A novel attention-based peptide language model for dipeptidyl peptidase-IV inhibitory peptides mining

Changge Guan[1,2*†], Jiawei Luo[2†], Shucheng Li[1†], Zheng Lin Tan[3], Jiahao Li[2], Zourun Wu[2], Yi Wang[1], Haihong Chen[4,5], Naoyuki Yamamoto[3], Chong Zhang[1,6], Junjie Chen[2*] and Xin-Hui Xing[1,4,5,6*]

[1]Key Laboratory for Industrial Biocatalysis, Ministry of Education of China, Department of Chemical Engineering, Tsinghua University, Beijing, China.
[2]Intelligent Computing Research Center, Harbin Institute of Technology, Shenzhen, China.
[3]School of Life Science and Technology, Tokyo Institute of Technology, Yokohama, Japan.
[4]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, China.
[5]Institute of Biomedical Health Technology and Engineering, Shenzhen Bay Laboratory, Shenzhen, China
[6]Center for Synthetic and Systems Biology, Tsinghua University, Beijing, China.

*Corresponding author(s). E-mail(s): xhxing@mail.tsinghua.edu.cn; junjiechen@hit.edu.cn; gcgtsinghua@gmail.com;
Contributing authors: luojiawei@stu.hit.edu.cn; lisc21@mails.tsinghua.edu.cn; william tan94@hotmail.com; lijiahao@stu.hit.edu.cn; zrwu@bliulab.net; wangyi671@tsinghua.edu.cn; n-yamamoto@bio.titech.ac.jp ; chongzhang@mail.thu.edu.cn;
†These authors contributed equally to this work.

Abstract

Dipeptidyl peptidase IV (DPP-IV) inhibitory peptides (DPP-IV-IP) are novel anti-diabetic drugs. However, limited number of DPP-IV-IP have been discovered due to lack of efficient tools for peptides mining. Thus, it is important to efficiently identify DPP-IV-IP from large-scale peptide dataset. In this study, we propose a peptide language model (PLM) to identify DPP-IV-IP and achieve state-of-the-art of accuracy (Acc) of 0.894. Visualization of model's attention has shown that the model can automatically learn information about physicochemical and structural properties from peptide sequences, i.e., DPP-IV cleavage sites, distinguish and represent amino acid sequences in high-dimensional spaces. For the first time, we demostrated that PLM could capture cleavage site information, and demonstrated the capability of PLM to guide biological experimental screening, which has not been reported. Results obtained from biological assays suggested that prediction accuracy of the model was 90% based on PLM-assisted DPP-IV inhibitory peptide screening. To explore diversity of DPP-IV-IP, we proposed a strategy based on the dipeptide repeat unit X-proline and showed that the strategy is feasible through modeling and biological experiments.

Keywords: DPP-IV inhibitory peptide, peptide language model, DPP-IV cleavage site, dipeptide repeat unit

1 Introduction

Dipeptidyl peptidase IV (DPP-IV) is a novel therapeutic target for anti-diabetic drugs[1], which the number of patients is estimated to be 537 million people and it has contributed to mortality of over 6.7 million people within age 20–79 in 2021 alone[2, 3]. Diabetes is a chronic disease and one of the fastest-growing global health emergencies of the 21st century (International Diabetes Federation Diabetes Atlas 10th edition). Therefore, it is crucial to develop new DPP-IV inhibitory drugs and functional foods.

DPP-IV (E.C.3.4.14.5) is a serine-type protease[4, 5], which preferentially cleaves substrate peptides at X-proline or X-alanine dipeptides from the N-terminus of polypeptides[6]. It involves in glucose metabolism by degrading incretin hormones, e.g., glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) that can be released into blood circulation within minutes, in response to glucose or fat uptake to stimulate insulin secretion and lower blood glucose levels[7, 8]. Inhibition of DPP-IV enzyme activity helps with glycemic control and weight reduction[1, 7, 9]. Thus, DPP-IV inhibitors represents a promising class of diabetes drugs for improving glucose metabolism in T2D patients[9]. To date, several DPP-IV inhibitors have been approved by the FDA for the treatment of diabetes [10, 11]. However, these chemically synthesized small molecule drugs result in various adverse effect in patients, e.g., headache, cystitis, and nasopharyngitis[12].

Fortunately, DPP-IV-IP derived from products of enzymatic hydrolyzed animal, plant, and macroalgal proteins exhibit DPP-IV inhibitory activity without inducing adverse effect in patients[13]. In addition, traditional enzymatic hydrolysate screening is still the main method used in this field, which has limited the efficiency of DPP-IV-IP screening. Synthetic peptide library screening methods have not been employed in DPP-IV-IP mining due to the lack of peptide library design strategies and efficient screening methods for DPP-IV inhibitory polypeptides[14]. Low screening efficiency has resulted in limited number of DPP-IV-IP discovered, which in turn limited the progress in development of DPP-IV-IP products.

However, conventional procedures in DPP-IV-IP screening are laborious, time consuming and costly (due to requirement of various type of column, large amount of materials and diverse variation of synthetic peptides and enzymes)[15, 16]. Furthermore, multiple steps of separation could lead to undesired peptide degradation and loss of peptide activities. So far, only 500 kinds of DPP-IV-IPs have been discovered. However, the DPP-IV inhibitory activities of the peptides discovered were too low for practical use. Therefore, it is an important task to develop a high-throughput method capable of rapid identification of effective anti-diabetic bioactive peptides that can be used in functional foods or medicine innovation[13, 15, 17]. Over the past decade, several computational approaches have been used to identify and extract novel peptide sequences from dietary proteins. Quantitative structureactivity relationship (QSAR) method has been developed to aid discovery of novel DPP-IV-IPs while elucidating in-depth understanding of DPP-IV inhibitory potency of these peptides[18–21]. QSAR generally encompasses the analysis of multivariate peptide features related to their biological activities with machine learning (ML) techniques[20, 22–25].

Although some progress were obtained with these models, predictive performance of these models is still limited. To improve predictive ability of model, several traditional ML-based predictors, e.g., support vector machine (SVM) and random forest (RF), have been established and employed for identification of DPP-IV-IPs[26–28]. However, the methods reported to date have the following major limitations: (1) performance of these models are heavily dependent on the quality of the features extracted by feature engineering, (2) they have poor representation ability for amino acid sequences, (3) they fail to capture information hidden in the amino acid sequence itself, render these models inefficient. Deep neural networks (DNNs) can overcome these limitations and their advanced architecture provides them with the capability to automate feature learning to extract discriminating feature representations with minimal human efforts[29]. In particular, deep learning methods, e.g., long-short term memory (LSTM), based on natural language processing (NLP) have been shown to be promising method in identifying functional peptides for extraction of discriminating feature representation, such as anti-microbial peptides[30–33].

However, deep learning (DL)-based models have been criticized for its uninterpretable and unexplainable characteristics, known as 'black-box'. Furthermore, classifiers for DPP-IV-IPs developed to date were not verified experimentally.

