear peptides that failed to test positive for anti-inflammatory cytokines were considered non-AIP. Importantly, peptides that yielded different results for the same cytokines were excluded due to ambiguity, unless they had additional unambiguous results. Additionally, we discarded the protein sequences containing blurred disabilities, such as those with amino acids 'X', 'Z', 'B', 'J', 'O', and 'U', as well as '*'. To avoid potential biases in the training process, we used the CD-HIT program [29] to eliminate protein sequences with high identity (>80 %). These procedures yielded a curated dataset containing 1759 AIPs and 3283 non-AIPs, which is comparable to the size of the datasets used in the recent studies (Table S2). It is noteworthy that employing lower thresholds of sequence identity (e.g., 30 % or lower) could potentially mitigate the bias arising from sequence homology and, in principle, yield more reliable and powerful trained models. However, given the limited size of the dataset in this study, especially AIP sequences, it was deemed imperative to use a higher threshold. In order to make a fair cross-sectional comparison of methods, the threshold used here is consistent with most previous studies [14,20]. For model training, we randomly selected 80 % of the data as the training dataset to fine-tune the model, and the rest of the data was taken as the test dataset to evaluate the performance of the model. Moreover, we obtained the training dataset derived from Gupta2017 [11] (named Gupta2017/training dataset here), which contained 690 AIPs and 1009 non-AIPs, as well as the benchmarking dataset from Manavalan2018 [14] (named Manavalan2018/benchmarking dataset here), which contained 1258 AIPs and 1887 non-AIPs (Table S2). To compare Antiinflam with our approach, we established independent dataset 1 by discarding sequences in our test dataset that shared high identity (>80 %) with Gupta2017/training dataset. Similarly, to compare our method with other existing predictors, we created independent dataset 2 by removing sequences consistent (>80 %) with Manavalan2018/benchmarking dataset from our test dataset. As a result, independent dataset 1 contained 215 AIPs and 488 non-AIPs, while independent dataset 2 contained 152 AIPs and 254 non-AIPs.

2.2. Model construction

Currently, multiple pre-trained BERT models with varying configurations have been released by the Google research team [24]. These models were pre-trained exclusively using a plain text corpus. Learning representations of language sentences can provide effective feature representations for protein sequences [27]. This study employed the BERT-base-uncased pre-trained model, which consists of 12 layers, 768 hidden units, and 12 attention heads, and requires 110 million parameters. We represented amino acids as text information by assigning each amino acid a one-word code. Since the BERT model accepts only sentences of fixed length as input, sequences containing less than 54 amino acids (the maximum sequence length) were padded at the end with the [PAD] token in order to maintain a consistent length of 54 amino acids. Besides, we marked the start and end of the sequences with [CLS] and [SEP] tokens, respectively, and separated amino acids with gaps. The BERT model was fed with the one-hot encoding of protein sequence as input, allowing each amino acid and its corresponding token to generate a contextualized word embedding vector, with a dimension of 768. Consequently, each input protein sequence was translated into a single vector comprised of 56 vectors of size 768 appended one after another. Each layer in the BERT pre-trained model acts as an encoder which takes in the output of the prior encoder layer. The first vector, which belongs to the [CLS] token, is a widely used "sentence vector" for classification tasks due to its ability to provide a summary of the other tokens via a self-attention mechanism that facilitates the intrinsic tasks of the pre-training [30]. Through fine-tuning of the model, the [CLS] token, located in the last layer, can further be optimized to capture additional semantically-relevant sentence-level context specific to the downstream task. As a result, in our study, we selected the [CLS] token's output as the feature vector, which had dimensions of 1 × 768, of the protein sequence. Following this, a dropout layer and a fully connected layer were employed to learn the information from the extracted features to classify the protein. Softmax function was used after the fully connected layer.

The BERT model was fine-tuned on the training dataset introduced in the section Data preparation. Initial parameters from the model were fine-tuned for up to ten epochs with the cross-entropy as a loss function, and the AdamW with default parameters as an optimizer. Five-fold cross-validation was carried out during model training to avoid overestimating or underestimating the real performance of the prediction model. The early stopping technique was deployed to terminate the process if the evaluation loss was no longer decreasing. The hyperparameters including the size of batch training (i.e. 8, 16, 32), and the learning rate (i.e. 1×10^{-5}, 2×10^{-5}, 5×10^{-5}) were optimized, respectively. The hyperparameter set with the highest performance on the test dataset was selected to develop BertAIP.

2.3. Feature visualization

t-distributed stochastic neighbor embedding (t-SNE) is a method developed primarily to visualize high-dimensional data by mapping them to a low-dimensional space [31]. The result is usually a clustering of similar data in the low-dimensional representation, and relations in the data can then be identified by visual inspection and comparisons with the original data [32]. For this study, we employed t-SNE to reduce the protein features obtained through the BERT model to two-dimension features, which were then visualized on a two-dimension (2D) representation. To determine the effect of the fine-tuning process on the performance of the BERT model, we also compared the results before and after the fine-tuning. R-package Rtsne was utilized to implement this process by setting dims = 2 and perplexity = 10.

