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

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], Yuan Lu[1], 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.

[†] These authors contributed equally to this work.

*Corresponding authors at:

Xin-Hui Xing

xhxing@mail.tsinghua.edu.cn

Junjie Chen

junjiechen@hit.edu.cn

Changge Guan

gcgtsinghua@gmail.com

Abstract

Dipeptidyl peptidase IV (DPP-IV) inhibitory peptides (DPP-IV-IPs) are next generation anti-diabetic drugs. Only a limited number of DPP-IV-IPs have been discovered because there are few efficient tools for peptide mining. In this study, we propose a peptide language model (PLM) to identify DPP-IV-IPs from large-scale peptide datasets and achieve state-of-the-art accuracy of 0.894. Visualization of the 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, and can distinguish and represent amino acid sequences in high-dimensional spaces. We demonstrated that this PLM could capture cleavage site information and can guide biological experimental screening, which has not been previously reported. Based on follow-up biological assays, the prediction accuracy of the PLM-assisted DPP-IV-IP screening was 90%. To explore the diversity of DPP-IV-IPs, we proposed a strategy based on the dipeptide repeat unit X-proline and showed that the strategy is feasible through modeling and biological experiments.

1. Introduction

DPP-IV is therapeutic target for next generation anti-diabetic drugs[1]. The number of patients affected by type II diabetes is estimated to be 537 million people, and contributed to mortality of >6.7 million people aged 20–79 years in 2021 alone[2,3]. Diabetes is a chronic disease and one of the fastest-growing global health emergencies of the twenty-first century[2]. 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] that preferentially cleaves substrate peptides at X-proline or X-alanine dipeptides from the N terminus of polypeptides[6]. It affects glucose metabolism by degrading incretin hormones, e.g., glucose-dependent insulinotropic polypeptide and glucagon-like peptide-1, which are rapidly released into the blood circulation after glucose or fat uptake to stimulate insulin secretion and lower blood glucose levels[7,8]. Inhibition of DPP-IV enzymatic activity helps with glycemic control and weight reduction by prolonging the effect of incretin *in vivo*[1,7,9]. Thus, DPP-IV inhibitors represent a promising class of oral hypoglycemics to improve glucose metabolism in patients with type II diabetes[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 effects in patients, e.g., headache, cystitis, and nasopharyngitis[12].

Fortunately, DPP-IV-IPs derived from products of enzymatically hydrolyzed animal, plant, and macroalgal proteins exhibit DPP-IV inhibitory activity without inducing adverse effects in patients[13]. However, conventional enzymatic hydrolysate screening is still the main method used, which has limited the efficiency of DPP-IV-IP screening. Synthetic peptide library screening methods have not been used for DPP-IV-IP mining due to the absence of peptide library design strategies and efficient screening methods for DPP-IV inhibitory polypeptides[14]. Low screening efficiency has resulted in a limited number of DPP-IV-IPs being discovered, which in turn has limited the progress in developing DPP-IV-IP products.

Furthermore, the conventional procedures for DPP-IV-IP screening are laborious, time-consuming and costly (due to the requirement for various type of columns, large amounts of materials, and diverse synthetic peptides and enzymes)[15,16]. Multiple steps in separation can also lead to undesired peptide degradation and loss of peptide activity. Although 500 kinds of DPP-IV-IPs had been discovered, their inhibitory activities are too low for practical use. Therefore, it is an important task to develop a high-throughput method capable of rapidly identifying 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. The quantitative structure-

activity relationship (QSAR) method was developed to aid the discovery of novel DPP-IV-IPs while elucidating an in-depth understanding of the 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[19,22–25].

Although progresses have been made with these modeling approaches, their predictive performances are still limited. To improve the predictive capability of these models, several traditional ML-based predictors, e.g., support vector machine (SVM) and random forest, have been established and used to identify DPP-IV-IPs[26–28]. However, the methods reported to date have some major limitations: (1) the performance of these models is heavily dependent on the quality of the features extracted by feature engineering, (2) they are poorly able to represent amino acid sequences, and (3) they fail to capture information hidden in the amino acid sequence itself, which renders these models inefficient. Deep neural networks can overcome these limitations, and their advanced architecture provides them with the capability of automating feature learning to extract discriminating feature representations with minimal human effort[29]. In particular, deep learning (DL) methods, e.g., long-short term memory, which is based on natural language processing (NLP), is a promising method in identifying functional peptides for extraction of discriminating feature representation, such as anti-microbial peptides[30–33]. However, DL-based models have been criticized for their interpretability and unexplainable characteristics, which are referred to as a ‘black box’. Furthermore, classifiers for DPP-IV-IPs developed to date have not been verified experimentally.

In this article, we propose a novel attention-based PLM-based DL model that combines an attention mechanism and pre-trained NLP bidirectional encoder representations from transformers (BERT) model to identify DPP-IV-IPs (Fig. 1). Introducing attention mechanisms in deep neural network 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 was fine-tuned based on a pre-trained BERT model, was able to automatically learn the characteristics and features contained in the original sequences and to distinguish and represent amino acid sequences in high-dimensional spaces. In addition, our model has good interpretability, which is more advantageous than existing machine learning algorithms for DPP-IV-IPs identification. Extensive analyses showed that our model can identify cleavage sites of DPP-IV enzymes on the substrate polypeptides by performing a vector representation learning of the DPP-IV substrate polypeptide sequences.

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

2. Results

2.1 Prediction of sequences of DPP-IV-IPs with BERT-DPPIV

Among NLP models, pre-trained models are more advantageous than direct training of models with respect to completing target task. They can leverage large unlabeled datasets to learn universal representations of data, which are beneficial for various downstream tasks. Therefore, we adopted a pre-training strategy for the construction of our model. Additionally, for model interpretability, a self-attention-based deep bidirectional language representation (i.e., BERT) model architecture was used as the initial model for pre-training. Protein sequences from the UniProt database, data from SWISS-PROT, TrEMBL, and PIR-PSD**,** were used as the 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, the peptide dataset of DPP-IV-IPs was leveraged to fine-tune the pre-trained BERT model to construct the PLM. Three models were constructed based on three kinds of subword tokenizer, kmer = 1, kmer = 2, and kmer = 3. Then, the test set was used to validate the three models. Highest accuracy (Acc) of 0.891 was achieved for kmer = 1, followed by 0.887 for kmer = 2, and 0.842 for kmer = 3 (Table 1). To exploit the potential of all three models, we have combined these three models to construct a new model, BERT-DPPIV, which resulted in an improved Acc of 0.894 (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) as described in each model was 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.