In this article, we proposed a novel attention-based peptide language model (PLM)-based DL model which combines the attention mechanism and pre-trained NLP Bidirectional Encoder Representations (BERT) model to identify DPP-IV-IPs (**Fig. 1**). Introducing attention mechanisms in DNNs provides a novel approach to neural network interpretability, and pre-trained NLP BERT models can be trained on large amounts of unlabeled protein sequence data to learn sequence representations. By leveraging these modules, our proposed model which has been fine-tuned based on a pre-trained BERT model could automatically learned the characteristics and features contained in the original sequences, distinguish and represent amino acid sequences in high-dimensional spaces and have a good interpretability, which is more advantageous than existing machine learning algorithms for identification of DPP-IV-IPs. Extensive analyses showed that our model can identify cleavage sites of DPP-IV enzyme on the polypeptides as substrates by performing a vector representation learning of the DPP-IV substrate polypeptide sequences.

In contrast to previous reports which the models were only validated on test set[26–28], we have conducted wet-laboratory experiment to validate our model, thus extending the computational model to practical application. By combining both PLM model and biological assays, we have proposed a novel DPP-IV inhibitory pentapeptides design strategy based on dipeptide repeat unit X-proline to provides a new idea for the design of DPP-IV-IPs. This design strategy has addressed the bottleneck in DPP-IV-IPs discovery due to lack of design strategies[14]. We define this tool for DPP-IV-IP screening as BERT-DPPIV.

2 Results

2.1 Construction and comparation of BERT-DPPIV for DPP-IV inhibitory peptides identification

In the NLP models, the pre-trained models have many advantages in contrast to direct training of model to complete target task. They can leverage the large unlabeled datasets to learn universal representations of data, which are beneficial for various downstream tasks. Therefore, we have adapted pre-train strategy in constructing our model. Additionally, for model interpretability, the self-attention-based deep bidirectional language representation (BERT) model architecture was used as the initial model for pre-training. The protein sequences from UniProt database integrating the data of SWISS-PROT, TrEMBL and pir-psd are used as dataset to pre-train the initial BERT model.

After pre-training, a classification layer was added to the structure of the pre-trained BERT model on its output layer. Then, peptide dataset of DPP-IV-IPs was leveraged to fine-tune pre-trained BERT model to construct peptide language model (PLM). 3 models were constructed by 3 kinds of subword

tokenizer including kmer=1, kmer=2, and kmer=3. Then, the test set was used to validate the 3 models. Highest Acc of 0.891 has been achieved for kmer=1, followed by 0.887 for kmer=2 and 0.842 for kmer=3 (**Table. 1**). In order to exploit the potentials of all three models, we combine the three models to construct a new model BERT-DPPIV. Acc has been improved to 0.894 for BERT-DPPIV (**Table. 1**).

To compare our model with previously reported DPP-IV discriminant models, i.e., iDPPIV-SCM[26], SVM[27] and StackDPPIV[28], the training set (with tuning set) provided by these models were used to fine-tune the pre-trained model and the corresponding testing set was used to test the resulting model. **Table. 1** summarizes the performance indicators of the models trained on the test set according to the different word segmentation methods. The first column is the name of the model to be evaluated and the second to sixth columns are values of Acc, Sn, Sp, Matthews correlation coefficient (MCC) and area under receiver operating characteristics curve (AUC), respectively.

Our models (kmer=1, kmer=2, and kmer=3) outperform the first sequence-based iDPPIV-SCM model with increases Acc and MCC by about 0.094 and 0.110, 0.090 and 0.114, and 0.045 and 0.094 respectively (**Table. 1**). These results suggest that our models can extract more information from peptide sequences, e.g., sequence features, physicochemical properties, structure information, etc. To further investigate the characteristics of our models, we have compared our models to sequence-based physicochemical property SVM model and the sequence- and structure-based StackDPPIV model (**Table. 1**). The result shows that our models (kmer=1, kmer=2) are better than SVM model with increasing Acc by 0.026 and 0.022 respectively. Our model 1 achieves the same accuracy with an Acc of 0.891 as the best model StackDPPIV and the model (kmer=2) is slightly lower than it with an Acc of 0.887. The best model was BERT-DPPIV combining 3 models with Acc 0.894. Based on these results, we speculated that the PLM-based BERT model with kmer=1 and kmer=2 can capture sequence features, physicochemical properties and structural features. However, it remained unknown what kinds of features from the peptide sequences were captured by our models.

2.2 Sequence representation and its properties learned by model

To investigate the information captured by our models, we kmer=1 for further study. First, we analyze the representation of the peptide sequences learned by peptide language model, as prior study has shown that NLP model can effectively learn the representation of the sequence[34]. The vector representation of the peptide sequences is extracted from model and analyzed by t-distributed stochastic neighbor embedding (t-SNE) during the training process of the model, then the result of analysis was visualized (**Supplementary Fig. 1**). **Supplementary Fig. 1** shows that the model can identify the DPP-IV-IPs by learning the vector representation of peptides.

Furthermore, we have analyzed the model with t-SNE to identify amino acid vector representation learned by the model, thus, illustrate the capability of the model to learn physicochemical properties of peptides. The vector representation learned by the model can well distinguish amino acids with different properties (**Fig. 2a**). Then, 10 kinds of physicochemical properties information from peptide sequences were extracted with modlAMP package[35], and extraction of the vector representation of them and t-SNE analysis were conducted. The results showed that the vector representation learned by the model can also distinguish sequences with different physical and chemical properties (**Fig. 2b,c**, **Supplementary Fig. 2**). To this point, we have shown that PLM can learn physicochemical properties of the peptide sequences.

2.3 Capturing the structure information and attention patterns by model

The attention mechanism provides a way to parse the black-box deep learning model and features learned [36, 37]. Next, we investigate the capability of our model to learn structural information of peptide sequences with the help of an attention mechanism, which is the interface connecting encoder and decoder. Firstly, the structure of the peptides with a length ≥ 5 from positive samples was predicted by APPTEST software[38], and 12 kinds of peptides with α-helix secondary structure were found (**Supplementary Fig. 3**). 3 peptides were randomly selected for further structural confirmation using AlphaFold2 and the presence of α-helix structures was confirmed[39]. Then, the attention of the model of each head at each layer of model was visualized (12 layers, 12 heads per layer) (**Supplementary Fig. 4**). As shown in **Fig. 1**, **Fig. 2d** and **Supplementary Fig. 5**, we can observe that attention of sites focus on their respective interacting site, suggesting that this model can capture the structural information of peptide sequences. Furthermore, analysis of 144 attention mechanism (**Supplementary Fig. 4**) showed that several learning mechanisms existed in the model, which include previous-word attention pattern(**Fig. 3a**), next-word attention patterns(**Fig. 3b**), delimiter-focused attention patterns(**Fig. 3c**), specific-word attention patterns(**Fig. 3d,e**), and related-word attention patterns (**Fig. 3f**).

2.4 Attentions of BERT-DPPIV divert to DPP-IV cleavage sites and proline

In most cases, the attentions of BERT-DPPIV focus on 2nd and 3rd positions of the N-terminus of the polypeptide sequence (**Fig. 3d-g**), which are the cleavage site of DPP-IV. This result suggested that our model can capture information about DPP-IV cleavage sites by learning the sequence of polypeptide substrate.