2.4. Method comparison

To make a comparison between BERT and state-of-the-art features, we incorporated several widely used protein descriptors, which comprise amino acid composition (AAC), dipeptide deviation from expected mean (DDE), dipeptide composition (DPC), and tripeptide composition (TPC) [33–35]. We utilized the random forest (RF) algorithm as a classifier to train on the training dataset. Moreover, we incorporated fastText [36], TextRNN [37] and ProtBert [38] in the comparative analysis, all of which were constructed on the same dataset used for BertAIP training. We optimized the hyperparameters of fastText and TextRNN, which were both implemented using pytextclassifier [39]. ProtBert model is part of the ProtTrans collection, which provides state of the art pre-trained models for proteins [38]. The specific hyperparameters optimized are detailed in Table S3. To maintain impartiality, we used the protein sequences of the test dataset as a referee, which had not been utilized for the development of BertAIP or any of these seven models, for evaluating the performance of these models.

We examined each of the ten previously reported predictors, and found that only five of them were working properly, as detailed in Table S1. These predictors were Antiinflam [11], AIPpred [14], PreAIP [15], PreTP-EL [19], and PreTP-Stack [13], and were included in comparison with the results of this study. We followed the default settings or thresholds for each predictor for the AIP predictions except Antiinflam and PreAIP. AntiInflam has two models, named less accurate and more accurate, which used different feature encodings and presented different performances [11]. The study of PreAIP provided three different levels of thresholds that can affect its final result [15]. Due to the difference in performance reported by the authors, we considered only the best-in-class performance methods for further comparison (Table S4).

2.5. Model evaluation

After fine-tuning, BertAIP remains fixed and unaltered. The output of BertAIP is a probability score between 0 and 1, indicating the likelihood of a protein being AIP. Higher values correspond to greater prediction probabilities for AIP, and proteins with prediction probabilities exceeding 0.5 are classified as AIP. To assess the predictive performance of the predictors, we utilized several traditional measurement metrics commonly used in binary classification tasks, including accuracy, Matthews correlation coefficient (MCC), sensitivity, and specificity. They are defined as follows:

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

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

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. Sensitivity measures the prediction ability of a predictor for positive samples, while specificity measures the ability of the predictor for negative samples. Accuracy and MCC are used to evaluate the overall performance of a predictor. These metrics are not only widely used in bioinformatics research, but have also been used in previous studies on AIP prediction.

2.6. Model interpretation

Integrated Gradients [40] (IG) is an axiomatic model interpretability algorithm that can explain the relationship between a model's predictions in terms of its features. IG represents the integral of gradients with respect to inputs along the path from a given baseline to input. In this study, two IG-based layer attribution algorithms were implemented using the Captum library [41] for model interpretation. The layer conductance algorithm [42] extends IG by examining the flow of IG attribution through the hidden neuron, and was employed to interpret the BERT layers. The resulting attributions were presented as a heat map, representing the distribution of attributions across all layers and tokens/amino acids. Moreover, the layer integrated gradients algorithm was utilized to compute the attributions with respect to the embedding layer, summarizing attributions for each amino acid in the sequence. These attributions were normalized by dividing them by the maximum value, resulting in scores within the range of $[-1, 1]$. The obtained attributions were then visualized as sequence logos using Logomaker [43]. Furthermore, a statistical analysis was conducted on the peptide sequences in the test dataset, calculating the frequency of the highlighted amino acids with summarized attribution scores greater than 0.4. Sequences lacking any emphasized amino acids were excluded from the analysis.

3. Results and discussion

The objective of this study is to devise an effective model that can identify potential AIPs using the pre-trained BERT language model. We present the development of a BERT-based method, BertAIP, that employs the BERT model to extract features from protein sequences and a fully connected feed-forward network for AIP classification. The performance of BertAIP was evaluated and compared with a variety of commonly used methods and existing AIP predictors. The reasoning process of BertAIP in predicting AIPs was further interpreted by using the layer attribution algorithm. In addition, BertAIP is available as a free standalone program for scientific researchers.

3.1. Construction of a BERT model for AIP prediction

BertAIP, a BERT-based approach, was developed to predict AIPs as demonstrated in Fig. 1. It extracts latent features using the BERT layer, which is then processed by the fully connected layer to determine if the input sequence is an AIP. To train this model, we constructed a curated dataset to maximize the use of current AIP sequences, rather than relying on outdated datasets from previous studies. By dataset splitting, the training dataset comprises 1407 AIP and 2626 non-AIP sequences, while the test dataset consists of 352 AIP and 657 non-AIP sequences. The class imbalance during model training and evaluation is an issue that usually needs attention. A previous study has shown that when the imbalance ratio is greater than 5, rebalancing is needed to obtain a useable model, with accuracy as the metric [44]. The ratio of the number of non-AIP to the number of AIP is less than 2, and the quantity difference is not enough to affect the result due to the imbalance. In addition, this study used Matthews correlation coefficient (MCC) as an alternative metric, which is more robust against imbalance compared to accuracy.