All our models (kmer = 1, kmer = 2, and kmer = 3) outperformed the first sequence-based model, iDPPIV-SCM, with increases in the Acc and Matthews correlation coefficient (MCC) of 9.4% and 19.0%, 9.0% and 18.1%, and 4.5% and 9.0%, respectively (Table 1). These results

suggested that our models can extract more information from peptide sequences, e.g., sequence features, physicochemical properties, structural information, etc. To further investigate the characteristics of our models, we compared our models to the sequence-based physicochemical property SVM model and the sequence- and structure-based StackDPPIV model (Table 1). Two of our models (kmer = 1, kmer = 2) were better than SVM in Acc, with an increase of 2.6% and 2.2%, respectively. Our model kmer = 1 achieved the same accuracy as the best model, StackDPPIV, and model kmer = 2 had slightly lower Acc (0.887). Based on these results, we speculated that a PLM-based BERT-DPPIV model that combined all our models would best 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 as learned by the kmer = 1 model

To investigate the information captured by our models, we selected kmer = 1 that exhibit highest Acc among our three kmer models for further study. First, we analyzed the representation of the peptide sequences learned by the PLM, as prior studies have shown that NLP models can effectively learn the representation of the sequence[30–33]. The vector representation of the peptide sequences was extracted from the model and analyzed by t-distributed stochastic neighbor embedding (t-SNE) during the training process of the model, and the results of the analysis were visualized (Supplementary Fig. 1). Supplementary Fig. 1 shows that the model was able to identify DPP-IV-IPs by learning the vector representation of peptides. In addition, we analyzed the model with t-SNE to identify amino acid vector representation learned by the model and illustrated the ability of the model to learn physicochemical properties of peptides. The vector representation learned by the model was well able to distinguish amino acids with different properties (Fig. 2a). We then used the modlAMP package[34] to extract 10 kinds of physicochemical property information from peptide sequences. The vector representations of these physicochemical properties were extracted, and t-SNE analyses were conducted. The vector representations learned by the model were able to distinguish among sequences with different physical and chemical properties (Fig. 2b,c, Supplementary Fig. 2). To this point, we have shown that the PLM can learn the physicochemical properties of the peptide sequences.

2.3 Capturing the structural information and attention patterns generated by the models

The attention mechanism provides a way to parse black-box DL model and features learned by it[35,36]. Next, we investigated the capability of our model to learn structural information about peptide sequences with the help of an attention mechanism, which is the interface connecting the encoder and decoder. First, the structure of the peptides with a length ≥ 5 amino acids from positive samples was predicted by APPTEST software[37], and 12 kinds of peptides with α-helix secondary structure were found (Supplementary Fig. 3). Three peptides were randomly selected from among these peptides for further structural confirmation using AlphaFold2[38]. The presence of α-helical structures was confirmed.

Then, the attention of the model for each head at each layer of the 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 the attentions focus on the interacting sites of peptides, suggesting that this model was able to capture the structural information of peptide sequences. Furthermore, analysis of 144 attention mechanisms (Supplementary Fig. 4) showed that several learning mechanisms existed in the model, which include previous-word attention patterns (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 attention data from BERT-DPPIV focused on the second and third positions of the N terminus of the polypeptide sequences (Fig. 3d-g), which are the cleavage sites of DPP-IV[14]. 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 results suggested that the attention of our model focused on the cleavage sites of DPP-IVs. Furthermore, a comparison of the changes in attention positions of the model before and after fine-tuning have supported that our model could capture cleavage site information of DPP-IVs.

We have further investigated the capability of our model to analyze the importance of each amino acid in the sequences. Statistical analyses have suggested that proline is the amino acid of major concern in our model, whereas cysteine is of the least concern. Increase in attention toward proline and decrease in attention toward cysteine were observed (Fig. 4c,d).

To investigate the importance of proline and cysteine in these amino acid sequences, we have visualized the attention of our model for each amino acid in a randomly selected sequence from α-lactalbumin, ”QNPHSSNICN”. We observed differences in all amino acids after training, which suggested that the model had directed its attention toward important amino acids, e.g., proline (which is an important amino acid for DPP-IV inhibition), whereas less important amino acids, e.g., cysteine, received less attention (Fig. 4e-h). To illustrate the accuracy of this result, we have statistically analyzed the proportion of our model’s attention that was focused on each of the 20 standard amino acids (Fig. 5, Supplementary Fig. 6, Supplementary Table 1). These results demonstrated that our model efficiently captured the amino acid types that play important roles in the amino acid sequences of a peptide.

2.5 Application of BERT-DPPIV for DPP-IV inhibitory tripeptide screening

As a proof of concept, we applied our model for the screening of DPP-IV inhibitory tripeptides. Prior work conducted by one of the authors (C.G.) classified DPP-IV inhibitory dipeptides into five classes, and, from one class of peptides, represented by VPX and IPX (where X represents one of the 20 amino acids), isolated nine tripeptides with efficient human DPP-IV (hDPP-IV) inhibitory activity[39]. However, it was unknown whether other kinds of tripeptides can efficiently inhibit hDPP-IV. In this study, one class of tripeptides that contains WPX, WAX, WRX, and WVX from five classes previously classified was selected for screening.