To further illustrate the capability of cleavage site prediction, we statistically analyzed the attention sites of the sequences of training set(**Fig. 4a,b**). The result suggested that the attention of our model focuses on the cleavage

sites of DPP-IV. Furthermore, comparison of the changes in attention position of model before and after fine-tuning have supported that our model could capture cleavage site information of DPP-IV.

We have further investigated the capability of our model to analyze the importance of amino acids in the sequences. Statistical analyses have suggested that proline is the amino acid of major concern in our model, while cystine is the least concerned. Slight increase in the attention toward proline, and slight decrease in attention toward cystine were observed (**Fig. 4c,d**). To investigate the importance of proline and cystine in the amino acid sequences, we have visualized the attention of our model based on each amino acid for sequence "QNPHSSNICN". Difference is observed in all amino acids after training, which suggested that the model has directed its attention toward important amino acids, e.g., proline, while less attention toward less important amino acid, e.g., cystine. (**Fig. 4e-h**). The result suggested that the model has been trained to direct its attention to proline which is an important amino acids in DPP-IV inhibition. To illustrate the accuracy of this result, we have statistically analyzed the proportion of our model's attention to each of the 20 standard amino acids (**Fig. 5**, **Supplementary Table. 1, Supplementary Fig. 6**). These results demonstrated that our model can efficiently capture the amino acid types that play important roles in the amino acid sequences of a peptide.

The above-discussed results have suggested that our model can learn biologically meaningful information from peptide sequences such as physicochemical properties, structure information, and cleavage site information, and has better interpretability. Model visualization allows us to study what our model can learn and illustrate that our model is helpful for understanding the biological function of polypeptides and DPP-IV.

2.5 Application of BERT-DPPIV in DPP-IV inhibitory tripeptides screening

As a proof of concept, we have applied our model in the screening of DPP-IV inhibitory tripeptides. Prior study conducted by one of the authors has classified DPP-IV inhibitory dipeptides into 5 classes, and from one class of the peptides, successfully screened from VPX and IPX (X represents one of the 20 kinds of amino acids), 9 kinds of tripeptides with efficient human DPP-IV inhibitory activity[40]. However, whether other kinds of tripeptides can efficiently inhibit human DPP-IV remained unknown. In this study, classes including WPX, WAX, WRX, and WVX from five classes were selected for screening.

We have predicted the DPP-IV inhibitory activity of 80 kinds of tripeptides with our models and found that these tripeptides possess DPP-IV inhibitory activity (**Supplementary Table. 2**). Then, the inhibitory activity of 80 chemically synthesized tripeptides were verified experimentally. IC_{50} were detected for the 72 predicted tripeptides, and the DPP-IV inhibitory activity of these tripeptides were confirmed (**Table. 2**). The Acc achieved in this study was

90%, which is in align with our prediction (**Table. 1**). The top 3 tripeptides for DPP-IV inhibitory activity were WAW, WAY, and WPN with IC_{50} values 103.66 µM, 117.40 µM and 128.59 µM, respectively. This result demonstrated that our model can predict DPP-IV-IPs with 90% accuracy.

2.6 Proposed a strategy to design DPP-IV inhibitory peptides based on the repeating dipeptide unit X-proline

Recently, development of protein synthesis technology has enabled screening of DPP-IV inhibitory activity from chemically synthesized peptides. These studies were conducted based on the property of preferential cleaving of X-proline or X-alanine from N-terminus of peptides by DPP-IV. However, the feasibility to design repetitive dipeptide unit X-proline remained unclear. To address this problem, we used our model to investigate the potential of this design strategy.

As VPI exhibit highest human DPP-IV (hDPP-IV) inhibitory activity, dipeptide unit VP was selected for further study[40]. A library of 20 pentapeptides (VPVPX) containing repeat VP was designed and 20 kinds of pentapeptides were predicted by our model. In addition, IPIPI and IPVPI were also predicted by our model for IPI exhibit the highest porcine DPP-IV inhibitory activity [40]. The model predicted that 14 pentapeptides exhibit the DPP-IV inhibitory activity, which has provided us with new insights on the properties of pentapeptides capable of inhibiting DPP-IV. .

To verify the properties of the peptides predicted, these peptides were chemically synthesized and IC_{50} values were measured (**Table. 3**, **Supplementary Table. 3**). 8 pentapeptides were identified as false negative while IC_{50} of other peptides were correctly predicted. The top 3 peptides with highest IC_{50} values were IPIPI, VPVPH, and VPVPC, which were 47.47 µM, 51.11 µM, and 54.77 µM respectively. The result suggests that pentapeptides with repetitive dipeptide unit VP exhibit higher DPP-IV inhibitory activity compared to tripeptides, and speculate that designing peptides with repeating units is a more efficient strategy.

To understand the DPP-IV inhibitory activity of these pentapeptides, the binding energy, which is inversely proportional to inhibitory activity was analyzed with MDockPeP[41]. Binding affinities of pentapeptides to human DPP-IV were calculated and we have found that the 3 pentapeptides with the lowest affinity were VPVPW, VPVPR, and VPVPY. However, difference was observed in the ranking of binding affinity to IC_{50}. The difference might probably due to (1) docking method can only reflect the first step of the interaction between the polypeptide and DPP-IV and cannot reflect the subsequent two consecutive VP cleavage processes, (2) docking affinity was calculated based on DPP-IV in static state, thus did not consider the dynamic of DPP-IV.

The results suggested that design strategy based on repeating X-proline units is effective and feasible. However, the efficiency of the peptides with repeating units of X-alanine have to be further investigated.

3 Discussion

In this study, we have successfully applied a pre-trained strategy-based BERT model to DPP-IV-IPs prediction and recognition, and obtained the peptide language models based on the BERT algorithm (**Fig. 1**). Our model, BERT-DPPIV, has demonstrated Acc 0.894, which is higher than those previously reported, e.g., StackDPPIV. Furthermore, this model can also capture more biologically relevant information from peptide sequences, notably the cleavage site information of DPP-IV and structure information of peptide sequences (**Fig. 2**, **Fig. 3**). Our model can automatically capture information from different levels of sequence, structure feature information from the peptide sequence, and the specific feature cleavage site information from the peptide dataset. This render our model more advantageous than StackDPPIV, which is the best model previously reported, which needs to extract sequence and structure feature information by feature engineering.

Although, our model has achieved good performance, there is still room for further improvement. As BERT model which uses the word segmentation method with overlap usually possess higher accuracy, e.g., DNABERT[42], our model which is based on non-overlapping word segmentation method might be further improved by employing overlapping segmentation method.

Attention-based BERT model adopted in this study overcome poor interpretability of deep learning methods. Visualizing the model's attention during learning revealed what the model has learned and how our model changes during training. Through this strategy, we have discovered that our model can capture the cleavage site information of DPP-IV in addition to the physicochemical properties and structural information of the peptide sequence (**Fig. 2**, **Fig. 3**), which has not been previously reported by a classifier. For the first time, we discovered that peptide language model BERT can discover enzyme cleavage site information from substrate information. This model is extremely helpful in understanding the biological implications DPP-IV. This feature can be extended to understand the functions of different enzymes, and to mine the cleavage site information of enzymes.