Hyperparameter optimization is a crucial stage in developing most ML models, particularly deep learning models, which can help them achieve the best performance [45]. The hyperparameters that were tuned for the BERT model included the batch size and the learning rate for training. Several combinations of these hyperparameters were tried out to identify the optimal set. To evaluate this hyperparameter optimization procedure, we conducted a five-fold cross-validation, which resulted in the best performance when using a batch size of 32 and a learning rate of 5E-5 (Table 1). For the test dataset, BertAIP exhibited an accuracy and MCC of 0.751 and 0.451,

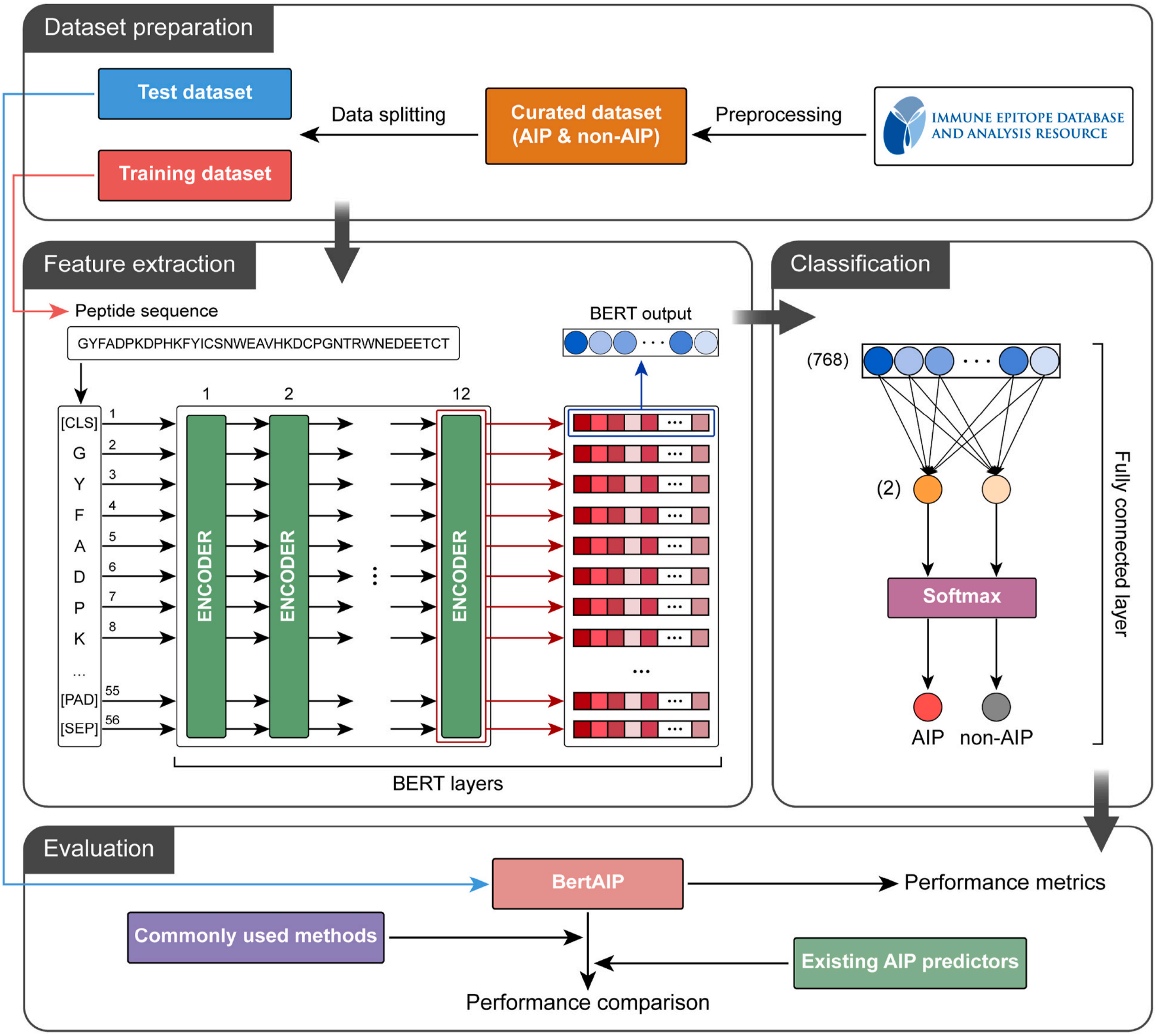

Fig. 1. The workflow chart of this study. A curated dataset consisting of AIP and non-AIP sequences is collected from the IEDB. The BERT layer and the fully connected layer extract features from sequences in textual form, resulting in the prediction of whether or not the given protein sequence is an AIP. The performance of the proposed method is compared with various commonly used methods and existing AIP predictors.

respectively (Table 2).