We predicted the DPP-IV inhibitory activity of the 80 kinds of tripeptides with our models and found that all these tripeptides are expected to possess DPP-IV inhibitory activity (Supplementary Table 2). Then, these tripeptides were chemically synthesized, and the inhibitory activity was verified experimentally based on the half maximal inhibitory concentration (IC_{50}), which is inversely correlated to inhibitory activity. IC_{50} were detected from 72 out of the 80 kinds of tripeptides, and the DPP-IV inhibitory activity of each of these tripeptides was confirmed (Table 2). The Acc achieved in this study was 90%, which is consistent with our prediction (Table 1). The top three tripeptides for DPP-IV inhibitory activity were WAW, WAY, and WPN with IC_{50} of 103.66 μM, 117.40 μM, and 128.59 μM, respectively. This result demonstrated that our model could predict DPP-IV-IPs with 90% accuracy.

2.6 A proposed strategy to design DPP-IV-IPs based on the repeating dipeptide unit X-proline

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

As VPI has the highest hDPP-IV inhibitory activity[39], the dipeptide unit VP was selected for further study. A library of 20 pentapeptides containing two VP repeats (VPVPX) was designed, and the inhibitory activity of these 20 kinds of pentapeptides were predicted by our model. In addition, IPIPI and IPVPI were also predicted by our model, as IPI has the highest porcine DPP-IV inhibitory activity[39]. The model predicted that 14 out of 22 pentapeptides exhibit inhibitory activity against DPP-IV (Supplementary Table 3), which has provided us with new insights into the properties of pentapeptides capable of inhibiting DPP-IV.

To verify the properties of the peptides predicted, these peptides were chemically synthesized, and their IC_{50} were measured (Fig. 6a, Supplementary Table 4). Eight pentapeptides were identified as false negatives, whereas the IC_{50} for the other peptides were correctly predicted. The top three peptides with the highest IC_{50} were IPIPI, VPVPH, and VPVPC, which were 47.47 μM, 51.11 μM, and 54.77 μM, respectively. Inhibitory activity of IPIPI was 2.18-fold higher than WAW, which is the tripeptides with highest inhibitory activity detected in this study. This result indicates that pentapeptides with the repeating dipeptide unit VP exhibit higher DPP-IV inhibitory activity as compared with tripeptides and thus suggests that designing peptides with repeating units is a more efficient strategy.

To understand the DPP-IV inhibitory activity of these pentapeptides, their binding energy, which is inversely proportional to inhibitory activity, was analyzed with MDockPeP[40]. We calculated the binding affinities of pentapeptides to hDPP-IV and found that VPVPW, VPVPR, and VPVPY had the lowest affinity (Fig. 6b, Supplementary Table 4). However, a difference was observed in their ranking with respect to binding affinity relative to that for IC_{50}. This difference might have resulted from (1) the docking method, which reflects only the first step of the interaction between the polypeptide and DPP-IV and cannot reflect the subsequent two consecutive VP cleavage steps, or (2) the docking affinity being calculated based on DPP-IV in the static state, which thus did not consider the dynamic of DPP-IV.

These results suggested that a design strategy based on repeating X-proline units is effective and feasible, although the efficiency of peptides with repeating units of X-alanine should be investigated.

3. Discussion

In this study, we have successfully applied a pre-trained BERT model to the prediction and recognition of DPP-IV-IPs and obtained a PLM based on the BERT algorithm (Fig. 1). Our model, BERT-DPPIV, had demonstrated Acc of 0.894, which is higher than those previously reported, e.g., StackDPPIV[28]. 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 various aspects of sequence, structural feature information from the peptide sequence, and the specific cleavage site features from the peptide dataset. This renders our model more advantageous than StackDPPIV, which has been the best model described to date, as the latter must extract sequence and structure feature information by feature engineering. Although our model has achieved good performance, there is room for further improvement. As BERT models that use the word segmentation method with overlap, e.g., DNABERT[41], typically show higher accuracy than those that do not, our model which is based on the non-overlapping word segmentation method, might be further improved by using an overlapping segmentation method.

The attention-based BERT model adopted in this study overcomes poor interpretability of DL methods. Visualizing the model's attention during learning revealed what the model had learned and how this model had changed during training. Through this strategy, we discovered that our model can capture cleavage site information for DPP-IV in addition to the physicochemical properties and structural information of the peptide sequence (Fig. 2, Fig. 3), which had not previously been reported as a classifier. This is the first use of the PLM BERT to discover enzyme cleavage site information from substrate information, and this model should be extremely helpful in understanding the biological implications of DPP-IV. This feature should be applicable to understanding the functions of different enzymes and to mining the cleavage site information of enzymes.

The previous classifiers for inhibiting DPP-IV peptides[26–28] were tested only on a dataset and were not verified by actual screening experiments. As a proof of concept, we used this 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 in the screening of functional peptides that inhibit DPP-IV.

The number of reported DPP-IV inhibitory polypeptides is still limited due to the absence of strategies for designing DPP-IV inhibitory polypeptide libraries. Therefore, we have also proposed a new peptide design strategy to design polypeptides that inhibit DPP-IV based on repeating dipeptide units to accelerate mining of DPP-IV inhibitory polypeptides. This design strategy has been verified and shown to be feasible through the predictions of our model and subsequent biological assays. Twenty-two pentapeptides containing VP or IP were discovered to inhibit DPP-IV efficiently. This design strategy can help researchers to efficiently explore peptides that inhibit DPP-IV. The accuracy of the model for pentapeptide prediction was 63.6% despite the lack of sequences with repeating dipeptide units in the training dataset. It is likely that the model can be improved by increasing the availability of data through biological assays. Verification of the efficiency of the model in predicting long peptide sequences is in progress.

In conclusion, we have constructed the first PLM-based DPP-IV classifier and have verified its performance experimentally. Furthermore, based on the assistance of the classifier, 72 peptides were revealed to inhibit DPP-IV. A novel design strategy for peptides with DPP-IV inhibitory activity was proposed and verified, based on which 22 pentapeptides were discovered to have DPP-IV inhibitory effects.