The previous classifiers for inhibiting DPP-IV peptides were only tested on the data set, and were not verified by actual screening experiments. As a proof of concept, we use the model to screen tripeptides with DPP-IV inhibitory activity, and showed experimentally that our model was 90% accurate. DPP-IV inhibitory activity was identified in 72 out of 80 tripeptides predicted (**Table. 2**). This result suggested that our model can be used to aid the screening of functional peptides that inhibit DPP-IV.

Reported DPP-IV inhibitory polypeptides is still limited due to the lack of strategies for DPP-IV inhibitory polypeptide library design. Therefore, we have also proposed a new peptide design strategy to design DPP-IV-inhibiting polypeptides based on repeating dipeptide units to accelerate mining of DPP-IV inhibitory polypeptides. The design strategy based on based on repeating dipeptide units has been verified and shown to be feasible through prediction of our model and biological assay. 22 Pentapeptides containing VP or

IP were discovered to be able inhibit DPP-IV efficiently. This design strategy can help researchers to efficiently explore peptides that inhibit DPP-IV. The accuracy of model for pentapeptides prediction was 63.6% despite the lack of sequences with repeating dipeptide units in the training dataset. The model can be improved by increasing the availability of data through biological assay. Verification of the efficiency of model in prediction of long peptides sequence is in progress.

In conclusion, we constructed the first PLM-based DPP-IV classifier and verified its performance by experiment. Furthermore, based on the assistance of the classifier, 88 peptides were revealed to inhibit DPP-IV. And a novel design strategy of peptides with DPP-IV inhibitory activity was proposed and verified, 22 pentapeptides were discovered to have DPP-IV inhibitory effects.

4 Method

4.1 Database

The pre-trained protein data was downloaded from UniProt and can be accessed on Google Drive at https://drive.google.com/file/d/1QeXWV5_-OIKgms7u5ShNfvPEKPBLOC3UT/view?usp=sharing. This dataset contain 556,603 protein sequences[43]. The dataset used for fine-tuning was as same as (Charoenkwan et al. 2020a) to train and test our proposed model and can be accessed via Github at https://github.com/Shoombuatong/Dataset-Code/tree/master/DPP-IV/. The benchmark dataset includes 532 DPP-IV-IPs and the same number of non-DPP-IV inhibitory peptides. The number of DPP-IV inhibitory and non-DPP-IV inhibitory peptides in the independent dataset is 133, respectively.

4.2 Reagents and Peptides

Human dipeptidyl peptidase-IV (hDPP-IV, > 200 U/ml) was obtained from the ATGen company (South Korea); Gly-Pro$-p$-nitroanilide (Gly-Pro-pNA) was obtained from the Cayman Chemical Company (Michigan, The United States of America); Peptides containing VPVPX, WPX, WAX, WRX, and WVX (X represent one of 20 kinds of amino acid) with purity $> 95\%$; were synthesized by Gen-Script (Suzhou, China);

4.3 Peptide language model construction

Bidirectional encoder representations from transformers (BERT)[44] is a pre-trained language model applied to natural language text by Google, and achieved state-of-the-art results on downstream NLP tasks through transfer learning. BERT learns effectively capturing the contextual information from an large-scale unlabeled corpus and generate informative and universal vector representations. We pre-trained 12-layer BERT-based language models on 556,603 protein sequences from UniProt dataset[43].

To adapt our model to different sequence lengths, different word segmentation was adapt including kmer=1, kmer=2 and kmer=3 to get three kinds of pre-trained BERT models. The pre-train process contains two pre-training tasks: masked language model (MLM) and next sentence prediction (NSP), so that our pre-trained BERT models can capture word-level and sentence-level representations and learn the common features of protein sequences. MLM are able to modeling complex relationships between amino acids and capture the evolutionary information for protein. In the MLM task, the original protein sequence x will be randomly masked 15% tokens to obtain the masked sequence, and the special token [MASK] is used to replace the masked word in 80% of the time, a random word is used in 10% of the time and the selected word keeps unchanged in 10% of the time. The model predicts the masked tokens through the context of the unmasked sequence for language modeling. For training, we minimize the negative log likelihood of the true amino acid at each of the masked positions, the formular is followed as Eq. (1).

$$L_{MLM} = -\sum_{\hat{x} \in m(x)} logp(\hat{x}|x_{\setminus m(x)}) \tag{1}$$

Where $m(x)$ and $x_{\setminus m(x)}$ denote the masked words from X and the rest words respectively. In the NSP task, the data are randomly divided into two parts and 50% of the sentence pairs are contextually continuous, while the remaining half are not. BERT is trained by identifying whether these sentence pairs are continuous. And Loss functions of NSP task is followed as Eq. (2):

$$L_{NSP} = -logp(t|x, y) \tag{2}$$

Where $t = 1$ if x and y are continuous sequence from corpus.

Then MLM and NSP are trained together, with the goal of minimizing the combined loss function of the two strategies, the formular is followed as Eq. (3):

$$L = L_{MLM} + L_{NSP} \tag{3}$$

For hyperparameters of pre-training, we take parameters including train steps of 10 million times, a learning rate of 2×10^{-5}, and each batch size of 32 to train the BERT model[43]. After the pre-train process, three kinds of pre-trained BERT models were obtained. In order to construct the peptide language model for a specific downstream task that identifies and predicts DPP-IV-IPs, we modify the pre-trained BERT model by adding a classification layer on top of the BERT output for the [CLS] token. We fine-tune the pre-trained model on a benchmark dataset containing DPP-IV-IPs without major architectural modifications. For hyperparameters of fine-tuning, we use a learning rate of 2×10^{-6}, each batch size of 32, a warm-up proportion of 0.1, and an average training of 50 epochs.

4.4 Model performance evaluation methods

We use four general quantitative indicators: sensitivity (Sn), specificity (Sp), Acc, and Matthews correlation coefficient (MCC) to evaluate our model, which are defined as for formula Eq. (4), Eq. (5), Eq. (6), and Eq. (7). TP is the number of correctly predicted DPP-IV-IPs, FN is the number of DPP-IV-IPs predicted to be non-DPP-IV inhibitory peptides, TN is the number of correctly predicted non-DPP-IV inhibitory peptides and FP is the number of non-DPP-IV inhibitory peptides predicted to be DPP-IV-IPs.

$$\text{Acc} = \frac{TP + TN}{(TP + FP + TN + FN)} \tag{4}$$

$$\text{Sn} = \frac{TP}{(TP + FN)} \tag{5}$$

$$\text{Sp} = \frac{TN}{(TN + FP)} \tag{6}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{7}$$

Sn and Sp, respectively, reflect the model's ability to recognize DPP-IV-IPs and non-DPP-IV inhibitory peptides, and Acc embodies the overall prediction effect of the model. The value range of the three is [0,1], and the larger the value, the more accurate the model's prediction. MCC is usually considered as a balanced indicator and can be used even if the sample is not balanced. Its value is between -1 and $+1$, reflecting the correlation between the true label of the sample in the testing set and the predicted result. The higher the value, the greater the correlation. When the value is close to 1, the classification performance of the model is excellent; when it is close to -1, the prediction result of the model is opposite to the actual result and when it is close to 0, the prediction result of the model is similar to the random prediction. In addition, the area under the receiver operating characteristic (AUC) was employed as another statistical metric. Considering these five evaluation indicators, the performance of the classification models can be better evaluated.