Furthermore, we assessed the ability of the BERT model to extract useful sequence features for AIP classification. The features extracted by the BERT model were respectively reduced to two dimensions with t-SNE algorithms for efficient visualization. Thus, the dissimilarities between AIP and non-AIP sequences in the high-dimensional space can be depicted by the proximities in the 2D space. We compared the differences between BERT models before and after fine-tuning, and observed that after fine-tuning, the distributions of AIP and non-AIP sequences become significantly separated in both training and test datasets (Fig. 2). This demonstrates that the extracted BERT features can provide valuable characteristics and patterns for AIP classification.

3.2. *Comparison between BERT and commonly used methods*

In the last few decades, a wide range of feature encoding methods or descriptors derived from protein and peptide sequence information have been suggested to predict protein function classes [35]. As shown in Table S1, many well-known descriptors have been frequently used in previous studies on AIP identification, such as AAC, DDE, DPC, and TPC. RF is the most commonly used classifier algorithm in these studies. Therefore, we re-generated these four descriptors and employed RF as the classifier, and contrasted them with the BERT method to see the difference in predictive performance. We also compared the performance of the BERT method with frequently used NLP models in the field of bioinformatics, namely fastText and TextRNN, as well as the protein pre-trained model ProtBert. The BERT method outperformed other approaches in terms of accuracy and MCC (Table 2), which suggests that the BERT-based feature can contribute to a better result. Among the well-known descriptors, DDE performed remarkably, which is in agreement with earlier studies [18]. In addition, TextRNN exhibited comparable performance to DDE. These results strongly reflect the advantages of the NLP-based approach, especially the BERT model, over the RF-based model in predicting AIPs.

3.3. *Comparison with previous works on AIP prediction*

Several ML-based tools have been developed to predict AIPs, some of which have achieved satisfactory results. To evaluate the effectiveness of BertAIP, it is necessary to compare its results with previously developed predictors. Currently, five of the AIP predictors including Antiinflam [11], AIPpred [14], PreAIP [15], PreTP-EL [19], and PreTP-Stack [13] are publicly available (Table S1). However, we cannot directly compare the result of our study with previous works due to the different datasets used for testing. Testing on the same dataset that has not been used to train any of these models would be a fair and unbiased comparison. Most of the previous studies used the datasets derived from Manavalan2018 [14], while the study of Antiinflam used an earlier dataset from Gupta2017 [11]. Therefore, we created two independent datasets to test these five AIP predictors separately, and compared their results with ours. As shown in Table 3, the accuracy and MCC of BertAIP reached 0.770 and 0.448 in independent dataset 1, and 0.719 and 0.402 in independent dataset 2, surpassing all other predictors. These results suggest that BertAIP has superior predictive performance compared to existing AIP predictors, demonstrating its significant role in the prediction of AIPs.

Among existing predictors, AIPpred produced the highest accuracy of 0.665 and an MCC of 0.398, only surpassed by BertAIP. However, Antiinflam did not perform well in the tests, which is consistent with a recent study [20]. This could be due to Antiinflam being an early-developed model that used a relatively small amount of training data, which has affected its ability to generalize previously unseen data. This finding highlights the need for dataset reconstruction in this study, and emphasizes the importance of recollecting up-to-date data in sequence identification study using ML techniques.

3.4. *Interpretation of the reasoning of BertAIP*

Due to the complexities and opaqueness of deep learning models, interpreting and understanding their predictions pose significant challenges. Therefore, in order to gain insights into how BertAIP predicts AIPs, we employed the IG-based layer attribution algorithms to explore and visualize the amino acid types that the model focuses on in the input peptide sequences, unveiling crucial information embedded in the input sequence. For this purpose, we randomly selected one AIP sequence and one non-AIP sequence as test subjects, and subsequently analyzed the distribution of attribution scores for each token/amino acid across all layers in the BERT model. As illustrated in Fig. 3A, it is apparent that the [CLS] token gained the highest attribution score in the last layer, which is consistent with its role in summarizing the other tokens. Furthermore, certain amino acids, such as K, L, and V in the tested AIP sequence and G, I, and

Table 1
Performance of BERT model according to hyperparameters based on five-fold cross-validation.