4. Methods

4.1 Database

The pre-trained protein data were downloaded from UniProt and can be accessed on Google Drive at https://drive.google.com/file/d/1QeXWV5_OIKgms7u5ShNfvPEKPBLOC3UT/view?usp=sharing. This dataset contains 556,603 protein sequences[42]. The dataset used for fine-tuning was previously described and was used to train and test our proposed model[26]. The benchmark dataset includes 532 DPP-IV-IPs and the same number of peptides with no inhibitory activity against DPP-IV. The number of IPs and non-inhibitory peptides in the independent dataset was 133 peptides each.

4.2 Reagents and peptides

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, US). Peptides containing VPVPX, WPX, WAX, WRX, and WVX (where X represents any one of 20 amino acids) with purity > 95% were synthesized by Gen-Script (Suzhou, China).

4.3 Training of the PLM

BERT[43] is a pre-trained language model developed by Google for natural language text applications and achieves state-of-the-art results on downstream NLP tasks through transfer learning. BERT learns effectively by capturing the contextual information from a large-scale unlabelled corpus and generates informative and universal vector representations. We pre-trained 12-layer BERT-based language models on 556,603 protein sequences from the UniProt dataset[42].

To adapt our model to different sequence lengths, we used different word segmentation lengths, kmer = 1, kmer = 2, and kmer = 3, to generate three pre-trained BERT models. The pre-training process consisted of two pre-training tasks: masked language model (MLM) and next sentence prediction (NSP), so that our pre-trained BERT models could capture word-level and sentence-level representations and learn the common features of protein sequences. MLM is able to model complex relationships between amino acids and capture the evolutionary information of proteins[44]. During the MLM task, 15% of the tokens among the original protein sequence were randomly masked to obtain the masked sequence, and a special token [MASK] was used to replace the masked token 80% of the time, a random token was used 10% of the time, and the selected token was unchanged 10% of the time. The model then predicted the masked tokens based on the context of the unmasked sequence for language modeling. For training, we minimized the negative log likelihood of the true amino acid at each of the masked positions using equation (1),

$$L_{\mathrm{MLM}} = -\sum_{\hat{x} \in m(x)} \log q\left(\hat{x} \mid x_{\setminus m(x)}\right) \#(1)$$

Where L_{MLM} is the loss function of MLM, $m(x)$ and $x_{\setminus m(x)}$ denote the masked tokens from the protein sequences x and the remaining tokens, respectively. During the NSP task, the data were randomly divided into two equal parts A and B: choosing the protein sequence segments A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A (labeled as IsNext), and 50% of the time it is a random sentence from the corpus (labeled as NotNext). BERT was trained by identifying whether these protein sequence segment pairs were continuous. The loss function of the NSP task, L_{NSP} was described as equation (2),

$$L_{\mathrm{NSP}} = -\log p(t \mid x, y) \#(2)$$

where $t = 1$ if x and y are continuous protein segments from protein corpus.

Then MLM and NSP were trained together, with the goal of minimizing the combined loss function of the two strategies as defined in Eq. (3).

$$L = L_{\mathrm{MLM}} + L_{\mathrm{NSP}} \#(3)$$

The pre-training hyperparameters included train steps of 10 million times, a learning rate of 2×10^{-5} and a batch size of 32 to train the BERT model[42]. After the pre-training process, we obtained three pre-trained BERT models. To construct the PLM for a specific downstream task that identifies and predicts DPP-IV-IPs, we modified the pre-trained BERT model**s** by adding a classification layer on top of the BERT output for the [CLS] token. We fine-tuned the pre-trained model on a benchmark dataset containing DPP-IV-IPs without major architectural modifications. For the fine-tuning hyperparameters, we used a learning rate of 2×10^{-6}, a batch size of 32, a warm-up proportion of 0.1, and an average training of 50 epochs.

4.4 Performance evaluation of the model

We used four general quantitative indicators to evaluate our model: Acc, sensitivity (Sn), specificity (Sp), and the MCC, each of which is defined by equation (4), equation (5), equation (6), and equation (7), respectively. TP (true positive) is the number of correctly predicted DPP-IV-IPs, FN (false negative) is the number of DPP-IV-IPs that were in fact predicted to be non-DPP-IV-IPs, TN (true negative) is the number of correctly predicted non-DPP-IV-IPs, and FP (false positive) is the number of non-DPP-IV-IPs that were in fact predicted to be DPP-IV-IPs.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})} \#(4)$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \#(5)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \#(6)$$

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

Sn and Sp reflect the model's ability to recognize DPP-IV-IPs and non-DPP-IV-IPs, respectively, 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 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 the opposite of the actual result; and when it is close to 0, the prediction result of the model is similar to a random prediction. In addition, the area under

the receiver operating characteristic (AUC) was used as another statistical metric. Considering these five evaluation indicators, the performance of the classification models can be better evaluated.

4.5 Visualization of the Model

Representation visualization was used to analyze the learning capacity of the model. To the original protein sequence, $x = [x_1, \ldots, x_n]$ will, we added the special start and end tokens, CLS and SEP, respectively, and got the final input sequence $x' = [\mathrm{CLS}, x_1, \ldots, x_n, \mathrm{SEP}]$. The resulting sequence representation h_i was obtained from hidden states in the ith layer of BERT-DPPIV.

Analyzing the information contained in the representation learned by the model, we used the vector corresponding to CLS in each layer of BERT-DPPIV to represent the analytical sequence and analyzed the representation at the residue level and sequence level by t-SNE. At the residue level, amino acids were divided into six categories, consisting of 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) residues. At the sequence level, 10 kinds of physicochemical property information were obtained from the modlAMP package[34].