4.5 Visualization of Model

Representation visualisation is used to analyze the learning capacity of the model. The original protein sequence $x = [x_1, ...x_n]$ will add the special start and end tokens respectively CLS and SEP used by BERT turn to result from input sequence $x^{'} = [CLS, x_1, ..., x_n, SEP]$. The result sequence representation h_i is got from hidden states in different layers of BERT-DPPIV.

Analyzing the information contained in the representation learned by the model, we use the vector corresponding to CLS in each layer of BERT- DPPIV to represent the analytical sequence and analyze the representation of residue level and sequence level by T-SNE. At the residue level, the amino acids are divided into six categories, including aromatic (W, F, Y), aliphatic (M, L, I, A,

V), positive (R, H, K), negative (D, E), polar neutral (Q, N, S, T), and special case (G, P, C). At the sequence level, 10 kinds of physicochemical properties information were obtained from the modAMP package.

Attention visualization is used to analyze the information specifically concerned by the model and provide interpretable analysis. Each attention head in a model layer produces an attention matrix α indicates the degree of correlation between token pairs, such as $\alpha_{i,j}$ is the attention from token i to token j.

$$\sum_{j=1}^{n} \alpha_{i,j} = 1 \tag{8}$$

We analyze the attention through a multiscale visualization tool for the Transformer model[36]. The attention-head view in this tool expresses self-attention matrix α in the form of lines connecting and displays the attention patterns generated by one or more attention headers in a given layer. The lines in the head view indicate how much of the hidden state information of the attending token (right) will flow to the attended token (left). And different colors of the lines are represented by different attention heads while the color depth of the line is related to the attention weight. The model view in the tool provides a global view of attention patterns across all layers and heads of the model. We then use a slightly simplified version of AlphaFold[39] to get the structure information of the polypeptide sequences and explore the relationship between sites of interaction in polypeptide structures and model attention patterns.

We use the attention matrix to calculate the importance of each token $d_j$in the peptide sequence.

$$d_j = \sum_{i=1}^{n} \alpha_{i,j} \tag{9}$$

We make statistics on 144 attention patterns including all layers and all heads of each layer. We remove sequences of length less than or equal to three. Then we calculate the importance of all tokens including special tokens (CLS, SEP), and take out the three most important tokens for each sequence in each attention pattern. We counted the distribution of the top three most important tokens in different patterns for each sequence by position and amino acid, respectively. We also selected a specific residue in the sequence and calculated the importance of this residue $d_{residue}$ in 144 attention modes, and visualized it using a heatmap. We then investigated the interaction between attention and particular amino acids[37]. We use the next formula for particular amino acids, and define an indicator function $f(i, j)$ that returns 1 if this amino acids

present in token j (e.g if the token j is Pro).

$$p_\alpha(f) = \frac{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} f(i,j) \cdot \mathbb{1}_{\alpha_{i,j} > \theta}}{\sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} \mathbb{1}_{\alpha_{i,j} > \theta}} \tag{10}$$

Where $p_\alpha(f)$ equals the proportion of attention that is directed to the particular amino acid. We computed the proportion of fine-tuned model's attention to each of the 20 standard amino acids.

4.6 DPP-IV Inhibitory Assay

DPP-IV inhibitory activity was measured according to a slightly modified version of the method described previously (Nongonierma et al. 2016)[21]. hDPP-IV was used for the triplicate measurements. Peptide solution (20 µl), substrate solution-Gly-Pro-pNA (2 mM; 20 µl), and Tris–HCl buffer (pH 8.0; 20 µl) were pre-mixed in a 96-well microplate. A reaction was initiated by adding 40 µl of hDPP-IV (final concentration 0.025 units/ml) in 100 µl of the above mixture. The samples were then incubated at 37 °C for 60 min. Absorbance at 405 nm was measured using a microplate reader. The negative control mixture included Tris–HCl buffer (pH 8.0), Gly-Pro-pNA, hDPP-IV, and PBS. The hDPP-IV inhibitory ratio of the sample was calculated according to the following Eq. (11):

$$\text{Inhibition ratio} = \frac{\text{Abs}_{\text{Control}} - \text{Abs}_{\text{Sample}}}{\text{Abs}_{\text{Control}}} \tag{11}$$

Where $\text{Abs}_{\text{Control}}$ is the absorbance of the control sample with PBS and $\text{Abs}_{\text{Sample}}$ is the absorbance of the peptide sample.

5 Tables

6 Figures

Declarations

- This work was founded by Shenzhen Science and Technology Innovation Commission (KCXFZ20201221173207022)
- The authors declare no competing interests
- Ethics approvalNot Applicable
- Consent to participate: Not Applicable
- All authors agree to publish this manuscript
- Availability of data and materials
- The code in this work was available and can be accessed at https://github.com/guanchangge/BERT-DPPIV.

- Authors' contributions: X.-H.X., N.Y., and C.G. conceived the project and design the study. C.G. designed the tripeptides and pentapeptides screening experiment research; J.C. and C.G. designed the model construction and analysis; C.G. and S.L. performed the tripeptides and pentapeptides screening experiment research; C.G. and J.L. performed model construction; C.G. conducted wet experiments; C.G., J.L. and Z.L.T. performed analysis. All authors contributed to writing the paper.

References

[1] De, S., Banerjee, S., Kumar, S.K.A., Paira, P.: Critical role of dipeptidyl peptidase iv: A therapeutic target for diabetes and cancer. Mini-Reviews in Medicinal Chemistry **19**(2), 88–97 (2019)

[2] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B.B., Stein, C., Basit, A., Chan, J.C., Mbanya, J.C., *et al.*: Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes research and clinical practice **183**, 109119 (2022)

[3] Ogurtsova, K., Guariguata, L., Barengo, N.C., Ruiz, P.L.-D., Sacre, J.W., Karuranga, S., Sun, H., Boyko, E.J., Magliano, D.J.: Idf diabetes atlas: Global estimates of undiagnosed diabetes in adults for 2021. Diabetes research and clinical practice **183**, 109118 (2022)

[4] Rasmussen, H.B., Branner, S., Wiberg, F.C., Wagtmann, N.: Crystal structure of human dipeptidyl peptidase iv/cd26 in complex with a substrate analog. Nature structural biology **10**(1), 19–25 (2003)

[5] Casrouge, A., Sauer, A., Barreira da Silva, R., Tejera-Alhambra, M., Sanchez-Ramon, S., ICAReB, Cancrini, C., Ingersoll, M., Aiuti, A., Albert, M.: Lymphocytes are a major source of circulating soluble dipeptidyl peptidase 4. Clinical & Experimental Immunology **194**(2), 166–179 (2018)