Batch size	Learning rate	Accuracy	MCC	Sensitivity	Specificity
8	1E-5	0.727	0.399	0.596	0.797
8	2E-5	0.734	0.404	0.565	0.825
8	5E-5	0.722	0.419	0.680	0.745
16	1E-5	0.727	0.404	0.620	0.784
16	2E-5	0.730	0.398	0.581	0.809
16	5E-5	0.732	0.401	0.561	0.823
32	1E-5	0.720	0.361	0.508	0.833
32	2E-5	0.725	0.405	0.632	0.775
32	5E-5	0.746	0.441	0.632	0.807

Table 2
Performance comparison between BERT and commonly used methods in AIP prediction.

Method	Accuracy	MCC	Sensitivity	Specificity
AAC	0.716	0.340	0.457	0.854
DDE	0.742	0.418	0.568	0.836
DPC	0.711	0.320	0.403	0.875
TPC	0.672	0.204	0.284	0.880
fastText	0.719	0.386	0.616	0.773
TextRNN	0.733	0.433	0.685	0.760
ProtBert	0.720	0.428	0.736	0.711
BERT	0.751	0.451	0.636	0.813

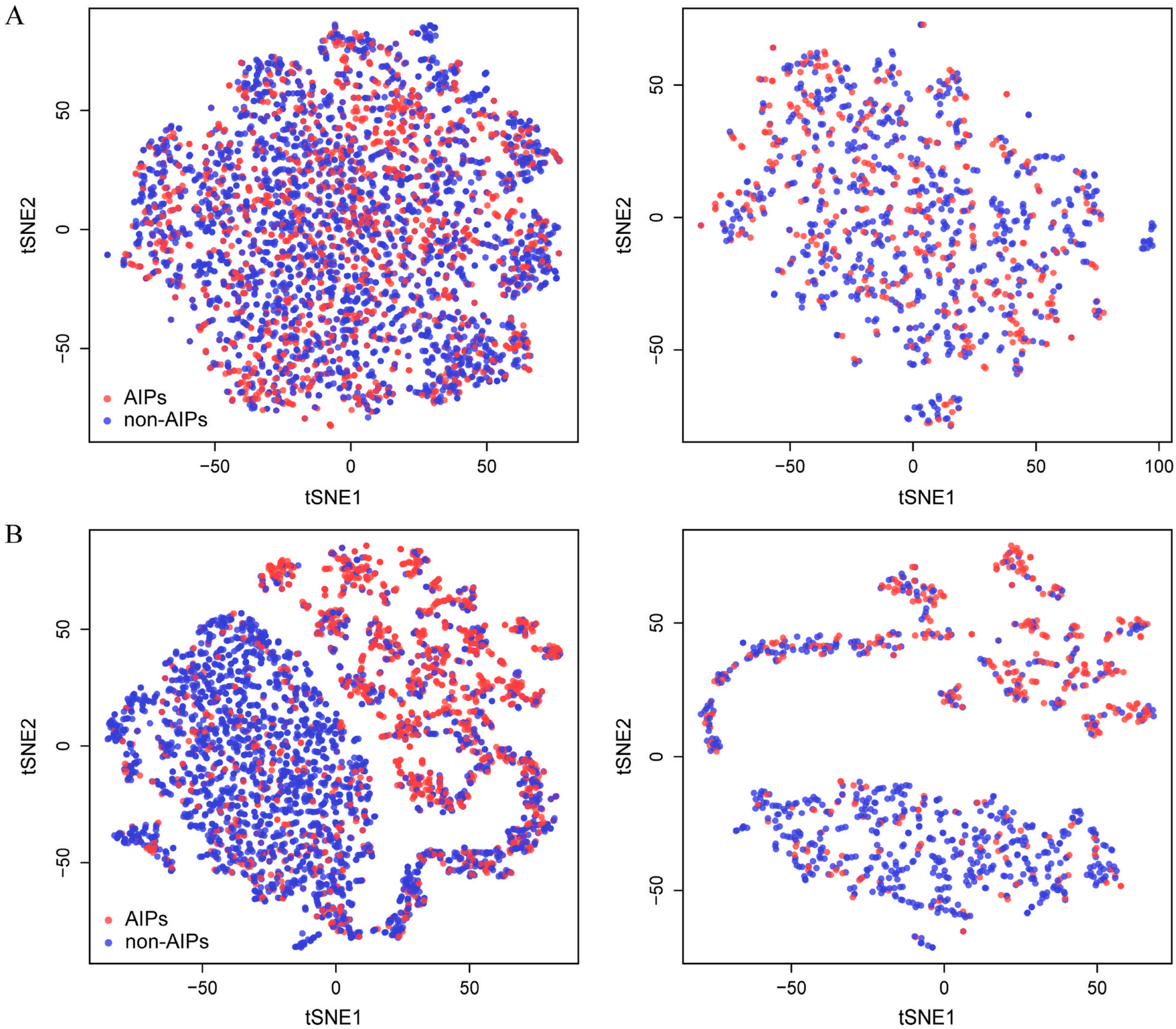

Fig. 2. t-SNE visualization of the protein sequences on the training (left) and test (right) datasets according to the features extracted by BERT models. (A) Before fine-tuning. (B) After fine-tuning. The red and blue dots represent AIPs and non-AIPs, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

R in the tested non-AIP sequence, were found to be of significant importance to BertAIP (Fig. 3A). To comprehensively understand the amino acid types that the model regarded as important in AIP prediction, we calculated the frequency of the highlighted amino acids in the sequences of the test dataset. The results revealed significant differences between AIP and non-AIP sequences (Fig. 3B). Notably, several amino acids in the AIP sequence garnered considerable attention, particularly amino acids C, K, and L which received attention

Table 3
Comparison between BertAIP and previous works using independent datasets.