Attention visualization was used to analyze the information about which the model was specifically concerned and provide interpretable analysis. Each attention head in a model layer produces an attention matrix, α, which indicates the degree of correlation between token pairs. For example, $\alpha_{i,j}$ indicates the attention from token i to token j, and the weight of token i for all tokens is 1, as shown in equation (8).

$$\sum_{j=1}^{n} \alpha_{i,j} = 1 \#(8)$$

We analyzed attention through a multiscale visualization tool for the Transformer model[36]. The attention-head view in this tool expresses the self-attention matrix α in the form of connected lines and displays the attention patterns generated by one or more attention heads 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) (Fig. 2d). The different colors of the lines indicate different attention heads, whereas 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[38] to

obtain the structure information of the polypeptide sequences and explore the relationship between sites of interaction in polypeptide structures and model attention patterns.

We then used the attention matrix to calculate the importance of each token, d_j, in the peptide sequence as shown in equation (9).

$$d_j = \sum_{i=1}^{n} \alpha_{i,j} \#(9)$$

We carried out a statistical analysis of 144 attention patterns including all layers and all heads of each layer. We removed sequences of less than or equal to three amino acids. Then we calculated the importance of all tokens including the special tokens (CLS, SEP) and selected the three most important tokens from each sequence for each attention pattern. We counted the distribution of the top three most important tokens across the different patterns for each sequence with respect to their position and amino acid type. We also selected a specific residue in the sequence and calculated the importance of this residue in 144 attention patterns and visualized the results using a heatmap. We then investigated the interaction between attention and particular amino acids[36]. We used equation (10) for amino acids and defined an indicator function, $f(i,j)$, that returns a value of 1 if this amino acid was present in the token j (e.g., if token j is a proline).

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

where $p_\alpha(f)$ equals the proportion of attention that is directed to the amino acid. We computed the proportion of attention for the fine-tuned model and the pre-trained model directed toward 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[20]. hDPP-IV was used for the triplicate measurements. Peptide solution (20 µl); substrate solution, consisting of 2 mM Gly-Pro-pNA (20 µl); and Tris-HCl buffer (pH 8.0; 20 µl) were pre-mixed in a 96-well microplate. The reaction was initiated by adding 40 µl of hDPP-IV (final concentration, 0.025 U/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 phosphate buffered saline (pH 7.4). The hDPP-IV inhibitory ratio of the sample was calculated according to equation (11),

$$\text{Inhibition ratio} = \frac{A_{\text{Control}} - A_{\text{Sample}}}{A_{\text{Control}}} \#(11)$$

where A_{Control} is the absorbance of the control sample with PBS and A_{Sample} is the absorbance of the peptide sample.

5. Declarations

- This work was funded by Shenzhen Science and Technology Innovation Commission (KCXFZ20201221173207022).
- The authors declare no competing interests
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: All authors agree to publish this manuscript
- The code in this work is publicly available and can be accessed at https://github.com/guanchangge/BERT-DPPIV.
- Authors' contributions: X.-H.X., N.Y., Y.L., and C.G. conceived the project and designed the study. C.G. designed the tripeptide and pentapeptide screening experiments; J.C. and C.G. designed the model construction and analysis; C.G. and S.L. performed the tripeptide and pentapeptide screening experiments; C.G. and J.Luo performed model construction; C.G. conducted wet experiments; C.G., J.Luo, and Z.L.T. performed data analysis and interpretation. C.G., J.Luo, J.C., and Z.L.T. drafted and revised the paper. J.Li, Z.W., H.C., Y.W. and C.Z. provided consultation for the project. All authors contributed to the revision of 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-Rev. Med. Chem.* **19**, 88–97 (2018).
2. Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **183**, 109119 (2022).
3. Ogurtsova, K. *et al.* IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res. Clin. Pract.* **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. *Nat. Struct. Biol.* **10**, 19–25 (2003).

5. Casrouge, A. *et al.* Lymphocytes are a major source of circulating soluble dipeptidyl peptidase 4. *Clin. Exp. Immunol.* **194**, 166–179 (2018).

6. Roppongi, S. *et al.* Crystal structures of a bacterial dipeptidyl peptidase IV reveal a novel substrate recognition mechanism distinct from that of mammalian orthologues. *Sci. Rep.* **8**, 2714 (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. *Eur. J. Biochem.* **214**, 829–835 (1993).

8. Yip, R. G. C. & Wolfe, M. M. GIF biology and fat metabolism. *Life Sci.* **66**, 91–103 (1999).

9. Demuth, H.-U., McIntosh, C. H. S. & Pederson, R. A. Type 2 diabetes—Therapy with dipeptidyl peptidase IV inhibitors. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1751**, 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. *Diab. Vasc. Dis. Res.* **3**, 159–165 (2006).

11. Deacon, C. F. A review of dipeptidyl peptidase-4 inhibitors. Hot topics from randomized controlled trials. *Diabetes Obes. Metab.* **20**, 34–46 (2018).

12. Jarvis, C. I., Cabrera, A. & Charron, D. Alogliptin: A New Dipeptidyl Peptidase-4 Inhibitor for Type 2 Diabetes Mellitus. *Ann. Pharmacother.* **47**, 1532–1539 (2013).

13. Barnett, A. DPP-4 inhibitors and their potential role in the management of type 2 diabetes: DPP 4 INHIBITORS AND MANAGEMENT OF TYPE 2 DIABETES. *Int. J. Clin. Pract.* **60**, 1454–1470 (2006).

14. Mulvihill, E. E. & Drucker, D. J. Pharmacology, Physiology, and Mechanisms of Action of Dipeptidyl Peptidase-4 Inhibitors. *Endocr. Rev.* **35**, 992–1019 (2014).

15. Nongonierma, A. B. & FitzGerald, R. J. Features of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides from dietary proteins. *J. Food Biochem.* **43**, e12451 (2019).

16. Wang, Y. *et al.* Functional discovery and production technology for natural bioactive peptides. *Chin. J. Biotechnol.* **37**, 2166–2180 (2021).

17. Duez, H., Cariou, B. & Staels, B. DPP-4 inhibitors in the treatment of type 2 diabetes. *Biochem. Pharmacol.* **83**, 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 Chem.* **165**, 489–498 (2014).

19. 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).

20. 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).

21. 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 Chem.* **244**, 340–348 (2018).