[6] Roppongi, S., Suzuki, Y., Tateoka, C., Fujimoto, M., Morisawa, S., Iizuka, I., Nakamura, A., Honma, N., Shida, Y., Ogasawara, W., *et al.*: Crystal structures of a bacterial dipeptidyl peptidase iv reveal a novel substrate recognition mechanism distinct from that of mammalian orthologues. Scientific reports **8**(1), 1–18 (2018)

[7] Mentlein, R., Gallwitz, B., Schmidt, W.E.: Dipeptidyl-peptidase iv hydrolyses gastric inhibitory polypeptide, glucagon-like peptide-1 (7–36)

amide, peptide histidine methionine and is responsible for their degradation in human serum. European journal of biochemistry **214**(3), 829–835 (1993)

[8] Yip, R.G., Wolfe, M.M.: Gif biology and fat metabolism. Life sciences **66**(2), 91–103 (1999)

[9] Demuth, H.-U., McIntosh, C.H., Pederson, R.A.: Type 2 diabetes—therapy with dipeptidyl peptidase iv inhibitors. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics **1751**(1), 33–44 (2005)

[10] Green, B.D., Flatt, P.R., Bailey, C.J.: Dipeptidyl peptidase iv (dpp iv) inhibitors: A newly emerging drug class for the treatment of type 2 diabetes. Diabetes and Vascular Disease Research **3**(3), 159–165 (2006)

[11] Deacon, C.F.: A review of dipeptidyl peptidase-4 inhibitors. hot topics from randomized controlled trials. Diabetes, Obesity and Metabolism **20**, 34–46 (2018)

[12] Jarvis, C.I., Cabrera, A., Charron, D.: Alogliptin: a new dipeptidyl peptidase-4 inhibitor for type 2 diabetes mellitus. Annals of Pharmacotherapy **47**(11), 1532–1539 (2013)

[13] Barnett, A.: Dpp-4 inhibitors and their potential role in the management of type 2 diabetes. International journal of clinical practice **60**(11), 1454–1470 (2006)

[14] Mulvihill, E.E., Drucker, D.J.: Pharmacology, physiology, and mechanisms of action of dipeptidyl peptidase-4 inhibitors. Endocrine reviews **35**(6), 992–1019 (2014)

[15] Nongonierma, A.B., FitzGerald, R.J.: Features of dipeptidyl peptidase iv (DPP-IV) inhibitory peptides from dietary proteins. Journal of food biochemistry **43**(1), 12451 (2019)

[16] Wang, Y., Li, S., Guan, C., He, D., Liao, X., Chen, H., Zhang, C., Xing, X.: Functional discovery and production technology for natural bioactive peptides. Sheng wu Gong Cheng xue bao Chinese Journal of Biotechnology **37**(6), 2166–2180 (2021)

[17] Duez, H., Cariou, B., Staels, B.: Dpp-4 inhibitors in the treatment of type 2 diabetes. Biochemical pharmacology **83**(7), 823–832 (2012)

[18] Nongonierma, A.B., FitzGerald, R.J.: An in silico model to predict the potential of dietary proteins as sources of dipeptidyl peptidase iv (DPP-IV) inhibitory peptides. Food Chemistry **165**, 489–498 (2014)

[19] Nongonierma, A.B., Paolella, S., Mudgil, P., Maqsood, S., FitzGerald, R.J.: Identification of novel dipeptidyl peptidase iv (DPP-IV) inhibitory peptides in camel milk protein hydrolysates. Food Chemistry **244**, 340–348 (2018)

[20] Nongonierma, A.B., Mooney, C., Shields, D.C., FitzGerald, R.J.: In silico approaches to predict the potential of milk protein-derived peptides as dipeptidyl peptidase iv (DPP-IV) inhibitors. Peptides **57**, 43–51 (2014)

[21] Nongonierma, A.B., FitzGerald, R.J.: Structure activity relationship modelling of milk protein-derived peptides with dipeptidyl peptidase iv (DPP-IV) inhibitory activity. Peptides **79**, 1–7 (2016)

[22] Nongonierma, A.B., FitzGerald, R.J.: Learnings from quantitative structure–activity relationship (qsar) studies with respect to food protein-derived bioactive peptides: a review. RSC advances **6**(79), 75400–75413 (2016)

[23] Shoombuatong, W., Prathipati, P., Owasirikul, W., Worachartcheewan, A., Simeon, S., Anuwongcharoen, N., Wikberg, J.E., Nantasenamat, C.: Towards the revival of interpretable qsar models, 3–55 (2017)

[24] Hellberg, S., Sjoestroem, M., Skagerberg, B., Wold, S.: Peptide quantitative structure-activity relationships, a multivariate approach. Journal of medicinal chemistry **30**(7), 1126–1135 (1987)

[25] Basith, S., Manavalan, B., Hwan Shin, T., Lee, G.: Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Medicinal research reviews **40**(4), 1276–1314 (2020)

[26] Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, M.M., Shoombuatong, W.: idppiv-scm: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase iv (DPP-IV) inhibitory peptides using a scoring card method. Journal of Proteome Research **19**(10), 4125–4136 (2020)

[27] Zou, H., Yin, Z.: Identifying dipeptidyl peptidase-iv inhibitory peptides based on correlation information of physicochemical properties. International Journal of Peptide Research and Therapeutics **27**(4), 2651–2659 (2021)

[28] Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Moni, M.A., Manavalan, B., Shoombuatong, W., et al.: Stackdppiv: A novel computational approach for accurate prediction of dipeptidyl peptidase iv (DPP-IV) inhibitory peptides. Methods (2021)

[29] Janiesch, C., Zschech, P., Heinrich, K.: Machine learning and deep

learning. Electronic Markets **31**(3), 685–695 (2021)

[30] Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H.: Signalp 5.0 improves signal peptide predictions using deep neural networks. Nature biotechnology **37**(4), 420–423 (2019)

[31] Youmans, M., Spainhour, C., Qiu, P.: Long short-term memory recurrent neural networks for antibacterial peptide identification. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 498–502 (2017). IEEE

[32] Yan, J., Bhadra, P., Li, A., Sethiya, P., Qin, L., Tai, H.K., Wong, K.H., Siu, S.W.: Deep-ampep30: improve short antimicrobial peptides prediction with deep learning. Molecular Therapy-Nucleic Acids **20**, 882–894 (2020)

[33] Wu, C., Gao, R., Zhang, Y., De Marinis, Y.: Ptpd: predicting therapeutic peptides by deep learning and word2vec. BMC bioinformatics **20**(1), 1–8 (2019)

[34] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences **118**(15) (2021)

[35] Müller, A.T., Gabernet, G., Hiss, J.A., Schneider, G.: modlamp: Python for antimicrobial peptides. Bioinformatics **33**(17), 2753–2755 (2017)

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[37] Vig, J.: A multiscale visualization of attention in the transformer model. arXiv preprint arXiv:1906.05714 (2019)

[38] Timmons, P.B., Hewage, C.M.: Apptest is a novel protocol for the automatic prediction of peptide tertiary structures. Briefings in bioinformatics **22**(6), 308 (2021)

[39] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)