Predictor	Dataset	Accuracy	MCC	Sensitivity	Specificity
Antiinflam	Independent dataset 1	0.653	0.004	0.112	0.891
BertAIP		0.770	0.448	0.591	0.848
AIPpred	Independent dataset 2	0.665	0.398	0.849	0.555
PreAIP		0.628	0.395	0.928	0.449
PreTP-EL		0.645	0.192	0.342	0.827
PreTP-Stack		0.658	0.362	0.796	0.575
BertAIP		0.719	0.402	0.632	0.772

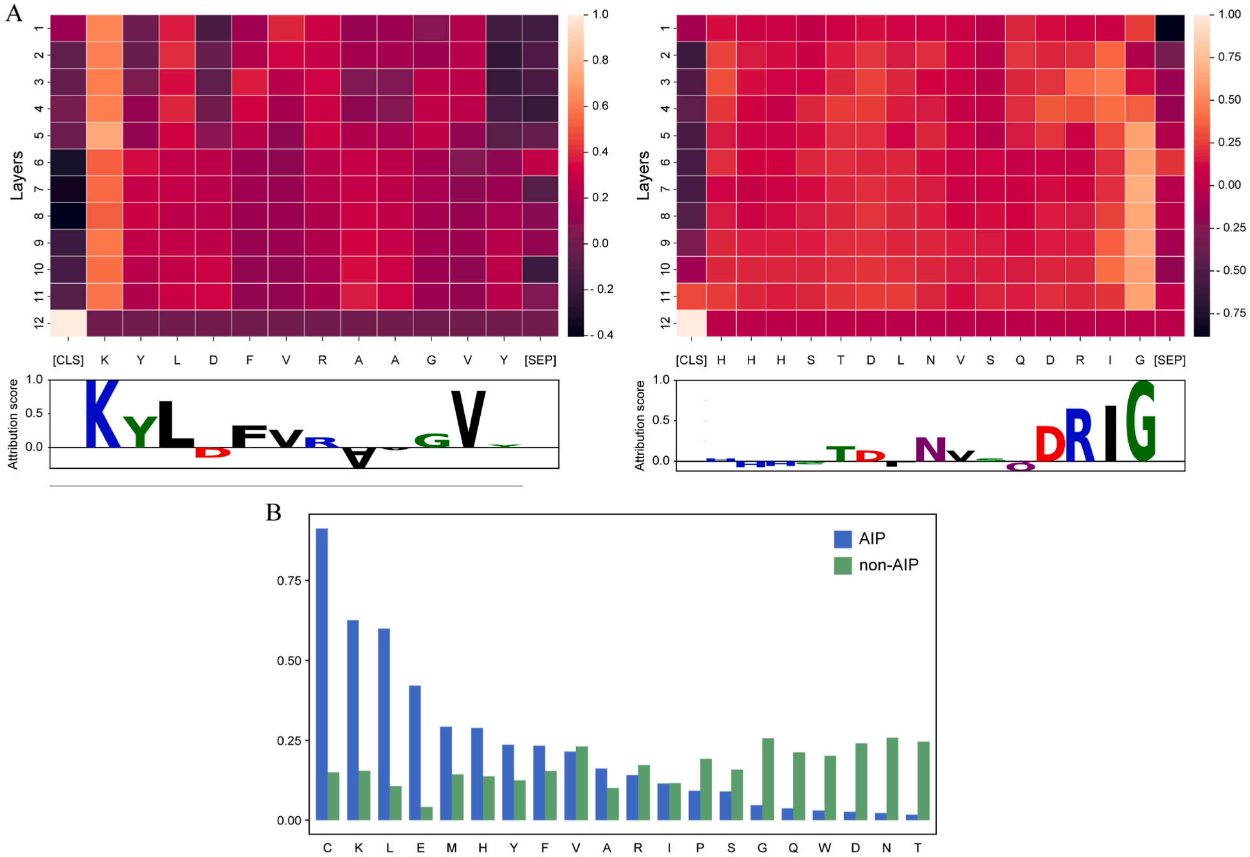

Fig. 3. Interpretation of the reasoning of BertAIP. (A) The distribution of amino acid attribution scores for two randomly selected AIP (left) and non-AIP (right) sequences. Shown are the heat map of attributions for each amino acid across all layers in the BERT model (upper), and the sequence logo of the summarized attributions for each amino acid (lower). (B) The frequency of different kinds of amino acids being concerned by the model.

frequencies exceeding 50 %. Conversely, non-AIP sequences lacked amino acids that exhibited a comparable degree of attention frequency. This underscores the ability of BertAIP to discern the characteristics of the AIPs from the sequence data without solely relying on memorizing the labels of the data. In addition, these findings suggest that the layer attribution algorithm can be utilized for interpretable analyses of our model's predictions.