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 Adv.* **6**, 75400–75413 (2016).

23. Shoombuatong, W. *et al.* Towards the Revival of Interpretable QSAR Models. in *Advances in QSAR Modeling* (ed. Roy, K.) vol. 24 3–55 (Springer International Publishing, 2017).

24. Hellberg, S., Sjoestroem, M., Skagerberg, B. & Wold, S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* **30**, 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. *Med. Res. Rev.* **40**, 1276–1314 (2020).

26. Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, Md. 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. *J. Proteome Res.* **19**, 4125–4136 (2020).

27. Zou, H. & Yin, Z. Identifying Dipeptidyl Peptidase-IV Inhibitory Peptides Based on Correlation Information of Physicochemical Properties. *Int. J. Pept. Res. Ther.* **27**, 2651–2659 (2021).

28. Charoenkwan, P. *et al.* StackDPPIV: A novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. *Methods* **204**, 189–198 (2022).

29. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **31**, 685–695 (2021).

30. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 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)* 498–502 (IEEE, 2017). doi:10.1109/BIBM.2017.8217697.

32. Yan, J. *et al.* Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther. - 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**, 456 (2019).

34. Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **33**, 2753–2755 (2017).

35. Vaswani, A. *et al.* Attention is all you need. in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (2017). doi:10.5555/3295222.3295349.

36. Vig, J. A Multiscale Visualization of Attention in the Transformer Model. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 37–42 (Association for Computational Linguistics, 2019). doi:10.18653/v1/P19-3007.

37. Timmons, P. B. & Hewage, C. M. APPTEST is a novel protocol for the automatic prediction of peptide tertiary structures. *Brief. Bioinform.* **22**, bbab308 (2021).

38. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

39. Guan, C., Iwatani, S., Xing, X. & Yamamoto, N. Strategic Preparations of DPP-IV Inhibitory Peptides from Val-Pro-Xaa and Ile-Pro-Xaa Peptide Mixtures. *Int. J. Pept. Res. Ther.* **27**, 735–743 (2021).

40. Xu, X., Yan, C. & Zou, X. MDockPeP: An *ab initio* protein–peptide docking server. *J. Comput. Chem.* **39**, 2409–2413 (2018).

41. 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**, 2112–2120 (2021).

42. Zhang, Y., Lin, J., Zhao, L., Zeng, X. & Liu, X. A novel antibacterial peptide recognition algorithm based on BERT. *Brief. Bioinform.* **22**, bbab200 (2021).

43. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019*

Conference of the North 4171–4186 (Association for Computational Linguistics, 2019). doi:10.18653/v1/N19-1423.

44. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans. Pattern Anal. Mach. Intell.* 1–1 (2021) doi:10.1109/TPAMI.2021.3095381.

Figure Captions

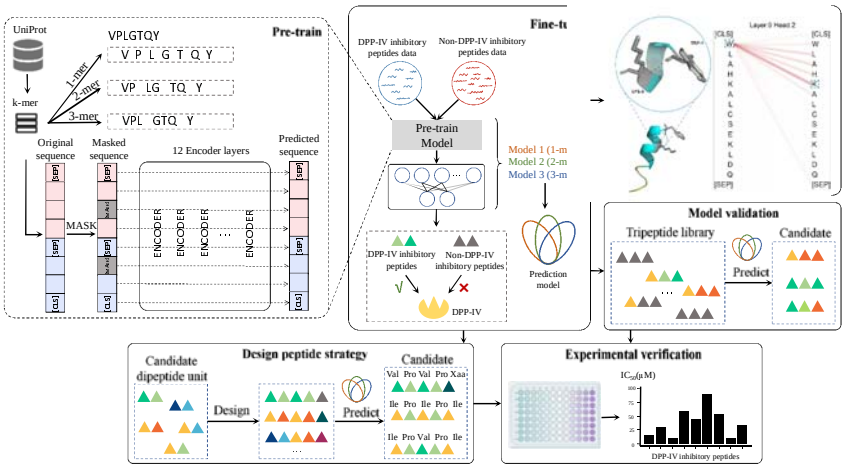

Fig. 1 Schematic representation of study workflow. We first collected sequences to pre-train a BERT model (upper left). We then used the DPP-IV-IP dataset for fine-tuning the pre-trained BERT model to obtain three kinds of PLMs. The three models were combined to construct the BERT-DPPIV (upper middle). To determine what the model had learned, we analyzed and visualized the model's attention (upper right**,** analysis module). We then used BERT-DPPIV to screen tripeptides that may 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 were verified by biological experiments (lower right). Having demonstrated the practical applicability of our model, we proposed a design strategy for DPP-IV-IPs based on dipeptide repeat units and demonstrated the feasibility of this strategy using models and experiments (lower left).