[40] Guan, C., Iwatani, S., Xing, X.-h., Yamamoto, N.: Strategic preparations of DPP-IV inhibitory peptides from val-pro-xaa and ile-pro-xaa peptide

mixtures. International Journal of Peptide Research and Therapeutics **27**(1), 735–743 (2021)

[41] Xu, X., Yan, C., Zou, X.: Mdockpep: An ab-initio protein–peptide docking server. Journal of computational chemistry **39**(28), 2409–2413 (2018)

[42] Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics **37**(15), 2112–2120 (2021)

[43] Zhang, Y., Lin, J., Zhao, L., Zeng, X., Liu, X.: A novel antibacterial peptide recognition algorithm based on bert. Briefings in Bioinformatics **22**(6), 200 (2021)

[44] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

Table 1 The classification results of BERT-DPPIV and other predictors

Model	ACC	Sn	Sp	MCC	AUC
Model with kmer=1	0.891	0.857	0.925	0.784	0.957
Model with kmer=2	0.887	0.880	0.895	0.775	0.961
Model with kmer=3	0.842	0.850	0.835	0.684	0.941
BERT-DPPIV	0.894	0.872	0.917	0.790	0.960
iDPPIV-SCM	0.797	0.789	0.805	0.594	0.847
SVM	0.865	0.874	0.856	0.731	0.939
StackDPPIV	0.891	0.857	0.925	0.784	0.961

Table 2 The results of tripeptides inhibiting DPP-IV

Peptides	IC_{50} (μM)	Peptides	IC_{50} (μM)	Peptides	IC_{50} (μM)	Peptides	IC_{50} (μM)
WPA	206.01	WAA	274.53	WRA	534.67	WVA	341.25
WPC	739.03	WAC	282.79	WRC	209.64	WVC	231.78
WPD	480.61	WAD	432.65	WRD	203.64	WVD	343.00
WPE	353.02	WAE	461.79	WRE	160.15	WVE	192.50
WPF	285.91	WAF	153.32	WRF	257.07	WVF	188.33
WPG	134.98	WAG	352.16	WRG	269.85	WVG	340.63
WPH	314.57	WAH	189.55	WRH	277.20	WVH	256.86
WPL	461.32	WAL	260.10	WRL	150.75	WVL	398.88
WPM	218.75	WAM	329.85	WRM	896.67	WVM	317.00
WPN	128.59	WAN	301.78	WRN	313.20	WVN	372.88
WPP	303.49	WAP	440.97	WRP	741.50	WVP	402.00
WPQ	196.18	WAQ	289.19	WRQ	285.50	WVQ	366.00
WPT	435.08	WAT	296.33	WRT	555.60	WVT	335.63
WPV	511.47	WAV	249.79	WRV	541.17	WVV	275.33
WPW	182.96	WAW	103.66	WRW	603.14	WVW	238.63
WPY	303.79	WAY	117.40	WRY	296.50	WVY	214.88
WPI	-	WAI	198.88	WRI	330.40	WVI	-
WPK	-	WAK	-	WRK	-	WVK	331.38
WPR	208.94	WAR	664.68	WRR	-	WVR	-
WPS	588.31	WAS	-	WRS	265.09	WVS	247.00

IC_{50} : half maximal inhibitory concentration.

- : not detected within 1 mM.

Table 3 The characteristic results of pentapeptides inhibiting DPP-IV

Pentapeptide	IC_{50} (μM)	Affinity (kcal/mol)	Pentapeptide	IC_{50} (μM)	Affinity (kcal/mol)
VPVPA	98.05	-76.4	VPVPM	99.92	-89.2
VPVPC	54.77	-79.4	VPVPN	88.74	-78.8
VPVPD	74.50	-77.4	VPVPP	323.70	-80.1
VPVPE	83.44	-85.7	VPVPQ	111.55	-84.8
VPVPF	81.87	-97.0	VPVPR	117.71	-101.2
VPVPG	84.91	-77.3	VPVPS	92.87	-79.3
VPVPH	51.11	-88.8	VPVPT	130.67	-76.7
VPVPI	103.19	-85.6	VPVPV	59.05	-78.2
VPVPK	104.29	-85.1	VPVPW	92.10	-108.2
VPVPL	129.62	-84.6	VPVPY	80.73	-97.3
IPIPI	47.47	-87.4	IPVPI	56.98	-93.0

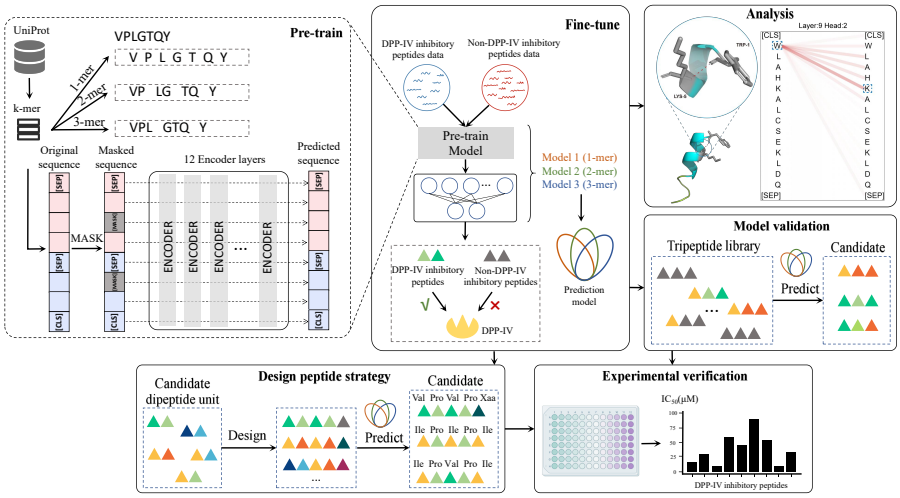

Fig. 1 Schematic representation of study workflow. In this study, we started by collecting sequences to pre-train a BERT model (upper left). We then used the DPP-IV-IP dataset for fine-tuning the pre-train BERT model to get three kinds of peptide language models. Then three models were combined for constructing the BERT-DPPIV (upper middle). To interpret what the model has learned, we analyze and visualize the model's attention (upper-right analysis module). Then we applied BERT-DPPIV to screen tripeptides that inhibit DPP-IV to illustrate the performance of our model inhibition (upper-right model validation module). After the model screening, the functional peptides screened out by the model are verified by biological experiments (down right). Having demonstrated the practical application capability of our model, we propose a design strategy for DPP-IV-inhibiting polypeptides based on dipeptide repeat units and demonstrate the feasibility of this strategy using models and experiments (down left).