4. Implementation of the BertAIP program

The exponential growth of big data, such as in genomics and proteomics, has created a pressing need for the analysis and processing of large-scale sequences. However, most of the reported tools are available in the form of online web servers, posing difficulties for processing large numbers of protein sequences and integrating them into genome analysis pipelines. These web servers may become inaccessible due to network connectivity issues or lack of maintenance. For this reason, as a first choice, BertAIP was developed as a stand-alone command-line program, which is freely available at https://github.com/ying-jc/BertAIP. This program enables scientific researchers to take any number of amino acid sequences as input, invoke the BertAIP model to make predictions, and obtain the estimated probability and binary classification of AIP for the given proteins. Additionally, we will also release a web server in the future to ensure that BertAIP is easy to use for individuals without programming or mathematical knowledge.

5. Conclusion

This study introduced a method called BertAIP, which identifies AIPs using the fully connected feed-forward network based on features extracted from the BERT model. Using the BERT pre-trained models, we are able to transfer the semantic and syntactic knowledge from the huge corpus of human language texts to biological data due to its state-of-the-art performance in a wide variety of NLP tasks as well as the similarities between genomic language and human language. In this study, the BERT contextual representation method was adopted to create vectors for each peptide sequence to capture more semantically relevant sentence-level context, making it possible to extract the meaningful hidden information from peptide sequences. The BERT method performed relatively well in comparison with methods such as DDE with RF classifier and TextRNN. And the proposed BertAIP has demonstrated its ability to predict potential AIPs from newly synthesized and discovered peptide sequences, superior to other existing AIP predictors. Given the problem of poor interpretability of deep learning models, we explored and visualized the crucial amino acids considered by BertAIP for predicting AIPs, thus demonstrating that the model has some degree of interpretability. In general, this study proposed a valuable tool for large-scale screening and identification of AIPs, which will complement the methods currently used for identifying AIPs and assist in follow-up research.

CRediT authorship contribution statement

Teng Xu: Data curation, Software, Writing – original draft. **Qian Wang:** Investigation, Visualization. **Zhigang Yang:** Writing – review & editing. **Jianchao Ying:** Formal analysis, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by grants from the Fundamental Research Funds for Wenzhou Medical University, China (Grant Number: KYYW201919) and the Natural Science Foundation of Zhejiang Province, China (Grant Number: LQ20H150004).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32951.

References

[1] L. Ferrero-Miliani, O.H. Nielsen, P.S. Andersen, S.E. Girardin, Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1beta generation, Clin. Exp. Immunol. 147 (2007) 227–235.
[2] S. Hannoodee, D.N. Nasuruddin, Acute Inflammatory Response, StatPearls, Treasure Island (FL), 2023.
[3] R. Pahwa, A. Goyal, I. Jialal, Chronic Inflammation, StatPearls, Treasure Island (FL), 2023.
[4] D.R. Germolec, K.A. Shipkowski, R.P. Frawley, E. Evans, Markers of inflammation, Methods Mol. Biol. 1803 (2018) 57–79.
[5] B.C. Wu, A.H. Lee, R.E.W. Hancock, Mechanisms of the innate defense regulator peptide-1002 anti-inflammatory activity in a sterile inflammation mouse model, J. Immunol. 199 (2017) 3592–3603.
[6] K. Dendoncker, C. Libert, Glucocorticoid resistance as a major drive in sepsis pathology, Cytokine Growth Factor Rev. 35 (2017) 85–96.
[7] S. La Manna, C. Di Natale, D. Florio, D. Marasco, Peptides as therapeutic agents for inflammatory-related diseases, Int. J. Mol. Sci. 19 (2018).
[8] E. Gonzalez-Rey, P. Anderson, M. Delgado, Emerging roles of vasoactive intestinal peptide: a new approach for autoimmune therapy, Ann. Rheum. Dis. 66 (Suppl 3) (2007) iii70–76.
[9] C. de la Fuente-Nunez, O.N. Silva, T.K. Lu, O.L. Franco, Antimicrobial peptides: role in human disease and potential as immunotherapies, Pharmacol. Ther. 178 (2017) 132–140.
[10] A.L. Tarca, V.J. Carey, X.W. Chen, R. Romero, S. Draghici, Machine learning and its applications to biology, PLoS Comput. Biol. 3 (2007) e116.
[11] S. Gupta, A.K. Sharma, V. Shastri, M.K. Madhu, V.K. Sharma, Prediction of anti-inflammatory proteins/peptides: an insilico approach, J. Transl. Med. 15 (2017) 7.
[12] S. Gaffar, M.T. Hassan, H. Tayara, K.T. Chong, IF-AIP: a machine learning method for the identification of anti-inflammatory peptides using multi-feature fusion strategy, Comput. Biol. Med. 168 (2024).
[13] K. Yan, H. Lv, J. Wen, Y. Guo, Y. Xu, B. Liu, PreTP-stack: prediction of therapeutic peptides based on the stacked ensemble learing, IEEE ACM Trans. Comput. Biol. Bioinf 20 (2023) 1337–1344.
[14] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[15] M.S. Khatun, M.M. Hasan, H. Kurata, PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features, Front. Genet. 10 (2019) 129.
[16] L.Y. Wei, C. Zhou, R. Su, Q. Zou, PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning, Bioinformatics 35 (2019) 4272–4280.
[17] J.H. Zhang, Z.H. Zhang, L.R. Pu, J.J. Tang, F. Guo, AIEpred: an ensemble predictive model of classifier chain to identify anti-inflammatory peptides, Ieee Acm T Comput Bi 18 (2021) 1831–1840.