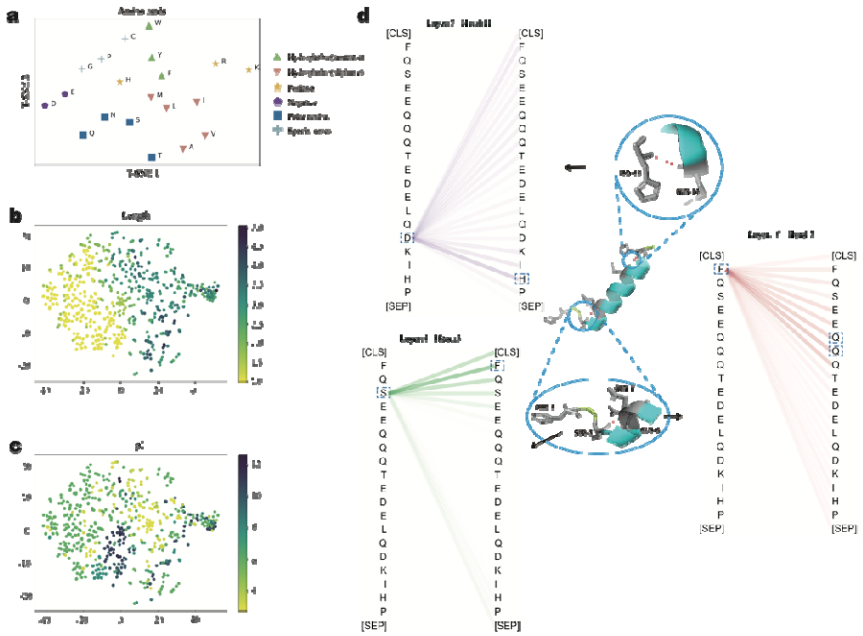

Fig. 2 Sequence representation visualization and attention visualization of structural information learned by the model. A vector representation of the peptide sequence and of individual amino acids was extracted from the model and analyzed by t-SNE. The amino acid representation learned by the model (a) and physicochemical properties contained in model-learned representations of peptide sequences (b and c) are shown. 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 relative to all other tokens in the sequence. High scores indicate that the model learned to put more weight on certain residue pairs (upper left, lower left, and lower right). The structure of the peptide shows that Ser-3 can interact with Phe-1, Gln-6, and Gln-7 and that 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.

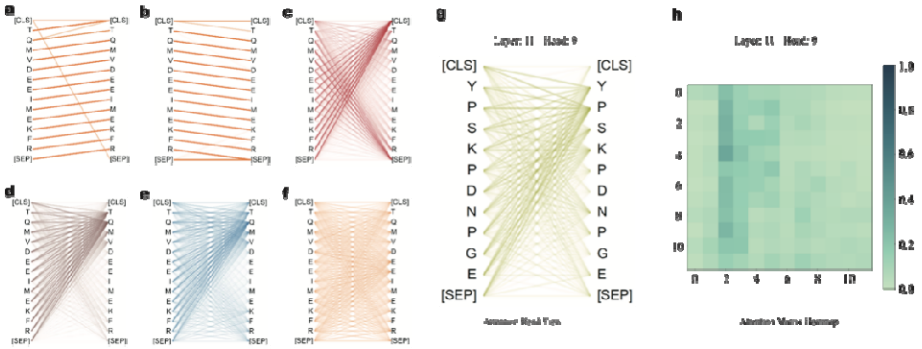

Fig. 3 Visualizing the attention patterns of the model. Previous-word attention patterns: virtually all attention is focused on the previous word in the input sequence (a). Next-word attention patterns: virtually all attention is focused on the next word in the input sequence (b). Delimiter-focused

attention patterns: most attention is focused on the [SEP] token or [CLS] token (c). Specific-word attention patterns: virtually all attention is focused on the specific site (d and e). Attention related–words attention patterns: attention is paid to identical or related words (f). Attention is focused mainly on the second and third position of the peptide sequence and the corresponding heatmap (g and h).

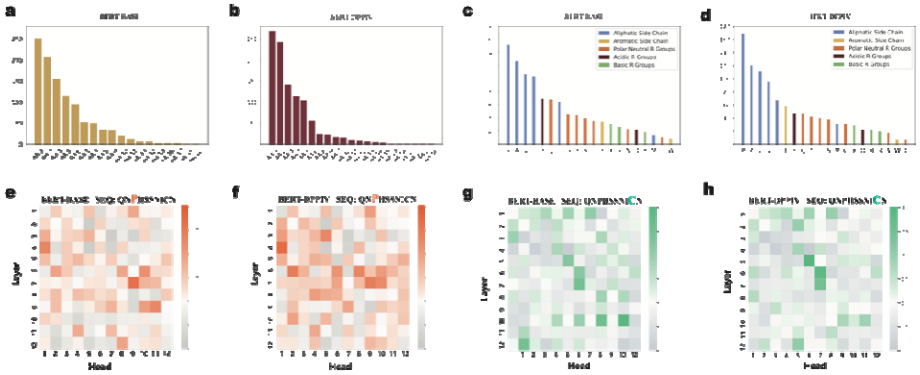

Fig. 4 Visualization of the position and importance of amino acid species. To study the importance of position and amino acid species, attention to positions and amino acids was statistically analyzed. BERT-BASE and BERT-DPPIV represent the models before and after fine-tuning, respectively. The statistical importance of the position of peptide sequence according to the attention of the model is shown (a and b). The statistical importance of amino acid species according to the attention of the model is shown (c and d). The attention of the model to proline and cysteine in the peptide sequence was visualized (e, f, g, and

h).