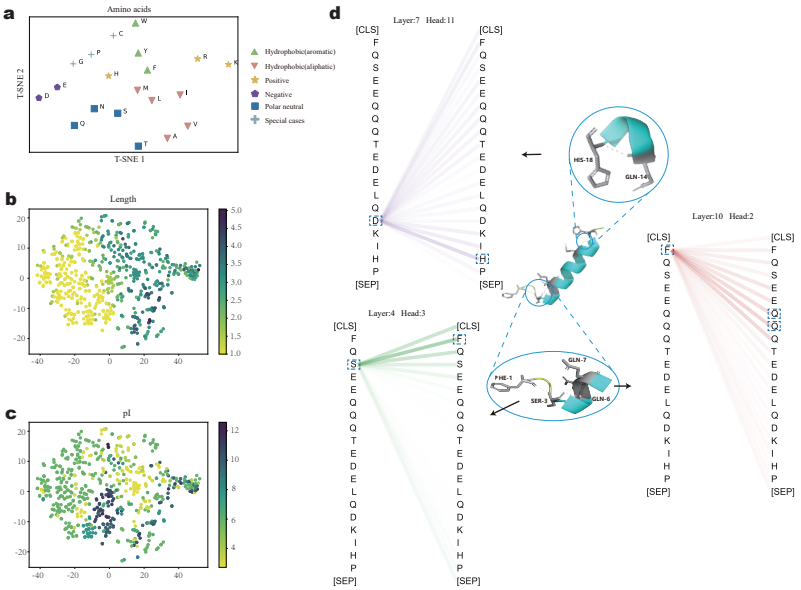

Fig. 2 Sequence representation visualization and attention visualization of structural information learned by the model. The vector representation of the peptide sequence and the amino acid was extracted from the model and analyzed by t-SNE. Amino acid representation learned by the model (a) and physicochemical properties contained in model-learned representations of peptide sequences (b and c). Here, we show how the inner workings of the model's attention heads can be used to analyze a single peptide in more detail. Each attention head performs internally an all-against-all comparison to compute weighted sums for each token over all other tokens in the sequence. High scores indicate that the model learned to put more weight on certain residue pairs (upper left, down left, and down right). The structure of the peptide was shown, that SER-3 can interact with PHE-1, GLN-6, and GLN-7 and GLN-14 can interact with HIS-18. When visualizing the attention weights for these sites, we observed that some of the attention heads can capture this interaction information.

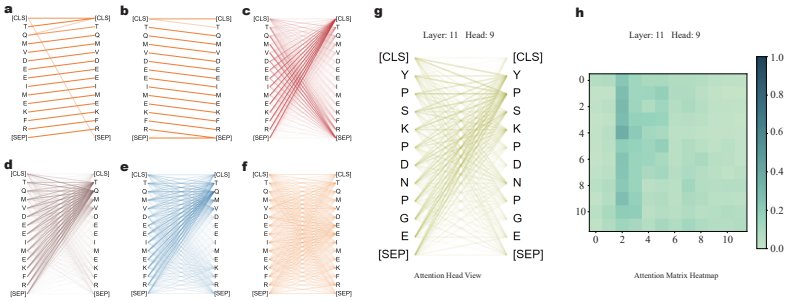

Fig. 3 Visualizing model's attention patterns. Previous-word attention pattern: virtually all the attention is focused on the previous word in the input sequence (a). Next-word attention patterns: virtually all the attention is focused on the next word in the input sequence (b). Delimiter-focused attention patterns: most attention is focused on the [SEP] token or [CLS] token (c). Specific-word attention patterns: virtually all the attention is focused on the specific site (d and e). Attention Related-words attention patterns: attention is paid to identical or related words (f). The attention mainly foucus on the 2th and 3th position of peptide sequence and the corresponding heatmap (g and h).

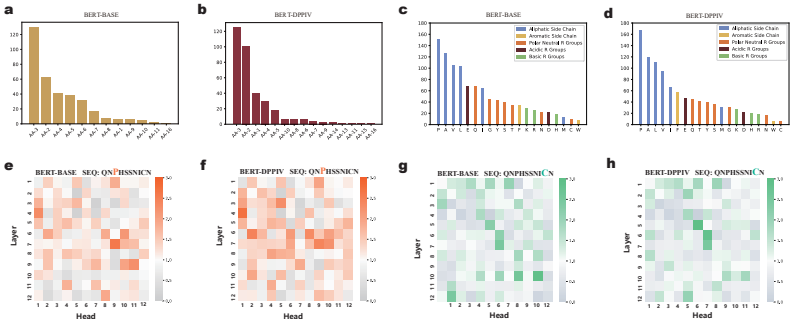

Fig. 4 Visualization of position and amino acid species importance. In order to study the importance of position and amino acid species, attention to positions and amino acids was statistical analysis. BERT-BASE and BERT-DPPIV represent the models before and after fine-tuning, respectively. The statistical importance of the position of peptide sequence according to the attention of the model (a and b). The statistical importance of amino acids according to the attention of the model (c and d). The attention of the model to Pro and Cys in the peptide sequence was visualized (e, f, g, and h).

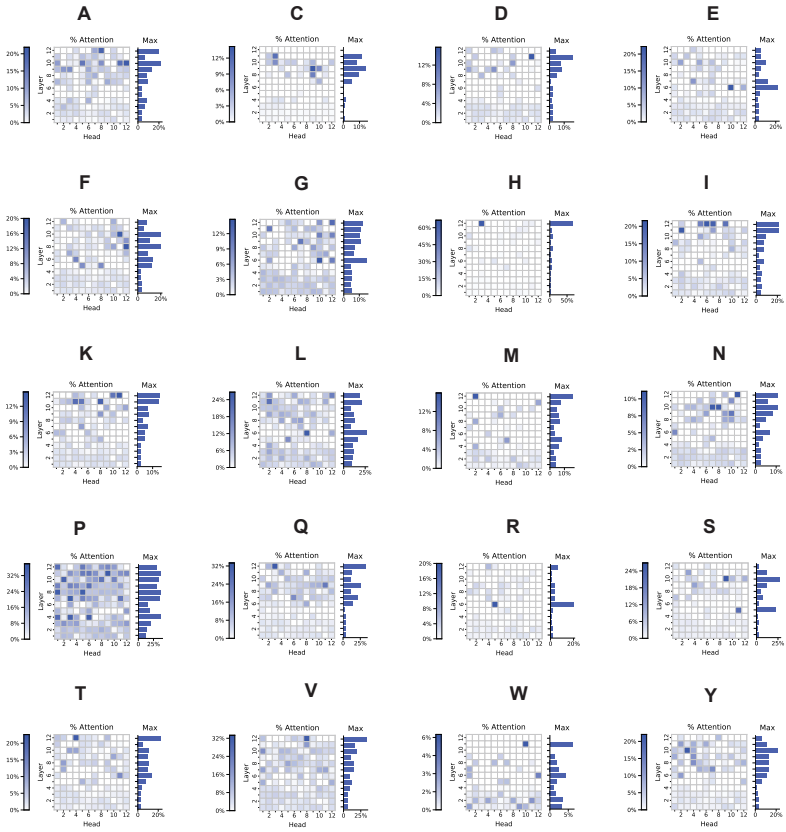

Fig. 5 Percentage of each head's attention focused on amino acids including 20 natural amino acids. We computed the proportion of our model's attention to each of the 20 standard amino acids. For 13 of the amino acids, there exists an attention head that focuses over 20% of attention on that amino acid, significantly greater than the background frequencies of the corresponding amino acids, which range from 0.56% to 17.7%. Compared with the the pre-trained model, the attention ratio of the fine-tuned model is more concentrated in the deep layers. Detailed results are presented in Supplementary Table. 1.