[18] D. Zhao, Z. Teng, Y. Li, D. Chen, iAIPs: identifying anti-inflammatory peptides using random forest, Front. Genet. 12 (2021) 773202.

[19] Y. Guo, K. Yan, H. Lv, B. Liu, PreTP-EL: prediction of therapeutic peptides based on ensemble learning, Briefings Bioinf. 22 (2021).

[20] H. Deng, C. Lou, Z. Wu, W. Li, G. Liu, Y. Tang, Prediction of anti-inflammatory peptides by a sequence-based stacking ensemble model named AIPStack, iScience 25 (2022) 104967.

[21] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[22] Q. Wang, T. Xu, K. Xu, Z. Lu, J. Ying, Prediction of transport proteins from sequence information with the deep learning approach, Comput. Biol. Med. 160 (2023) 106974.

[23] P. Charoenkwan, C. Nantasenamat, M.M. Hasan, B. Manavalan, W. Shoombuatong, BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides, Bioinformatics 37 (2021) 2556–2562.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[25] N.Q.K. Le, Q.T. Ho, T.T. Nguyen, Y.Y. Ou, A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information, Briefings Bioinf. 22 (2021).

[26] Y. Zhang, J. Lin, L. Zhao, X. Zeng, X. Liu, A novel antibacterial peptide recognition algorithm based on BERT, Briefings Bioinf. (2021) 22.

[27] S.W. Taju, S.M.A. Shah, Y.Y. Ou, Identification of efflux proteins based on contextual representations with deep bidirectional transformer encoders, Anal. Biochem. 633 (2021) 114416.

[28] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D.K. Wheeler, A. Sette, B. Peters, The Immune epitope Database (IEDB): 2018 update, Nucleic Acids Res. 47 (2019) D339–D343.

[29] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[30] H. Choi, J. Kim, S. Joe, Y. Gwon, Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5482–5487.

[31] L.v.d. Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[32] M. Svantesson, H. Olausson, A. Eklund, M. Thordstein, Get a new perspective on EEG: convolutional neural network encoders for parametric t-SNE, Brain Sci. 13 (2023).

[33] M. Bhasin, G.P. Raghava, Classification of nuclear receptors based on amino acid composition and dipeptide composition, J. Biol. Chem. 279 (2004) 23262–23266.

[34] V. Saravanan, N. Gautham, Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor, OMICS A J. Integr. Biol. 19 (2015) 648–658.

[35] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J. Song, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.

[36] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 427–431.

[37] Z. Wang, B. Yang, Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT, in: 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2020, pp. 562–568.

[38] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 7112–7127.

[39] M. Xu, Pytextclassifier: Text classifier toolkit for NLP, https://github.com/shibing624/pytextclassifier, 2022.

[40] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks, 2017.

[41] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A Unified and Generic Model Interpretability Library for PyTorch, 2020.

[42] K. Dhamdhere, M. Sundararajan, Q. Yan, How important is a neuron?, abs/1805, CoRR (2018) 12233.

[43] A. Tareen, J.B. Kinney, Logomaker: beautiful sequence logos in Python, Bioinformatics 36 (2020) 2272–2274.

[44] F.M. Megahed, Y.J. Chen, A. Megahed, Y. Ong, N. Altman, M. Krzywinski, The class imbalance problem, Nat. Methods 18 (2021) 1270–1272.

[45] N.Q.K. Le, T.T. Huynh, E.K.Y. Yapp, H.Y. Yeh, Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles, Comput. Methods Progr. Biomed. 177 (2019) 81–88.