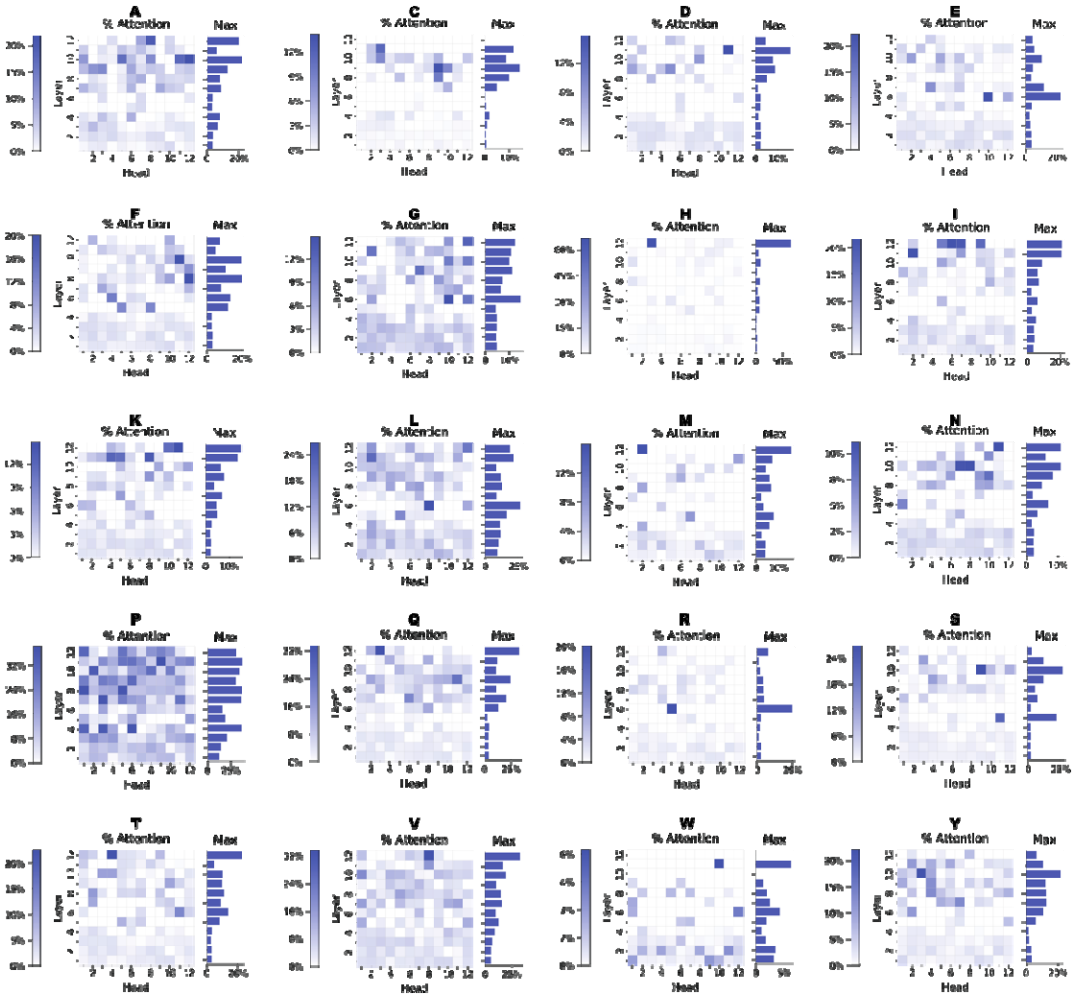

Fig. 5 Percentage of each attention head that is focused on the 20 natural amino acids.

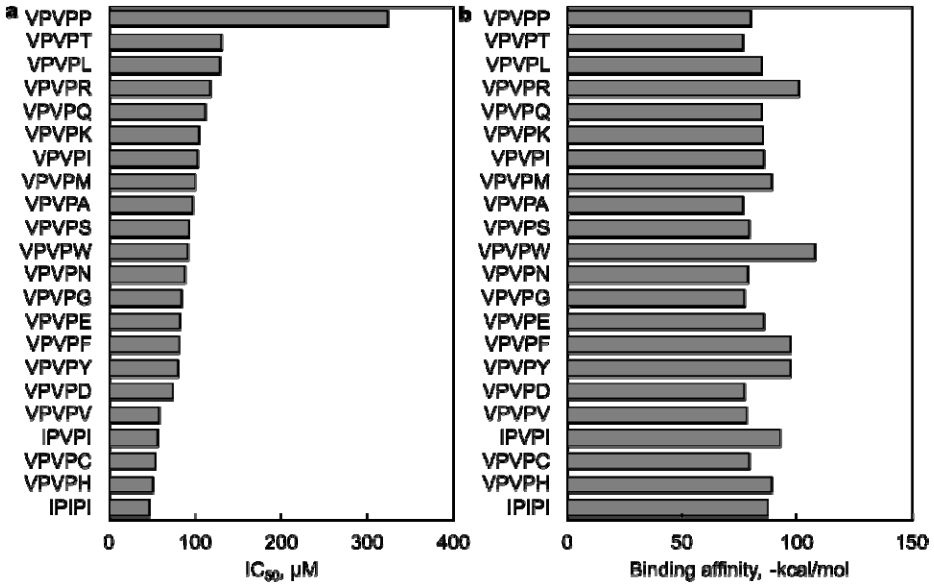

Fig. 6 Characteristics of the activity of DPP-IV inhibitory pentapeptides. The pentapeptides were characterized by (a) IC_{50} measured experimentally and (b) binding affinity based on modelling.

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

Acc, accuracy; Sn, sensitivity; Sp, specificity; MCC, Matthews correlation coefficient; AUC, area under receiver operating characteristics curve

Table 2 Experimental characterization of tripeptides that inhibit DPP-IV

Peptide	IC_{50}, μM	Peptide	IC_{50}, μM	Peptide	IC_{50}, μM	Peptide	IC_{50}, μM
WRM	896.67	WPE	353.02	WAC	282.79	WPQ	196.18
WRP	741.50	WAG	352.16	WRH	277.20	WVE	192.50
WPC	739.03	WVD	343.00	WVV	275.33	WAH	189.55
WAR	664.68	WVA	341.25	WAA	274.53	WVF	188.33
WRW	603.14	WVG	340.63	WRG	269.85	WPW	182.96
WPS	588.31	WVT	335.63	WRS	265.09	WRE	160.15
WRT	555.60	WVK	331.38	WAL	260.10	WAF	153.32
WRV	541.17	WRI	330.4	WRF	257.07	WRL	150.75
WRA	534.67	WAM	329.85	WVH	256.86	WPG	134.98

WPV	511.47	WVM	317.00	WAV	249.79	WPN	128.59
WPD	480.61	WPH	314.57	WVS	247.00	WAY	117.40
WAE	461.79	WRN	313.20	WVW	238.63	WAW	103.66
WPL	461.32	WPY	303.79	WVC	231.78	WPI	-
WAP	440.97	WPP	303.49	WPM	218.75	WPK	-
WPT	435.08	WAN	301.78	WVY	214.88	WAK	-
WAD	432.65	WRY	296.50	WRC	209.64	WAS	-
WVP	402.00	WAT	296.33	WPR	208.94	WRK	-
WVL	398.88	WAQ	289.19	WPA	206.01	WRR	-
WVN	372.88	WPF	285.91	WRD	203.64	WVI	-
WVQ	366.00	WRQ	285.50	WAI	198.88	WVR	-

IC_{50} : half maximal inhibitory concentration.
- : not detected within 1